ICNS 2013

The Ninth International Conference on Networking and Services

ISBN: 978-1-61208-256-1

LMPCNA 2013

The Fifth International Workshop on Learning Methodologies and Platforms used in the Cisco Networking Academy

March 24 - 29, 2013

Lisbon, Portugal

ICNS 2013 Editors

Eugen Borcoci, University 'Politehnica' Bucharest, Romania

Steffen Fries, Siemens, Germany

Sandra Sendra Compte, Universidad Politécnica de Valencia, Spain

Mak Sharma, Birmingham City University City, UK

ICNS 2013

Foreword

The Ninth International Conference on Networking and Services [ICNS 2013], held between March 24 - 29, 2013 in Lisbon, Portugal, continued a series of events targeting general networking and services aspects in multi-technologies environments. The conference covered fundamentals on networking and services, and highlighted new challenging industrial and research topics. Network control and management, multi-technology service deployment and assurance, next generation networks and ubiquitous services, emergency services and disaster recovery and emerging network communications and technologies were considered.

IPv6, the Next Generation of the Internet Protocol, has seen over the past three years tremendous activity related to its development, implementation and deployment. Its importance is unequivocally recognized by research organizations, businesses and governments worldwide. To maintain global competitiveness, governments are mandating, encouraging or actively supporting the adoption of IPv6 to prepare their respective economies for the future communication infrastructures. In the United States, government's plans to migrate to IPv6 has stimulated significant interest in the technology and accelerated the adoption process. Business organizations are also increasingly mindful of the IPv4 address space depletion and see within IPv6 a way to solve pressing technical problems. At the same time IPv6 technology continues to evolve beyond IPv4 capabilities. Communications equipment manufacturers and applications developers are actively integrating IPv6 in their products based on market demands.

IPv6 creates opportunities for new and more scalable IP based services while representing a fertile and growing area of research and technology innovation. The efforts of successful research projects, progressive service providers deploying IPv6 services and enterprises led to a significant body of knowledge and expertise. It is the goal of this workshop to facilitate the dissemination and exchange of technology and deployment related information, to provide a forum where academia and industry can share ideas and experiences in this field that could accelerate the adoption of IPv6. The workshop brings together IPv6 research and deployment experts that will share their work. The audience will hear the latest technological updates and will be provided with examples of successful IPv6 deployments; it will be offered an opportunity to learn what to expect from IPv6 and how to prepare for it.

Packet Dynamics refers broadly to measurements, theory and/or models that describe the time evolution and the associated attributes of packets, flows or streams of packets in a network. Factors impacting packet dynamics include cross traffic, architectures of intermediate nodes (e.g., routers, gateways, and firewalls), complex interaction of hardware resources and protocols at various levels, as well as implementations that often involve competing and conflicting requirements.

Parameters such as packet reordering, delay, jitter and loss that characterize the delivery of packet streams are at times highly correlated. Load-balancing at an intermediate node may, for example, result in out-of-order arrivals and excessive jitter, and network congestion may manifest as packet losses or large jitter. Out-of-order arrivals, losses, and jitter in turn may lead to unnecessary retransmissions in TCP or loss of voice quality in VoIP.

With the growth of the Internet in size, speed and traffic volume, understanding the impact of underlying network resources and protocols on packet delivery and application performance has assumed a critical importance. Measurements and models explaining the variation and interdependence of delivery characteristics are crucial not only for efficient operation of networks and network diagnosis, but also for developing solutions for future networks.

Local and global scheduling and heavy resource sharing are main features carried by Grid networks. Grids offer a uniform interface to a distributed collection of heterogeneous computational, storage and network resources. Most current operational Grids are dedicated to a limited set of computationally and/or data intensive scientific problems.

Optical burst switching enables these features while offering the necessary network flexibility demanded by future Grid applications. Currently ongoing research and achievements refers to high performance and computability in Grid networks. However, the communication and computation mechanisms for Grid applications require further development, deployment and validation.

ICNS 2013 also featured the following workshop:

- LMPCNA 2013, The Fifth International Workshop on Learning Methodologies and Platforms used in the Cisco Networking Academy

We take here the opportunity to warmly thank all the members of the ICNS 2013 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 ICNS 2013. 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 ICNS 2013 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that ICNS 2013 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the fields of networking and services.

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

ICNS 2013 Chairs:
Pedro Andrés Aranda Gutiérrez, Telefónica I+D - Madrid, Spain
Carlos Becker Westphall, Federal University of Santa Catarina, Brazil
Eugen Borcoci, University 'Politehnica' Bucharest, Romania
Jaime Lloret Mauri, Polytechnic University of Valencia, Spain
Sathiamoorthy Manoharan, University of Auckland, New Zealand
Yoshiaki Taniguchi, Osaka University, Japan
Toan Nguyen, INRIA - Grenoble - Rhone-Alpes, France

Abdulrahman Yarali, Murray State University, USA
Emmanuel Bertin, France Telecom R&D - Orange Labs, France
Steffen Fries, Siemens, Germany

LMPCNA 2013 Workshop Chair

Mak Sharma, Birmingham City University City, UK

ICNS 2013

Committee

ICNS Advisory Chairs

Pedro Andrés Aranda Gutiérrez, Telefónica I+D - Madrid, Spain
Carlos Becker Westphall, Federal University of Santa Catarina, Brazil
Eugen Borcoci, University 'Politehnica' Bucharest, Romania
Jaime Lloret Mauri, Polytechnic University of Valencia, Spain
Sathiamoorthy Manoharan, University of Auckland, New Zealand
Yoshiaki Taniguchi, Osaka University, Japan
Toan Nguyen, INRIA - Grenoble - Rhone-Alpes, France
Abdulrahman Yarali, Murray State University, USA
Emmanuel Bertin, France Telecom R&D - Orange Labs, France
Steffen Fries, Siemens, Germany

ICNS 2013 Technical Program Committee

Johan Åkerberg, ABB AB - Corporate Research - Västerås, Sweden
Ryma Abassi, Higher School of Communication of Tunis /Sup'Com, Tunisia
Ferran Adelantado i Freixer, Universitat Oberta de Catalunya, Spain
Javier M. Aguiar Pérez, Universidad de Valladolid, Spain
Rui L.A. Aguiar, University of Aveiro, Portugal
Basheer Al-Duwairi, Jordan University of Science and Technology, Jordan
Ali H. Al-Bayatti, De Montfort University - Leicester, UK
Mario Anzures-García, Benemérita Universidad Autónoma de Puebla, Mexico
Pedro Andrés Aranda Gutiérrez, Telefónica I+D - Madrid, Spain
Patrick Appiah-Kubi, Towson University, USA
Bourdena Athina, University of the Aegean, Greece
Mohamad Badra, Dhofar University, Oman
Mohammad M. Banat, Jordan University of Science and Technology, Jordan
Javier Barria, Imperial College of London, UK
Mostafa Bassiouni, University of Central Florida, USA
Michael Bauer, The University of Western Ontario - London, Canada
Carlos Becker Westphall, Federal University of Santa Catarina, Brazil
Tarek Bejaoui, University of Carthage, Tunisia
Mehdi Bennis, University of Oulu, Finland
Luis Bernardo, Universidade Nova de Lisboa, Portugal
Emmanuel Bertin, France Telecom R&D - Orange Labs, France
Alex Bikfalvi, Madrid Institute for Advanced Studies in Networks - Madrid, Spain
Eugen Borcoci, University "Politehnica"of Bucharest (UPB), Romania
Fernando Boronat Seguí, Polytechnic University of Valencia, Spain
Kalinka Branco, University of São Paulo, Brazil
Jens Buysse, Ghent University/IBBT, Belgium
Maria Calderon Pastor, Universidad Carlos III, Madrid, Spain
Maria Dolores Cano Baños, Polytechnic University of Cartagena - Campus Muralla del Mar, Spain

Tarik Caršimamovic, BHTelecom, Bosnia and Herzegovina
José Cecílio, University of Coimbra, Portugal
Patryk Chamuczynski, Technisat Digital R&D, Poland
Bruno Chatras, Orange Labs, France
Wei Cheng, University of California - Davis, USA
Jun Kyun Choi, KAIST, Korea
Victor Clincy, Kennesaw State University, USA
Hugo Coll Ferri, Universidad Politecnica de Valencia, Spain
Todor Cooklev, Indiana University - Purdue University Fort Wayne, USA
Alejandro Cordero, Amaranto Consultores, Spain
Noelia Correia, Universidade do Algarve, Portugal
Taiping Cui, Inha University - Incheon, Korea
Carlton Davis, École Polytechnique de Montréal, Canada
João Henrique de Souza Pereira, University of São Paulo, Brazil
Wei Ding, New York Institute of Technology, USA
Zbigniew Dziong, ETS - Montreal, Canada
Giuseppe Durisi, Chalmers University of Technology - Göteborg, Sweden
El-Sayed El-Alfy, King Fahd University of Petroleum and Minerals, Saudi Arabia
Safwan El Assad, Institut d'Electronique et des Télécommunications de Rennes || École d'ingénieurs de l'université de Nantes, France
Fakher Eldin Mohamed Suliman, Sudan University of Science and Technology, Sudan
Issa Elfergani, Instituto de Telecomunicações - Aveiro, Portugal || University of Bradford, UK
Juan Flores, University of Michoacan, Mexico
Steffen Fries, Siemens, Germany
Sebastian Fudickar, University of Potsdam, Germany
Michael Galetzka, Fraunhofer Institute for Integrated Circuits - Dresden, Germany
Alex Galis, University College London, UK
Ivan Ganchev, University of Limerick, Ireland
Elvis Eduardo Gaona G., Universidad Distrital Francisco José de Caldas, Colombia
Abdennour El Rhalibi, Liverpool John Moores University, UK
Stenio Fernandez, Federal University of Pernambuco, Brazil
Gianluigi Ferrari, University of Parma, Italy
Miguel Garcia Pineda, Universitat Politecnica de Valencia, Spain
Rosario Garroppo, Università di Pisa, Italy
Sorin Georgescu, Ericsson Research, Canada
Mikael Gidlund, ABB, Sweden
Marc Gilg, University of Haute Alsace, France
Debasis Giri , Haldia Institute of Technology, India
Ivan Glesk, University of Strathclyde - Glasgow, UK
Ann Gordon-Ross, University of Florida, USA
Victor Govindaswamy, Texas A&M University-Texarkana, USA
Dominic Greenwood, Whitestein, Switzerland
Jean-Charles Grégoire, INRS - Université du Québec - Montreal, Canada
Vic Grout, Glyndwr University - Wrexham, UK
Ibrahim Habib, City University of New York, USA
Go Hasegawa, Osaka University, Japan
Hermann Hellwagner, Klagenfurt University, Austria
Enrique Hernandez Orallo, Universidad Politécnica de Valencia, Spain

Zhihong Hong, Communications Research Centre, Canada
Per Hurtig, Karlstad University, Sweden
Naohiro Ishii, Aichi Institute of Technology, Japan
Arunita Jaekel, University of Windsor, Canada
Tauseef Jamal, University Lusofona - Lisbon, Portugal
Peter Janacik, University of Paderborn, Germany
Imad Jawhar, United Arab Emirates University, UAE
Ravi Jhawar, Universitàdegli Studi di Milano - Crema, Italy
Sudharman K. Jayaweera, University of New Mexico - Albuquerque, USA
Ying Jian, Google Inc, USA
Fan Jiang, Tuskegee University, USA
Eunjin (EJ) Jung, University of San Francisco, USA
Enio Kaljic, University of Sarajevo, Bosnia and Herzegovina
Georgios Kambourakis, University of the Aegean - Karlovassi, Greece
Hisao Kameda, University of Tsukuba, Japan
Nirav Kapadia, Public Company Accounting Oversight Board (PCAOB), USA
Georgios Karagiannis, University of Twente, The Netherlands
Masoumeh Karimi, Technological University of America, USA
Aggelos K. Katsaggelos, Northwestern University - Evanston, USA
Sokratis K. Katsikas, University of Piraeus, Greece
Razib Hayat Khan, NTNU, Norway
Bithika Khargharia, Cisco Systems, Inc., USA
Dong Seong Kim, University of Canterbury, New Zealand
Younghan Kim, Soongsil University - Seoul, Republic of Korea
Mario Kolberg, University of Stirling - Scotland, UK
Lisimachos Kondi, University of Ioannina, Greece
Jerzy Konorski, Gdansk University of Technology, Poland
Elisavet Konstantinou, University of the Aegean, Greece
Kimon Kontovasilis, NCSR "Demokritos", Greece
Andrej Kos, University of Ljubljana, Slovenia
Igor Kotenko, St. Petersburg Institute for Informatics and Automation, Russia
Evangelos Kranakis, Carleton University, - Ottawa, Canada
Francine Krief, University of Bordeaux, France
Suk Kyu Lee, Korea University at Seoul, Republic of Korea
DongJin Lee, Auckland University, New Zealand
Leo Lehmann, OFCOM, Switzerland
Ricardo Lent, Imperial College London, UK
Alessandro Leonardi, AGT Group (R&D) GmbH - Darmstadt, Germany
Qilian Liang, University of Texas at Arlington, USA
Wen-Hwa Liao, Tatung University - Taipei, Taiwan
Fidel Liberal Malaina, University of Basque Country, Spain
Thomas Little, Boston University, USA
Giovanni Livraga, Università degli Studi di Milano - Crema, Italy
Jaime Lloret Mauri, Polytechnic University of Valencia, Spain
Edmo Lopes Filho, Algar Telecom, Brazil
Albert Lysko, Meraka Institute/CSIR- Pretoria, South Africa
Zoubir Mammeri, ITIT - Toulouse, France
Sathiamoorthy Manoharan, University of Auckland, New Zealand

Vasco Soares, Instituto de Telecomunicações / University of Beira Interior / Polytechnic Institute of Castelo Branco, Portugal
José Soler, Technical University of Denmark, Denmark
Gritzalis Stefanos, University of the Aegean, Greece
Akira Takura, Jumonji University, Japan
Yoshiaki Taniguchi, Osaka University, Japan
Olivier Terzo, Istituto Superiore Mario Boella - Torino, Italy
Christian Timmerer, Alpen-Adria-Universität Klagenfurt, Austria
Petia Todorova, Fraunhofer Institut FOKUS - Berlin, Germany
Binod Vaidya, University of Ottawa, Canada
Hans van den Berg, TNO / University of Twente, The Netherlands
Ioannis O. Vardiambasis, Technological Educational Institute (TEI) of Crete - Branch of Chania, Greece
Dario Vieira, EFREI, France
Bjørn Villa, Norwegian Institute of Science and Technology, Norway
José Miguel Villalón Millan, Universidad de Castilla - La Mancha, Spain
Manuel Villén-Altamirano, Universidad Politécnica de Madrid, Spain
Demosthenes Vouyioukas, University of the Aegean - Karlovassi, Greece
Arno Wacker, University of Kassel, Germany
Bin Wang, Wright State University - Dayton, USA
Mea Wang, University of Calgary, Canada
Tingkai Wang, London Metropolitan University, UK
Michelle Wetterwald, EURECOM - Sophia Antipolis, France
Ouri Wolfson, University of Illinois - Chicago, USA
Feng Xia, Dalian University of Technology, China
Homayoun Yousefi'zadeh, University of California - Irvine, USA
Vladimir S. Zaborovsky, Polytechnic University/Robotics Institute - St.Petersburg, Russia
Sherali Zeadally, University of the District of Columbia, USA
Tao Zheng, Orange Labs Beijing, China
Yifeng Zhou, Communications Research Centre, Canada
Ye Zhu, Cleveland State University, USA
Piotr Zuraniewski, University of Amsterdam (NL), The Netherlands /AGH University of Science and Technology, Poland

LMPCNA 2013 Workshop Chair

Mak Sharma, Birmingham City University City, UK

LMPCNA 2013 Technical Program Committee

Nalin Abeysekera, Open University of Sri Lanka, Sri Lanka
Ron Austin, Birmingham City University, UK
Irfan Awan, University of Bradford, UK
Rehan Bhana, Birmingham City University, UK
Giancarlo Bo, Technology and Innovation Consultant – Genova, Italy
Maiga Chang, Athabasca University, Canada
Pavel Cicak, Slovak University of Technology, Slovakia
Giuseppe Cinque, Consorzio ELIS - Rome, Italy
Dumitru Dan Burdescu, University of Craiova, Romania

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

Automated Dynamic Topology Configuration

An Innovative Approach to Online Rack Renting

Vimukthi S. M. W. T. Mudiyanselage
School of Science & Technology
University of Northampton
Northampton, United Kingdom
vimukthilk@ieee.org

Rashmi Dravid
School of Science & Technology
University of Northampton
Northampton, United Kingdom
rashmi.dravid@northampton.ac.uk

***Abstract*— The facility to rent access-time to computer network hardware on the Internet has been available for a number of years. Provided as a service targeted at academics engaging in the study of Computer Networks, it has been a viable solution successfully filling the void created by numerous obstacles encountered in procuring physical equipment due to various factors. However, current business models offer limited flexibility to end users because the online labs are offered as a predetermined selection of devices prewired to a topology deemed suitable by the service provider; consequently the end user, although paying for this service, has no control over the composition or topology of the labs they hire. This paper introduces a new improved model and the underlying technological implementation, which features a paradigm shift in the way online labs are defined, configured and ultimately offered to the end user. It aspires to eliminate the above limitations through automated and dynamic configuration of the network topology, allowing the end user to select the composition and topology of the labs they hire, thus unleashing their true potential.**

Keywords-online rack rental; dynamic topology configuration; Layer 2 Protocol Tunneling; VLAN

I. Introduction

The facility to rent access-time to computer network hardware on the Internet (also called online rack rental) has been available for a number of years. Provided as a service targeted mainly at students, academics and professionals engaging in the study of Computer Networks, it has been a viable solution successfully filling the void created by various impracticalities in economics and logistics of procuring physical networks equipment.

The demand on modern education systems to provision around-the-clock access to IT resources is conspicuous and the ability to meet that demand is no more a nicety but a necessity. Online rack rental systems are therefore an ideal platform to provide students with the means to configure and test network configurations without having to worry about economical or logistical confines. However, existing online rack rental systems suffer from some restrictions which limit the flexibility they offer to end users. Service providers offer a predetermined selection of devices prewired to a specific topology that they deem suitable. Therefore when renting a lab, the user would not only hire access-time to the devices but also the topology that comes with it.

Generally, a prospective client would browse the available labs and select one or more that best fits their requirements. Consequently, a lab may not be an exact match of the user's requirements; it may be comprised of too few or too many devices, and/or may not offer the topology they require. In the former case, the user could be paying for redundant devices, which could otherwise have been rented out to other users or shutdown providing a reduced environmental footprint. Most service providers offer a full or partial mesh topology on their labs where possible, to work around the latter [1].

Root cause analysis conducted to find the underlying cause(s) for the above limitations have yielded some interesting results, as shown in Figure 1. It is evident that the limitations discussed above stem from the presence of a prewired topology. Therefore, if it is possible to eliminate the presence of a prewired topology, such a solution would minimize, if not eradicate, the above limitations and provide a better experience to both end users and the service provider. The authors have not come across any previous work which has identified these limitations with an online rack rental system. Therefore, this research is characteristically novel in its field. As a result, implementation of the fundamental technical concepts has been empirical in nature.

This research paper introduces a new approach to online rack renting by moving the definition of a lab from one which is static and predetermined by the service provider, to dynamic composition of network devices selected by the end user where they only pay for what they use and are able to dynamically configure a topology of their choice. Users who hire labs from the same service provider are able to collaborate notwithstanding geographical boundaries, by networking their individual labs, provided the labs are on the same platform. By adapting their business model to accommodate an automatic, dynamically configurable lab platform for the provision of online rack rental systems, the service provider makes substantial gains as a result of increased revenue opportunities by retaining a more satisfied client base, optimum utilisation of merchandise, reduced electricity costs and promoting a greener business ethos. These lab models could be sold as service packages by

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-256-1

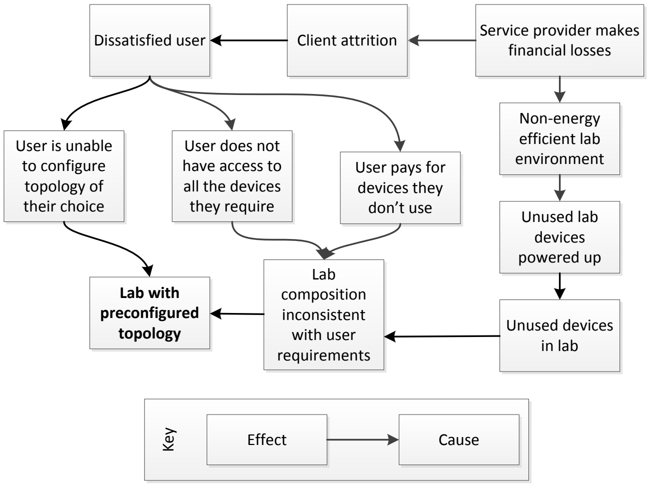

Dissatisfied user
Client attrition
Service provider makes financial losses
User is unable to configure topology of their choice
User does not have access to all the devices they require
User pays for devices they don't use
Non-energy efficient lab environment
Unused lab devices powered up
Lab with preconfigured topology
Lab composition inconsistent with user requirements
Unused devices in lab
Key
Effect
Cause

Figure 1. Root cause analysis

service providers to academic institutions through cloud hosting, helping institutions save on initial capital expenditure and recurring expenses on maintenance contracts.

A. Problem Domain

Normally, a service provider would have an assortment of devices similar to that given in Figure 2 (usually, although not essentially, on a larger scale). Consider two users X and Y (see Table 1) who have different device requirements. User X may hire either lab, but will be unable to source both the routers required, unless he/she hires both labs. However, by doing so they would also be paying for a redundant switch, and as each lab is intended to be rented out separately, they are self-contained and do not offer networking between them.

Either lab would address User Y's requirements; however they would also inevitably be paying for a redundant router. Ultimately neither user's requirements are fully met despite the service provider being in possession of an adequate number of devices to be able to do so.

TABLE 1. EXAMPLE USER REQUIREMENTS

User	Device requirement	
	Routers	*Switches*
X	2	1
Y	0	1

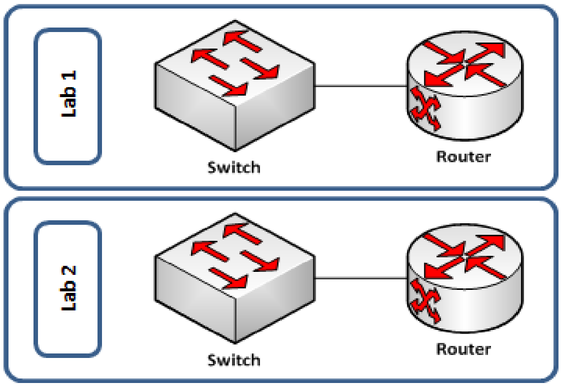


Figure 2. Example setup of current online rack rental system

B. The Proposition

Having identified the root cause of the problem, the principal research focus of this paper was on developing a mechanism which would enable service providers to offer labs independent of a prewired topology, where the end user would be able to dynamically configure a topology as they wish. A prospective solution must satisfy some fundamental requirements to qualify as a successful solution. It should (a) remain transparent to the end user; (b) not interfere with the devices being offered in a way which impedes their normal operation; (c) allow users to determine the composition of and dynamically alter their lab both in terms of its constituent devices and topology; and (d) require minimal involvement from the end user to setup and manage.

As we seek a solution where the devices being offered to users (hereinafter referred to as user devices) are not prewired to each other in the normal fashion, naturally this prompted us to explore various network devices which would be configurable by the service provider, thereby connecting user devices as and when required and severing those connections when they are not required. Our device of choice should meet the requirements identified above. Further, the authors were primarily interested in developing a solution which would benefit an academic audience. Therefore, in addition to the primary requirements, our solution should enable this facility to be provisioned as a service package to academic institutions and, where such institutions choose to offer online racks to its students, our solution should facilitate the option to provide the online labs as collaborative networking platforms to their students.

In the Methodology section we will explore several approaches which were considered when selecting a viable solution, the chosen method and justification to the same. The Implementation section contains a detailed description of the technical application of the new model along with a comprehensive example. In this section we will also discuss how a new business model could evolve around the new technical capabilities the new model offers and its relevance to an academic audience. In the discussion section, we will evaluate our solution's fit-for-purpose and acknowledge its limitations offering workarounds where possible. This paper concludes by identifying avenues which may lead to future work.

II. METHODOLOGY

The fundamental concept proposed by this paper is to eliminate inter-device connections by connecting all user devices to a Central Device (CD). The service provider would then configure the CD to restrict communication between user devices connected to its ports by setting up and severing connections between them. Figure 3 helps visualise this concept. Two user devices would be able to communicate with each other only if the CD permits. By altering the configuration, permission can be granted or denied, therefore dynamic. By delegating a bespoke application to monitor user requirements, script and forward the appropriate configuration commands to the CD, the above process can be automated. Therefore, the end result would be a system capable of automated dynamic topology configuration on an online rack rental platform.

A. *Physical segmentation*

Several candidates were considered for the role of CD. Network segmentation devices (see [2]) considered were bridges, hubs, switches and routers. A bridge normally has two ports. This makes it an ideal medium to interconnect two Local Area Networks, but not so much to interconnect more than two devices. A hub, on the other hand, has several ports, but by nature they forward traffic out on all ports bar the ingress port; this would not be suitable as we need to be able to restrict traffic between devices. A router is characteristically a device used to segment networks. It sections broadcast domains. Each network segment connected to a router would be a subnet on its own and would normally have its own IP addressing scheme. Each port will need to be addressed with an address from their respective subnet address pools. This is not merely a feature of a router, but also a requirement. Two interfaces on a router cannot have addresses on the same subnet. Therefore a routing protocol would need to be employed to facilitate inter-device communication. Invariably this would introduce routing table lookup delays on traffic traversing the CD.

OSI Layer 1, 2 and 3 switches were considered. Physical Layer switches have been used for network testing purposes for a number of years. Commonly referred to as “wire-once infrastructure”, they replace the manual patch panel and allows users to program a connection from any port to any other port within the system using a non-blocking switching matrix [3]. Justification for both initial and recurring investment is the return on investment the device provides. In an industrial test environment physical layer switches offer an array of advanced features conducive to test lab automation [4], which are far beyond the requirements of an online rack rental system designed for students of Computer Networks. For a fraction of the cost, a Layer 2 or Layer 3 switch can be obtained and maintained. On the basis that a switch (a) can offer complete Link layer segregation to (b) a relatively large number (dependent on number of available ports) of connected devices; (c) does not require assignment of an IP addressing scheme and thus (d) offers up to wire-speed data transfer rates through its ASICs without routing table lookup delays, a switch was deemed the most suitable candidate for the role of a CD.

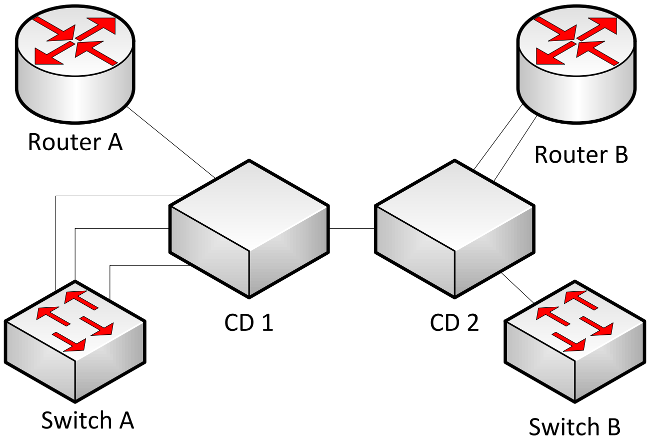


Figure 3. Example setup of proposed solution

User devices are added to the platform by connecting them to the CD. A user device can have multiple connections to the CD. The number of user device – CD connections is only limited by the number of available ports on the CD switch and on a particular user device. However, it is possible to hook up additional switches to the original CD-switch by daisy-chaining them, as shown in Figure 3, to which more user devices could then be connected.

B. *Logical segmentation*

Having physically segmented the network rack, the next step was to evaluate logical segmentation technologies. Virtual Local Area Network (VLAN)s are a commonly used technology in contemporary networks, which allows isolation of ports at Layer 2 and above on the device this technology is configured [5]. Inter-VLAN communication is denied by default. By configuring two ports on a network device to be on the same VLAN, we allow exclusive communication between them.

Layer 2 Protocol Tunneling (L2PT) (not to be confused with Layer 2 Tunneling Protocol) allows Internet Service Providers to carry traffic from multiple customers across their core network while preserving VLAN and other Layer 2 protocol information without impacting other customers; enabling customers to operate a consistent VLAN implementation across a Wide Area Network. L2PT tunnels Layer 2 Protocol Data Units by encapsulating them. Numerous, but not all, Layer 2 protocols can be tunneled. For instance protocols supported by Cisco Systems Inc. and Juniper Networks Inc. are given in [6] and [7], respectively. VLANs are available on both Layer 2 and 3 switches. However, L2PT is not available on Layer 2 switches. Therefore a Layer 3 switch was selected as the most suitable candidate for the role of CD.

The CD should be configured by the service provider so that when for instance, a user requests a specific port on a user device connected to a specific port on the same or another user device, the CD allows communication exclusively between the two ports on the CD(s) to which the respective ports on the user device(s) are connected, while remaining transparent. To remain transparent, the CD must preserve and relay information at Layer 2 and above between the device(s). Our solution will employ L2PT to communicate Layer 2 protocol information between any two ports on the same VLAN. Each port on the CD connected to a user device will have L2TP configured. A unique VLAN ID will be assigned to each pair of ports on the CD to which the user device ports are connected, when communication should be allowed between them. By functioning in unison, VLAN and L2PT protocols render the CD completely transparent to the end user while allowing exclusive communication between the two devices.

III. Implementation

The new model will consist of bespoke front-end and back-end applications to support its delivery. Patrons of

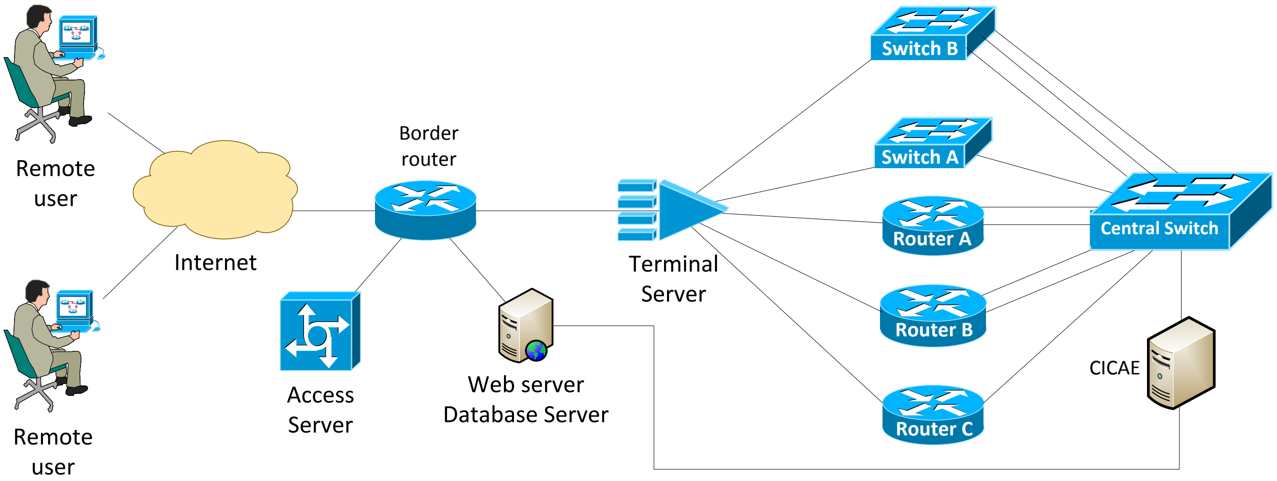


Figure 4. Graphical representation of new online rack rental model

online rack rental systems are familiar with using a web interface to make and manage reservations for rack-time. Therefore, in the interest of user familiarity, we will retain a web portal as the front-end, which will feature the ability to select devices and the topology to be used, allowing end users to build a customized lab of their choice, in addition to making and managing reservations.

To be a scalable model, the back-end is required to have some form of automation for the underlying processes. An application server (identified in Figure 4 as CICAE (Cisco IOS Command Automation Engine)) will be responsible for managing (a) a collection of CDs; (b) a collection of user devices; (c) mappings of user device port to CD port connections; (d) CD port VLAN pairs. The CICAE is not an off-the-shelf application, and was designed and developed by the authors on the Microsoft .NET platform, using the C# programming language. This application interfaces with the database server and web server to keep track of user requests and provide users access to specific devices at specific times, by scripting and issuing commands to the CD. The database server will serve as the repository for user registration/login details and lab reservation information. The terminal server provides console access to remote users. The access server is used to authenticate and authorise remotely connecting users. The border router is the gateway to and from the Internet. Figure 4 is a graphical representation of the complete solution.

The following example gives a detailed illustration of how the new model may be deployed by a service provider. This particular implementation was successfully exhibited by the author for the University dissertation presentation and thus has been practically tested and verified to be a working model. Assume ACME Online Rack Rentals Limited (a fictitious organisation) is a provider of online rack rental solutions who have implemented the new model as shown in Figure 4. After the network rack is setup and the devices connected as shown, details of the user device-CD connections are entered on to the CICAE. Each port on the CD in a connection with a user device is assigned a unique VLAN ID. It should be unique across multiple CDs. The CICAE has been programmed to assign VLAN IDs according to the following algorithm to ensure this.

- VLANs 1 – 9 reserved for administration purposes
- Add 100 for CD 1, 200 for CD 2 and so on
- Add 10 for port Fa0/1, 20 for port Fa0/2 and so on

For instance, a connection on port Fa0/16 on CD 2 would be assigned to VLAN 200 + 160 = 360.

When two ports are to be mapped to each other, the lower of the two VLAN IDs is assigned to the other port, thereby allowing traffic exclusively between the interfaces connected to that pair of ports. When the mapping is no longer required, the VLAN assignment is reset according to the above algorithm.

Two users based geographically distant from each other would like to collaborate to work on a project which involves configuring a networked system. Assume the users (User X and User Y) and their requirements are identical to those in Table 1. Both users would visit ACME's web portal where they would register their details and make a reservation each for the devices they would be working on, specifying the date/time they wish to have access to their labs depending on availability. They have agreed to work on the devices as shown in Figure 5.

The request would be stored on the database. At the

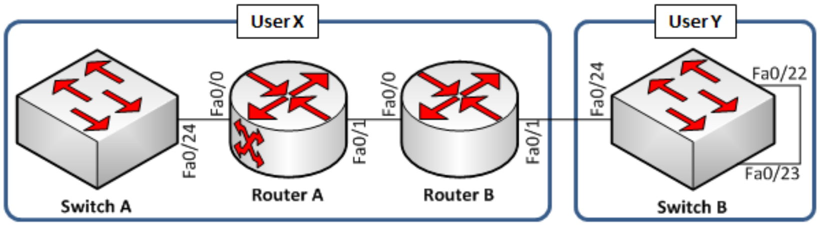


Figure 5. Example setup for collaborative working

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-256-1

date/time the users had requested access, the CICAE application would issue commands to the CD to setup the topology requested. Note that the date/time for each user may or may not be the same if they are working on them separately. However, if they wish to work collaboratively, they would need to agree on a mutually convenient time when reserving the devices.

The CD would be configured so that the ports connecting Switch A to Router A would be in a different VLAN to the ports connecting Router A to Router B and so on. User Y has not requested Switch B to be connected to any other device at this stage; however they have requested a link between two ports on the switch. Unique VLAN IDs would be worked out by the application before the commands are pushed through. Assume the devices are connected to the CD as shown in Table 2 and the given VLAN IDs have been worked out by the application. The following commands configured on a Cisco 3560 switch (used as the CD) would be for the connection between Router A and Router B, and exemplify the commands sent to the CD for all the other connections.

TABLE II
EXAMPLE SETUP OF SERVICE PROVIDER NETWORK RACK

CD port	User device	User device port	VLAN ID
Fa0/1	Switch A	Fa0/24	10
Fa0/2	Router A	Fa0/0	10
Fa0/3	Router A	Fa0/1	30
Fa0/4	Router B	Fa0/0	30
Fa0/5	Router B	Fa0/1	50
Fa0/6	Switch B	Fa0/24	60
Fa0/7	Switch B	Fa0/23	70
Fa0/8	Switch B	Fa0/22	70

```
interface FastEthernet0/3
description CONNECTION TO ROUTER A FA0/1
interface FastEthernet0/4
description CONNECTION TO ROUTER B FA0/0

interface range FastEthernet0/3-0/4

!Assign ports to VLAN 30
switchport access vlan 30
!Establish a tunnel between the ports
switchport mode dot1q-tunnel

!Specify the Layer 2 protocols to be
tunneled
l2protocol-tunnel cdp
l2protocol-tunnel stp
l2protocol-tunnel vtp
```

User X would be granted console access to Switch A, Router A and Router B, while User Y would be granted console access to Switch B once they have been successfully authenticated. Once the users have completed their individual tasks and wish to conjoin their individual labs they indicate their intention to do so to the service provider. Once ACME has received corresponding requests from both users who also indicate, which device and port (note that a single device may be connected to CD via multiple ports) they wish to use to connect to the other user's lab, the ports on the CD to which the two devices are connected are configured with the same VLAN IDs. In this example, port Fa0/6 on the CD would be assigned to VLAN 50, thereby configuring it to be in the same VLAN as Fa0/5 on the CD; thus allowing communication between the two labs. Now User X and User Y are able to network between their individual labs, allowing them to collaborate to complete the project.

C. Adapting an existing environment on to the new model

The proposed model fundamentally suggests how current rack rental systems can be improved to offer a more flexible and cost effective service to end users. We have also looked at how adopting the new model may be in the interest of the service provider. Service providers are able to re-configure labs which they currently offer on to the new model as follows.

Primarily the network rack would need to be rewired to a hub-spoke topology as shown in Figure 4 with any CDs as hubs and user devices as spokes. Depending on the number of user devices available and how many CD-user device connections they wish to offer, additional CD-switches may need to be procured to connect all user devices. Details of the CDs, user devices and how they are connected to the CDs would then need to be added to the CICAE server via a graphical user interface. Most rack rental systems employ a webserver, database server and an AAA server of some flavor, all of which can be retained and reconfigured. The web application required for the new model offers users the ability to select the composition and topology of the lab. The same or a similar interface should be served off the webserver. The database would need to be restructured to represent the various entities such as CD, user device, CD-user device connection, VLAN mapping etc. in addition to user details. The AAA server would require no additional configuration. Existing network configurations such as NAT,

TABLE III.
COST OF HIRING FIXED LABS VS. INDIVIDUAL DEVICES

Service Provider	Avg. cost per session per lab (USD)	No. of devices offered in lab	Avg. cost per device in lab (USD)	Avg. cost for 3 devices (USD)
A	15	16	0.9	2.8
B	17.50	23	0.8	2.4
C	16	13	1.2	3.6

load-balancing, Layer 2 and Layer 3 redundancy, network management and firewalls will require either little or no changes to accommodate the new model.

D. The New Business Model

Current business models are built around a fixed topology lab model, which primarily cater to the likes of CCIE (Cisco Certified Internetwork Expert) aspirants. Therefore, both the caliber and number of devices they offer, as well as the topologies they feature, are aimed at satisfying the requirements of advanced Cisco certification tracks. Naturally, the costs associated with hiring these labs are representative of this. Table 3 gives an indication of prices charged by 3 service providers in the present rack rental market. The pricing model is per session per lab. Sessions typically last from between 4 to 8 hours. The table identifies the cost per session per lab; the total number of devices offered in their CCIE Routing & Switching certification track labs; the average cost per device; and the cost to a user if the service provider were to offer individual devices and the user hired 3 devices. The average costs are rough estimates and does not take into account economics of scale etc., but gives a good indication of how offering the end user the ability to determine the composition of their lab is cost-effective from the user's perspective.

By re-configuring their labs according to the new lab model, a service provider can reengineer their business model around this to offer a more customized solution to a much larger client base. They will be able to offer an array of devices to an advanced user and a single device to a client who requests a standalone device. The service provider may then adapt their pricing model to reflect this, making their offering more attractive to prospective clients. Moreover, they can boast the ability for users to collaborate as part of their service offering. This encourages peer recommendation.

From the perspective of an academic institution, despite the increasing demand on education systems to provide ubiquitous learning resources to support evolving delivery models which cater for internationalisation and distant learning, it may not be feasible to offer separate conventional labs to students individually in the face of increasing budget constraints. However, where an institution wishes to provide its students this facility, by implementing the proposed solution, a tutor is not only able to allocate specific devices between their students to work on remotely from a single lab platform, but can also encourage students to collaboratively work on completing lab assignments. Figure 6 is a visual representation of the proposed business model; a flow diagram showing the various interactions between user, back-end (CICAE, database server, access server and terminal server) and the network rack (CD and user devices).

IV. DISCUSSION & FUTURE WORK

From the outset, we established that an improved solution to online rack renting should meet the following requirements to be considered successful. (a) It was important that intermediate devices should remain transparent to the end user; we have been able to achieve this through the use of L2PT technology. (b) It must not interfere with the normal operation of the user devices; all configurations are done on the CD, and its role is merely restricting traffic through the use of VLANs and relaying traffic between devices intra-VLAN, thus remaining indifferent to the state of the user device or the nature of traffic. (c) The new system needs to be flexible and highly customisable by the end user to suit their individual requirements; this has been achieved by eliminating inter-

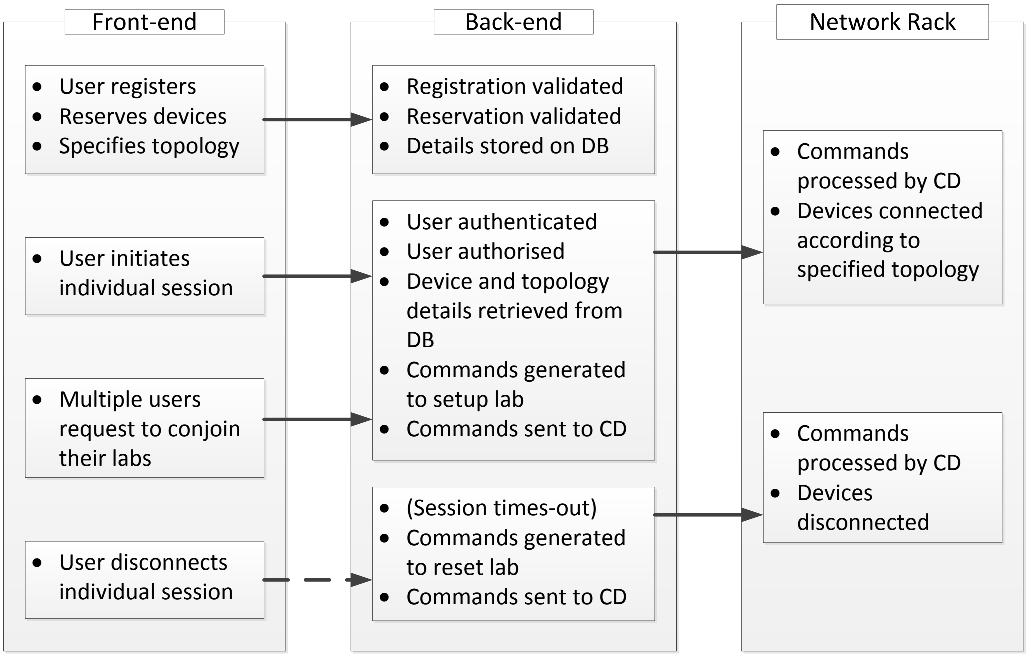


Figure 6. New business model. Dashed arrow indicates optionality; a lab is reset when the user's allocated time comes to an end (session times-out), or when the user initiates a request to terminate their session, whichever precedes.

device connections and introducing a CD in its stead. End users are able to select the composition of their labs and determine its topology as they wish. (d) It should require minimal involvement from the end user to setup and manage; this model retains the web interface users are accustomed with to make and manage reservations for devices and specify the lab topology of their choice. Finally, as we are interested in the academic relevance of this solution, we resolved that (e) it must be able to serve as a collaborative learning tool; using an elaborate example we have looked at how this model achieves this objective. The proposed solution meets all the above requirements.

This model is however not without some limitations. Firstly, the system has at present only been tested on Ethernet, Fast Ethernet, Gigabit Ethernet and Fibre Optic interfaces. Technical limitations dictate that VLANs cannot be configured on serial interfaces. However, it is possible to incorporate serial connections as prewired connections on a hybrid setup. Alternatively, future work could research the possibility of incorporating frame relay switches to which serial links from devices with serial ports could be connected and, by enabling the system to automatically configure frame relay circuits this limitation could potentially be overcome.

Secondly, the CD will normally transition its ports into a “disabled” state in case certain errors are detected on them. Although they can be configured to return to their functional state automatically, this may not eliminate the reason the ports were disabled in the first place.

Thirdly, shutting down an interface on a device which has been mapped to another device does not automatically shut down the interface on the other end of that mapping as the device interface status reflects the status of the port on the CD it is connected to. Physical layer switches we discussed in the Physical Segmentation section overcome this limitation by shutting down the corresponding interface through software intervention. Therefore, it is a capability which could potentially be introduced on to the CICAE, but has not been tested.

A further limitation is that in a solution such as this where a single central device has all peripheral devices connected to it is implemented, a single point of failure is introduced. If the central device fails or becomes compromised, although users will still have remote access to the peripheral devices via the terminal server, they will no longer have connectivity between them.

One of the opportunities we discussed, offered by introducing this model, was the ability for the service provider to have knowledge of which devices on their lab platform will not be used in upcoming sessions. The advantage here is that these devices can be identified and shutdown, saving on operating costs. Technology enabling remotely power-cycling devices has been available for a number of years. Remote Power Management (RPM) solutions offer just that [8]. They are deployed in industry primarily to enable network administrators to recover locked-up devices. There a number of vendors who offer RPM solutions and a majority of them offer the ability to configure via command-line. Therefore, future work could explore the integration of an RPM device on to the current model. The CICAE could be enhanced to script the necessary commands to power-cycle unused devices on the lab platform thus reducing power consumption.

V. Conclusions

This paper introduced a new approach to online rack renting. The improvements suggested in this paper build on the success of prevalent online rack rental systems, which have been rendering an indispensable service to academics engaging in the study of Computer Networks across the globe. By reengineering the way labs are defined and offered, we identified a number of opportunities to add value to both end users and service providers. We discussed how current business models could adapt to accommodate the proposed improvements and the potential opportunities the new model offers to academic institutions by enabling collaborative learning.

Finally, we reflected on some limitations and investigated ways of overcoming these limitations. The authors have identified a number of avenues for future work and encourage and invite interested parties to engage.

References

[1] McGahan, B, bmcgahan@ine.com, 2011. *Dynamic topology rack rental - pros and cons?*. [E-mail] Message to V S M W T Mudiyanselage. Sent Oct 7 2011, 17:49. Available at:https://learningnetwork.cisco.com/thread/35586 [Accessed 21 February 13].

[2] Intel Corporation, (2004) Using Segmentation to Increase Network Performance, http://www.intel.com/network/connectivity/resources/doc_library/white_papers/30514101.pdf [Accessed 23 September 2012].

[3] T. Smith and J. Alnwick, "Wire-once infrastructure for optimal test lab efficiency," AUTOTESTCON, 2008 IEEE , vol., no., pp.421-427, 8-11 Sept. 2008

[4] MRV Optical Communications Systems. 2011. Test Automation : Media Cross Connect™ (MCC) Chassis. [ONLINE] Available at: http://www.mrv.com/product/MRV-MCC-Chass/. [Accessed 07 February]

[5] J. E. Hom and R. Little, (2011). VLAN guide for networking professionals, http://searchnetworking.techtarget.co.uk/tutorial/VLAN-guide-for-networking-professionals,[Accessed 21 February 13]

[6] Cisco Systems, Inc. (2004) Transparent Layer 2 Protocol Tunneling and PDU Filtering. [Online]. Available: http://www.cisco.com/en/US/docs/ios/12_0s/feature/guide/l2pt.html [Accessed 21 February 13]

[7] Juniper Networks, Inc.(2012) layer-2-protocol-tunneling. [Online]. Available:http://www.juniper.net/techpubs/en_US/junos12.1/topics/reference/configuration-statement/layer2-protocol-tunneling-edit-vlans-l2pt-ex-series.html, [Accessed 21 February 13]

[8] Lantroix Inc. (2011) *Remote Power Management: The Key to Maximizing Network and Server Uptime,*California

ISBN: 978-1-61208-256-1

Cisco Academy Support Center and Distance Education Course

Khondkar R. Islam
Department of Applied IT
Volgenau School of Engineering
George Mason University
Fairfax, Virginia, U.S.A.
kislam2@gmu.edu

Louis R. D'Alessandro
Department of Applied IT
Volgenau School of Engineering
George Mason University
Fairfax, Virginia, U.S.A.
ldalessa@gmu.edu

Abstract- **Because of the enormous growth in Information Technology (IT) over the past 20 years, an abundance of job opportunity requirements for a variety of IT specialists in high technology corporations and the U. S. Federal Government is apparent and is continually growing. George Mason University (Mason), with over 33,000 students enrolled located in Northern Virginia, has become the focal point for educating IT professionals to fill this need. Recognizing this opportunity, in 2005, Mason became a Cisco Regional Academy not only because of prestige and recognition, but also to add value to the networking concentration of the undergraduate IT degree of the Department of Applied IT (AIT). Now, the Academy has 45 active Cisco Academies from Virginia, Maryland and Washington, DC areas. This year Cisco shared their upcoming Academy restructuring initiative with Mason, and asked whether the university is receptive to this transition in becoming a Cisco Academy Support Center (ASC) and Instructor Training Center (ITC). Mason has been successful in the role of a Regional Academy, and was positive toward Cisco's offer. This year, Mason has officially become an ASC that will serve 55 Cisco Academies in the State of Virginia and is in the process of becoming an ITC in May 2013. In this paper, we give a background of our partnership with Cisco Networking Academy, describe how we became a successful partner by blending Academy courses with AIT undergraduate degree curriculum and mentoring Local Academies, the transition process involved in becoming an ASC, and the added coverage and challenges Mason faces in the future. We also discuss the role of distance education (DE) on two important data communications courses of the undergraduate IT degree curriculum, because most of the materials of Cisco Certified Network Associate (CCNA) part I and II are taught in these courses.**

Keywords- Cisco; Academy; networking; Mason; curriculum; education; support.

I. INTRODUCTION

There has been tremendous growth in the IT job sector. This is particularly true for the Washington, DC metropolitan area. Mason is located in the high-tech Northern Virginia corridor, and its Department of Applied IT (AIT) is always busy updating its curriculum and courses to meet the challenging and changing demands of the industry, to enable its graduates secure good jobs and move forward with a promising career. This is why the AIT department, during the summer of 2004, agreed to the proposition of the Cisco National Initiative Manager to sponsor a Cisco Regional Academy. Being part of the successful Engineering School, AIT department has to produce graduates with solid technical and hands-on skills to meet the stringent needs of the practical working environment. The department envisioned this to be an effective partnership that would enrich its curriculum and enhance its visibility to the high school, and two-year college transfer technical students. The School Dean saw this as an opportunity and was receptive to this collaboration, and extended his support with School resources, including a new full-time position to coordinate this Academy. An experienced networking professional with a long career in the telecommunications industry joined the AIT faculty to dedicate 50% of his time to create and manage Mason's Cisco Regional Academy. With a seasoned fellow networking instructor, the new instructor began formal instructor certification studies and coordinated the recruitment of Local Academies with the Area Cisco Academy Manager. Mason built a template and foundation for its Regional Academy with support from the Cisco Academy Training Center Manager [1].

At that time, the Bachelor of Science degree in Applied IT had about 900 students with about 200 annual graduates. During Academic Year 2011-2012, the program had about 1,200 students with over 300 graduates. A large percentage of the students select Networking and Telecommunications (NTEL) concentration curriculum, out of the five concentration areas of the degree program. Students are required to complete five courses from a list of courses in the concentration area of their choice. In NTEL, the second of the two networking courses of the degree program, Advanced Networking Principles (IT 445), is on the list. The first networking course is Data Communications and Networking Principles (IT 341), and is a core and required course of the degree program. Mason also has a thriving Cisco Local Academy primarily due to the alignment of these two networking courses with the Academy. At first, students were able to enroll to Mason's Local Academy on a voluntary basis, but that did not achieve positive results because from fall semester of 2007 to the fall semester of 2008, 89 students enrolled in

ISBN: 978-1-61208-256-1

the Cisco networking curricula and only 41 completed the course.

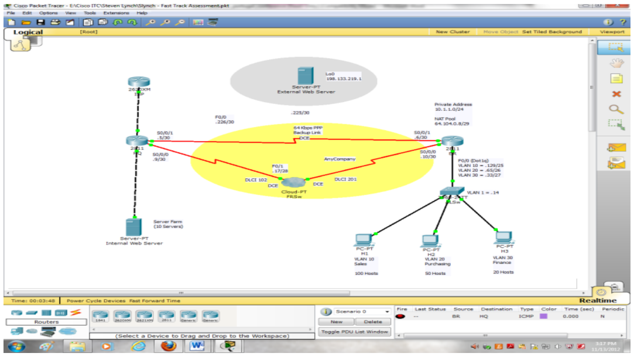
II. PROGRESS

Since becoming a Cisco Regional Academy in 2005, and until becoming an Academy Support Center (ASC) in May 2012, Mason had 16 Local Academies in its jurisdiction that includes Arlington, Loudon, Fairfax, Shenandoah, Warren, Prince William, and Frederick counties in Virginia, Howard University in Washington, DC, and Marymount University in Virginia. There were 14 Local Academies with Mason's Regional Academy until the addition of two new Local Academies in 2012. They are Northern Virginia Community College System's Manassas and Alexandria campuses. The growth in the Regional Academy has been supported by the demand for IT system engineers and design specialists with AS and BS degrees and industry certifications.

It was determined, at first, to offer the Cisco Networking course as a separate Mason Local Academy course where students could voluntarily enroll with the objective of becoming a CCNA [5]. However, due to low voluntary enrollment in the Local Academy because the students were burdened with other AIT courses, we blended considerable material of CCNA part I to the core IT 341 networking course content, and made compulsory enrollment of IT 341 students in Mason's Local Academy Exploration Fundamentals course. This expanded the horizon for the students because they now had access to the powerful virtual network configuration software PacketTracer [2]. Here is a brief background on the reasons for the shift toward the virtual lab configuration exercises. Mason's network lab is equipped with 40 workstations, and kits having two routers and one switch each with all the necessary cables. Four students share each kit. This lets the students build the network from scratch by configuring the routers and switch from the console port, and use straight-through and crossover cables for the interfaces. Due to the unanticipated demand of the BS in IT degree program, enrollment grew rapidly, which prompted the need for several sections of IT 341. Each class section has a capacity of 38 students, and to run simultaneous sections it was a toll on the network equipment. Students of each section had to configure the routers and switches, and save their configurations on a flash drive, because students of the next section would configure the equipment erasing the configuration of the previous class. The wear and tear of the equipment became apparent with the Cisco Internet Operating System (IOS) failing frequently. Unnecessary amount of time was spent to redundantly configure the network at the beginning of each lab, because the students already did the set-up configuration during their first lab session. Further time was spent to dismantle the network, and pack up the kit for the next class. The instructors and teaching assistants were also spending a good amount of time trouble shooting equipment failures and reinstalling the IOS on routers and switches. The following photo of Fig. 1 depicts students configuring a router.

Figure 1: Students busy configuring a router

It was determined Cisco's PacketTracer, a virtual client software application, could replace the use of actual hardware and could address the previously noted problems. The desktop software makes it possible to build complex networks by merely clicking and dragging components onto a desktop configuration as specified in the laboratory manual and then configuring each component using the Cisco's command line interface (CLI) syntax. By making enrollment in the Local Academy Exploration Fundamentals course a part of the curriculum course, students could download and install this virtual network configuration software to work on their lab exercises, after completion of configuration exercises using physical equipment from the kit during the initial three lab sessions. The kits are still in use in the first three lab sessions because they give the students hands-on configuration opportunity in preparation for the real world environment. Fig. 2 depicts a typical PacketTracer desktop.



Fire	Last Status	Source	Destination	Type	Color	Time (sec)	Periodic
●	--	BR	HQ	ICMP	🟪	0.000	N

Figure 2: A virtual network of PacketTracer desktop

TABLE I. STUDENT TRAINING 2007-2008

Course	Semester	Students Enrollment at Start	Students Successfully Completed	Students % Successful	Students Incomplete
CCNA-1	Fall 2007	27	5	18.52	22
CCNA-2	Fall 2007	5	4	80.00	1
CCNA-3	Fall 2007	9	7	77.78	2
CCNA-4	Fall 2007	7	3	42.86	4
CCNA-3	Spring 2008	4	4	100.00	0
CCNA-4	Spring 2008	4	4	100.00	0
Exploration					
Network Fundamentals	Spring 2008	25	6	24.00	19
Routing Protocols & Concepts	Spring 2008	6	6	100.00	0
LAN Switching & Wireless	Spring 2008	1	1	100.00	0
Accessing the WAN	Spring 2008	1	1	100.00	0
Network Fundamentals	Fall 2008	25	2	8.00	23
Totals		**114**	**43**	**37.72 %**	**71**

TABLE II. STUDENT TRAINING 2009-2012

Course	Semester	Students Enrollment at Start	Students Successfully Completed	Students % Successful	Students Incomplete
Exploration					
Network Fundamentals	Spring 2009	76	73	96.05	3
Routing Protocols & Concepts	Spring 2009	6	6	100.00	0
LAN Switching & Wireless	Summer 2009	1	1	100.00	0
Network Fundamentals	Summer 2009	19	19	100.00	0
Network Fundamentals	Fall 2009	85	82	96.47	3
Network Fundamentals	Spring 2010	100	93	93.00	7
Network Fundamentals	Fall 2010	113	110	97.35	3
Routing Protocols & Concepts	Fall 2010	13	10	76.92	3
Network Fundamentals	Spring 2011	120	120	100.00	0
Network Fundamentals	Summer 2011	25	25	100.00	0
Network Fundamentals	Fall 2011	120	120	100.00	0
Routing Protocols & Concepts	Fall 2011	19	19	100.00	0
LAN Switching & Wireless	Fall 2011	10	10	100.00	0
Network Fundamentals	Spring 2012	100	100	0.00	0
LAN Switching & Wireless	Spring 2012	4	0	0.00	4
WAN	Spring 2012	4	0	0.00	4
Routing Protocols & Concepts	Spring 2012	10	4	40.00	6
Network Fundamentals	Fall 2012	120	120	100.00	0
Totals		**945**	**912**	**96.5%**	**33**

Table 1 data makes it evident that voluntary student enrollment was low. Refer to Table 2 for the significant jump in enrollment numbers once this was made mandatory. The course completion success rate more than doubled, which is an indication that the students took the course seriously. This is because with students completing the Exploration Fundamentals segment [3], the Local Academy course completion became almost certain. The effect was quite similar with instructor training, because there was a direct relationship between student enrollment and number of instructors that needed training. After the inception of the Regional Academy, many high school visits were made to meet the instructors of the Local Academies, and it was discovered that there was a need to train new instructors because many of the existing educators left or were leaving to work for the industry.

During those visits, we noticed the lab equipment in the Local Academies were dated and needed replacement soon. The visits to the Local Academies helped Mason's Regional Academy [4] initiate a Local Academy Instructors Training Program.

TABLE III. INSTRUCTOR TRAINING 2007-2008

Course	Semester	Instructors Enrollment
Orientation for Instructors	Fall 2007 to Spring 2008	7
CCNA-1	Spring 2007	3
CCNA-1	Fall 2007	2
CCNA-1	Fall 2007	1
CCNA-1	Spring 2008	1
Exploration		
Network Fundamentals	Fall 2007	2
Routing Protocols & Concepts	Spring 2008	1
LAN Switching & Wireless	Spring 2008	2
Accessing the WAN	Spring 2008	2
Discovery		
Networking for Home and Small Business	Fall 2007	2
Networking at a Small-to-Medium or ISP	Spring 2008	2
Introducing Routing & Switching in the Enterprise	Fall 2008	2
Designing and Supporting Computer Networks	Fall 2008	0
Totals		**27**

TABLE IV. INSTRUCTOR TRAINING 2009-2012

Course	Semester	Instructors Enrollment at Start	Instructors Successfully Completed	Instructors % Successful	Instructors Incomplete
Fast Track					
IT 341	Spring 2009	4	4	100.00	0
IT 341	Spring 2010	5	5	100.00	0
IT 341	Spring 2011	2	2	100.00	0
IT 341	Summer 2011	2	2	100.00	0
IT 341	Summer 2012	2	2	100.00	0
Marymount	Summer 2011	1	1	100.00	0
Exploration					
Network Fundamentals	Fall 2007	5	5	100.00	0
Routing Protocols & Concepts	2008-2010	3	3	100.00	0
LAN Switching & Wireless	2008-2010	3	3	100.00	0
Accessing the WAN	2008-2010	3	3	100.00	0
Discovery					
Designing & Supporting Computer Networks	Fall 2008	1	0	0.00	1
Totals		**31**	**30**	**96.77%**	**1**

The Training Program also made recommendations to the school administrators for upgrading their lab equipment. Table 3 presents the data of Instructor Training Program for the spring 2007 to fall 2008 period. Table 4 results show there was a rise in the number of instructor training with increased enrollments, because it was necessary to have more qualified instructors at Mason to teach the Academy Program. To accomplish this, we set up Fast Track instructor courses for additional instructors and teaching assistants for the Regional Academy, and graduated all to support our IT 341 and IT 445 courses since the spring semester of 2009. CCNA certification adds tremendous value to our graduates of AIT degree with NTEL concentration.

It would be overwhelming for the students if IT 341 alone would cover most of the content of CCNA part I and II. This prompted us to develop the second networking course, which is better known as the Advanced Networking Principles (IT 445) course. This covers most of the lecture and lab materials of CCNA part II, and it is not a required class like IT 341 for the AIT majors, because this is one of the NTEL concentration courses that the students have the option to take. The students who have interest with the networking career enroll in IT 445 because their goal is to become CCNA certified. Enrollment of IT 445 has been gradually growing since it was first offered in fall 2007. Due to popular demand, we have offered a second section of IT 445 in spring 2013, which is a distance education (DE) course.

III. DISTANCE EDUCATION (DE)

There are two general categories of DE delivery methods: 1) *Asynchronous* and 2) *Synchronous*. With asynchronous, some instructors choose to record lectures that are stored in a server or prepare lessons as web pages. Students access the server at their convenience to retrieve the lectures. Home assignments, exams and other class materials are also uploaded to the server. Synchronous distance learning is similar to in-class sessions. This is because students attend online classes during the class time. They participate in lectures, view slide presentations and interact with the instructor and other students via the Internet. This creates an environment where the students feel they are attending a live classroom without having to actually go to a classroom. It is worth noting, video streaming is generally not mandatory since synchronous video with DE delivery has several tradeoffs and challenges [12]. High capacity network services are required for reliable video stream [13]. Further, audio and video are sometimes not synchronized which lead to confusion since lip movement and audio being heard is not always the same. Also, low video resolution that is required to conserve network capacity and small display screens of Learning Management Systems (LMS) and Synchronous Distance Education Tools (SDET) do not show clear view of facial expressions that enable better understanding, which is the main argument for video in the first place [14].

Research shows that problems arise when students do not get the opportunity to interact with the instructor and other students while they are in an asynchronous learning environment. Some students are confused about the assignments and course objectives, and feel frustrated and isolated. On the other hand, despite the challenges associated with synchronous education, it approximates face-to-face dialog and promotes a sense of community. Overall student outcomes also are better with synchronous education over asynchronous learning. This is because students are motivated since synchronous education makes the courses more engaging [12]. Characteristics of synchronous and asynchronous DE delivery are presented in Table 5.

DE LMS and SDET must offer a user-friendly graphical user interface, simple navigation options, and have enhanced security features to deter unauthorized access to the system and files. Course creation and management has to be easy, and the system must support common file types. There has to be an option to reuse course contents so instructors are able to reuse contents in other sections of the same course or during another semester with minor modifications.

TABLE V. SYNCHRONOUS AND ASYNCHRONOUS DE DELIVERY [15]

Characteristics	Synchronous DE	Asynchronous DE
Positive	Increases psychological arousal	Increases cognitive participation
Negative	Does not increase cognitive participation	Personal participation is low

Early research suggests web users need to be provided with an effective usable environment because it drives substantial savings and achieves better performance. In academia, effective LMS and SDET need little instructor time to set up and manage the course, improving the learning experience of students. It is important for the LMS and SDET to be not cluttered with too many appealing design options as that may integrate with features in course design, which can be confusing for students and the instructor. Only features that meet course objectives and are relevant to a sound-learning environment for designing an effective course should be included in the LMS and SDET. Since usability is critical, the LMS and SDET must be easy-to-use and learn, and offer options that are easy-to-remember. Web usability requires having web pages that are easy-to-navigate and display information in an organized manner so users do not have to struggle to find what they are looking for. Pedagogical usability ensures users learn effectively and retain the skills and knowledge, and is integrated with technology usability, which is referred to ease-of-use and usefulness of the technology [16]. Students do not have a high degree of pedagogical usability when technology usability is poor.

To comply with the directives of leadership, all Colleges, Schools and Departments started offering at least one DE section of the live in-class sections of a course. We were in the forefront in implementing this initiative by offering two asynchronous DE sections of IT 341 alongside two live in-class sections. We hope to add synchronous lectures to these courses in the upcoming semesters. The Cisco Academy website [3] has been supplemental to the DE and in-class live sections, where students take online Exploration Fundamentals segment exams that are graded on a real-time basis and recorded in Blackboard (Bb) [6] *gradebook*. Bb is widely used as a LMS and SDET by many course sections university-wide as we do in our AIT department. The home assignments and lab exercises are submitted online in Bb. Camtasia Studio [7] is used to video the lecture and lab session recordings, which are posted in Bb for students to view during their time of convenience. The *Discussion Board* of Bb is heavily used to make the DE class interactive. All lecture and lab assignment, and exam release and due dates are announced via the *Announcement* feature of Bb, and also communicated via Mason email system with the students. At present, students are required to come to campus for the midterm and final examinations, but we will implement online exams for the DE sections in fall 2013.

IV. TRANSITION TO ACADEMY SUPPORT CENTER (ASC)

This section covers a brief background of why Mason decided to become an ASC, and its roles and responsibilities as an ASC. It was envisioned by becoming an ASC, Mason would support 55 current Cisco Local Academies in the state of Virginia. As a Regional Academy, we were supporting only 16 Local Academies, now named Cisco Academies as changed by the Evolution Program. Our university is also a partner in the 4-VA Initiative that was initiated by our president and the presidents of James Madison University, University of Virginia, and Virginia Tech University. 4-VA was established in 2011 in response to the Governors' Higher Education Commission recommendations to find methods to collaborate to meet for higher quality and affordable education focusing on the Science, Technology, Engineering, and Mathematics (STEM) programs. Cisco Systems, Inc. is a solid partner in this program by providing its TelePresence Systems at many sites on the four university campuses. TelePresence will be used to achieve the goals to improve communications efficiency. This will enhance student success, sharing the delivery of course strategies to improve sharing course strategies to Virginia's economic development, and increasing each university's research competiveness. During the transition period in becoming an ASC, we realized distance education at our proposed ASC will be improved by reaching out to the Cisco Academies' audio-visual facilities utilizing our TelePresence facilities where our agendas will include Academy teaching and technology updates, conduct seminars, share student's success stories and course experiences. If some Academies do not have audio-visual facilities, we planned to use WebEx [8] communications in a point-to-multipoint configuration. In May of 2012, we officially became a Cisco ASC, and to date we have 45 Cisco Academies throughout Virginia.

As a member of an ASC, Mason specializes and excels in preparing and enhancing the success and sustainability of the Network Academies in the Commonwealth of Virginia. Our efforts have a positive effect on academy administrators, instructors, and students. We provide essential operational support to academies in a relevant format. Localized operational support is essential throughout an Academy's engagement starting with onboarding and throughout their lifecycle.

Our ASC provides Cisco Academies in a number of ways, including but not limited to the following major services:

- In-person visits for lectures, consultation and/or support.
- Remote consultation, troubleshooting and monitoring via telephone, email, and/or other technology.
- Access to training using TelePresence, webinars, and/or presentations.
- Access to ASC Information Portal.
- Invitation to an Annual Meeting to sharpen teaching skills and disseminate new customer programs.
- Support with continuation of Cisco Membership Agreement and the responsibilities described in the Agreement.

There are two required roles at an ASC: 1) Academy Support Center Contact (one required); and (2) Support Advisor (two required). A person can be both an Academy Support Center Contact and a Support Advisor, or two separate individuals can fulfill these roles. These ASC roles primarily interact with Academy Contacts and Academy Success Leads at Cisco Academies. They also interact with the following Cisco roles:

- Area Academy Manager (AAM)
- Cisco Quality Manager (CQM)
- Global Support Desk CSR

The ASC Contact is responsible for managing the annual membership. This is achieved by securing appropriate institution administrator to sign the online ASC membership. The individual is also responsible for updating the ASC profile, and ensure compliance with Cisco policies and minimum standards as outlined in the *Membership Guide*. Other roles are to develop the *Annual ASC Plan*, document any support focus areas, review feedback from Academies Mason supports, address plans to improve any unsatisfactory performance areas, and ensure services and support are marketed/advertised using technologies provided. Academies require support throughout their lifecycle. The type of support varies depending on the maturity of the Academy. The Support Advisors are knowledgeable about all areas of Academy operation. They are the channels that Cisco uses to ensure critical operational messages are received and understood by the Academies. Cisco provides ongoing educational opportunities for Support Advisors to ensure ASCs have the information they need to be successful.

As a result of the Cisco Academy Evolution Program, Mason's former Regional Academy responsibilities are now being undertaken in our new status as an ASC. To become an ASC we were required to submit our application along with a Business Plan to Cisco for review. Approval was granted in May of 2012. In addition, Mason will become an Instructor Training Center (ITC) in May of 2013. This is important because we have a large number of students enrolled in the networking courses that require several teaching assistants (TAs) to assist the professors in the delivery of the courses. It would facilitate the process of training and proctoring exams for the TAs and our local Cisco academy instructors in becoming Cisco Networking Academy instructors on an on-going basis without having to depend on other ITCs. We plan to have three Instructors take Cisco's new rigorous Instructor Training Program. Instructor trainers are required to take a *pre-test* and two days of professional *in-person* training.

V. RESEARCH

Since the inception of becoming a Regional Academy in 2005 we had a major obstactle in offering the Cisco CCNA courses into an academic university environment. Our first approach was to offer the Cisco courses as they stand to our academic accreditation committee. This effort was rejected based on the premise that the courses were of an apprentice hands-on experience more likely to be offered at a technical school level. To amiliorate this objection, we researched The Association of Technology, Management, and Applied Engineering [11] to compare the CCNA course material to their standards and found that the Cisco courses matched the standard IT components. With this knowledge, we decided to insert components of the Cisco courses into our IT341 and IT445 courses and academically enhance each by inserting essay exercises to ensure that the students were understanding the concepts.

Still to be researched, is the effect of the Evolution Program organizational structure change and the transition from the Academy Connection to NetSpace. Our initial observations and research, not documented, found that academies had the usual objections to change. The former Cisco Academy Training Center (CATC), Regional, and Local Academy structure appeared to be quite adequate in delivering the Cisco course material. The CATCs that were formerly finacially supported by Cisco and Regional Centers found that they needed either to convert to an Instructor Training Center (ITC) charging a fee for training courses and/or an Academy Support Center (ASC) charging an annual support fee to Local Academies. Complicating this issue, is the competitive aspect of former Regional Centers that converted to Academy Support Centers (ASC) are now in competition with each other to enroll academies to capture the annual fees. Each ASC is now charging fees as a business center to cover their costs.. Almost all Local Academies, now renamed Cisco Academies, never paid for support before and found this financially onerous. In addition, this expenditure was not forecast in their annual 2012 operational school budgets. This is resulting in the school system's administrations' seeking an ASC with the lowest fee.

The change from the Academy Connection to NetSpace concurrent with the Evolution Program is now being undertaken. Complications are prevalent in that both the Academy Connection and new NetSpace learning sites are operational at the same time with a phase out plan for the Academy Connection by mid year 2013. Instructors are now offering courses in both systems and learning how to use NetSpace.

The intended research to be accomplished will determine if these changes prove to enhance IT knowledge, and increase the number of students achieving Cisco Certifications resulting in successful exceptional challenging occupations and career promotional opportunities for them. The methodology to perform this research may use surveys, interviews, and a compilation of data and statistics.

VI. CONCLUSION

In summary, our graduates receiving the BS in Applied Information Technology benefit from the School's adaptation to include most of the content of the Cisco Exploration course in our IT 341 and IT 445 courses. To satisfy the academic requirements of a university, our students are required to provide twelve technical essays for these two courses. Our students also receive strong courses in the study of IT wireless and Internet Protocol (IP) telephony. We are considering adding Cisco Certified Network Professional (CCNP) and network security content to our Master of Science degree program networking and security courses in the near future. A student who has completed the NTEL concentration can take the CCNA examination at our certification testing center. To promote certification, our School is a Pearson View Certification Testing Center where our students and faculty can take certification examinations at a much reduced fee. The ultimate goal of Mason is to prepare our students for careers in this most abundant IT job opportunity domain of Metropolitan Washington, DC. For an example of Virginia IT occupational opportunities refer to Table 6 [9, 10].

Table VI: VIRGINIA OCCUPATION PROJECTIONS [9, 10]

Virginia Occupation Projections	Employment		Average Annual Openings	Occupational Employment as of May[2] 2009
	2008	2018		
Computer Support Specialists	19,115	23,302	948	18,840
Computer Systems Analysts	36,518	47,978	1,933	35,030
Network and Computer Systems Administrators	18,407	25,626	1,029	18,460
Network Systems and Data Communications Analysts	16,981	28,651	1,472	13,650
Computer and Information Systems Managers	12,726	16,179	552	12,320

REFERENCES

[1] L. D'Alessandro and D. Gantz, "Combining academic studies with IT certifications: Becoming a Cisco regional academy," in *Proceedings of the 10th ACM conference on SIG-information technology education*, Fairfax, Virginia, 2009, pp. 182–188

[2] *Cisco Systems* | Cisco Packet Tracer, (accessed November, 2012). http://www.cisco.com/web/learning/netacad/course_catalog/PacketTracer.html.

[3] *Cisco Networking Academy* | Academy Connection Instructor Home, (accessed November, 2012). https://cisco.netacad.net/cnams/dispatch.

[4] *Cisco Academy at Mason* | Department of Applied Information Technology, (accessed November 2012). https://ait.gmu.edu/cisco/index.htm.

[5] *Cisco Career Certifications & Paths* | CCNA, (accessed November, 2012). http://www.cisco.com/web/learning/le3/le2/le0/le9/learning_certification_type_home.html.

[6] *Blackboard* | Education Technology, (accessed November, 2012). http://www.blackboard.com.

[7] *Camtasia Studio* | Screen Recording and Video Editing, (accessed November, 2012). http://www.techsmith.com/camtasia.html.

[8] *Cisco WebEx* | Web Conferencing and Collaboration, (accessed November, 2012). http://www.webex.com.

[9] *Virginia Workforce Connection* | VWC, (accessed February 2, 2013). http://www.vawc.virginia.gov/analyzer/default.asp.

[10] *Bureau of Labor Statistics, May 2009 State Occupational Employment and Wage Estimates* | Department of Labor, (accessed January 29, 2013). http://stat.bls.gov/oes/current/oessrcst.htm.

[11] *The Association of Technology, Management, and Applied Engineering*, (accessed January 30, 2013). http://www.atmae.org/index.php/accreditation-10.

[12] S. Smith, "Examining the impact of synchronous video on distance education delivery and outcomes,". Rensselaer Polytechnic Institute, 2004.

[13] M. Hentea, M. J. Shea, and L. Pennington, "A perspective on fulfilling the expectations of distance education," in *Proceedings of the 4th conference on Information Technology Curriculum*, Lafayette, Indiana, 2003, pp. 160–167.

[14] R. Anderson, T. VanDeGrift, and F. Videon, "Videoconferencing and presentation support for synchronous distance learning," *presented at the 33rd ASEE/IEEE Frontiers in Education Conference*, Boulder, Colorado, 2003.

[15] S. Hrastinski, "Asynchronous and synchronous distance learning," *Educause Quarterly*, vol. 31, no. 4, pp. 51–55, 2008.

[16] Z. Unal and A. Unal, "Evaluating and comparing the usability of web-based course management systems," *Journal of Information Technology Education*, vol. 10, pp. 19–38, 2011.

Technical Points in the Implementation of the Support System for Operation and Management of DACS System

Kazuya Odagiri
Yamaguchi University
Yamaguchi, Japan
odagiri@yamaguchi-u.ac.jp
kazuodagiri@yahoo.co.jp

Shogo Shimizu
Gakushuin Women's College
Tokyo, Japan
shogo.shimizu@gakushuin.ac.jp

Naohiro Ishii
Aichi Institute of Technology
Aichi, Japan
nishii@acm.org

***Abstract*—As the work for managing a whole LAN effectively without limited purposes, there are works of Policy-based network management (PBNM). The existing PBNM is defined in some organizations including the Internet Engineering Task Force (IETF). However, it has structural problems. For example, it is necessary to add and exchange the mechanism called the PEP located between network servers and clients. That is, it is needed to exchange the network system configuration. To improve the problems, we have been studying next generation PBNM called Destination Addressing Control system (DACS) Scheme. The DACS Scheme controls the whole LAN through communication control by the client software as PEP which locates on a client computer. We have been directly studied the essential part to realize the DACS Scheme. That is, we have not been examined the support system for operation and management of DACS system. In this paper, technical points in the implementation of the support system are examined.**

Keywords-policy-based network manageme; support system.

I. INTRODUCTION

In enterprise computer networks, because network policies and security policies are well defined and are observed forcibly, network management is relatively easy. On the other hand, in campus-like computer networks, network management is quite complicated. Because a computer management section manages only a part of the campus network, there are some user support problems. For example, when mail boxes on one server are divided and relocated to some different servers by a system change, it is necessary for some users to update client's setups. Most of users in campus computer networks are students. Because students do not check frequently their e-mail, it is hard work to make all students aware of necessity of settings update. As the result, because some users inquire for the cause that they cannot connect to a mail server, a system administrator must cope with it. For the system administrator, individual user support is a stiff part of the network management.

As the works on network management, various kind of works such as the server load distribution technology [1][2][3], VPN (Virtual Private Network) [4][5] are listed. However, these works are performed forward the different goal, and don't have the purpose of effective management for a whole LAN. As the work for managing the whole network, works on Opengate [6][7] are listed. This is a kind of Policy-based network management (PBNM). Frameworks of PBNM are defined in various organizations such as Internet Engineering Task Force (IETF) and Distributed Management Task Force (DMTF). However, the PBNM has some structural problems. First problem is communication concentration on a communication control mechanism called PEP (Policy Enforcement Point). Second problem is the necessity of the network updating at the time of introducing the PBNM into LAN. Moreover, third problem is that it is often difficult for the PBNM to improve the user support problems in campus-like computer networks explained above.

To improve these problems of the PBNM, we showed a next generation PBNM. We call it Destination Addressing Control system (DACS) Scheme. As the works of DACS Scheme, we showed the basic principle of the DACS Scheme 20], and security function [21]. In addition, we showed new user support realized by use of the DACS Scheme [22]. Then, the DACS system to realize the DACS Scheme was implemented [23]. We have been directly studied the essential part to realize the DACS Scheme, and have not been examined the support system for operation and management of DACS system. In this paper, technical points in the implementation of the support system are examined.

The rest of paper is organized as follows. Section II shows past works of the network management including the existing PBNM. In Section III, we describe the mechanisms and effectiveness of the DACS scheme. In Section IV, technical points in the implementation of the support system for operation and management of DACS system are shown.

II. MOTIVATION AND RELATED WORKS

As the works on existing network management, various works such as authentication [24][25], the server load distribution technology [1][2][3], VPN [4][5] and quarantine network [26] are listed. However, these works are performed forward the different goal. Realization of effective management for a whole LAN is not a purpose. These works are performed for the specific purpose, and don't have the purpose of managing a whole LAN. As the work for managing a whole LAN, there is the work of Opengate [6][7], which controls Web accesses from LAN to internet. This work is a kind of PBNM. In PBNM, the whole LAN is managed through various kinds of communication controls such as access control and Quality of Service (QOS) control, communication encryption. The principle of PBNM is described in Figure 1. To be concrete, in the point called PDP (Policy Decision Point), judgment such as permission and non-permission for communication pass is performed based on policy information. The judgment is notified and transmitted to the point called the PEP which is the mechanism such as VPN mechanism, router and firewall located on the network path between servers and clients. Based on that judgment, the control is added for the communication that is going to pass by.

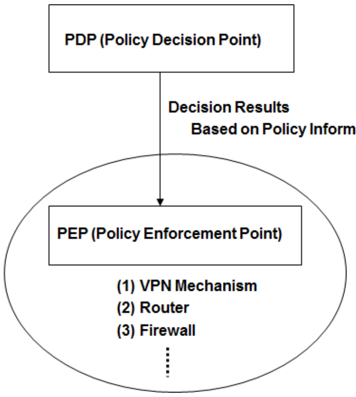

PDP (Policy Decision Point)
Decision Results
Based on Policy Information
PEP (Policy Enforcement Point)
(1) VPN Mechanism
(2) Router
(3) Firewall

Figure 1. PBNM in IETF

The PBNM's standardization is performed in various organizations. In IETF, a framework of PBNM [8] was established. Standards about each element constituting this framework are as follows. As a model of control information stored in the server storing control information called Policy Repository, Policy Core Information model (PCIM) [9] was established. After it, PCMIe [10] was established by extending the PCIM. To describing them in the form of Lightweight Directory Access Protocol (LDAP), Policy Core LDAP Schema (PCLS) [11] was established. As a protocol to distribute the control information stored in Policy Repository or decision result from the PDP to the PEP, Common Open Policy Service (COPS) [12] was established. Based on the difference in distribution method, COPS usage for RSVP (COPS-RSVP) [13] and COPS usage for Provisioning (COPS-PR) [14] were established. RSVP is an abbreviation for Resource Reservation Protocol. The COPS-RSVP is the method as follows. After the PEP having detected the communication from a user or a client application, the PDP makes a judgmental decision for it. The decision is sent and applied to the PEP, and the PEP adds the control to it. The COPS-PR is the method of distributing the control information or decision result to the PEP before accepting the communication.

Next, in DMTF, a framework of PBNM called Directory-enabled Network (DEN) was established. Like the IETF framework, control information is stored in the server storing control information called Policy Server which is built by using the directory service such as LDAP [15], and is distributed to network servers and networking equipment such as switch and router. As the result, the whole LAN is managed. The model of control information used in DEN is called Common Information Model (CIM), the schema of the CIM（CIM Schema Version 2.30.0）[17] was opened. The CIM was extended to support the DEN [16], and was incorporated in the framework of DEN.

In addition, Resource and Admission Control Subsystem (RACS) [18] was established in Telecoms and Internet converged Services and protocols for Advanced Network (TISPAN) of European Telecommunications Standards Institute (ETSI), and Resource and Admission Control Functions (RACF) [19] was established in International Telecommunication Union Telecommunication Standardization Sector (ITU-T).

However, all the frameworks explained above are based on the principle shown in Figure 1. As problems of these frameworks, two points are presented as follows.

(1) Communications sent from many clients are controlled by the PEP located on the network path. Processing load on the PEP becomes very heavy.
(2) The PEP needs to be located between network servers and clients. Depending on the network system configuration, updating for adding the PEP is needed.

To improve these problems of the PBNM, we have been proposed a next generation PBNM called the DACS Scheme. However, the DACS Scheme has some troublesome points in doing operation and management practically. The points are described as follows.

(Point 1)
Because the tool which easily performs registration and deletion of DACS rules does not exist, it is necessary for the

network administrator to register it with a database using an SQL language directly.

(Point 2)

When the network administrator grasps the IP address of the client which the user who logged in uses, it is necessary to refer to the table with the information in the direct database.

(Point3)

The tool to easily send a message to the client-side does not exist. Though the possibility of sending the message was confirmed by the functional experiment [22], the tool which each network administrator can use was not implemented.

III. Existing DACS Scheme

A. Basic Principle of the DACS Scheme

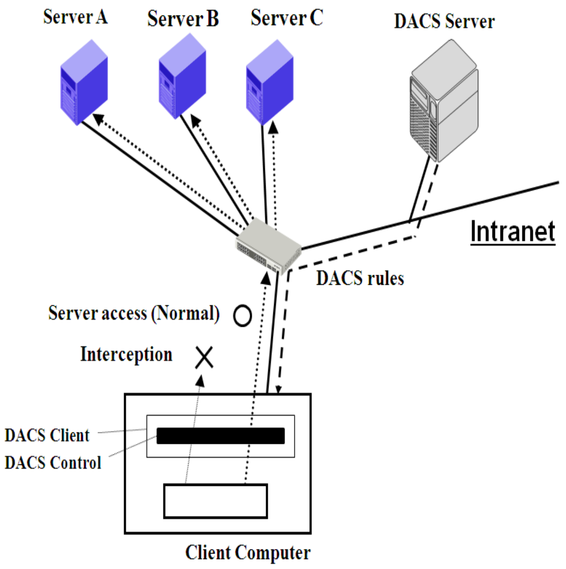


Figure 2. Basic Principle of the DACS Scheme

Figure 2 shows the basic principle of the network services by the DACS Scheme. At the timing of the (a) or (b) as shown in the following, the DACS rules (rules defined by the user unit) are distributed from the DACS Server to the DACS Client.

(a) At the time of a user logging in the client.

(b) At the time of a delivery indication from the system administrator.

According to the distributed DACS rules, the DACS Client performs (1) or (2) operation as shown in the following. Then, communication control of the client is performed for every login user.

(1) Destination information on IP Packet, which is sent from application program, is changed.

(2) IP Packet from the client, which is sent from the application program to the outside of the client, is blocked.

An example of the case (1) is shown in Figure 2. In Figure 2, the system administrator can distribute a communication of the login user to the specified server among servers A, B or C. Moreover, the case (2) is described. For example, when the system administrator wants to forbid an user to use MUA (Mail User Agent), it will be performed by blocking IP Packet with the specific destination information.

In order to realize the DACS Scheme, the operation is done by a DACS Protocol as shown in Figure 3. As shown by (1) in Figure 3, the distribution of the DACS rules is performed on communication between the DACS Server and the DACS Client, which is arranged at the application layer. The application of the DACS rules to the DACS Control is shown by (2) in Figure 3. The steady communication control, such as a modification of the destination information or the communication blocking is performed at the network layer as shown by (3) in Figure 3.

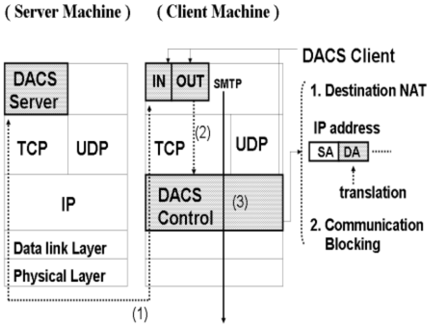


Figure 3. Layer Setting of the DACS Scheme

B. Communication Control on Client

The communication control on every user was given. However, it may be better to perform communication control on every client instead of every user. For example, it is the case where many and unspecified users use a computer room, which is controlled. In this section, the method of communication control on every client is described, and the coexistence method with the communication control on every user is considered.

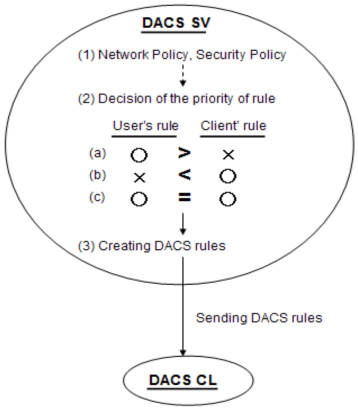


Figure 4. Creating the DACS rules on the DACS Server

When a user logs in to a client, the IP address of the client is transmitted to the DACS Server from the DACS Client. Then, if the DACS rules corresponding to IP address,

ISBN: 978-1-61208-256-1

is registered into the DACS Server side, it is transmitted to the DACS Client. Then, communication control for every client can be realized by applying to the DACS Control. In this case, it is a premise that a client uses a fixed IP address. However, when using DHCP service, it is possible to carry out the same control to all the clients linked to the whole network or its subnetwork for example.

When using communication control on every user and every client, communication control may conflict. In that case, a priority needs to be given. The judgment is performed in the DACS Server side as shown in Figure 4. Although not necessarily stipulated, the network policy or security policy exists in the organization such as a university (1). The priority is decided according to the policy (2). In (a), priority is given for the user's rule to control communication by the user unit. In (b), priority is given for the client's rule to control communication by the client unit. In (c), the user's rule is the same as the client's rule. As the result of comparing the conflict rules, one rule is determined respectively. Those rules and other rules not overlapping are gathered, and the DACS rules are created (3). The DACS rules are transmitted to the DACS Client. In the DACS Client side, the DACS rules are applied to the DACS Control. The difference between the user's rule and the client's rule is not distinguished.

C. Security Mechanism of the DACS Scheme

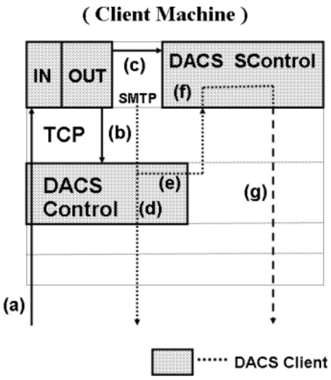


Figure 5. Extend Security Function

In Figure 5, the DACS rules are sent from the DACS Server to the DACS Client (a). By the DACS Client that accepts the DACS rules, the DACS rules are applied to the DACS Control in the DACS Client (b). The movement to here is same as the existing DACS Scheme. After functional extension, as shown in (c) of Figure 5 the DACS rules are applied to the DACS SControl. Communication control is performed in the DACS SControl with the function of SSH. By adding the extended function, selecting the tunneled and encrypted or not tunneled and encrypted communication is done for each network service. When communication is not tunneled and encrypted, communication control is performed by the DACS Control as shown in (d) of Figure 5. When communication is tunneled and encrypted, destination of the communication is changed by the DACS Control to localhost as shown in (e) of Figure 5. After that, by the DACS STCL, the communicating server is changed to the network server and tunneled and encrypted communication is sent as shown in (g) of Figure 5, which are realized by the function of port forwarding of SSH. In the DACS rules applied to the DACS Control, localhost is indicated as the destination of communication.

D. Specifications of DACS System

Technical points for implementation of the DACS Scheme are described form (a) to (c).

(a) Communications between the DACS Server and the DACS Client

The Communications between the DACS Server and the DACS Client such as sending and accepting the DACS rules were realized by the communications through a socket in TCP/IP.

(b) Communication control on the client computer

In this study, the DACS Client working on windows XP was implemented. The functions of the destination NAT and packet filtering required as a part of the DACS Control were implemented by using Winsock2 SPI of Microsoft. As it is described in Figure 6 Winsock2 SPI is a new layer which is created between the existing Winsock API and the layer under it.

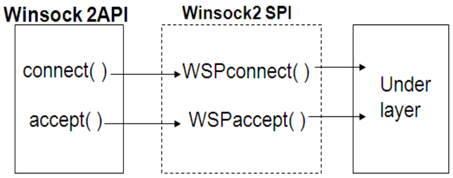


Figure 6. Winsock2 SPI

To be concrete, though connect() is performed when the client application accesses the server, the processes of destination NAT for the communication from the client application are built in WSP connect() which is called in connect(). In addition, though accept() is performed on the client when the communication to the client is accepted, the function of packet filtering is implemented in WSPaccept() which is called in accept().

(c) VPN communication

The client software for the VPN communication, that is, the DACS SControl was realized by using the port forward function of the Putty. When the communication from the client is supported by the VPN communication, first, the destination of this communication is changed to the localhost. After that, the putty accepts the communication, and sends the VPN communication by using the port forward function.

IV. Technical Points in the Implementation of Support System for Operation and Management

In this section, to overcome three troublesome points, the support system for operation and management in DACS system is shown. The functions the support system must have are as follows.

(1) Function the network administrator register, change and delete the DACS rules to DACS Server.

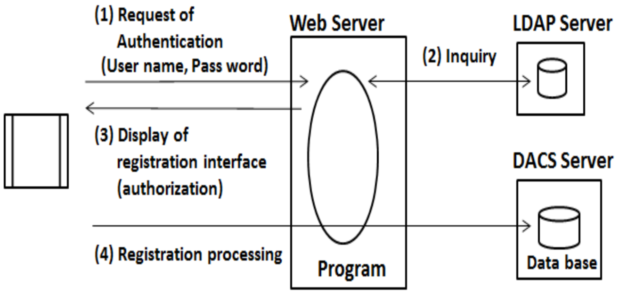

(1) Request of Authentication (User name, Pass word)
Web Server
LDAP Server
(2) Inquiry
(3) Display of registration interface (authorization)
DACS Server
(4) Registration processing
Program
Data base

Figure 7. Function of DACS rule's registration

Figure 7 shows the function of registering the DACS rules. Process (1) is the request process of User Authentication. After the network administrator inputs a user name and pass word into the demand interface of them, they are sent to the program on the Web Server which is implemented in this research. After it, the inquiry processing for user authentication is performed between the program and a LDAP Server which has the information of user account (2). After the authentication is authenticated, registration interface is displayed on the Web Browser on the client (3). When the system administrator inputs the registration information for DACS rules, it is sent to the data base of the DACS Server through the program on the Web Server. In other functions of changing and deleting the DACS rules, the processes from (a) to (c) in Figure 7 are same processes. Only the process (4) replaces changing processing or deleting processing.

The reason of implementation based on http (or https) protocol is why the DASC system will be extended to the direction of Internet management. Therefore, the Web system is more convenient than C/S system.

(2) Function to grasp the correspondence relationship of the user name that logged in the client and the IP address of it

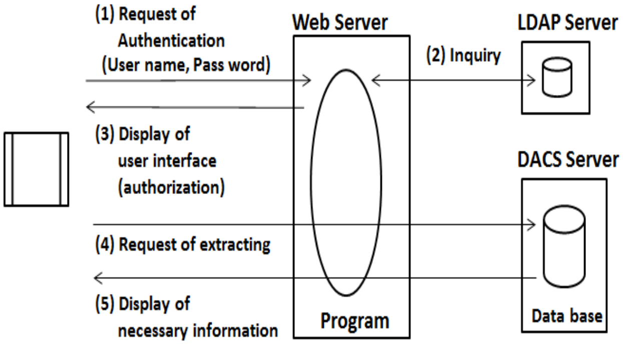

(1) Request of Authentication (User name, Pass word)
Web Server
LDAP Server
(2) Inquiry
(3) Display of user interface (authorization)
DACS Server
(4) Request of extracting
(5) Display of necessary information
Program
Data base

Figure 8. Function of extracting the correspondence relationship of user name and IP address

Figure 8 shows the function of extracting the correspondence relationship of the user name that logged in the client and the IP address of it. Process (1) and process (2) are same as those in Figure 7. After the authentication is authenticated, user interface for inputting extraction conditions are displayed on the Web Browser on the client (3). When the system administrator inputs the extraction conditions, it is sent to the program on the Web Server. The program performs an inquiry on the table which stores the user name and IP address. The correspondence relationship records of user name and IP address are extracted and displayed on the Web Browser through the program on the Web Server.

(3) Function that a system administrator transmits a message to a client-side

Figure 9 sows the function of sending a message to the client side. Process (1) and process (2) are same as those in Figure 7. After the authentication is authenticated, the lists of the user name that logged in the client and the IP address are acquired (3). Based on the lists, the user interface for sending a message to the client side is displayed on the Web Browser on the client. The user interface has the function of selecting destination users or clients, and the function of inputting the message sentences. After the network administrator selects and inputs them, a request of sending the message is sent to the program on the Web Server. The message is sent through the program, and displayed on the Web Browser at the client side.

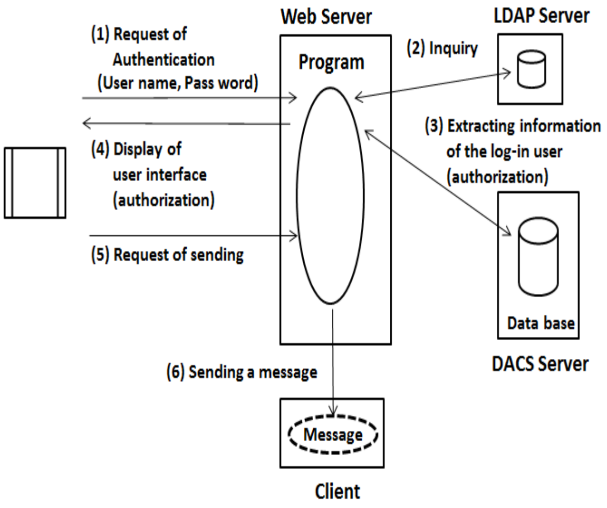

Web Server
LDAP Server
(1) Request of Authentication (User name, Pass word)
Program
(2) Inquiry
(3) Extracting information of the log-in user (authorization)
(4) Display of user interface (authorization)
(5) Request of sending
Data base
DACS Server
(6) Sending a message
Message
Client

Figure 9. Function of sending a message to the client side

As a first point, authentication processes are needed to be implemented by use of encrypted communication based on https protocol and public key infrastructure (PKI). To be concrete, processes of (1)(3)(4) in Figure 7 and those of (1)(3)(4)(5) in Figure 8, those of (1)(4)(5) in Figure 9 need to be encrypted. Next, as means of the security measures of Web Server, access control by client authentication needs to be adapted. Client authentication is a method of access control on the Web Server that is realized by using the private key which a system administrator holds. Therefore, security level becomes higher than simple authentication method using user name and password.

This is because it is necessary for future extensibility of the DACS system to be considered. The DACS system is going to be expanded to operate on the Internet. Therefore, implementation using that is used https and PKI on the Internet normally, is necessary. In addition, LDAP server is adapted as an authentication server. Because it is used on the Windows Server and Unix/Linux Server normally, it is adopted in many organizations. To be concrete, Active Directory is used on a Windows server, OpenLDAP is used on UNIX/Linux servers.

Then, when a Web Server and the LDAP server are located on the different server machine, process (2) in Figure 7,8 and 9 need to be also encrypted. Similarly, when a Web Server and the LDAP server are located on the different server machine, communication processes between the Web Server and a DACS Server need to be also encrypted. Then, as a Certificate Authority (CA) which is used for secure and certain key management, the CA in the high-integrity organization needs to be selected.

VI. CONCLUSION

The DACS system is for the realization of the effective network management based on the DACS Scheme which is one of the policy-based network management schemes. In this paper, we showed technical points in the implementation of the support system for operation and management of the DACS system. To be concrete, we showed three problem of the DACS Scheme on operation and management, and technical points in the implementation of the functions. Because the DASC scheme will be extended into the Internet management system, these functions are realized a web-based application based on http (or https) protocol to meet it. As a near future work, implementation of the support system proposed in this paper will be performed.

REFERENCES

[1] S.K. Das, D.J. Harvey, and R. Biswas,"Parallel processing of adaptive meshes with load balancing," IEEE Tran.on Parallel and Distributed Systems, vol. 12, No. 12, pp. 1269-1280, Dec 2002.

[2] M.E. Soklic,"Simulation of load balancing algorithms: a comparative study," ACM SIGCSE Bulletin, vol. 34, No. 4, pp. 138-141, Dec 2002.

[3] J. Aweya, M. Ouellette, D.Y. Montuno, B. Doray, and K. Felske,"An adaptive load balancing scheme for web servers," Int.,J.of Network Management., vol. 12, No. 1, pp. 3-39, Jan/Feb 2002.

[4] C. Metz, "The latest in virtual private networks: part I," IEEE Internet Computing, Vol. 7, No. 1, pp. 87–91, 2003.

[5] C. Metz, "The latest in VPNs: part II," IEEE Internet Computing, Vol. 8, No. 3, pp. 60–65, 2004.

[6] Y. Watanabe, K. Watanabe, E. Hirofumi, and S. Tadaki,"A User Authentication Gateway System with Simple User Interface, Low Administration Cost and Wide Applicability," IPSJ Journal, Vol. 42, No. 12, pp. 2802-2809, 2001.

[7] S. Tadaki, E. Hirofumi, K. Watanabe, and Y. Watanabe,"Implementation and Operation of Large Scale Network for User' Mobile Computer by Opengate," IPSJ Journal ,Vol. 46, No. 4 pp. 922-929, 2005.

[8] R.Yavatkar, D. Pendarakis, and R. Guerin, "A Framework for Policy-based Admission Control", IETF RFC 2753, 2000.

[9] B. Moore, E. Ellesson, J. Strassner, and A. Westerinen, "Policy Core Information Model -- Version 1 Specification", IETF RFC 3060, 2001.

[10] B. Moore.,"Policy Core Information Model (PCIM) Extensions", IETF 3460, 2003.

[11] J. Strassner, B. Moore, R. Moats, and E. Ellesson, " Policy Core Lightweight Directory Access Protocol (LDAP) Schema", IETF RFC 3703, 2004.

[12] D. Durham, Ed., J. Boyle, R. Cohen, S. Herzog, R. Rajan, and A. Sastry,"The COPS (Common Open Policy Service) Protocol", IETF RFC 2748, 2000.

[13] S . Herzog, Ed., J. Boyle, R. Cohen, D. Durham, R. Rajan, and A. Sastry,"COPS usage for RSVP", IETF RFC 2749, 2000.

[14] K. Chan et al.,"COPS Usage for Policy Provisioning (COPS-PR)", IETF RFC 3084, 2001.

[15] CIM Core Model V2.5 LDAP Mapping Specification, 2002.

[16] M. Wahl, T. Howes, and S. Kille,"Lightweight Directory Access Protocol (v3)", IETF RFC 2251, 1997.

[17] CIM Schema: Version 2.30.0, 2011.

[18] ETSI ES 282 003: Telecoms and Internet converged Services and protocols for Advanced Network (TISPAN); Resource and Admission Control Subsystem (RACS); Functional Architecture, June 2006.

[19] ETSI ETSI ES 283 026: Telecommunications and Internet Converged Services and Protocols for Advanced Networking (TISPAN); Resource and Admission Control; Protocol for QoS reservation information exchange between the Service Policy Decision Function (SPDF) and the Access-Resource and Admission Control Function (A-RACF) in the Resource and Protocol specification", April 2006.

[20] K. Odagiri, R. Yaegashi, M. Tadauchi, and N. Ishii, "Efficient Network Management System with DACS Scheme : Management with communication control," Int. J. of Computer Science and Network Security, Vol. 6, No. 1, pp. 30-36, January 2006.

[21] K. Odagiri, R. Yaegashi, M. Tadauchi, and N. Ishii, "Secure DACS Scheme, "Journal of Network and Computer Applications, Elsevier, Vol. 31, Issue 4, pp. 851-861, Nov 2008.

[22] K. Odagiri, R. Yaegashi, M. Tadauchi, and N. Ishii, "New User Support in the University Network with DACS Scheme," Int. J. of Interactive Technology and Smart Education.

[23] K. Odagiri, S. Shimizu, R. Yaegashi, M. Takizawa, and N. Ishii, "DACS System Implementation Method to Realize the Next Generation Policy-based Network Management Scheme," Proc. of Int. Conf. on Advanced Information Networking and Applications (AINA20010), Perth, Australia, Japan, IEEE Computer Society, pp. 348-354, May 2010.

[24] K. Wakayama, Y. Decchi, J. Leng, and A. Iwata, "A Remote User Authentication Method Using Fingerprint Matching," IPSJ Journal, Vol. 44, No. 2, pp. 401-404, 2003.

[25] Seno, Y. Koui, T. Sadakane, N. Nakayama, Y. Baba, and T. Shikama, "A Network Authentication System by Multiple Biometrics," IPSJ Journal, Vol. 44, No. 4, pp. 1111-1120, 2000.

[26] Trusted Computing Group, TNC Architecture for Interoperability Version 1.4, Revision 4, 2009.

Infrastructure Optimization in a Transmission Network

Mary Luz Mouronte
Departamento de Ingeniería y Arquitecturas Telemáticas
Universidad Politécnica de Madrid
Madrid, Spain
mouronte.lopez@upm.com

***Abstract*—This paper presents an algorithm to optimize the number of necessary resources in a transmission network; this procedure could be executed by the operators during the maintenance and fulfillment tasks. It allows the reduction of the investment, improvement of the resource utilization and achievement of the high resilience. We have developed an experimental prototype and have executed it over different transmission networks saving resources up to a percentage of 30%.**

***Keywords*-Transmission network; Optimization**

I. Introduction

This paper shows a method to lower the usage of resources in a transmission network. The algorithm achieves significant savings.

A transmission network consists of end-to-end circuits strictly designed as ring, mesh, bus and other motifs connecting equipments with different link capacities (Mbps/Gbps). This network has several components:

- Regenerators or equipments carry out the regeneration of the signals.
- Terminal multiplexers combine the plesionchronous and synchronous input signals into higher bitrate signals. They are the network elements that originate and terminate the signals.
- **A**dd-**D**rop **M**ultiplexers (ADM)/**O**ptical **A**dd-**D**rop **M**ultiplexers (OADM) combine several lower-bandwidth streams of data into a single beam of light. An ADM also has the capability to add lower-bandwidth signals to an existing high-bandwidth data stream, and at the same time can extract or drop other low-bandwidth signals, removing them from the stream and redirecting them to some other network path.
- **D**igital **C**ross-**C**onnects/ **O**ptical **C**ross-**C**onnects exchange traffic between different fiber routes. The key difference between the DCC/OCC and the ADM/OADM is that the DCC/OCC provides a switching function, whereas the ADM/OADM performs a multiplexing function. The DCC/OCC moves traffic from one facility route to another.
- Cards:
  - Aggregate cards provide the line interface. They connect the ADM to the exchange via fibre optical cables.
  - Tributary cards collect customer traffic and pass them, via the common cards, to the aggregate cards. Tributary cards are used to provide interfaces to one or more lower speed devices.
  - Other types of cards.

Transmission networks are managed by means of **N**etwork **E**lement **M**anagers (NEMs), **S**ub**N**etwork **M**anagers (SNMs) and, in some cases, also by a **N**etwork **M**anagement **S**ystem (NMS).

Link connection, trails and others entities for the tansmission network are defined in [1]:

- A link connection is a transport entity provided by the client/server association. It is formed by a near-end adaptation function, a server trail and a far-end adaptation function between connection points.
- A trail is a transport entity in a server layer which is responsible for the integrity of the transfer of characteristic information from one or more client network layers and between server layer access points. It defines the association between access points in the same transport network layer. It is formed by combining a near-end trail termination function, a network connection and a far-end trail termination function.

The proposed algorithm reduces the usage of resources in a transmission network. The resource release is obtained by the optimization of:

- Port usage in tributary cards.
- Occupancy rate in trails. The aim is the reduction of the hop number in the trail from the origin to the destination point, by increasing the occupancy rate in link connections.
  The analysis is done independently over the different network layers. Additionally, when a protected trail is examined by the procedure, it is verified that no optimized trail is matched with the current backup trail.

This optimization method has been included as a software module in a NMS, which manages different vendors (Nortel, Alcatel-Lucent, Ericsson and Huawei) and technologies (**S**ynchronous **D**igital **H**ierarchy (SDH), Ethernet over SDH, **W**avelength **D**ivision **M**ultiplexing (WDM)). This fact verified the operation results.

The rest of the paper is organized as follows: Section II details related works, Section III gives an overview about the transmission network, algorithm is described in Section IV, Section V summarizes the results of applying a prototype on a life transmission network, and Section VI describes the main conclusions.

II. Related works

There are several works on optimization in a network, for example:

In [2], refering to the traditional multicast IP network, authors describe a method based on ant colony algorithms to minimize unnecessary overhead while achieving the desired throughput in a multicast scenario. In this type of network, the intermediate nodes take care of replicating the packet to reach multiple receivers only when necessary, so it is difficult for the network to achieve the maximum transmission rate. However, not all intermediate nodes are required for network coding operations.

In [3], authors present an algorithm with two optimization modes in a network architecture with a three-layer IP/MPLS over SDH over WDM: in optimization mode 1, the service blocked in the upper layer can be transmitted to the lower layer by its idle resources; in optimization mode 2, authors regard the three-layer network as an integrated network and search route for each service in the integrated network.

In [4], the author's research addresses the problem of Routing and Wavelength Assignment (RWA) for survivable networks with the objective of optimizing the needed wavelength links and the number of optical/electrical devices.

The main differences between the aforementioned methods and the presented procedure are:

- It maximizes the tributary cards with zero occupation rate and the server trails with minimum number of hops for a specific occupation rate.
- It works in an multiprovider and multitechnology environment. In an operative environment, NEMs have analysis tools which optimize subnetworks within each manufacturer domain. However, networks are composed of different technologies and vendors. They have isolated management domains where it is not possible to analyze and optimize with a whole network vision.

III. Transmission Network

Transmission networks are increasingly demanding greater capacity and more effective communications to support telecommunications services.

Standards agencies define a set of International Telecommunication Union (ITU) recommendations. [1], [5],[6], [7], [8], [9], [10], and [14] are designed to build telecommunication networks that allow greater flexibility and interconnection capacity between different technologies and equipment manufacturers.

ITU-T recommendations define a transport network for different technologies. Currently, transmission networks are set up mainly with SDH and OTH (Optical Transport Hierarchy).

According to ITU-T definitions, transmission networks are divided into independent layers where each layer has a server-client relationship with the adjacent layers. Each layer is also divided to reflect the internal structure and enable its management.

On a physical level, these networks are set up by nodes and connections between them. The nodes consist of racks containing cards in charge of performing different tasks in the network.

- Tributary cards: they introduce the client signal into the network.
- Aggregate cards: they add client signals to a server signal which is transmitted to other nodes through a physical trail.
- Matrix: it cross-connects signals to drive their route. This matrix can be electrical (SDH, OTH electric layer) or optical (OTH electric layer).
- Control card: it is responsible for node control.
- Transponder: it receives client signals and adapts them to optical signals for the Optical Transport Network (OTN).
- Muxponder: it is a hybrid card, with tributary ports, aggregate ports and electric cross-connect capacity (they are common in OTN technology).
- Amplifiers, Filters: they adapt the signal to physical transmission media.

The SDH network structure is defined in the ITU-T recommendation G.783 (Characteristics of synchronous digital hierarchy (SDH) equipment functional blocks) [1]. This document describes its layered structure and interfaces between different network layers.

- Physical layer SDH.
- RS: Regenerator Section layer.
- MS: Multiplex Section layer.
- HO: High Order trail layer.
- LO: Low Order trail layer.

The OTN network structure is defined in the ITU-T recommendation G.809/Y1331 (Interfaces for the Optical Transport Network (OTN)) [14]. This document describes the layered structure and interfaces among different network layers.

- OTS: Optical Transmission Section.
- OMS: Optical Multiplex Section.
- OCh: Optical Channel.
- OTU: Optical channel Transport Unit. There are different units (OTUk), where k can take values 1,2,3,4 and represents the signal speed: 2,5 Gb/s, 10 Gb/s, 40 Gb/s, 100 Gb/s.
- ODU: Optical channel Data Unit. There are different

ISBN: 978-1-61208-256-1

units (ODUk), where k can take values 0,1,2,3,4 and represents the signal speed: 1,2 Gb/s, 2,5 Gb/s, 10 Gb/s, 40 Gb/s, 100Gb/s. A lower level ODU can be multiplexed into a high level ODU using multiplexing TDM (Time Division Multiplex).
- OPU: Optical Payload Unit. Layer for adapting the client signal to the ODU channel payload. Like the OTU and ODU channels, there are different OPUk values depending on channel speed.

In their inception, networks are planned in order to optimize resources such as nodes, cards, ports and link connections. The aim is to reduce the infrastructure costs taking into account the forecast traffic growth due to the emergence of new services. Transmission networks are constituted as interconnected islands, each island belonging to a single vendor. Network design and planning is carried out over these management islands without an end to end vision over the whole network.

Network operation results in several processes which add, modify and remove services may cause a non optimal usage of elements such as:

- Tributary cards with low occupancy.
- Server trails with low occupancy.
- Network layers fragmentation that results in low efficiency trails.

The network operators need to manage the usage of resources by means of end to end optimization mechanisms, which will avoid the network capacity degradation. Our algorithm, which is included as a software module within NMS, can be a very useful tool.

IV. Algorithm

Our optimization procedure uses the following input information relating to the network structure, which is obtained from NMS:

- The trails that support other low order layer trails, the available capacity in them and the equipments where they end.
- Ports:
  - The tributary ports that perform termination function on low order trails in SDH or WDM equipments.
  - Aggregated ports or line ports.
- Cards.
- Layer trails to analyze:
  - Client layer trails with source and destination points in the same node, in order to group traffic and to release resources in cards with tributary ports.
  - Server layer trails aiming to look up alternative trails which allow reducing the hop number and releasing resources in the transmission media and in cards with aggregated ports.
- The algorithm also receives the occupancy rate in each server layer, which is established by the operator.

Once the information is stored, the procedure performs the two following analysis:

- *Analysis type 1. Method for tributary card optimization in a exchange office*: It calculates the tributary card occupancy rate and redistributes the input ports in the cards with low occupation in other busier cards, thus some cards are released.

  The details of tributary card optimization procedure are outlined in Fig 1. The following parameters are estimated:
  - OR_{Ci}: Occupation rate of tributary card Ci

$$OR_{Ci} = \frac{TotalOfBusyPortsInCi}{TotalOfPortsInCi} \quad (1)$$

  - M : Number of tributary cards where $OR_{Ci} = 0$.
  - Goal to achieve or maximum M value ($Max[M]$).
- *Analysis type 2. Method for trail optimization*:
  - It groups the client trails with the same origin and destination and analyzes the server trails in each group:
    - ∗ The server trail with the least number of hops between source and destination points is searched applying the Bellman-Ford algorithm [15], [16].
    - ∗ If these server trails have not reached the maximum occupancy rate set by the operator (who can reserve resources for future network deployments), the client trails are set-up in them. When the maximum occupancy rate is reached, the procedure looks up the next shortest server trail. The process is repeated for all client trails with the same origin and destination.
  - By reducing the server trail size, it is possible to release resources which will be available for setting up other client trails.

  The details for the server trail optimizing procedure in each layer are outlined in Fig 2. This procedure is applied on each server layer recursively until specified by the operator. The following parameters are estimated:
  - G : Set of server trails with the same origin and destination.
  - OR_j: Occupation rate of j, where j : is the server trail j.

$$OR_j = \frac{TotalOfBusyLinkConnectionsInj}{TotalOfLinkConnectionsInj} \quad (2)$$

  - E : Set of server trails in G with the least number of hops and where occupation rate is $= ID$, ID is the occupation rate for server trails set by the operator.
  - Goal to achieve or maximum E value ($Max[E]$).

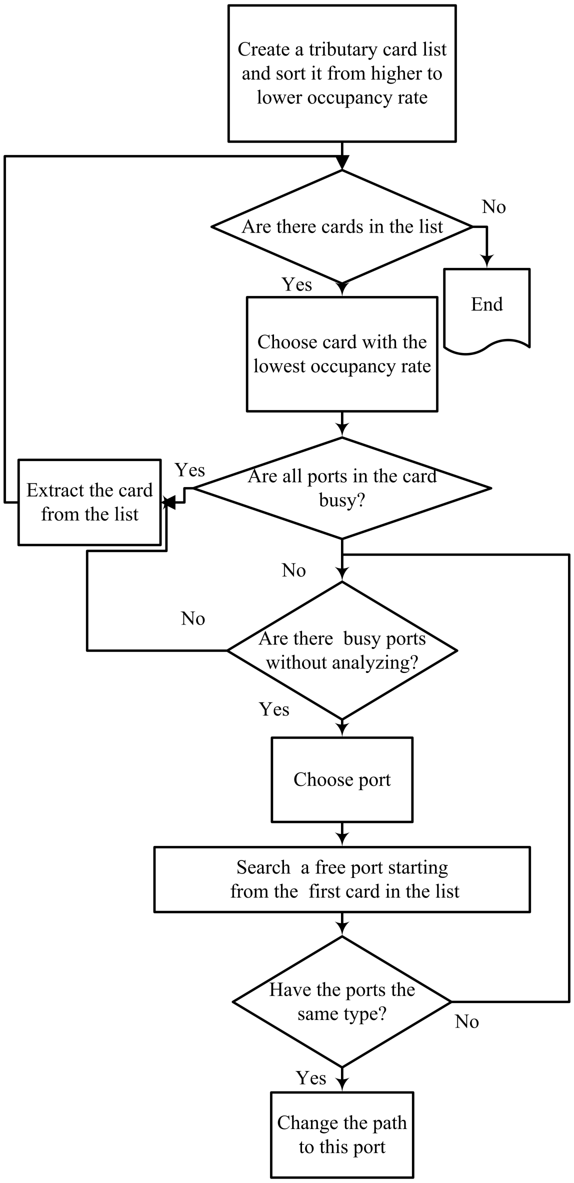

Create a tributary card list and sort it from higher to lower occupancy rate
Are there cards in the list
No
End
Yes
Choose card with the lowest occupancy rate
Are all ports in the card busy?
Yes
Extract the card from the list
No
No
Are there busy ports without analyzing?
Yes
Choose port
Search a free port starting from the first card in the list
Have the ports the same type?
No
Yes
Change the path to this port

Figure. 1 Procedure for tributary cards optimization.

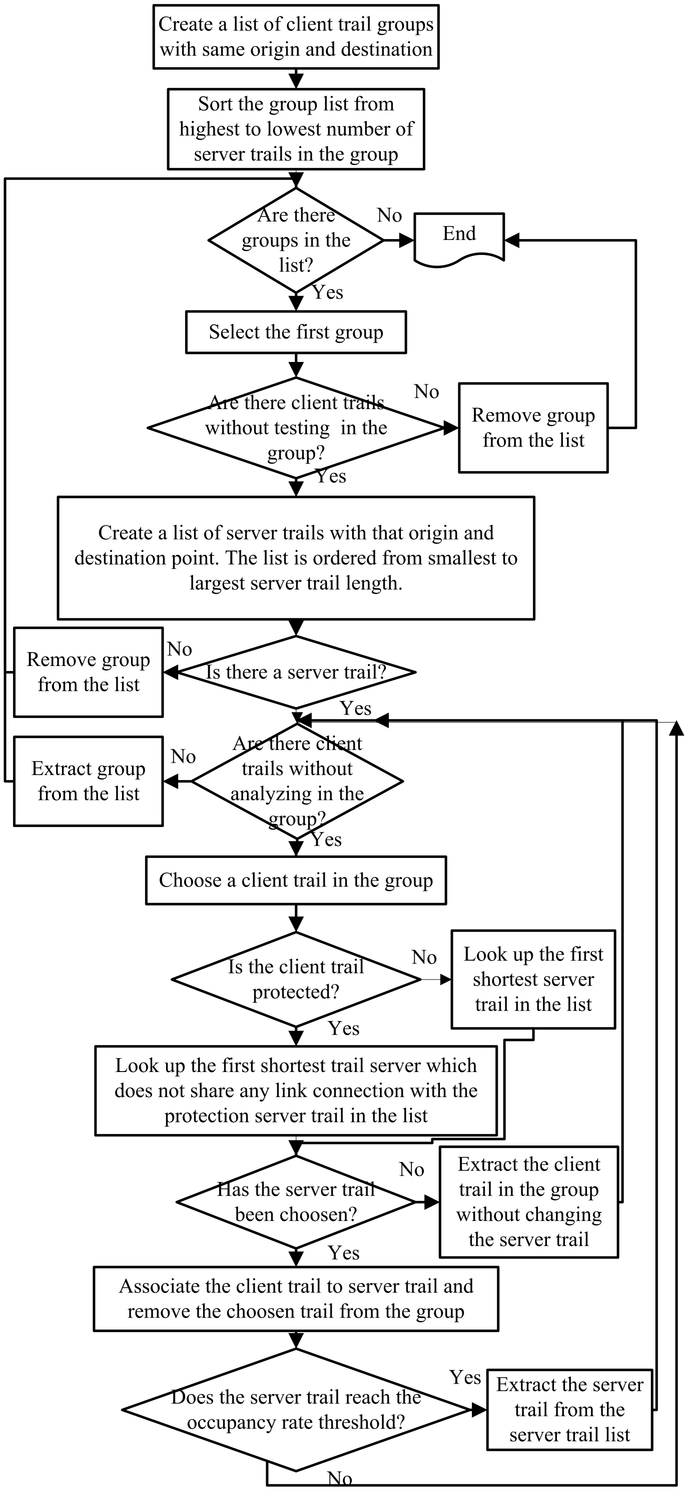

Create a list of client trail groups with same origin and destination
Sort the group list from highest to lowest number of server trails in the group
Are there groups in the list?
No
End
Yes
Select the first group
Are there client trails without testing in the group?
No
Remove group from the list
Yes
Create a list of server trails with that origin and destination point. The list is ordered from smallest to largest server trail length.
Remove group from the list
No
Is there a server trail?
Yes
Extract group from the list
No
Are there client trails without analyzing in the group?
Yes
Choose a client trail in the group
Is the client trail protected?
No
Look up the first shortest server trail in the list
Yes
Look up the first shortest trail server which does not share any link connection with the protection server trail in the list
Has the server trail been choosen?
No
Extract the client trail in the group without changing the server trail
Yes
Associate the client trail to server trail and remove the choosen trail from the group
Does the server trail reach the occupancy rate threshold?
Yes
Extract the server trail from the server trail list
No

Figure. 2 Procedure for server path optimization.

Fig 3 shows a network situation where the algorithm can be applied:

- Equipments: A, B, C, D, E
- Cards:
  - Equipment A: 2 cards with 2 tributary ports, and 2 cards with 2 aggregated ports.
  - Equipment B: 2 cards with 2 tributary ports, and 2 cards with 2 tributary ports.
  - Equipment C: 2 cards with 2 tributary ports, and 2 cards with 2 aggregated ports.
  - Equipment D: 2 cards with 2 tributary ports and 2 cards with 2 aggregated ports.
  - Equipment E: 2 cards with 2 tributary ports and 2 cards with 2 aggregated ports.
- Physical trails: A-B, A-E, B-C, B-D, B-E, C-D, C-E, D-E
- Client trails: A-B, A-B-C,A-E-B-C, B-D, E-B-C, E-B-

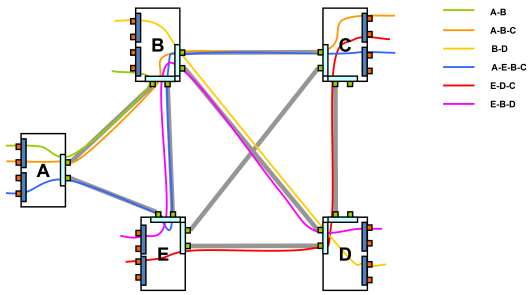


Figure. 3 Network situation to be optimized.

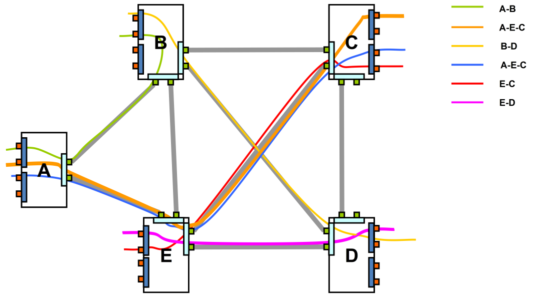


Figure. 4 Optimized network by means of our algorithm.

D

Fig 4 shows the obtained results. 3 cards are released by grouping ports in cards (6 cards were used previously) and 3 physical trails (only one was available before) are emptied by restoration of client trails.

V. RESULTS

An experimental prototype aimed to test and validate the optimization improvement in a network has been prepared.

Our optimization method is developed in PL/SQL and C++ language. These programs are included as a module in a NMS of a Telecommunication Operator where they interact with its inventory to obtain and update the network information. The programs are executed in the NMS by the operator when maintenance or fulfillment tasks are carried out.

We also implemented the necessary programs to compare with Shortest trail First (SPF) and Constrained Shortest trail First (CSPF) algorithms.

The NMS is a unified network management solution which provides end to end view and homogenous functions across different vendors. Furthermore, it executes all business processes related to the transmission network and its services: network fullfilment, circuit provisioning, network supervision and performance monitoring. This NMS manages SDH, Ethernet over SDH and WDM networks.

The NMS is a system constructed around a standards based network model, supported by an ORACLE DBMS, with a business logic layer that allows interaction with the core applications through a CORBA bus. The NMS works as a centralized system, with a primary machine of 16 CPUs, another one of 4 for the mediation with the plant and three more for users access.

In our validation test, 30 large networks (more than one province), 50 mid-size networks (provinces) and 125 small networks (provincial subdivisions) were evaluated. The networks had different topologies: point-to-point links, rings (single and dual), fully connected meshes, and the following characteristics:

- Big size network (average values): equipments: 20,248; circuits: 11,233; trails: 22,538; cards: 328,443; ports: 523,277 (tributary ports: 446,741; aggregate ports: 76,536).
- Medium size network (average values): equipments: 4,193; circuits: 5,274; trails: 16,064; cards: 90,343; ports: 118,427 (tributary ports: 108,806; aggregate ports: 10,341).
- Small size network (average values): equipments: 3,237; circuits: 1,233; trails: 1,823; cards: 11,243; ports: 15,456 (tributary ports: 9,235; aggregate ports: 6,221).

VI. Conclusions

In this research, we show a method to check and improve the element usage in the trails with an entire network vision (multi-provider and multi-technology environment); up to now the telecommunication operators just have tools to optimize subnetworks within each manufacturer domain. Besides, the algorithm obtains good results in an operative environment.

The two previous features are the principal novelty in our algorithm, which offers the following benefits:

- Significant resource savings by releasing cards with tributary ports, delivering link connections and redistributing trails.
- Higher quality deployment due to redistribution of resources. Setup, modification and removal operations do not usually take into account the whole network (with several vendors and technologies), so they do not use resources that could be employed with a different allocation.
- Resilience improvement. Minor resources to support a specific traffic demand are necessary (i.e. there is lower error probability in the network).

References

[1] ”ITUT Recommendation G.803: Architectures of Transport Networks Based on the Synchronous Digital Hierarchy (SDH)”. March 2000. http://www.itu.int/rec/T-REC-G.803/en, May 2011.

[2] J. L. Yun Pan, ”Research of network coding resources optimization based on ant colony optimization”, International Conference on Computer Application and System Modeling (ICCASM 2020), IEEE Xplore, vol. 11, pp. 135-138, November 2010.

[3] Z- Wei, L. San-yang and Q- I. Xiao-gang, ”Integrated resources optimization in three-layer dynamic network”, Journal of Convergence Information Technology, vol. 5, No. 6, pp. 40-46, August 2010.

[4] K. Wang, ”Resource optimization and QoS for WDM optical networks”, Doctoral Dissertation, University of Nebraska at Lincoln, NB, USA, 2006.

[5] ”ITUT Recommendation G.805: Generic functional architecture of transport networks”, March 2000. http://www.itu.int/rec/T-REC-G.805/en, May 2011.

[6] ”ITUT Recommendation G.707: Network node interface for the synchronous digital hierarchy (SDH)”, January 2007. http://www.itu.int/rec/T-REC-G.707/en, May 2011.

[7] ”ITU-T Recommendation G.783: Characteristics of synchronous digital hierarchy (SDH) equipment functional blocks”. March. 2006. http://www.itu.int/rec/T-REC-G.783/en, May 2011.

[8] ”ITUT Recommendation G.841: Types and characteristics of SDH network protection architectures”, October 1998. http://www.itu.int/rec/T-REC-G.841/en, May 2011.

[9] ”ITUT Recommendation G.842: Interworking of SDH network protection architectures”, April 1997. http://www.itu.int/rec/T-REC-G.842/en, May 2011.

[10] ”ITUT Recommendation G.872: Architecture of optical transport networks”, November 2001. http://www.itu.int/rec/T-REC-G.872/en, May 2011.

[11] ”ITUT Recommendation G.709: Interfaces for the Optical Transport Network (OTN)”, December 2009. http://www.itu.int/rec/T-REC-G.709/en, May 2011.

[12] ”ITUT Recommendation G.798: Characteristics of optical transport network hierarchy equipment functional blocks”, October 2010. http://www.itu.int/rec/T-REC-G.798/en, May 2011.

[13] ”ITU-T Functional architecture of connectionless layer networks”, March 2003, http://www.itu.int/rec/T-REC-G.809/en, May 2011.

[14] V. V. Georgievskii and E. G. Davydov, ”A method of constructing independent routes and sections of network models”, Journal of Minning Science, vol. 11, No. 6, pp. 711-715, November-December 1975.

[15] D. Walden, ”The Bellman-Ford algorithm and Distributed Bellman-Ford”, May 2003. http://www.walden-family.com/public/bf-history.pdf, pp. 1-12, May 2011.

[16] D. Torrieri, ”Algorithms for finding an optimal set of short disjoint paths in a communication network”, Military Communications Conference (MILCOM 91), IEEE Xplore, vol. 40, pp. 1698-1702, August 1992.

Virtual Network Topologies Adaptive to Large Traffic Changes by Reconfiguring a Small Number of Paths

Masahiro Yoshinari, Yuichi Ohsita, Masayuki Murata
Graduate School of Information Science and Technology
Osaka University
Osaka, Japan
{m-yoshinari, y-ohsita, murata}@ist.osaka-u.ac.jp

***Abstract*—A virtual network reconfiguration is one efficient approach to accommodate the traffic that changes significantly. By reconfiguring the virtual network, the network accommodates the traffic even when the traffic pattern changes significantly. The reconfigure has a large impact on the traffic passing the reconfigured paths. Thus, the number of reconfigured paths should be minimized. The number of reconfigured paths depends on the virtual network topology before the reconfiguration, and some topology requires a large number of reconfigured paths to handle the traffic changes. In this paper, we investigate the virtual network topology, which can handle significant traffic changes by reconfiguring only a small number of paths. To investigate the virtual network, we propose a index based on the evolution model in the changing environments. We evaluate our index through simulation, and clarify that our index indicates the adaptability of virtual networks. We also compare our index with betweenness centrality, and clarify that our index identifies the virtual network with high adaptability to traffic changes more accurately.**

Keywords*—*Traffic Change; Traffic Engineering; Topology; Optical Network; Reconfiguration

I. Introduction

In recent years, various new applications have been deployed over the Internet. Such application leads the increase of the traffic amount and the unpredictable traffic changes [1]. A network must accommodate such a time-varying traffic efficiently. However, accommodating time-varying traffic efficiently is difficult, because even if a backbone network suitable for the current traffic is constructed, the backbone network becomes no longer suitable to traffic after the traffic change.

One approach to accommodate such a large time-varying traffic is to reconfigure the virtual network. Several methods to reconfigure the virtual network have been proposed [2], [3], [4]. In these methods, a virtual network is constructed over the optical network, which is constructed of the optical cross connects (OXCs) and IP routers. In this optical network, each outbound port of an edge IP router is connected to an OXC port. Lightpaths (hereafter called optical paths) are established between two IP routers by configuring OXCs along the route between the routers. A set of routers and optical paths between the routers forms a virtual network. Traffic between two routers is carried over the virtual network using IP layer routing. In this network, the virtual network is reconfigured dynamically by adding or deleting optical paths so as to suite the current traffic.

In case of significant traffic change, a large number of optical paths may be required to be added to accommodate the traffic after the change. However, adding a large number of optical paths may take a large overheadbecause we require setting OXCs for each optical path.

One approach to avoid adding a large number of optical paths is to construct an adaptive virtual network, which can handle any traffic change by adding a small number of optical paths in advance. Thus, we investigate the adaptive virtual network.

There are several metrics of the network topology. The betweenness centrality of a link [5] indicates the probability that traffic from a source node to a destination node passes the link. The link criticality [6], [7] is obtained by dividing the betweenness centrality by the bandwidth of the link. The link whose betweenness centrality or link criticality is high may be passed by a large amount of traffic. Thus, the topology with links with high betweenness centrality or link criticality is easy to be congested. However, these indices do not indicate the number of optical paths required to be added when congestions occur.

In this paper, we propose an index that identifies the adaptive virtual network that can handle significant traffic changes by adding only a small number of optical paths. To propose the index, we are inspired by the natural lifeforms that survive and evolve in the case of significant environmental changes. The natural lifefroms are modeled [8] as a suite of functions, which synthesizes products from environmental resources. In this model, each individual dies if enough products for survival cannot be generated, while it duplicates itself and evolves if enough products are generated.

We model the virtual network by the similar way to the natural lifeforms, and propose an index called *flow inclusive relation modularity* (FIRM).

The rest of this paper is organized as follows. In Section II, we explain the characteristics of the lifeforms, which survive and evolve under significant environmental changes. Then, inspired by the characteristics of the natural lifeforms, we propose an index that identifies the adaptive virtual network. In Section III, we mention the steps to evaluate our index. In Section IV, we show the result of the evaluation. Finally, we conclude and mention about future work in Section V.

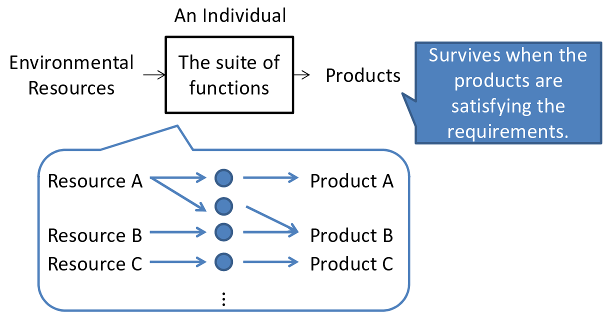


Figure 1: Lifeform Model

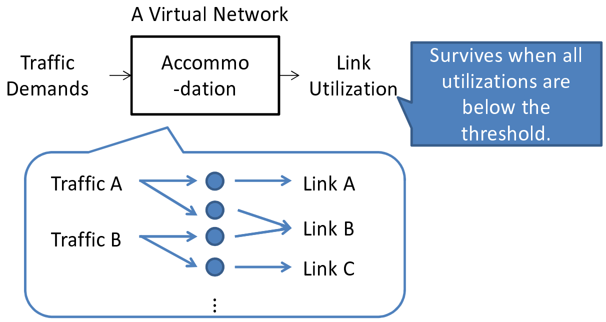


Figure 2: Virtual Network Flow Model

II. Adaptability to Environmental Changes and Modularity

A. Environmental Changes and Modularity in Lifeforms

Lipson et al. [8] clarified one of the characteristics of the lifeforms that survive and evolve under significant environmental changes through simulations. They modeled individuals as a suite of functions, which synthesizes products from environmental resources as shown in Fig. 1. In this model, each individual dies if enough products for survival cannot be generated, while it duplicates itself and evolves if enough products are generated. Through the simulation using this model, they investigated the characteristics of lifeforms that can survive and evolve in significant environmental changes.

Among the characteristics of the lifeforms, they focus on the relationship between functions. Each function consumes some resources and generate products. The relationship between functions exists when those functions consumes the same resources or when those functions carries out or blocks the production of the same product. By using the relationship, they define the index called *modularity*, which is defined by the number of groups after dividing the functions into groups that includes the related functions. According to the results of Lipson et al. [8], the lifeforms with higher modularity survive and evolve. In addition, the lifeforms evolves so as to have higher modularity.

When the modularity is high, functions belonging to the different modules have only a small impact on each other. Thus, the environmental changes on the functions in a module do not affect the other functions in the other modules. As a result, the individuals with the large modularity survive and evolve in the environmental changes.

B. Traffic Changes and Modularity in Virtual Network Operation

Inspired by the lifeforms that survive and evolve in the significant environmental changes, we model the virtual network, and propose an index that identifies the virtual network that can handle significant traffic changes by adding a small number of optical paths.

1) Functions in Virtual Network: The function of the virtual network is to accommodate traffic. We model this function of the virtual network, as shown in Fig. 2. In this model, a virtual network accommodates traffic demands by assigning traffic demands with links. When the utilizations of all links are less than the threshold, we regard the virtual network as being operated properly.

The model shown in Fig. 2 is similar to the model of lifeforms shown in Fig. 1. The traffic demands of the model in Fig. 2 correspond to the resources of the model in Fig. 1.

Therefore, applying the results of the lifeforms, a virtual network whose functions have large modularity may be adaptive to the significant traffic changes. Thus, in this paper, we define the modularity of the functions of the virtual network, and investigate the relationship between the modularity and the number of optical paths required to be added to accommodate significant traffic change.

2) Relationships between Functions in Virtual Network: To define a modularity of a virtual network, we need to define the relationship among the functions in the virtual network. In this paper, the function of the virtual network is modeled as the suite of the function that accommodates each flow passing between source and destination nodes. In this subsection, we define the relationship between functions. There are several approaches to define the relationship between functions. For example, one approach is to regard the functions related to the flows passing the same link as the related functions. In this paper, we focus on the close relationship between the functions. We define the relationship as follows; the functions for flow A and for flow B is regarded to have the relationship if the all links passed by the flow A are also passed by the flow B. Hereafter, we call this relationship *flow inclusive relation* (FIR).

As shown in Fig. 3, FIR is described as a graph where a vertex is defined for each flow. The vertices are connected with edges if their corresponding functions have FIR. Hereafter, we call this graph *flow inclusive relation graph*, and call its vertices *flow nodes.*

3) Flow Inclusive Relation Modularity: Applying the results of Lipson et al. [8] to the virtual network, the virtual network with high modularity has strong adaptability to environmental changes. In this paper, we define the modularity by the modularity of the flow inclusive graph calculated by the method proposed by Newman [9]. Hereafter, we call this modularity the *flow inclusive relation modularity (FIRM).*

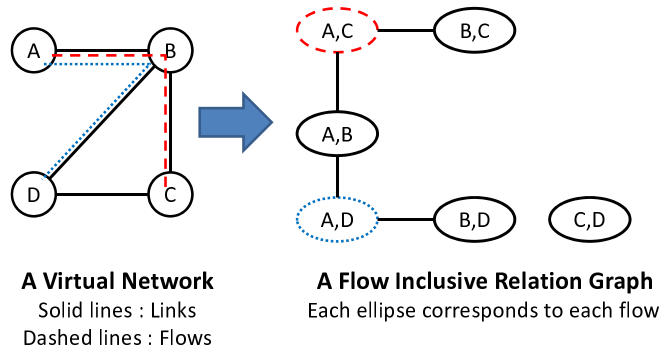


Figure 3: Flow inclusive relation model

A modularity of a graph is defined as

$$Q = \sum_{g \in G} \left[\frac{1}{2m} \sum_{i,j \in N_g} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \right], \qquad (1)$$

where $A_{ij}$is the number of edges between node i and node j, $k_i$is the degree of node i, $m = \frac{1}{2}\sum_i k_i$ is the total number of edges, G is the set of modules and N_g is a set of nodes which satisfy$g \in G$.

In (1), $\frac{k_i k_j}{2m}$ indicates an expected value of the total number of edges in the group in a random network having the same number of nodes and the same number of edges. $\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)$ indicates the difference between the total number of edges in the group and the expected value of the total number of edges in the corresponding group in a random network. The modularity Q is a normalized value of $\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)$ by multiplying $\frac{1}{2m}$ so that the maximum value of Q is 1. As Q approaching 1 closer, the structure has denser inner-module edges and sparser inter-module edges.

Newman [9] proposed a method to divide a given network into modules so as to achieve higher modularity. This method recursively divides a network into two modules so as to maximize the modularity until the division no longer increases the modularity.

In this paper, we obtain a flow inclusive relation modularity of the virtual network by applying this method [9]. The obtained flow inclusive relation modularity indicates whether the functions of the network are divided into groups so that each group includes the functions closely related to each other.

In the network with the large flow inclusive relation modularity, several flows are closely related. If the congestion occurs, the congestion is mitigated by adding an optical path to the node pair which are the source and destination node of a flow passing the congested links, and changing the route of the flows. Moreover, the congestion of the other links may also be mitigated, because adding the optical path enables the route change of the other flows belonging to the same module. As a result, the congestion of all links may be mitigated by adding a small number of paths.

III. Evaluation Method

A. Overview of the Evaluation Method

In this paper, we evaluate the relationship between the flow inclusive relation modularity of a virtual network and the number of optical paths required to be added to accommodate significant traffic changes. In our evaluation, we perform the following steps. First, we prepare some initial virtual networks having various flow inclusive relation modularities. We calculate the flow inclusive relation modularities of initial virtual networks. Then we generate the traffic changes by randomly generating the traffic, and reconfigure the virtual network so as to accommodate the traffic. We count the number of optical paths added to accommodate the traffic.

In this evaluation, we generate 10 patterns of traffic matrices for each initial virtual network.

In the rest of this section, we describe the details of the generation method of initial virtual networks and traffic matrices, and the reconfiguration method of virtual network used in the evaluation.

B. The Generation Method of Initial Virtual Networks

In order to generate initial virtual networks with various FIRMs, we use a method to generate topologies with various modularity proposed by Hidaka [10]. This method uses the number of groups n and probability parameter p as inputs, and generates the topology by the following steps.

First, this method generates n groups and locate one node in each group. The nodes are connected so as to form a ring. Then, this method adds the nodes. When adding a node, the group of the node is selected randomly. The additional node is connected to one node randomly selected from the nodes in the group. Furthermore, an edge between the additional node and the node which belongs to other group is added with probability p or an edge between the additional node and the node in the same group is added with probability $(1 - p)$. This method generates various topologies depending on the value p.

In this paper, we generate 255 initial virtual networks by changing the parameter p from 0.00 to 1.00 at 0.02 intervals. We set the number of nodes to 49 and the number of groups to 5.

C. The Generation Method of Traffic Matrices

Antoniou et al. [11] monitored the traffic in ISPs and clarifies that the traffic between source and destination router pairs follows a log normal distribution. Thus, in this paper, we generate traffic matrices so as to follow the lognormal distrubution, whose parameters are set to the same value as the results by Antoniou et al. [11].

D. The Reconfiguration Method of Virtual Netowrks

In this paper, we use a reconfiguration method based on the method proposed by Gençata et al. [4]. This method continues to add optical paths until the utilizations of all links become lower than the threshold Th.

ISBN: 978-1-61208-256-1

In this paper, we use a method that accommodates the traffic by a small number of optical paths. To make the number of optical paths required to be added small, we add the optical paths where we can minimize the maximum link utilization.

The reconfiguration method performs the following steps.

1) Calculate all the utilizations of links. Denote the maximum link utilization as L.
2) If $L \leq Th$ the reconfiguration is over. Otherwise go 3.
3) For each node pair, calculate the maximum link utilization when the optical path between the pair is added.
4) Add the lightpath between the node pair that minimizes the maximum link utilization. Then go back to 1.

In the above steps, we calculate the routes over the virtual network by CSPF. To avoid a large overhead, when adding an optical path, we change only the routes of the flows passing the link whose utilization is larger than Th.

IV. Evaluation Result

A. Relation between Flow Inclusive Relation Modularity and the Number of Added Paths

Fig. 4 shows the relation between the flow inclusive relation modularity and the number of added paths. In this figure, the horizontal axis indicates FIRMs of each virtual network, and the vertical axis indicates the number of added paths of each virtual network. Each circle indicates the average number of added paths and each error bar indicates the 68.27% confidence interval of the added paths.

From Fig. 4, there are negative correlation between the FIRM and the number of added paths except for 2 virtual networks, which do not contain links with utilization larger than Th. This is because adding an optical path between a source and destination nodes of a flow whose corresponding function belongs to a module mitigates not only the congestions of the links passed by the flow but also the congestions of the links passed by the flows whose corresponding functions belong to the same module.

The modules in the flow inclusive relation graph correspond to the cohesion of flows passing the same links. If the FIR in the module is close, by adding an optical path between the source node and destination node of a flow in the module, we change not only the routes of the flow but also the routes of the other flows in the same module. As a result, by adding a small number of optical paths, all of the congestions may be mitigated. On the other hand, if the FIRM is low, a large number of optical paths are required to be added because the number of flows whose routes can be changed by adding optical paths is small.

B. Comparison with Betweenness

In this subsection, we compare the FIRM with the betweenness centrality. The betweenness centrality of a link indicates the probability that the congestion occur at the link. In this subsection, we investigate the maximum betweenness centrality among all links. Fig. 5 shows the relation between the maximum betweenness centrality and the number of optical paths required to be added. In this figure, the horizontal axis indicates the maximum betweenness centrality of each virtual network, and the vertical axis indicates the number of added paths of each virtual network.

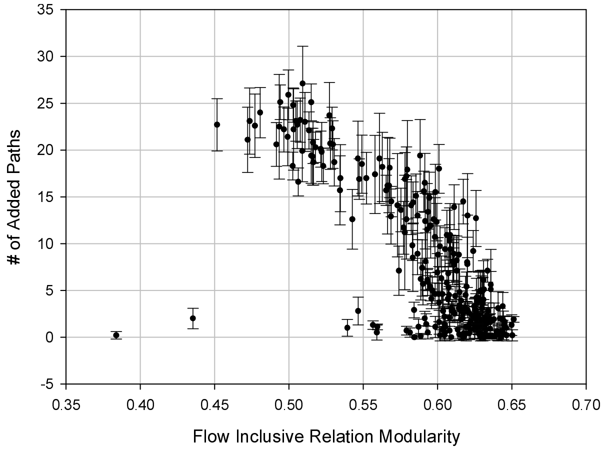


Figure 4: Flow Inclusive Relation Modularity and the Number of Added Paths

From Fig. 5, there are the positive relation between the maximum betweenness centrality and the number of optical paths required to be added. This is because the virtual network with the smaller maximum betweenness centrality has less possibility to congest. Therefore, in the virtual network with the small maximum betweenness centrality, few links are congested, and the number of optical paths required to be added is small.

However, the above discussion does not indicates that the virtual network with the smaller maximum betweenness centrality is adaptive to any traffic changes by adding only a small number of optical paths. If the traffic changes more significantly, the number of the congested links becomes large. Even in this case, the virtual network should accommodate traffic by adding only a small number of optical paths.

Therefore, we focus on virtual networks having the multiple congested links. In this comparison, we use the virtual networks whose maximum betweenness centralities are from 0.4 to 0.5. In this comparison, we also exclude the virtual networks with the larger maximum betweenness centralities than 0.5, because the virtual network with the large betweenness centrality should not be constructed, because it is too easy to be congested.

Fig. 6 shows the relations between the flow inclusive relation modularity or the maximum betweenness centrality and the number of optical paths required to be added, among the virtual networks whose maximum betweenness centralities are from 0.4 to 0.5. There is clearly observable negative correlation in Fig. 6a. On the other hand, several virtual networks have the similar maximum betweenness centralities, but have various numbers of added optical paths. This is because the maximum betweenness centrality identifies only the virtual network easy to congest and cannot identify the virtual networks that handle traffic changes by adding only a

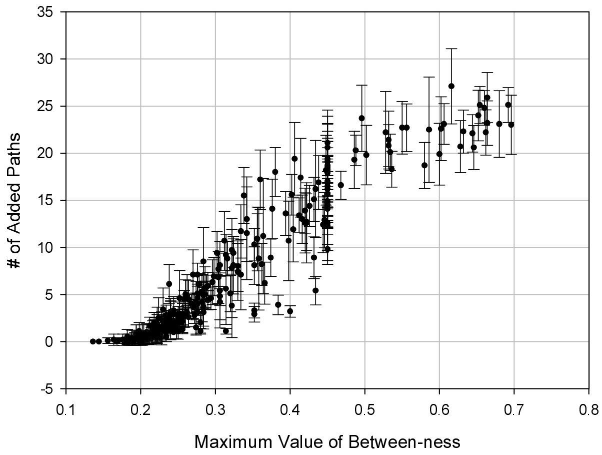


Figure 5: Maximum Betweenness and the Number of Added Paths

small number of optical paths.

C. Evaluation Result with False Negative Rate and False Positive Rate

To evaluate the relation between the flow inclusive modularity and the number of optical paths required to be added more clearly, we investigate the accuracy of the method to identify the virtual network that only a small number of additional optical paths to handle traffic changes. In this evaluation, we use the following method to identify the virtual networks based on the flow inclusive relation modularity or the maximum betweenness centrality. In the case of FIRM, we set a threshold to the FIRM, and identify a network with FIRM higher than the threshold as the network that requires only a small number of additional optical paths. In the case of maximum betweenness centrality, we identify a network with the maximum betweenness centrality lower than the threshold as the network that requires only a small number of additional paths.

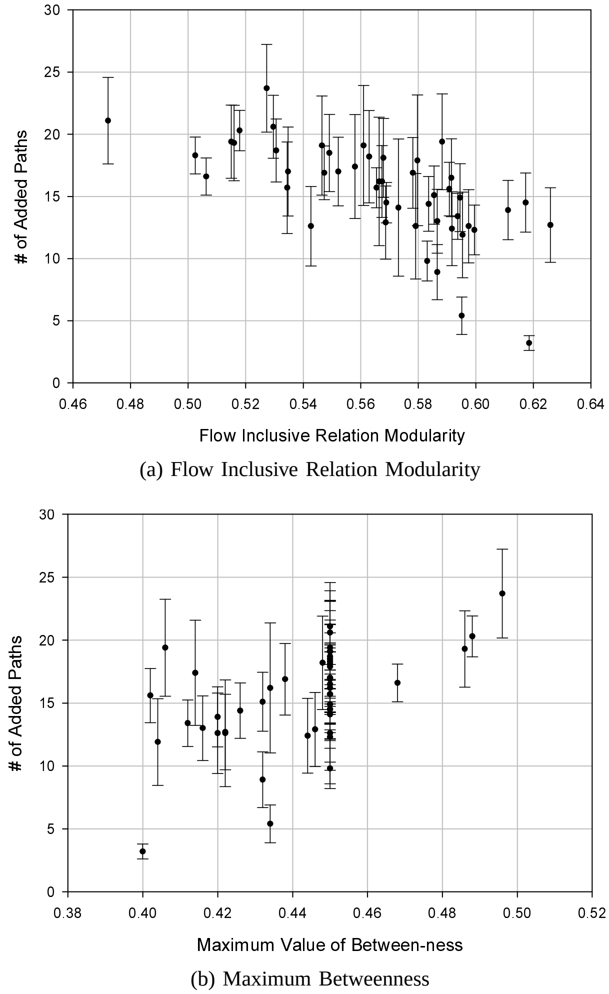
In this paper, we use the false negative rate (FNR) and the false positive rate (FPR) as metrics to evaluate the accuracy. FNR is defined by

$$FNR = {}^{m_{fn}}/_{m_p}, \tag{2}$$

where m_p is the number of virtual networks whose average numbers of additional paths are less than a certain threshold R_{goal}, and m_{fn} is the number of virtual networks which are identified as the virtual networks that require more than R_{th} additional optical paths but require only less than R_{th} additional optical paths. Similarly, the false positive rate (FPR) is defined by

$$FPR = {}^{m_{fp}}/_{m_n}, \tag{3}$$

where m_n is the number of virtual networks whose average numbers of additional paths are more than a certain threshold R_{th}, and m_{fp} is the number of virtual networks which are identified as the virtual networks that require less than R_{th} additional optical paths but require more than R_{th} additional optical paths.



(a) Flow Inclusive Relation Modularity

(b) Maximum Betweenness

Figure 6: Partial Comparison of Virtual Networks with Their Maximum Betweenness from 0.4 to 0.5

In this section, we investigated the relationship between FNR and FPR of virtual netwoks with the maximum betweenness centrality from 0.4 to 0.5, changing the threshold for each index. Fig. 7 shows the relationship between FNR and FPR in the case of FIRM. In Fig. 7, R_{th} is set to 10. The horizontal axis indicates the FNR, and the vertical axis indicates the FPR.

In the same manner, Fig. 8 shows the relationship between FNR and FPR in the case of maximum betweenness centrality. In Fig. 8, R_{th} is set to 10. The horizontal axis indicates the FNR, and the vertical axis indicates the FPR.

Comparing Fig. 7 with Fig. 8, the method using FIRM achieves both lower FNR and lower FPR at the same time. This result means that FIRM identifies virtual networks, which can accommodate significant traffic changes with less additional paths more accurately. Therefore, to construct the adaptive virtual network that can handle significant traffic changes by adding only a small number of optical paths, we

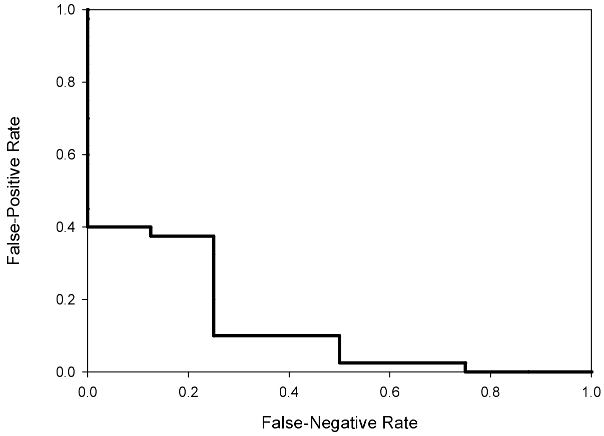


Figure 7: The Relationship between FNR and FPR Using FIRM

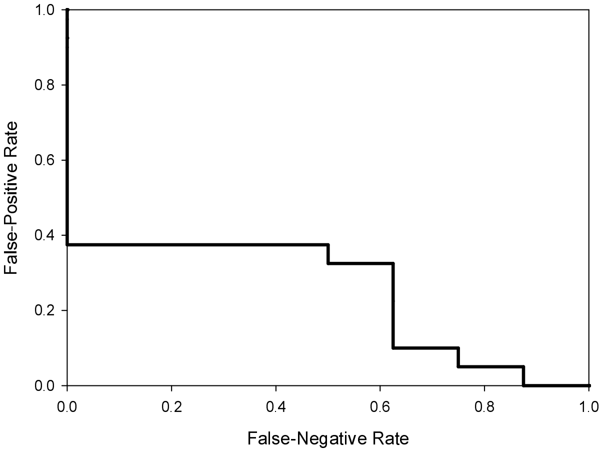


Figure 8: The Relationship between FNR and FPR Using Maximum Betweenness

should construct the virtual network whose FIRM is large.

V. Conclusion and Future Work

In this paper, we proposed the flow inclusive relation modularity (FIRM) as an index to identify the virtual network, which can handle significant traffic changes by reconfiguring only a small number of optical paths. Through the evaluation of relationship between the FIRM and the number of optical paths required to be added, we clarified that the FIRM identifies the virtual networks, which can accommodate any traffic changes with a small number of additional paths. Our future work includes a method to construct virtual networks with the large FIRM. For example, if the virtual network has low adaptability and the link utilization are increasing, some optical paths should be added to increase the FIRM so as to increase the adaptability to future traffic increases. On the other hand, if the link utilization is sufficiently low, some optical paths should be deleted to release the resources so as to be used by the future reconfiguration. When deleting the optical paths, by considering the FIRM, we keep the adaptability to traffic changes.

Acknowledgement

This work was partially supported by JSPS Grant-in-Aid for Young Scientists (B)23700077.

References

[1] Ministry of Internal Affairs and Communications, “2012 White Paper Information and Communications in Japan,” ” Jul. 2012.

[2] B. Ramamurthy and A. Ramakrishnan, “Virtual topology reconfiguration of wavelength-routed optical WDM networks,” in Proceedings of Globecom, vol. 2, Nov. 2000, pp. 1269 –1275.

[3] Y. Ohsita et al., “Gradually reconfiguring virtual network topologies based on estimated traffic matrices,” IEEE/ACM Transactions on Networking, vol. 18, no. 1, Feb. 2010, pp. 177 –189.

[4] A. Gençata and B. Mukherjee, “Virtual-topology adaptation for WDM mesh networks under dynamic traffic,” IEEE/ACM Transactions on Networking, vol. 11, Apr. 2003, pp. 236–247.

[5] L. C. Freeman, “A set of measures of centrality based on betweenness,” Sociometry, vol. 40, no. 1, Mar. 1977, pp. 35–41.

[6] A. Tizghadam and A. Leon-Garcia, “Autonomic traffic engineering for network robustness,” IEEE Journal on Selected Areas in Communications, vol. 28, Jan. 2010, pp. 39–50.

[7] A. Bigdeli, A. Tizghadam, and A. Leon-Garcia, “Comparison of network criticality, algebraic connectivity, and other graph metrics,” in Proceedings of SIMPLEX. ACM, Jul. 2009, pp. 4:1–4:6.

[8] H. Lipson, J. B. Pollack, and N. P. Suh, “On the origin of modular variation,” Evolution, vol. 56, no. 8, Aug. 2002, pp. 1549–1556.

[9] M. E. J. Newman, “Modularity and community structure in networks,” Proceedings of the National Academy of Sciences, vol. 103, no. 23, Jun. 2006, pp. 8577–8582.

[10] N. Hidaka, “A topology design method for sustainable information networks,” Master’s thesis, Graduate School of Information Science and Technology, Osaka University, Feb. 2009.

[11] I. Antoniou, V. Ivanov, V. V. Ivanov, and P. Zrelov, “On the log-normal distribution of network traffic,” Physica D: Nonlinear Phenomena, vol. 167, no. 1-2, Jul. 2002, pp. 72 – 85.

Comparison of Contemporary Solutions for High Speed Data Transport on WAN 10 Gbit/s Connections

Dmitry Kachan, Eduard Siemens
Department of Electrical, Mechanical and Industrial Engineering
Anhalt University of Applied Sciences
Köthen, Germany
{d.kachan, e.siemens}@emw.hs-anhalt.de

Vyacheslav Shuvalov
Department of Transmission of Discrete Data and Metrology
Siberian State University of Telecommunications and Information Sciences
Novosibirsk, Russia
shvp04@mail.ru

Abstract **– This work compares commercial fast data transport approaches through 10 Gbit/s Wide Area Network (WAN). Common solutions, such as File Transport Protocol (FTP) based on TCP/IP stack, are being increasingly replaced by modern protocols based on more efficient stacks. To assess the capabilities of current applications for fast data transport, the following commercial solutions were investigated: *Velocity* – a data transport application of BitSpeed LLC; *TIXstream* – a data transport application of Tixel GmbH; *FileCatalyst Direct* – a data transport application of Unlimi-Tech Software Inc; *Catapult Server* – a data transport application of XDT PTY LTD; *ExpeDat* – a commercial data transport solution of Data Expedition, Inc. The goal of this work is to test solutions under equal network conditions and thus compare transmission performance of recent proprietary alternatives for FTP/TCP within 10 Gigabit/s networks where there are high latencies and packet loss in WANs. This research focuses on a comparison of approaches using intuitive parameters such as data rate and duration of transmission. The comparison has revealed that of all investigated solutions *TIXstream* achieves maximum link utilization in presence of lightweight impairments. The most stable results were achieved using *FC Direct*. *ExpeDat* shows the most accurate output.**

Keywords-high-speed data transport; transport protocol; WAN acceleration, Managed File Transport.

I. Introduction

The growing demand for the fast exchange of huge amounts of data between distant locations has led to the emergence of many new commercial data transport solutions that promise to transport huge amounts of data many times faster than conventional FTP/TCP solutions. Currently, most common solutions for reliable data transport in IP networks are based on the TCP protocol, which was developed in 1970s. A number of papers describe how TCP, with some tuning, can perform reasonably on Local Area Networks (LAN) with a high available bandwidth [1]. However, it is well known that TCP has a very limited performance when used in long distance networks with a high bandwidth - called “Long Fat Pipe Network (LFN)” [2]. For example, a test with *iperf* using the topology described in Fig 2 on an end-to-end 10 Gbit/s link with a 50 ms round trip time delay (RTT) and in the presence of at least 0.1% packet loss rate shows a data rate of about 40 Mbit/s. Even after increasing socket buffers and windows sizes to 128 MiBytes, the performance of TCP and, accordingly, of most of solutions based on it (SCP, *rsync*, FTP), does not reach more than 60 Mbit/s. Comparable measurements of TCP over 10 Gbit/s were also performed by Wu et al. [1]. In their work, the authors obtained a data rate of less than 1 Gbit/s even in the presence of a loss rate of 0.001% and an RTT of 120 ms. They show a significantly decreasing trend in a data rate with growing packet loss rate. Another example of TCP weaknesses over long distances is described by Armbrust et al. in [3], where the transmission of 10 TBytes of data from Berkeley, California to Seattle, Washington via a common TCP connection takes about 45 days, whereas transmission of 10 TBytes hard drive takes less than one day. A similar solution is described by Armbrust et al. in [4]. Nevertheless, many scenarios of remote collaboration (e.g cloud computing) demand data transport with maximum synchronization times for huge data sets from a few minutes to hours. As a result, many large companies, for which the exchange of huge amounts of data is often critical, avoid using legacy TCP-based transport solutions and either prefer commercial high speed approaches based on both TCP and UDP transport protocols [5] [6] or develop their own solutions based on an open source fast protocol stacks such as UDT [7] and RBUDP [8].

II. Related work

The main goal of our work is to assess the capabilities of transport solutions in a 10 Gbit/s network. Of interest is the maximal possible end-to-end application data rate on such networks in the presence of impairments such as packet losses and high round-trip times. Currently, there are a few different performance measurements that have been used to assess these impairments in open source and freeware solutions. For example, in [9] Grossman et al. present the performance evaluation of UDT [7] through a 10 Gbit/s network. The article shows how using UDT and in the presence of 116 ms of RTT, this network has a maximum throughput of 4.5 Gbit/s within a single data stream. Two parallel streams achieve together about 5 Gbit/s and within 8 parallel streams about 6.6 Gbit/s are achieved. Further, a performance result for data transmission using RBUDP was presented at the CineGrid 3rd Annual International Workshop [10]. While the disk access speed limited the data transport speed to 3.5 Gbit/s, on the link from Amsterdam to San Diego only 1.2 Gbit/s was reached. The distance of that path is about 10 000 km through optic fiber, which corresponds to about 100 ms of RTT.

Most other data transport performance results are presented for 1 Gbit/s networks e.g. three rate based transport protocols have been evaluated by Wu et al. in [11]: RBUDP, SABUL and GTP. The overall data rate of applications based on these protocols was compared for all three protocols and for "standard unturned TCP". The experiment was performed on a real network in the presence of 58 ms of RTT and a loss rate of less than 0.1%. The results showed that all solutions utilize the 1 Gbit/s link approximately 90%. These test results show that for open source data transfer solutions, even those using parallel streams, it is quite hard to achieve full, or even close to full, utilization of 10 Gbit/s links.

For commercial closed source solutions, the situation differs significantly. There are several published performance results of commercially available solutions, provided by the manufacturers themselves: *Velocity* [12], *TIXstream* [13], *FC Direct* [14] and *Catapult Server* [15] – all of whom report perfect results. However these results are mainly providing commercial information to attract potential customers and the investigative conditions vary. To overcome this deficit, the main idea behind our work is to place all investigated solutions under equal conditions within the same environment.

III. BACKGROUND

For IP networks, packet loss behavior depends on many factors such as quality of transmission media, CPU performance of intermediate network devices, presence of cross traffic etc. It is therefore impossible to use one universal value of packet losses for all cases. The best way to assess the approximate values of packet losses is through empirical measurements. In [16] V. Paxson discusses the heterogeneity of packet losses and shows that, even in 1994, the value of packet loss rate in experiments between 35 sites in 9 countries was about 2.7%. He also shows that, within one year, the value of packet losses increased up to 5%. Probably, such packet loss values are not relevant to the current Internet; however the author pointed out that distribution of packet losses is not uniform. Thus, for some connections, ACK packet loss was not observed at all. Nevertheless, relative values of all lost IP packets in both directions for all experiments were approximately equal. Recent views on the packet loss ratio are presented by Wang et al. in [17]. In this research, tests were made across 6 continents between 147 countries, involving about 10 000 different paths. The authors show that across all continents for more than 70% of continental connections, packet loss rate is less than 1%, in Europe and North America this value is even on about 90% of connections. The authors also highlighted that for intercontinental connections, packet loss value in general is lower than for intra-continental – across the entire world, packet loss rate is lower than 1% for about 75% of the connections.

In [18] Settlemyer et al. use a hardware emulator to emulate 10 Gbit/s paths, and they compare throughput measurement results of the emulated paths with real ones. The maximal RTT of a real link used in the research is 171 ms. The authors have shown that differences between profiles of both kinds of paths - emulated and real ones - are not significant, and they conclude that using emulated infrastructure is a much less expensive way to generate robust physical throughput.

IV. TESTBED TOPOLOGY DESCRIPTION

In this work, the following solutions have been investigated: *Velocity*, *TIXstream*, *FileCatalyst Direct*, *Catapult Server*, *ExpeDat*. Manufactures of all these solutions claim that their transport solutions are able to handle data transmission via WAN in the most efficient way.

Since all these solutions are commercial and closed source, it was necessary to get in touch with the support team of each manufacturer for both obtaining trial licenses of their products and consulting them about achieved results. Unfortunately, not all manufactures were interested in such investigations. Thus, for example, it would have been interesting to test Aspera's solution for fast data transport. However we received no answer from this vendor.

To avoid unexpected inaccuracies, the scheme of test topology is kept simple. Fig. 1 presents the typology. The core of the test environment was the WAN emulator *Apposite Netropy 10G* [19], which allows the emulation of WANs under various conditions such as packet delay, packet loss rate and jitter in different variations, with an accuracy of about 20 ns. By comparison, software emulators, such as *NetEm*, provide an accuracy of about tens of milliseconds and this value is greatly dependent on the hardware and operating system [20]. Moreover, software emulators are very limited in their maximum achievable data rates. *Apposite 10G*, for example, enables a transmission through Ethernet traffic with an overall throughput of up to 21 Gbit/s on both copper and fiber optic links.

The testbed topology used here contains two servers, connected via the 10 Gbit/s Ethernet switch Extreme Networks Summit x650 and via the WAN Emulator. The typology was implemented by means of fiber optics with a 10 Gbit/s bandwidth, see Fig. 2.

There is no background traffic on the path since this investigation focuses on the pure applications' performance, not on the fairness aspects of the protocols. The setup corresponds to the case when a L2-VPN is used for big data transmission and another application's traffic is isolated by means of QoS.

Each server is equipped as follows:

- *CPU: Intel Xeon X5690 @3.47GHz;*
- *RAM: 42 GiBytes (speed 3466 MHz);*
- *OS: Linux CentOS 6.3;*
- *NIC: Chelsio Communications Inc T420-CR, 10Gbit/s.*

Operating system socket buffers were extended up to:

- */proc/sys/core/net/wmem_max – 64MiBytes*
- */proc/sys/core/net/rmem_max – 64MiBytes*

The MTU size of all network devices along the path was set to 8900 Bytes.

For sending and receiving data with a rate of 10 Gbit/s, it is necessary to have a storage device that can read on the sender side and write on the receiver side with a sustained rate not less than 1 250 MByte/s (corresponding to 10 Gbit/s). Off-the-shelf hard drives provide read and write rates of up to 100 MByte/s, so in the investigated case, data transfer rate would have been limited by the hard drives. To circumvent this limitation, storage systems such as RAID arrays with write/read rates not less than the expected transport rate must be used.

In the presented experiments, both storage write and read bottlenecks and inefficient file access implementations were avoided by using a RAM-based file system on both servers. In comparison to common hard drives, the read rate of *RAMdisk*, as obtained in several test runs during these investigations, was not less than 4 500 MiBytes/s; the write rate of *RAMdisk* was not less than 3 000 MiBytes/s. Therefore, the servers used for tests were equipped with 42 GiBytes of RAM onboard, but due to the operating system's RAM requirements, it was only possible to use 30 GiBytes of space on *RAMdisk* for test purposes.

Under ideal conditions, a transmission of 30 GiBytes through the network with a bandwidth of 10 Gbit/s without impairments, as explained in (1), should take about 26 seconds.

$$T = \frac{S}{R_i} = \frac{30 \times 1024^3 \times 8 / 10^6}{10 \times 10^3} = 25.76\ s, \qquad (1)$$

where T – time of transmission; S – Size of data, R_i – ideal data rate.

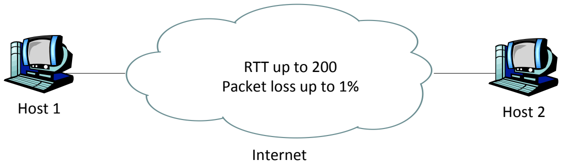


Figure 1. Logical view of topology

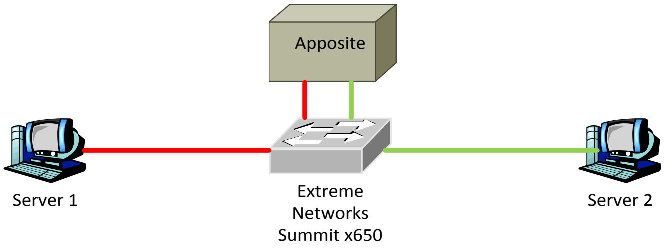


Figure 2. Technical view of topology

However, this calculation disregards L2-L4 headers along with some proprietary protocol headers and the overhead for connection management and retransmissions.

So, under real conditions, for some packet overhead and retransmission handling, each solution needs a certain amount of time for connection initialization and the releasing of the network path. Besides this, in high-performance implementations, initialization of the protocol stacks and the internal buffers often takes a significant amount of time, which is also investigated during this research.

V. EXPERIMENTAL RESULTS

The experiment on each data transport solution under consideration consists of 25 consecutive tests. Each test comprises the transfer of a 30 GiBytes file from one server to another through the network emulator. The RTT latency range is varied from 0 to 200 ms in steps of 50 ms and the packet loss rate takes the values 0; 0.1; 0.3; 0,5 and 1 %. Since one km of fiber optics delays the signal by about 5 µs, the maximum RTT in this test corresponds to 20 000 km of fiber channel in both the forward and the backward directions. The RTT is configured symmetrically across the forward and backward paths of the emulator; thus 200 ms of RTT would delay data by 100 ms and acknowledgments or other control information in the backward direction by another 100 ms. The packet losses are randomly injected according to a normal distribution, whereby the set loss ratio is applied to both forward and backward direction. Such packet loss behavior is easier to reproduce, and it is more complicated for protocols to handle than typical packet loss behavior on the Internet [16]. An attempt was made to configure each solution so that the maximum possible data rate and the minimum possible overall transmission time were achieved. The tuning of the operating system and the configuring of parameters are described below for each solution. All the tests were repeated 4 times to avoid inaccuracies, and the best result of each series is presented on the plots.

The results of each test contain two parameters: data rate and transfer duration. The first parameter is average data rate i.e. the average speed of data transportation shown by the application during the experiment. The second parameter is independent of the solution output and represents the time interval. This time interval was collected by means of the operating system and shows the period of time between the launching of the send command and the time of completion of this command. This time interval contains not only the time of actual data transmission but also the time for service and retransmission overhead.

A. *Velocity*

This solution was developed in the USA. It is a TCP-based file transfer application, and, according to the vendor's web site, it allows the available path capacity to be fully utilized. *Velocity ASC* is also available with on-the-fly data encryption of up to 24 Gbit/s and AES encryption of up to 1 600 Mbit/s. The supported platforms are *Windows*, *Mac OSX*, *Linux* and *Solaris*. According to the user manual, this solution automatically adapts its parameters to network conditions and chooses the optimal parameters for data transmission. Fig. 3 shows the behavior of the transport rate. The results of tests in the presence of delays of more than 0.1 % are not shown since the data rate here was lower than 100 Mbit/s.

Increasing latencies do not significantly affect *Velocity*'s data rate behavior, slowing it down to only 8 Gbit/s. The solution performs reasonably in the presence of small packet loss rates without any emulated delay (back-to-back RTT latency in the testbed is about 0.15 ms). Thus it achieves a data rate of 9 Gbit/s in the presence of 0.1 % of packet loss, and this value decreases down to 500 Mbit/s with a packet loss of 1%. However, this configuration does not correspond to the situation in Wide Area Networks. In the presence of 0.1 % packet loss and at least 50 ms RTT, the data rate is reduced to 250 Mbit/s.

By default *Velocity* uses multi-streaming TCP. It opens 7 TCP sockets on every single test on each side. When the number of streams is manually set to 1, the data rate in presence of 200 ms RTT without packet loss is about 2.2 Gbit/s. The transfer durations of the solution are shown in Fig. 4. The numbers on the plot are obtained for two cases: without latency and with a latency of 200 ms. A result worth noting was obtained at a loss rate of 0.1% and an RTT of 0 ms. Under these conditions, the data rate in the presence of losses is, with 800 Mbits/s, less than the value without loss rate. However, the transfer duration in the latter case is higher by only 0.1 ms. This behavior was observed in several experiments.

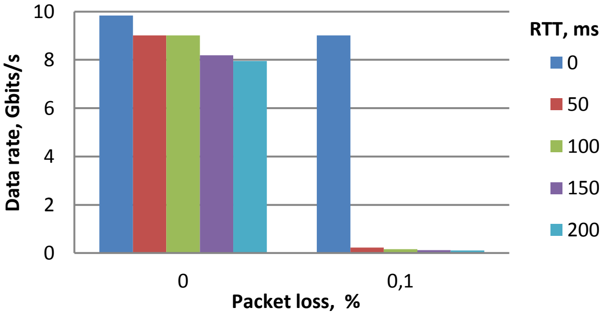


Figure 3. Behavior of data rate of *Velocity*

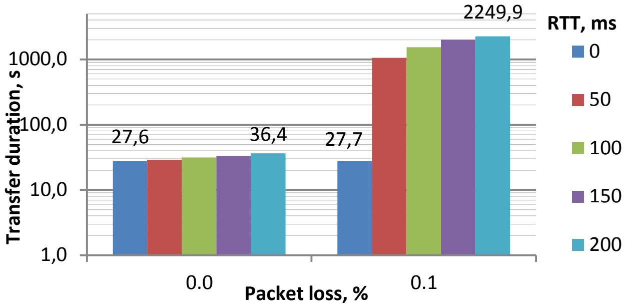


Figure 4. Data transfer duration of Velocity

B. *TIXstream*

This transfer engine has been developed by Tixel GmbH, Germany, which spun off from Technicolor Corporate Research in 2010. The core of *TIXstream* is Tixel's proprietary Reliable WAN Transfer Protocol (RWTP) [21], which provides high-performance data transmission between two hosts in the network using only one UDP-socket on each host. The application works under Linux OS.

TIXstream 3.0 (the latest version) provides up to 20 Gbit/s end-to-end performance. It has a platform-independent web-based user interface for the management of data transmission between remote SAN- and NAS systems. *TIXstream* also provides on-the-fly AES-256 encryption of data without any effect on data rate [22]. *TIXstream* has a peer-to-peer architecture and uses one TCP socket for control communication and one UDP socket for data transmission connection on both sides.

Parameters of application:

- *RWTP Buffer size – 4362076160 Bytes (4 GiBytes)*
- *MSS = 8800 Bytes*
- *Receiver buffer size (on both sides) = 1073741824 Bytes (1GiByte)*
- *Sender buffer size (on both sides) = 1073741824 Bytes (1GiByte)*

The behavior of *TIXstream*'s data rate as a function of network impairments is shown in Fig. 5.

There is no visibly decreasing effect on data rate behavior till 100 ms RTT and till 0.3 % of packet loss. The

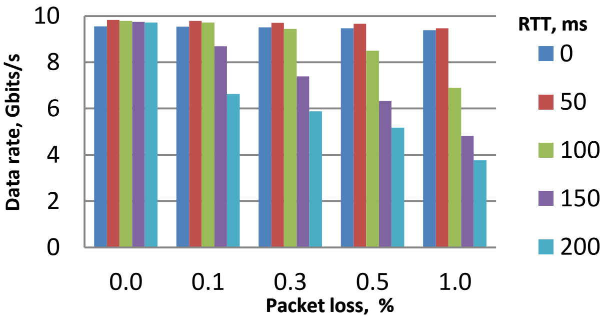


Figure 5. Behavior of data rate of *TIXstream*

solution achieves not less than 9.7 Gbit/s (97% of capacity) with these impairments. With higher delays in the presence of heavy packet losses, the figure shows decreasing data rates down to 3 750 Mbit/s, as on a path with 200 ms of RTT and 1% of packet losses. However, with modest impairments that correspond to fairly normal WAN links, for example RTT 150 ms and packet loss rate 0.1 %, *TIXstream* achieves a data rate of about 8 700 Mbit/s; an 87 % utilization of a 10 Gbit/s link. It is worth noting that in the presence of 50 ms of latency, *TIXstream* performs better than without any latency for all values of loss rate. This behavior was found in several experiments.

Fig. 6 shows that the transfer duration quite accurately corresponds to the behavior of the data rates. However, the theoretically minimum time of transmission calculated in Section IV, with a data rate of 8700 Mbit/s, is 29.62 s versus the 37.25 s that was measured in the experiment. These 7.63 seconds were spent on connection initialization, and the establishing and releasing of the control channel. Since no packet loss shall occur on this link, time for packet retransmission shall be neglected.

C. *FileCatalyst Direct*

FileCalatyst Direct was developed by Unlimi-Tech Software Inc., a Canadian based company. Like *TIXstream*, it transmits data via UDP and implements packet loss management, rate and congestion control in the user layer. The application obeys a client-server architecture and the solution operates under *Windows*, *Mac OSX*, *Linux* and *Solaris* operating systems. The data sheet on the vendor's website shows that this solution provides data rates of up to 10 Gbit/s [23] and that there is an option to use AES encryption for secure transmission. *FC Direct* provides both, graphical and command line user interfaces for server and client applications.

Parameters of application:

- *Start rate = 9000000 (9Gbit/s)*
- *MSS = 8800 Bytes*
- *Buffers* = 3840000000 *Bytes (3,58 GiBytes)*
- *Number of send sockets = 10*
- *Number of receiver sockets = 4*

As shown on Fig. 7, *FC Direct* achieves 90 to 94 % link utilization, even under high network impairments. Data rate behavior is fairly immune to growing latency and packet loss ratio. The data rate of *FC Direct* shows values between 9 Gbit/s and 9.4 Gbit/s for all the tests. During the tests, the Linux system monitor reveals that each data transmission opens 10 UDP sockets on the sender side and 4 UDP sockets on the receiver side and one TCP socket on each side. In this mode, maximal data rates can be achieved. Data packets from ten sender sockets are not uniformly distributed over all four receiver sockets, but according to a special proprietary distribution rule. The vendor does not call it multi-streaming but "more intelligent resource management". However, with this behavior, significant firewall transversal issues are to be expected.

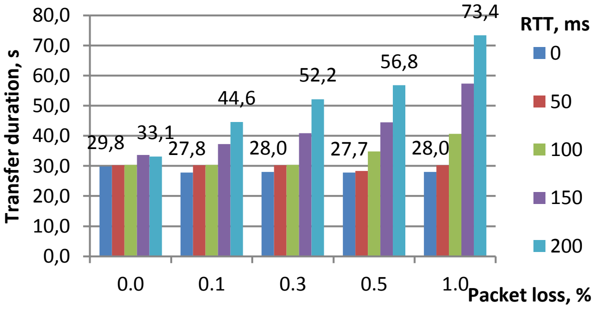


Figure 6. Data transfer duration of *TIXstream*

The distribution of session time durations showed on Fig. 8 has slightly monotonically increasing behavior with rising latencies.

D. *Catapult Server*

The *Catapult Server* is TCP-based and was developed by XDT PTY LTD, Australia. This solution follows a client-server architecture and, according to the vendor's web site, provides up to 8 Gbit/s on the 10 Gbit/s link. The solution functions under the *Windows*, *MAC OSX* and *Linux* operating systems. The vendor positions the solution as a high data rate transmitting tool for networks with a high level of latency but without any packet losses. To prepare the operating system for high speed transmissions, the vendor suggests using a shell script to change network parameters. By default, this script extends the TCP buffers to 32 Mbytes. However, for our tests, the 64 Mbytes setting was chosen since better performance had been reached with

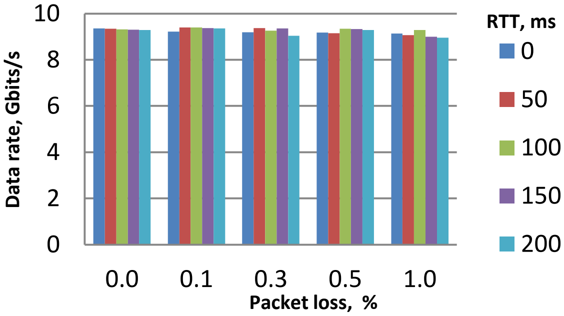


Figure 7. Behavior of send rate of FC Direct

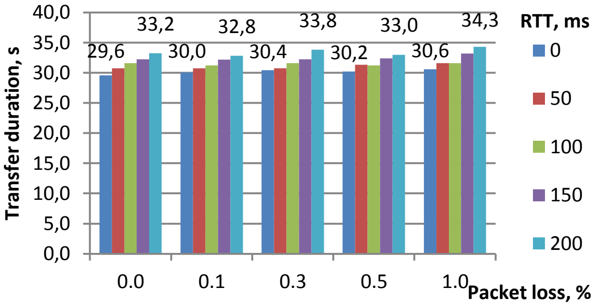


Figure 8. Data transfer duration of FC Direct

this setting. The script changes are:

- *tcp_congestion_control=htcp*
- *net.ipv4.tcp_rmem=4096 87380 67108864*
- *net.ipv4.tcp_wmem=4096 65536 67108864*
- *net.ipv4.tcp_no_metrics_save=1*
- *net.core.netdev_max_backlog=250000*
- *net.core.rmem_max=67108864*
- *net.core.wmem_max=67108864*

To improve the behavior of this solution in the presence of packet losses, the manufacturer’s support team also suggested applying the following configurations:

- *net.ipv4.tcp_timestamps=1*
- *net.ipv4.tcp_sack=1*

Note that the command line client of XDT - *sling shot copy*, which was used for the tests, shows the data rate as GB/s, probably it means GiByte/s. Furthermore the value 1.1 GB/s immediately follows the value 1.0 GB/s, without any intermediate values. However, the solution shows a transfer duration with an accuracy of up to milliseconds. Therefore, the data rate was calculated as

$$R_{XDT} = \frac{S}{T_o}, \tag{2}$$

whereby R_{XDT}– is the data rate of XDT, which is used for result presentation; S – data size (30 GiByte), T_o – transfer duration from the output of client application.

Fig. 9 represents the data rate of *Catapult Server* dependent on network impairments. The presence of packet losses on the link makes the transmission ineffective, so the data rate is reduced to less than 100 Mbit/s. However, in the presence of 150 ms RTT without packet loss, transmission is about 8300 Mbit/s

Fig. 10 shows the transfer durations for Catapult technology.

E. ExpeDat

ExpeDat is a UDP-based data transport solution developed by Data Expedition Inc., USA. The core of this application comprises the Multipurpose Transaction Protocol (MTP) [24], developed by the founder of Data Expedition. *ExpeDat* supports *Windows*, *Mac OSX*, *Linux* / *Solaris*, *NetBSD* / *FreeBSD*, *AIX* and *HP-UX* platforms. According to the company’s web site, *ExpeDat* allows transmission of data with 100 % utilization of allocated bandwidth and in the presence of on-the-fly AES encryption [25]. It implements the transport protocol logics on a UDP channel, and uses a single UDP socket on each side of the connection for both data transmission and control information.

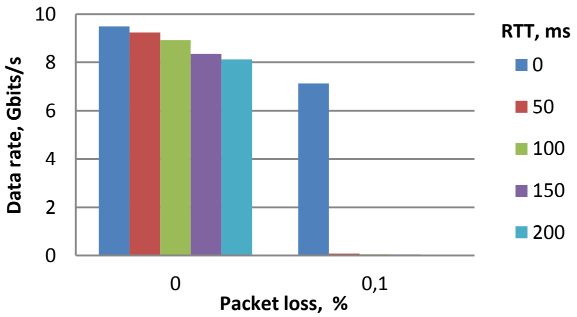


Figure 9. Behavior of data rate of XDT Catapult

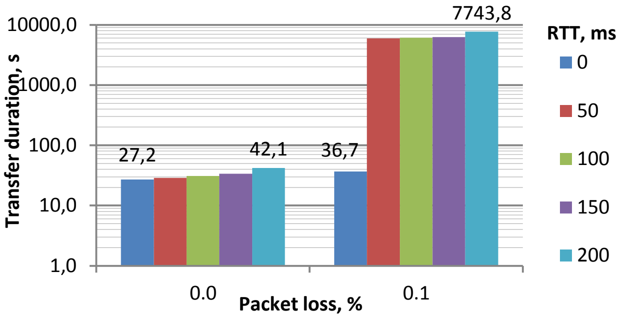


Figure 10. Data transfer duration of XDT Catapult

Though the product web site [25] claims that the solution has “zero-config installation”, the significant increase of data rate, namely from 2 Gbit/s up to 9 Gbit/s, even without impairments (RTT=0 ms, packet loss = 0%), was obtained only after application of configuration changes as follows:

- *MSS – 8192 Bytes*
- *Environment variable MTP_NOCHECKSUM=1*

With high values of packet loss on the channel, the higher results were achieved using the following option on the command line:

- *-N 25*

The use of this option shows that heavy packet loss rate is introduced in a channel.

In Fig. 11, the data rate values of *ExpeDat* tests are presented. The plot shows that network latencies lead to a much higher reduction of the transmission rate than packet loss.

The distribution of transfer times is presented in Fig. 12.

VI. COMPARISON OF THE SOLUTIONS

Since not all of the investigated solutions perform well in the presence of heavy packet losses, the comparison of data rates was split into two stages. The first stage is

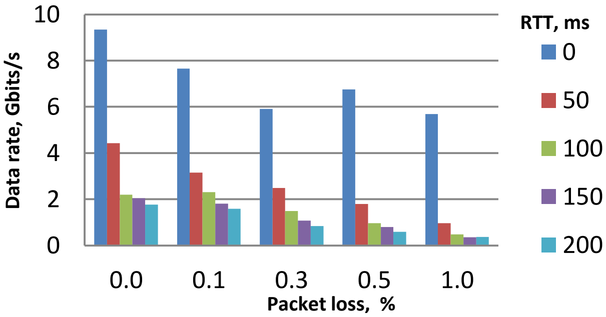


Figure 11. Behavior of data rate of *ExpeDat*

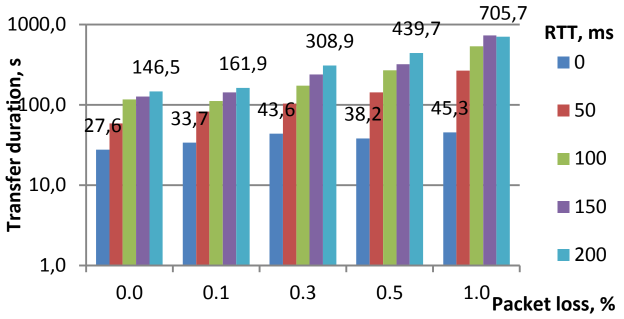


Figure 12. Data transfer duration of *ExpeDat*

dedicated to a comparison of all presented solutions on the networks with different packet latencies and without any packet loss. In the second stage, the solutions are compared under harder conditions for terrestrial networks - with packet loss of 1% and the whole range of investigated RTT. Only solutions that show a data rate higher than 1% from maximal channel capacity (100 Mbit/s) have been considered in the second stage.

A comparison between the distribution of transfer duration without packet losses and the ideal value shows which solution spends more time on service needs such as the initialization and releasing of a channel.

A further comparison shows the discrepancy between the actual time of transmission and the calculated time from the output of all solutions. This analysis is also split into two stages as described above and shows how the values from the output of the particular solution correspond to reality.

Fig 13 shows a consolidated diagram of transmission data rates of all tested solutions in the presence of increasing latency and without any packet loss on the path. The first set of bars shows how fast a large set of data can be transmitted in a back-to-back connection. For this case, the highest result was achieved by *Velocity*. However all solutions showed results of not less than 9.3 Gbit/s. With increased latencies, *Velocity* performs worse, and of all the remaining cases, *TIXstream* shows the best performance with up to 9.8 Gbit/s (98% channel utilization) without any significant decrease at higher RTTs. *FC Direct* also shows very stable results. For all presented impairments, its data rate lies between 9.2 and 9.3 Gbit/s. All other solutions show decreasing data rates on increasing round-trip-times.

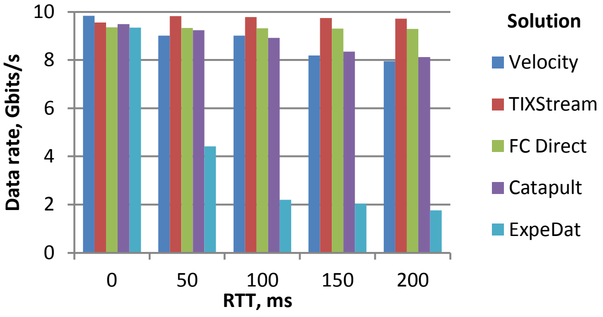


Figure 13. Comparison of data rates of investigated solutions; packet loss = 0

TCP-based solutions obviously cannot cope efficiently with the presence of packet losses on the path. Although for all solutions except *Velocity*, the support teams of the respective manufacturers were involved, those solutions did not provide adequate results in the presence of packet loss. Fig. 14 represents the behavior of solutions of stage two in the presence of 1% of packet loss.

With an RTT of 0 and 50 ms, *TIXstream* shows the best results. However, starting at 100 ms, throughput is decreased whereas *FC Direct* shows nearly constant data rates. The *ExpeDat* data rate abruptly decreases down to 5.7 Gbit/s on zero-delay links, and with increasing latencies *ExpeDat*'s results are lower than 1 Gbit/s.

As pointed out in Section IV, the theoretical minimum transfer duration for the transport of 30 GiBytes of data via a 10 Gbit/s WAN is 25.76 seconds. Fig. 15 compares, for each solution, the ideal value with the lowest transfer durations obtained during the experiments.

The minimum transfer duration was achieved by XDT *Catapult server*. The time is only 1.4 seconds longer than the theoretical minimum. This means that XDT *Catapult* initializes the software stack along with protocol buffers, and establishes and closes the connection within less than 1.4 seconds. The time overhead of *Velocity*, *TIXstream* and *ExpeDat* is slightly higher but still less than 2 seconds. The worst result was obtained when using *FC Direct*: it needed about 3.8 s. for ramping-up and finishing the application.

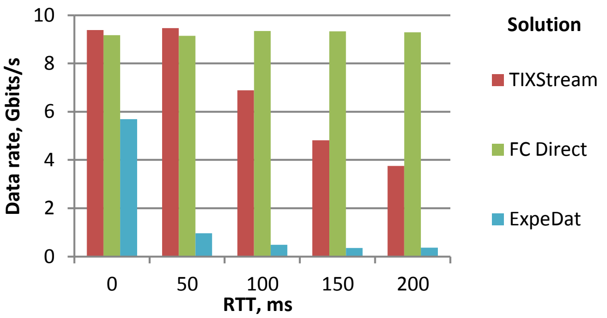


Figure 14. Comparison of data rate of tested solutions; packet loss = 1%

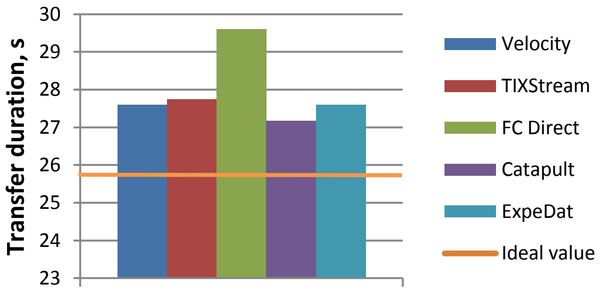


Figure 15. Comparison of transfer durations with theoretical minimum

Also of interest is the accuracy of the performance outputs of the transport solutions. During the experiment, the actual data rate was obtained from the output of the running application, and program run time was also measured by means of the operating system. Transfer durations with a transmission of 30 GiByte with a data rate from output are calculated as

$$T_t = \frac{S,[Bits]}{R_o,[\frac{Bits}{s}]} \qquad (3)$$

where T_t – is the calculated transfer duration; R_o– data rate from output; S – data size (30 GiByte).

The differences between calculated transfer durations and real program run time for the tests performed without packet loses are presented in Fig. 16. The discrepancy of the values generally has an increasing trend at higher latencies. *Velocity* showed the lowest value of discrepancy along the tests without packet loss. The second TCP-based solution – XDT *Catapult* - shows good results on RTTs below 200 ms. However, with 200 ms RTT, this solution shows the worst of all results. *ExpeDat* almost always has the lowest discrepancy values for all cases. In the presence of RTTs of 100 ms and 150 ms without loss rate, the solution showed negative result, meaning that the actual transfer duration was lower than the calculated one. It is evident that the output of some solutions show the average achieved data rate including service processes such as connecting, initializing and releasing the link, and some of the solutions show the average data rate of the transmission process only.

Similar to Fig. 16 but with a packet loss of 1%, Fig. 17 shows the differences of calculated transfer durations and real program run time. A comparison of these two figures shows that FC Direct has almost the same discrepancies for all RTT cases except for the 200 ms and 150 cases, where the discrepancy is higher in the presence of packet loss than without it. The results of *TIXstream* have a decreasing trend and *ExpeDat* shows again the lowest values – the actual times of transmission are almost equal to the calculated ones.

VII. Conclusion

This work compares the state of the art of commercial solutions for reliable fast data transport via 10 Gbit/s WAN IP networks in the presence of high delays and varying packet loss rates. The main problem of such research is that the vendor companies usually hide the technology used for the accelerated data transport. The protocol used in *ExpeDat* solution – MTP - is covered by some US patents. However this does not mean that *ExpeDat* does not use any algorithms besides the ones described in those patents. The only independent method to assess these commercial solutions is to externally observe the solutions during tests under well-defined conditions.

All investigated solutions position themselves as reliable high-speed transfer applications designed to provide alternatives to FTP/TCP and overcoming the pure TCP performance on 10 Gbit/s-WAN connections by orders of magnitude. Two of them, *Velocity* and XDT *Catapult*, exploit the TCP stack of the *Linux* OS and the rest - *FC Direct*, *TIXstream* and *ExpeDat* use UDP sockets and implement the protocol logics in the user-level.

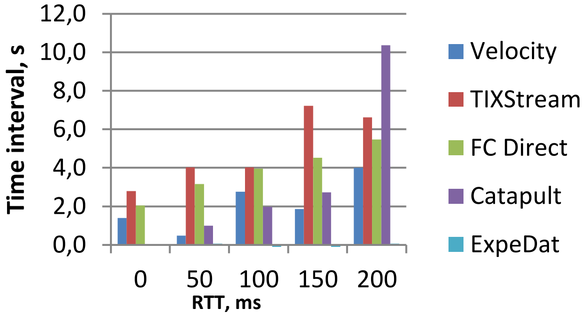


Figure 16. Difference between calculated and actual transfer durations; Packet loss=0

The results obtained show that solutions based on TCP inherit its native problems on 10 Gbit/s links – a significant decrease of data rate down to 1% of the link capacity in the presence of packet loss on the path. The commercial solutions achieve a higher speed by increasing TCP window size or by establishing multiple parallel TCP streams. However, the experiments show that this solution only works on links without any packet loss. However, even the known STCP [26] on WAN networks with a low loss rate show a reasonable result of about 5 Gbit/s [1]. Although in that paper, the authors tested pure protocol performance, their results show that it is possible to achieve good results by only tuning the TCP on such networks.

UDP-based solutions show a good utilization of a 10 Gbit/s path even under bad network conditions such as a loss rate of 1% in the presence of RTTs of up to 200 ms. The best link utilization at the highest impairment value was achieved by FileCatalyst *Direct* – the values were never lower than 93% for all performed tests. For the loss ratio up to 0.3% and RTT up to 100 ms, *TIXstream* shows a better utilization of about 97 %.

Transmission duration measurements were primarily intended to prove that the solutions show accurate data

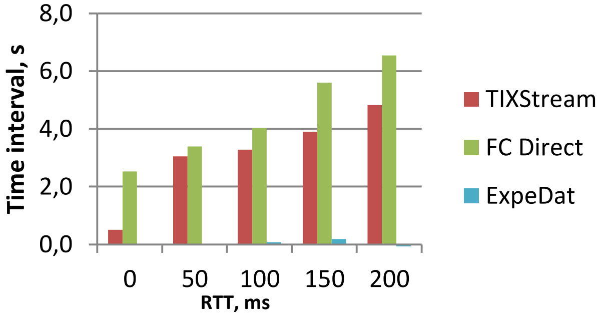


Figure 17. Difference between calculated and actual transfer durations; Packet loss =1

transport numbers in their outputs. The comparison showed that the lowest transfer duration of each solution is fairly close to the ideal one and that the discrepancy of the obtained output values are close to reality for all solutions.

Each solution uses some time for the allocation of system resources and the initialization of network resources. This time cannot be neglected, at least not in sessions with up to 30 GiBytes data transport. The comparison presented in Fig. 15 attempts to assess this service time. Probably, the time overhead is also due to the solutions not fully utilizing the bandwidth. It was found that the data rate of *FC Direct* is not the lowest one, but the transfer duration is higher than for all other solutions under test conditions. This result is possibly due to known java performance bottlenecks, because *FC Direct* is the only solution written purely in java.

When the solutions work under very light network impairment conditions (for example back-to-back) and the data rate achieves maximal value, the CPU usage is fairly high. For example, the maximum achieved data rate of *ExpeDat* seems to be due to CPU limitations. The system monitor showed 99% CPU usage in the *ExpeDat* process, and it also showed that one core of twelve is used in 99%. Other solutions, e.g. *TIXstream*, showed a CPU usage of about 150%, the usage of two used cores was about 70 % and 80 % respectively on the sender side, and on the receiver side 3 cores were used with a usage of 40%, 90%, 30%. This solution distributes the performance among several cores to maximally use the available bandwidth when possible.

One more significant point of resources management is the socket use. As shown in Section V.C, *FC Direct* uses different numbers of sockets on the sender and receiver sides. This use causes no problems for corporate LANs or simple back-to-back connections. However, for data transport using more complex structures, like real Internet connections, this use could cause problems on such devices such as firewalls. Such problems are well known even in simple multi streaming cases. This is an even worse situation in which each sender socket is sending data to different destination ports, so at least M x N port pairs must be tunneled in the firewall. It is very likely that intrusion detection systems can consider such behavior as violent traffic.

VIII. Future work

The analyzed solutions were tested on their abilities in the presence of high values of latency and packet losses. However, delay jitter is also a common network impairment and measurements using different values and different jitter patterns would also be of interest.

The present research shows the behavior of the solutions in an empty path such as VPN. Further investigations could be made into the behavior of the solutions in the presence of back-ground traffic.

The testbed topology was simplified to get a first representation of the presented solutions. An extension of the experimental topology makes sense for in-depth investigations.

During the experiments, only the performance of data transfer for commercial applications was investigated. To get a deeper understanding of only the telecommunication part, it would be of interest to make tests with the technology cores (e.g. protocol stacks without any wraps as e.g. file system).

The questions of system resource consumptions were addressed very briefly here. It would also be interesting to research this topic more extensively.

IX. References

[1] Y. Wu, S. Kumar, and S.-J. Park. "Measurement and performance issues of transport protocols over 10 Gbps high-speed optical networks". Computer Networks, vol 54. 2010, pp. 475-488. doi:10.1016/j.comnet.2009.09.017

[2] H. Kamezawa, M. Nakamura and M. Nakamura. "Inter-Layer Coordination for Parallel TCP Streams on Long Fat Pipe Networks". Proc. of the 2004 ACM/IEEE conference on Supercomputing. Pittsburg, PA, USA. 2004, pp. 24 -34.

[3] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia. "Above the Clouds: A Berkeley View of Cloud Computing". 2009. pp. 19-25. Tec. Rep. No. UDB/EECS-2009-28.2009.

[4] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia. "A view of Cloud Computing". Communications of the ACM. ACM New York, NY, USA, April 2010, Vol. 53, Issue 4, pp. 50-58. doi:10.1145/1721654.1721672

[5] S. Höhlig. "Optimierter Dateitranfer über 100 Gigabit/s". 100 Gigabit/s Workshop of the DFN, Mannheim. Sept. 2011.

[6] Aspera. Aspera. "custumer Deluxe Digital Studios". [retrieved: 11, 2012] http://asperasoft.com/customers/customer/view/Customer/show/deluxe-digital-studios/.

[7] Y. Gu and R. L. Grossman. "UDP-based datatransfer for high-speed wide area networks". Computer Networks. Austin, Texas, USA. May 2007, Vol. 51, issue 7, pp. 1465-1480. doi:10.1016/j.comnet.2006.11.009,

[8] E. He, J. Leigh, O. Yu, and T. A. DeFanti. "Reliable Blast UDP : Predictable High Performance Bulk Data Transfer". Proc. of IEEE Cluster Computing. Chicago, USA. Sept. 2002, pp. 317-324.

[9] R. L. Grossman, Y. Gu, X. Hong, A. Antony, J. Blom, F. Dijkstra, and C. de Laat. "Teraflows over Gigabit WANs with UDT". Journal of Future Computer Systems. Volume 21, 2005, pp. 501-513. doi:10.1016/j.future.2004.10.007

[10] L. Herr and M. Kresek. "Building a New User Community for Very High Quality Media Applications On Very High Speed Networks". CineGrid. [retrieved: 02, 2013] http://czechlight.cesnet.cz/documents/publications/network-architecture/2008/krsek-cinegrid.pdf.

[11] X. Wu and A. A. Chien. "Evaluation of rate-based transport protocols for lambda-grids". High performance Distributed Computing, 2004. Proc. of 13th IEEE International Symposium on High Performance Distributed Computing.Honolulu, Hawaii, USA. 2004, pp. 87-96.

[12] Bitspeed LLC. "From Here to There - Much Faster". Whitepaper. [retrieved: 10, 2012.] http://www.bitspeed.com/wp-content/uploads/2011/10/BitSpeed-White-Paper-From-Here-to-There-Much-Faster.pdf.

[13] Tixel GmbH. Tixstream: Overview. [retrieved: 10, 2012] http://www.tixeltec.com/ps_tixstream_en.html.

[14] File Catalyst. "Accelerating File Transfers". Whitepaper. [retrieved: 10, 2012] http://www.filecatalyst.com/collateral/Accelerating_File_Transfers.pdf.

[15] XDT PTY LTD. "High-Speed WAN and LAN data transfers". XDT. [retrieved: 10, 2012.] http://www.xdt.com.au/Products/CatapultServer/Features.

[16] V. Paxson. "End-to-End Internet Packet Dynamics". Networking, IEEE/ACM Transactions. 1999, Vol. 7, Issue 3, pp. 277-292. doi: 10.1109/90.779192

[17] Y. A. Wang, C. Huang, J. Li, and K. W. Ross. "Queen: Estimating Packet Loss Rate between Arbitrary Internet Hosts". Proc. of the 10th International Conference on Passive and Active Network Measurement. Seoul, Korea. 2009, pp. 57-66.

[18] B. W. Settlemyer, N. S. V. Rao, S. W. Poole, S. W. Hodson, S. E. Hick, and P. M. Newman. "Experimental analysis of 10Gbps transfers over physical and emulated dedicated connections". Proc. of Computing, Networking and Communications (ICNC). Maui, Hawaii, USA. 2012, pp. 845-850.

[19] Apposite Technologies. "Apposite". [retrieved: 10, 2012] http://www.apposite-tech.com/index.html.

[20] A. Jurgelionis, J.-P. Laulajainen, M. I. Hirvonen, and A. I. Wang. "An Empirical Study of NetEm Network Emulation Functionalities". 20. ICCCN. Maui, Hawaii, USA, 2011, ISBN 978-1-4577-0637-0, pp. 1-6. doi: 10.1109/ICCCN.2011.6005933

[21] Tixel GmbH. White Papers and Reports. Tixel. [retrieved: 11, 2012] http://www.tixeltec.com/papers_en.html.

[22] Tixel GmbH. Tixel news. tixel.com. [retrieved: 10, 2012] http://www.tixeltec.com/news_en.html.

[23] FileCatalyst. FileCatalyst. "Direct". [retrieved: 10, 2012] http://www.filecatalyst.com/collateral/FileCatalyst_Direct.pdf.

[24] Data Expedition, Inc. Data Expedition. "Difference". [retrieved: 10, 2012] http://www.dataexpedition.com/downloads/DEI-WP.pdf.

[25] Data Expedition, Inc. Overview. Data Expedition, Inc. [retrieved: 10, 2012.] http://www.dataexpedition.com/expedat/.

[26] R. Stewart, Q. Xie, Motorola, K. Morneault, C. Sharp, Cisco, H. Schwarzbauer, Siemens, T. Taylor, Nortel Networks, I. Rythina, Ericsson, M. Kalla, Telcordia, L. Zhang, UCLA, V. Paxson and ACIRI. "Stream Control Transmission Protocol". IETF, RFC 2960. [retrieved: 01, 2013] http://www.ietf.org/rfc/rfc2960.txt.

Performance of Cloud-based P2P Game Architecture

Victor Clincy and Brandon Wilgor
Computer Science Department
Kennesaw State University
Kennesaw, GA, USA

Abstract **- The traditional multiplayer video game architecture requires costly investment in physical game servers and network infrastructure. The peer-to-peer network model alleviates some of these concerns, but makes cheat prevention, software updates, and system monitoring far more difficult for the game publisher. Recent advancements in Infrastructure as a Service (IaaS) cloud platform providers such as Amazon Web Services and Microsoft Azure offers video game companies the option to host virtual game servers in the cloud. These services now allow gamers to build custom game servers in the cloud. This paper explores the performance of a cloud-based First Person Shooter game server compared to established performance metrics.**

Keywords – Cloud; Gaming; Game Architecture; Network Performance.

I. INTRODUCTION

Traditional game server models utilize the familiar client-server architecture to host multiplayer games. This model represents the majority of multiplayer game systems and supports millions of game sessions every day. Here, the server is responsible for maintaining game state information between clients, synchronization, and communications [1]. This model is popular for many reasons such as cheating and piracy prevention, reliability and performance, and centralized control. However, the client-server architecture does suffer from high bandwidth requirements and infrastructure cost and scalability problems.

Conversely, the Peer-to-Peer (P2P) model synchronizes games states directly between hosts without necessarily requiring a central game server. Although this approach has excellent scalability and extremely low costs associated with the Client-Server model, the lack of an authoritative central server introduces a number of key problems [2]. Among these are poor access control, limited cheat prevention, and non-uniform state synchronization.

However, many game studios are turning to the cloud to reduce some of the costs associated with the client-server model. For example, Microsoft's 343 Industries [10] recently utilized the Microsoft Azure [9] cloud computing platform to support the release of "Halo 4," the latest release in Microsoft's tremendously popular video game franchise [3]. 343 cited cost and scalability as one of the key factors in deciding to host the multiplayer game on the cloud. Previously, game studios were forced to make a massive investment in server and network infrastructure to support the huge spike in players associated with a game's release. However, as games age, the player population typically drops rapidly, leaving a high number of unutilized servers. However, Azure allows 343 to dynamically and efficiently adjust server capacity to support the player base at a significant cost savings [3].

The cloud model also offers advantages to P2P game architectures. Gamers can now host 24x7 "peer servers" on the cloud, rather than locally on their machine or by renting commercial game server space. This offers great advantages in reliability, performance, security, and most of all, excellent cost savings. Gamers have long hosted games on their own computer, acting as a de facto game server. This enables the gamer maintain high levels of control over game parameters, access control, performance, and other factors. However, the huge associated bandwidth and security vulnerabilities put this method out-of-reach for many casual gamers. Cloud-based service providers such as Amazon Web Services now offer these gamers the option of building custom game servers on the cloud.

Iosup, et al., [4] explored the performance variability of cloud service providers, such as Amazon's web services (called AWS), through the use of "performance indicators." One example of said indicator is the response time of a "resource acquisition operation" provided by the Amazon EC2 Service. Iosup, et al. also investigated various performance metrics associated with so-called "social games" such as Farm Town and Mafia Wars. However, the study did not include First-Person Shooters.

A First-Person Shooter (FPS) is a prominent type of game in which gameplay generally focuses on weapon-based combat from a first-person perspective. Popular examples of FPS include the Doom, Half-Life, Halo, and the Call of Duty series. Players of FPS games have been shown to be especially sensitive to network conditions relative to other genres such as role playing games (RPG) or real-time strategy (RTS) games. For example, one study finds that while online RTS games are unaffected by latencies as high as 1000ms, the relatively faster-paced FPS requires a latency of less than 100ms [5].

Barker and Shenoy performed a gamer server case study wherein a popular First Person Shooter (Quake 3) dedicated server was installed and tested in a lab-based virtual machine [6]. The test included an evaluation of both map loading times and server latency metrics. However,

ISBN: 978-1-61208-256-1

there are no known studies evaluating the performance of an FPS game server hosted by a cloud provider.

This study will explore performance parameters of a cloud-based FPS game server compared to established performance requirements of traditional client-server architecture.

Section two of this paper will describe the setup of the study and the various configurations. Section two will also present some response time results. The third section will present the analysis of the results. Sections four and five will cover future work and the conclusion, respectively.

II. THE STUDY

Response time is a widely recognized measure of performance in a First Person Shooter [5]. Response times of 200 milliseconds or less are generally considered the benchmark for acceptable Quality of Experience in a FPS game [7]. For this study, a client-side response time will be measured by a player of a popular FPS game. Specifically, the virtual game server will be hosted on a cloud-based service provider. Client-side response times will be measured in a single-player (light server load), 8-player (moderate load), and 16-player (high server load) death match games.

The first step was to configure a game server on a cloud environment. The Source Dedicated Server platform functions as a dedicated virtual game server for Source-Engine games, such as the popular First Person Shooter “Half-Life” [8]. For this study, a Half-Life srdcs Game Server was configured on an Amazon EC2 instance. The server was installed on a “free usage tier” 32-bit Amazon Linux-based machine instance. The specific game server installed in this study can host up to 16 AI “bots” or human players.

The game server is initially configured for a single-player “deathmatch” game against fifteen AI-controlled bots. This will ensure data throughput consistent with a typical free-for-all type game session. However, the server load, with respect to network traffic, will be minimal. The client side response times are shown in Figure 1 below.

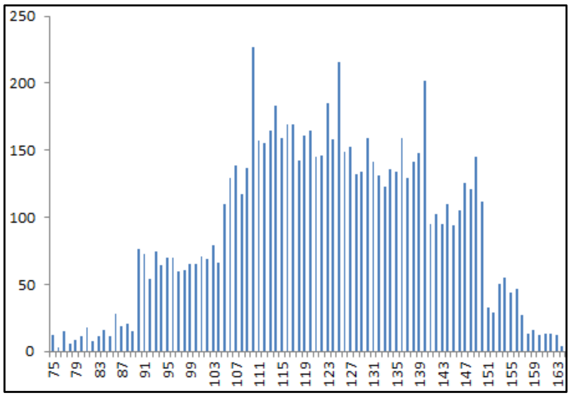


Figure 1. 1 Player Game Response Times

The next test session will simultaneously host 8 clients in a death match game. The remaining eight players will consist of AI bots. This test session will represent a moderate level of server network load. Response time measurements will again be taken client-side. These measurements are shown below in Figure 2.

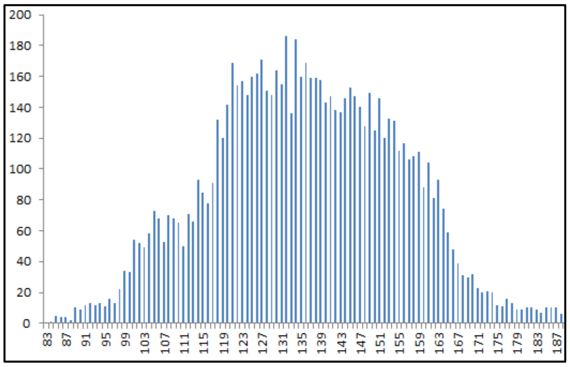


Figure 2. Eight Player Game Response Times

The final test session will consist of sixteen human players. This represents a maximum level of network load under typical death match conditions. Response time measurements will again be taken client-side. These measurements are shown below in Figure 3.

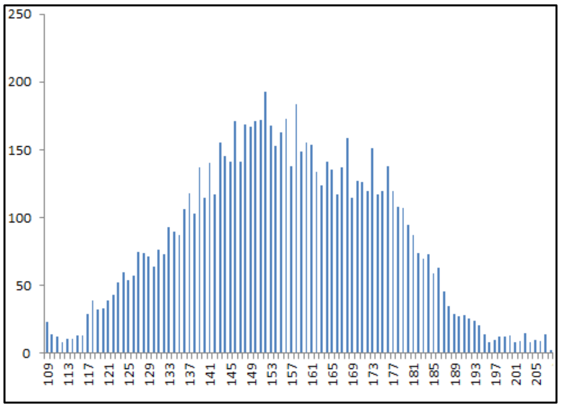


Figure 3. Sixteen Player Game Response Times

III. ANALYSIS

A single player game session showed response times well within established acceptable limits. As seen in Figure 1, this session saw a maximum response time of approximately 164 ms, with a vast majority of response within 90-150 ms. Figure 2, representing an eight-player game, also shows very good performance. However, the higher network traffic associated with an increase in human players shows a general increase in response times.

Figure 3 above represents relatively high server and network loads, featuring a maximum of sixteen simultaneous human players. During this test session, the client-side response times increased significantly from the single player baseline test. Compared to the single player game, average response times increased by well over 10%. In fact, approximately 25 responses exceeded the established threshold of 200 ms ideal for FPS games. The occurrence of these instances was relatively low however, accounting for approximately 1% of all traffic.

Although the sixteen-player test showed a significant increase in client-side response times from the single-player baseline test, approximately 99% of server responses were 200 ms or less. This represents an acceptable user Quality of Experience according to typical measures of FPS game performance. It is unclear whether the increased network activity of 16 simultaneous network connections, or the associated increase in server processing requirements, caused the increase in response times.

IV. FUTURE WORK

This study utilized a simple Source Dedicated Server hosted on an Amazon free usage tier EC2 instance. Although performance was acceptable up to full (sixteen-player) server and network loads, a general increase in response times was seen as the number of human players was increased. Future work may investigate the effects of thirty-two-player games of even Massively-Multiplayer Online (MMO) games to further explore the capabilities of cloud-based game servers. More advanced FPS games, such as Call of Duty or Crysis, will also increase the processing requirements of the server. Finally, network analysis or application performance monitoring may be used server-side in order to truly gauge game performance across multiple clients.

V. CONCLUSION

Traditionally, the client–server multiplayer video game architecture requires costly investment in physical game servers and network infrastructure. The peer-to-peer network model alleviates some of these concerns, but makes cheat prevention, software updates, and system monitoring far more difficult for the game publisher. However, modern cloud platform providers such as Amazon Web Services and Microsoft Azure offers video game companies the option to host virtual game servers in the cloud. These providers also give individual gamers the option to build and maintain custom game servers in the cloud. This study established the viability of a cloud-based FPS game server with respect to established performance parameters.

REFERENCES

[1] A. M. Khan, I. Arsov, M. Preda, S. Chabridon, and A. Beugnard. “Adaptable client-server architecture for mobile multiplayer games,” *Proceedings of the 3rd International ICST Conference on Simulation Tools and Techniques* (SIMUTools '10). Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Brussels, Belgium, Article 11 , 2010. DOI=10.4108/ICST.SIMUTOOLS2010.8704 http://dx.doi.org/10.4108/ICST.SIMUTOOLS2010.8704 [retrieved: October, 2012]

[2] T. Hampel, T. Bopp, and R. Hinn. “A peer-to-peer architecture for massive multiplayer online games,” *Proceedings of 5th ACM SIGCOMM workshop on Network and system support for games* (NetGames '06). ACM, New York, NY, USA, , Article 48, 2006 . DOI=10.1145/1230040.1230058 http://doi.acm.org/10.1145/1230040.1230058 [retrieved: October, 2012]

[3] Microsoft. “Meet the ‘Plumbers’ Powering ‘Halo 4’ Infinity Multiplayer,” [Press Release]. October, 31, 2012. Retrieved from http://www.microsoft.com/en-us/news/features/2012/oct12/10-31halo4.aspx. [retrieved: September, 2012]

[4] A. Iosup, N. Yigitbasi, and D. Epema. “On the Performance Variability of Production Cloud Services,” *Proceedings of the 2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing* (CCGRID '11). IEEE Computer Society, Washington, DC, USA, 2011, pp. 104-113.

[5] M. Claypool and K. Claypool. “Latency and player actions in online games,” Communications ACM 49, November 11, 2006, pp. 40-45.

[6] S. K. Barker and P. Shenoy. “Empirical evaluation of latency-sensitive application performance in the cloud,” Proceedings of the first annual ACM SIGMM conference on Multimedia systems (MMSys '10). ACM, New York, NY, USA, 2010, pp. 35-46. DOI=10.1145/1730836.1730842 http://doi.acm.org/10.1145/1730836.1730842 [retrieved: September, 2012]

[7] W. Wu, A. Arefin, R. Rivas, K. Nahrstedt, R. Sheppard, and Z. Yang. “Quality of experience in distributed interactive multimedia environments: toward a theoretical framework,” Proceedings of the 17th ACM international conference on Multimedia (MM '09). ACM, New York, NY, USA, 2009, pp. 481-490.

[8] "Source Dedicated Server Resource for Source-Engine Games," Source Dedicated Server (srcds). N.p., n.d. Web. 30 November 2012. <http://www.srcds.com/>. [retrieved: November, 2012]

[9] Microsoft Azure is Microsoft’s cloud computing platform for building and managing applications and services. 2013. http://www.windowsazure.com/en-us/ [retrieved: January, 2013]

[10] 343 Industries is a video game development company headquartered in Kirkland, Washington. 343 Industries’ parent company is Microsoft Studios. 2013. http://www.halowaypoint.com/en-US/. [retrieved: January, 2013]

Improving the Quality of Life of Dependent and Disabled People through Home Automation and Tele-assistance

Carlos Rivas Costa, Miguel Gomez Carballa, Luis E. Anido Rifon, Sonia Valladares Rodriguez, Manuel J. Fernandez Iglesias
Telematics Engineering Department
University of Vigo
Vigo, Spain
{carlosrivas, miguelgomez, lanido, soniavr, manolo}@det.uvigo.es

***Abstract*— Lack of mobility in certain groups of dependents forces them to spend a lot of time at home. In many cases, this limitation makes these people to stay most of the time in a specific room in their houses such as the bedroom or living room, where the only means of entertainment and information gathering is the TV set. Most of present-day households have a personal computer, but the digital divide and lack of adaptation produces certain rejection in this population group. This paper discusses a proposal that leverages the familiar TV set to be used as the user interface for a complete tele-assistance system and control centre of home automation devices. For this, the system makes use of a Home Theatre Personal Computer (HTPC) connected to the TV and offers the features like the monitoring and remote monitoring of a wide range of vital signs, intelligent adaptation of services and interfaces according to the level and type of disability, and centralized control of home automation devices installed at home.**

Keywords— e-Health; HTPC; TV; teleassistance; home automation.

I. INTRODUCTION

In industrialized countries, care provision to dependent individuals is becoming a priority. The increase in the quality of life fosters an increase in life expectancy, and therefore their longevity [1]. As a consequence, it is a major challenge for industrialized countries to maintain the quality of life of these groups, and to adapt health policies to the new population demographics.

One of the aspects that is getting special attention to tackle this problem is the application of new technologies adapted to improve the quality of life of these groups. Systems and platforms proliferate in the quest for novel technological solutions are numerous, in most cases providing services for the care and the improvement of the quality of life of these individuals [2], [3].

While all of these platforms have advantages and benefits for the target population's everyday lives, in many cases users are sceptical or even reject new technological systems, mainly because they feel unprepared for understanding their behaviour, and therefore to adequately interact with them.

This paper introduces an open source platform that provides a complete tele-assistance and home automation control system of through a simple and familiar user interface as the TV set. Relying on this interface allows target user groups, especially the elderly, to overcome their initial reluctance to use new technologies. TV becomes the user interface, and through it all interaction between the user and the platform takes place.

To provide the needed interactivity on a TV-based platform we rely on a home theatre PC (HTPC). Thanks to its PC architecture, the platform has the required modularity and processing power to provide the services implemented. Besides, the PC architecture facilitates the integration of new communication interfaces to support the communication with home automation devices or devices intended to monitor biomedical parameters.

Among the services offered by this platform are medical services, such as the monitoring of vital signs, rehabilitation games, educational videos providing information on a wide range of diseases or disabilities and videos with rehabilitation exercises; social services such as medication reminders and alarms and access to social networks like Facebook or Twitter; and home automation services, through which any home automation device may be controlled from the TV set.

The heterogeneity of this proposal's target population and the particular needs of individual users require paying special attention to user interfaces. Our platform's user interfaces are automatically and transparently tailored to the specific needs of the individual accessing the system. To achieve this, the platform implements an automatic adaptation layer for the user interfaces and services offered, which simplifies the usage and control of the system by taking into account the particular characteristics and needs of the user.

Along the rest of the paper we will discuss the most relevant elements of the proposed solution. Section II briefly introduces the general aspects of the platform. Then, Section III describes the medical parameter's monitoring system, and Section IV discusses the home automation control system. Finally, we will outline the main conclusions drawn from this research in Section V.

II. TELE-ASSISTANCE PLATFORM

As stated above, the main objective of this proposal is to create a platform for tele-assistance and home automation control through a familiar device as the home TV. Thus, the television will become the control centre of all remote assistance services provided, and of all home automation devices in the user's home.

The use of television as a user interface minimizes the initial rejection to the use of new technologies by some groups of dependent individuals. Rendering information and services through a familiar and easy to use device as the TV set, simplifies learning and therefore access to services and information. For example, most users are accustomed to interact with the TV via the classical TV remote. Therefore, the TV remote no longer becomes a new device but a familiar one, which fosters an early adaptation to the new platform, and the perception of new services as extensions of the functionality of the TV instead of perceiving them as new services they have to learn to use.

The design and architecture of the TV set is corresponds to a completely passive device, in which the user is a mere spectator of the contents displayed by the screen. To provide the TV with the required interaction capabilities, our proposal is based on an HTPC connected to the TV, which will provide the necessary interaction features, a high degree of modularity, and support for interconnecting additional peripheral devices, both wired and wireless. Through this low cost PC, users can access the full potential offered by state-of-the-art Internet services, applications and platforms just by changing the TV channel. The final system is fully modular and extensible, and supports the adaptation and integration of the functionalities offered according to the type and degree of dependence of final users.

The need to adapt the platform to different user needs, doesn't affect graphical interfaces. This platform goes a step further, and based on the ability of HTPC for integrating different communication protocols, we defined new adapted control mechanisms. The appearance on the market of devices that integrate gyroscopes and accelerometers that can detect movements, as like Nintendo's WiiMote [5] offers new possibilities for motion-based user interaction. This remote control device can be connected to the HTPC via the common Bluetooth. Therefore, besides the classic control through push buttons, control can be performed via the detection of movements performed by the user. This form of adaptive control, allows on the one hand a reduction in the number of buttons required in the remote control and therefore in control complexity. On the other hand, users who have some limitations [6] in the use of conventional remote controls can access and interact with the platform through motion detection.

The simplicity of the interconnection with external devices also enables the incorporation of a broad range of vital sign measurement devices. Besides, the HTPC's PC architecture supports the development of specific communication interfaces to virtually all monitoring devices currently on the market. Thus, the proposed platform becomes an extensible system that can be easily adapted to incorporate new measurement and patient monitoring solutions, together with other systems such as smart card readers, home automation systems, etc.

All data collected in the HTPC are sent over an Internet connection to the control centre. The platform is designed around a client-server Service Oriented Architecture (SOA, [11]). The platform makes use of distributed network services both to transmit and receive information. Storing data remotely allows users to access their accounts from any deployed home platform. There is no need to manually customize client platforms, but the platform automatically personalizes interfaces and services once the user is identified.

Besides, having data stored remotely presents some advantages, but also requires special attention to protect the transmission of information. Many of the transmitted data will be of medical nature, and therefore very sensitive to the data protection laws in many countries (e.g., Spain's Data Protection Act). All sensitive data will be transmitted via an SSL-encrypted secure communications channel from the client computer and the central server.

From the users' point of view, the management and transmission of information is a completely transparent process. A software module based on the XBMC media centre [10] will be responsible for performing each of the required actions on behalf of the user. XBMC is licensed under GNU / GPL, and has been modified to add new functionality to the native functionality of a media centre, which is focused on multimedia playback and graphic event management.

Among the new features are the management of the specific communications services developed for this platform. For this, we relay on the standard Extensible Messaging and Presence Protocol (XMPP), which has been integrated within the platform. A central messaging server based on OpenFire performs the routing tasks for the information transmitted by each of the services to their intended recipient. The integration of the XMPP protocol for the exchange of information supports the enrichment of the messages transmitted and their customization for each of the services implemented. The XMPP protocol relies on the exchange of XML files. These XML files follow a common base structure that can be easily extended with new labels to represent any messaging requirement. Thus, services may add new fields with the required information to the messages conveyed, where only the corresponding service is able to capture and understand the information transmitted.

XMPP's presence control support enables the platform to be continuously aware of the state of each user. When users access the system or simply change their state, an associated event is collected by the communications server and transmitted to all users or services that are authorized to detect changes in their status. For example, with this functionality a service may detect when a given user is accessing a specific service, and therefore it will be able to establish a synchronization mechanism to send the required information.

III. Measurement of Biomedical Parameters

Tele-monitoring care systems are among the ones most benefiting of technological progress. Monitoring has become a basic functionality in a tele-assistance platform, and many state-of-the-art tele-assistance platforms include functionality for monitoring and measuring several biomedical parameters.

The application of information and communication technologies to home monitoring provides a great capacity for gathering and transmitting information, which in turn enables a more complete and continuous monitoring of patient evolution, and therefore enhances the ability of health professionals to perform a better analysis and diagnosis of patients' situations.

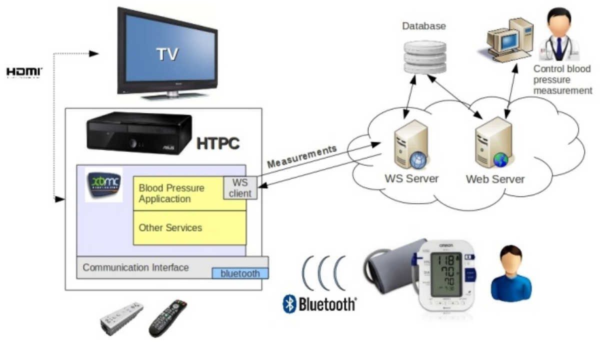


Fig. 1: Remote blood pressure control

The flexibility and extensibility provided by the HTPC architecture allows the integration of the vast majority of devices for the measurement and analysis of biomedical parameters in the market, and even the integration of any custom-made device. In our case, the HTPC serves as a communications gateway between measurement sensors and the data centre (cf. Fig. 1). Besides, the availability of both wired and wireless communication interfaces facilitates the adaptation to almost any sensor available.

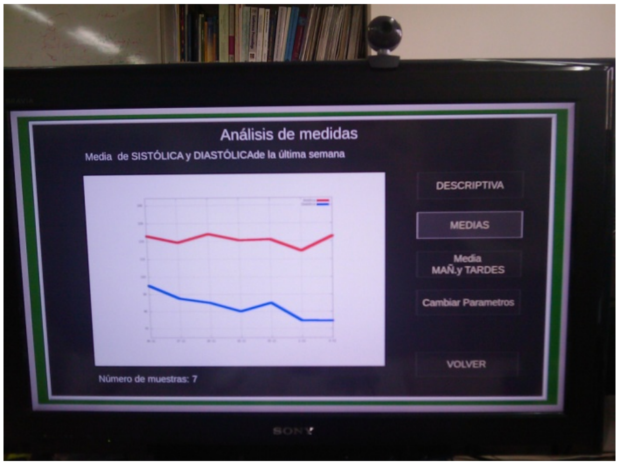
A. Measuring Biomedical Parameters

All measurements are made at patients' premises to be automatically sent to the data centre through the residential gateway. To perform a measurement, the platform implements two modes, namely on-demand and remote warning.

By operating on-demand, the system acts as a passive device where users take the initiative to carry out a particular action. For example, for a blood pressure measurement performed under this mode of operation, users will navigate through the service menus to select blood pressure monitoring. Once the sphygmomanometer takes the measurement, values are transmitted to the wireless communication gateway. The residential gateway performs some error checking and connects to the biomedical parameters' storage server to transmit the measurement results.

The transmission of the blood pressure measurement taken at the patient's home is performed via the SOAP protocol. An instantiation of a Tomcat Web server performs all necessary operations related to authentication and communication with the residential gateway. As this service manages personal and medical data, users have to authenticate at the residential gateway to access the blood pressure measurement service. Two authentication modes have been implemented respectively based on a username and password pair, and on a smart card. Smart card authentication no longer requires users to remember complex passwords, as authentication is performed through the credentials stored in the personal smart card inserted into a smart card reader.

When the user inserts the smart card into the reader, the residential gateway reads the credentials stored in the card and establishes a communication through a Web Service with central user management server. The server will check the authenticity of the credentials submitted and will generate a session token, which will be returned to the residential gateway. From this moment on, this session token will be used for all transfers made from the patient's home to the medical data servers. In this case, the duration of the session is determined by the state of the smart card reader. If the reader detects the card removal or the disconnection of the reader, it will automatically disconnect the user and therefore deny any access to services handling private information such as biomedical parameters' measurement services.


Análisis de medidas

Media de SISTÓLICA y DIASTÓLICAde la última semana

Número de muestras: 7

Fig. 2: PA analysis

In the remote warning mode, the platform notifies the user and executes the appropriate actions on the user's behalf. A message on the TV screen warns the user about a pending measurement of some biomedical parameter. The user just acknowledges the message by pressing an OK button on the interface to run the corresponding service. After the measurement, the process continues in the same way as in the case of an on-demand measurement. Obviously, this operation mode is valid only for users who are already authenticated by the platform.

Figure 2 below illustrates a blood pressure analysis performed at the patient's premises.

B. Access to remotely stored data

All biomedical data are stored remotely in the data centre to be accessed by authorized medical staff. Medical staff may use any Web browser to access the information of any of their patients (cf. Fig. 3). When doctors access the web platform they are prompted for a username and password. Once they are authenticated, a list with all their patients is displayed. Using this web tool, doctors may access the historical measurements performed by any of their patients, which in turn improves the diagnosis and treatment processes.

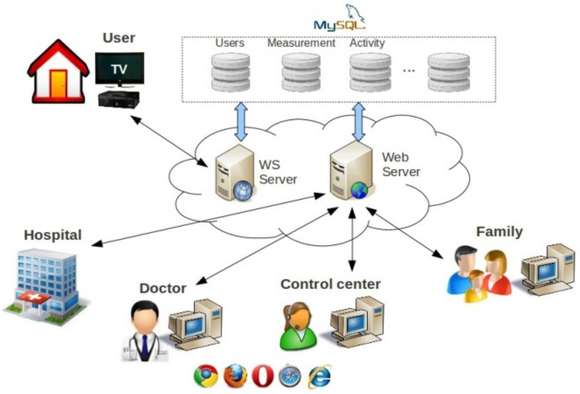


Fig. 3: Access to patients' information

Once the data have been analysed by the medical staff, they may schedule new measurements or update the instructions given to patients. They may also schedule new remote-mode measurements, which will be notified to patients the next time they are authenticated on the platform.

IV. Home Automation System

In recent years we are experiencing and ever increasing impact of home automation on our lives. This is a particularly interesting field for dependent people, as many tasks they cannot perform, may be performed when assisted by home automation devices installed at home. A task as simple as turning on and off of the lights can be a very difficult task to be performed by a person with certain physical disabilities, or even impossible for a bedridden person. Thus, home automation has great development potential for these population groups.

In many occasions, the simplicity of certain tasks when performed through home automation devices contrasts with the excessive complexity of the controls of these devices, and the need to have different controls to cover all brands of home devices installed in a home.

Although we can find in the market several communication protocols and control systems for home automation, manufacturers use to provide their own custom solutions, so that the integration in a single installation of devices from different manufacturers becomes a difficult task in many cases. To overcome this situation, the platform discussed in this paper acts as a middleware for different home automation systems installed in the user's home. For this, the residential gateway integrates into a single device all the communication protocols needed to interact with products from different vendors. As discussed above, the use of a HTPC as the hardware platform provides the required flexibility to integrate all those different wired and wireless communication interfaces (Fig. 4).

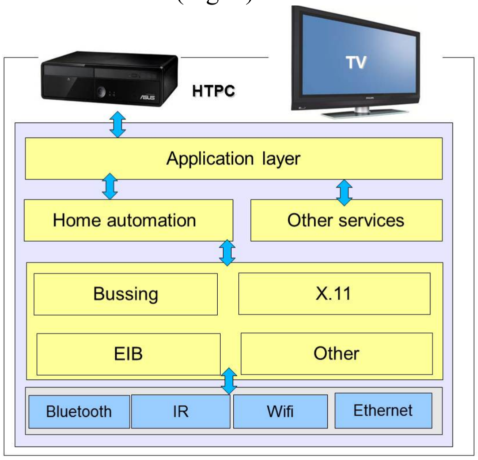


Fig. 4: Home automation architecture

By selecting the appropriate interface, information can be routed to each specific home automation device, and the TV set becomes the control centre of the complete home automation infrastructure. Users may send commands to perform the desired actions through their customized graphical user interface, which will be automatically and transparently routed to one or several physical devices. We describe below the operating mode of a particular home automation protocol, namely the Busing protocol. This example can be extrapolated to other home automation protocols installed at the user's premises.

A. Busing Protocol

This is a proprietary home automation communication protocol. The Busing system uses a four-wire bus for transmitting information among devices. Each home automation device is connected to the bus, having a unique identification number. Commands are transmitted to a data bus together with a unique identifier representing the destination device. Then, the destination device will capture the transmitted information and execute the action associated with the command. All information sent to the bus is packed into data frames, and these frames are transmitted by a specific device called ETHBus. This device has an Ethernet interface and a proprietary Busing interface, so the ETHBus behaves as an Ethernet-Busing bridge.

The proposed tele-assistance and home automation platform connects to the home automation network through an ETHBus connected to the HTPC's Ethernet interface. Users' TV screen interactions are translated into Busing commands to the appropriate devices and sent to the home automation bus for execution. When an order is executed or the home automation system detects a change (e.g., the presence of a person in a surveyed area) it sends that information to the bus to be detected by the ETHBUS, which will convert the data package into an Ethernet frame to be sent to the HTPC. Eventually, this information will be conditioned to be presented to the user as a notification or alarm displayed on the TV screen.

The control of electrical and electronic devices is performed via a 6E6S device. This device consists of 6 on-off relays. Each of these relays may control an electric appliance, and they may be enabled and disabled manually or automatically.

Manual activation is performed through each of the six input gates in the 6E6S device. These terminals may be connected to on-off buttons or switches that the user may operate manually. Automatic activation and deactivation is performed through the transmission of Busing commands as discussed above.

Although this device may present a very basic functionality, it is one of the most used devices in present-day smart home installations. Simple on/off actuations may cover a broad range of electrical and mechanical operations, such as opening and closing motorized window blinds, switching lights on an off, opening and closing windows, doors and gates, etc.

In addition to sending commands to smart home infrastructure, the platform also has the ability to receive and process data sent by other devices such as motion detectors. These devices are placed in the user's home to detect movement in any room. They use infrared beams to detect notify presence in their range of detection, and when the remote assistance platform detects motion in any of the areas of the house where motion detection is installed, a notification is displayed on the TV. While these devices are typically used for motion detection in certain areas of the house, such as the front door, they may also be combined with advanced algorithms to detect specific user behaviours, and therefore to detect abnormal situations that may occur at home, or even the lack of activity.

As with other platform services, user access to home automation control requires authentication. Once the user has been identified, the user interface displays all installed automated systems in the house making transparent their types and models. Thus, home automation systems are integrated and centralized at the TV set (cf. Fig. 4).

V. Conclusions

Industrialized countries are facing demanding situations in relation to care providing to dependent and disabled individuals. Population ageing and the inversion of the population pyramid are leading many governments to rethink their current welfare systems. More and more resources are needed to fund the required welfare policies aimed at these groups of people. This is one of the reasons why tele-assistance and the introduction of ICT at home are seen as a promising way to tackle a problem that is relevant at a greater or lesser extent to all industrialized countries.

This paper proposes a low cost and open I tele-assistance solution based on the HTPC hardware architecture connected to the TV. This approach allows an easy integration of novel tele-assistance system and the centralization in a single device of all home control and remote assistance systems installed in the user's premises.

The fact that a large number of dependent and disabled people spend much of their time alone at home makes the TV set to become their only method of entertainment, and therefore the only option to interact with other people. The ubiquity of the TV leads us to think about this appliance as a perfect medium to access not only to conventional television programmes, but also to a service portfolio that may dramatically increase their life quality.

Our system aims to facilitate typical tele-assistance services such as the monitoring of different biometrical parameters, and the user-friendly control of home automation devices. The possibility to integrate a comprehensive home automation system in a totally transparent way makes this to be seen as a simple and easy to use solution, avoiding the initial rejection attitude common in certain groups of dependent people, especially our elders.

Using the TV as the unifying centre for ICT services as well as reducing the initial rejection to new technologies, also influences deployment costs. Virtually all households in industrialized countries have at least one TV set, so the acquisition costs will be dramatically reduced. On the other hand, the HTPC's modular architecture and its peripheral interconnection capabilities make this platform easy to extend and adaptable to the needs of each individual user.

Finally, we would like to note that the open nature of most of this platform's components facilitates the integration of third-party developments to increase the functionalities provided.

Acknowledgment

This work has been performed under the support of: (1) the European Regional Development Fund (ERDF) and the Galician Regional Government under projects “Improving the quality of life of people in situations of dependency or disabilities through the implementation of Home–Telecare system” (10SEC012E) and "RedPlir" (CN 2012/319); and (2) the Spanish Ministry of Science and Innovation under grant “Methodologies, Architectures and Standards for adaptive and accessible e-learning (Adapt2Learn)” (TIN2010-21735-C02-01).

References

[1] (2010) World Health Organization. [Online] Available: http://who.int [retrieved: 11, 2012].

[2] O. Onyimadu, F. Harding, J Briggs,”Designing a telecare product for eledery”, 5th International Conference on Pervasive Computing Technologies for Healthcare, Newbury , May 2011, pp. 336 – 339

[3] (2012) Philips Motiva [Online] Available: http://www.healthcare.philips.com/es_es/products/telehealth/Products/motiva.wpd [retrieved: 11, 2012]

[4] A. Tesanovic, G.Manev, M. Pechenizkiy, E. Vasilyeva, "eHealth personalization in the next generation RPM systems", 2009 22nd IEEE International Symposium on Computer-Based Medical Systems, pp.1-8, Aug. 2009, doi:10.1109/CBMS.2009.5255383

[5] F. Iacopetti, L. Fanucci, R. Roncella, D. Giusti, A. Scebba, "Game Console Controller Interface for People with Disability", International Conference on Complex, Intelligent and Software Intensive Systems, 2008, pp.757-762

[6] Y.-J. Chen, "Using real-time acceleration data for exercise movement training with a decision tree approach", 2009 International

Conference on Machine Learning and Cybernetics", doi:10.1109/ICMLC.2009.5212632, July 2009, pp. 3005-3010.

[7] (2012) OpenTV [Online] Available: http://opentv.com/ [retrieved: 11, 2012]

[8] (2012) MHP [Online] Available www.mhp.org/ [retrieved: 11, 2012]

[9] E.J.W. Van Someren, "Actigraphic monitoring of movement and rest-activity rhythms in aging, Alzheimer's disease, and Parkinson's disease", IEEE Transactions on Rehabilitation Engineering, vol. 5, pp.394-398, doi:10.1109/86.650297

[10] (2012)The XBMC website. [Online]. Available:http://www.xbmc.org [retrieved: 11, 2012]

[11] L. White et al., "Understanding interoperable systems: Challenges for de maintance of SOA applications", Conference on system science (HICSS), Hawaii, january 2012, pp. 2199 – 2206

[12] B. Xuefu, and Y. Ming, "Design and implementation of web instant Message System based on XMPP", 3rd International conference on software engineering and service science (ICSESS), Dalian, June 2012, pp. 83-84.

Customized and Geolocated TV Channels about The Way of Saint James

Sonia Valladares Rodríguez, Manuel J. Fernández Iglesias, Roberto Soto Barreiros, Luis E. Anido Rifón, Carlos Rivas Costa
Department of Telematics Engineering
University of Vigo
Vigo, Spain
{soniavr, manolo, rsoto, lanido, carlosrivas}@det.uvigo.es

***Abstract*—The technical and procedural state of the art in video distribution and computer-mediated collaboration is mature enough to revisit the traditional television concept. The evolution of the state of the art made possible the concept of personal television, understood as viewers configuring what they wish to watch assisted by a smart recommendation system that proposes content adapted to their personal profile. We introduce a personalized interactive service platform intended to enhance the experience of completing the Way of Saint James. This platform can be accessed using devices like TV sets, mobile terminals or personal computers, and it is technologically supported by a blend of interactive TV, mobile communications, geolocation and semantic recommendation technologies. This proposal enhances the online provision of multi-technology services, the final users' experience by providing accurate, personalized and geolocated multimedia content on the Way of Saint James as well as the promotion of tourism initiatives related to the Way, a key aspect in the promotion of Galice and Spain as touristic destinations.**

Keywords- TV Channels; customization; geolocation; The Way of Saint James; recommendation semantic model.

I. Introduction

The Way of St. James as a framework for personal and spiritual development has not been conveniently addressed from a technological point of view. Along the Way, pilgrim-targeted services are mainly focused on weather forecast, alternateroutes, difficulties to overcome, places to visit, etc. In turn, the platform discussed in this paper is intended to provide interactive, personalized online information services according to the profile and location of individual pilgrims. More specifically, it facilitates the creation, management and visualization of customized content channels fed from a broad range of content providers. In line with the very concept of completing the Way of St. James, this proposal tackles some aspects with relevant social impact, like accessibility, personalisation and ubiquity.

The development of this service platform was initiated with a thorough analysis of the state of the art, which is discussed in Sect. II. This analysis was focused on the technologies supporting the development of the system, namely Internet TV, mobile technologies, geolocation, and semantic technologies. Then, we discuss in Sect. III the application of semantic technologies and smart recommendation techniques to develop personalized services and to provide access to geolocated content using mobile terminals, which will guarantee the ubiquity of the services provided. We also discuss there collaborative tagging of resources as a tool to implement automated personalized content programming, which basically consists on utilizing a collectively generated knowledge base to develop algorithms to generate a personalized programming grid targeted to final users, namely pilgrims and pilgrim hostels along the Way. A content revision system has also been implemented to address issues like the verification of author rights, age classification of content, or tag validation.

Section IV is devoted to introduce the platform's main features, architecture, services deployed, and some implementation feedback. Geolocated and personalized multimedia channels may be accessed using mobile devices or smart TV sets in hostels. In this later case, the hostel management will define the actual programming grid. We will also discuss the main characteristics of the system back-end, that is, a multimedia content repository fed from different sources, including official content providers (e.g., public administrations, public broadcasters); hostels along the Way providing news, information about hostel facilities or specific recommendations; and the pilgrims themselves, as they have available a personal space to socialise and share videos according to the Web 2.0 concept.

Finally, Section V provides some conclusions and lessons learnt from the implementation of the platform and the provision of multiple services.

In a nutshell, this platform illustrates the application of state-of-the art and novel information and communication technologies to contribute to the advance of tourism-related services, a strategic sector for the Galician and Spanish economy.The development of a semantic recommendation engine in this context is the most relevant novel contribution of the research presented in this paper.

II. State of the Art

There are several information systems available related to the Way of St. James. Most of them are web portals like the Pilgrim's Web [1], caminodesantiago.com [2] or xacobeo.es [3]. All these systems provide as their main feature information about the Way, including route alternatives, points of interest, or weather forecast information. In some cases, they also provide functionalities to facilitate communication among pilgrims though discussion fora, comments or ratings. The information provided is mostly static and is never personalized according to the pilgrims' profile or their location. Thus, we can conclude that the pilgrims' experience could be dramatically improved by providing a multimedia-based, multi-platform information service. For this, we will rely on three base technologies to build the foundation of this proposal:

1) Internet-TV: it facilitates access to Internet services through TV sets. This technology blends information transmission and multimedia content broadcasting through IP networks, and supports the visualisation of content using a broad range of devices, including smartphones, tablets, game consoles, Internet-enabled TV sets, or HTPCs, that is, small, low-cost computers embedding a media centre (e.g., Mediaportal [4], Windows Media Center [5], MooVida [6] or XBMC [7]) to facilitate multimedia content visualisation, Web access, and additional Internet TV services.

2) Mobile and Geolocation technologies: they providegeographically contextualized information to final users. The present-day market of mobile devices is an especially dynamic one, and technological innovations from major players in this sector (e.g. Samsung, Apple, Nokia, HTC, etc.) are constantly being introduced. Besides, the integration of GPS and movement detection support enables the development of novel geolocation applications and services like Google Latitude [8], Fire Eagle [9]orPlazes[10].

3) Semantic technologies: they support the semantic description and processing of content, resources and services, their adaptation to users' profiles, and the provision of personalized recommendations. As this is the main technological contribution of this work, the application of semantic technologies in the framework of this project is discussed in more detail below.

III. Semantic Model for Recommendation

A. General Overview

Knowledge representation techniques support the explicit declaration, using sentences and logic rules, of the factual and tacit knowledge in a given domain of interest. This knowledge may be manipulated and semantically queried using an inference engine, which in turn facilitates its automated processing and the extraction of new knowledge from already existing one (i.e., to perform inference processes). In turn, this enhances the quality of the recommendations to final users.

The formal foundations of knowledge representation are being addressed in the framework of Artificial Intelligence since the '60, but their mainstream usage in practical applications was only possible after the introduction of the Semantic Web concept [11]around ten years agoby the World-Wide Web Consortium (W3C**¡Error! No se encuentra el origen de la referencia.**).Presently, W3C supports the standardisation of logic sentence declaration languages like Resource Description Framework Schema (RDF/S [12])andWeb Ontology Language (OWL [13]), logic rules (Semantic Web Rule Language, SWRL [14]) and semantic queries (SPARQL Protocol and RDF Query Language, SPARQL [15]). These languages are supported by several software development frameworks (e.g., Jena [16]) and by several inference engines like Racer [17], Fact [18]or Pellet [19]. Racer and Pellet are used in this proposal.

B. Ontology Development Process

To construct the semantic models that eventually will support the smart programming systems to provide content recommendations according to pilgrims' profiles, we followed a methodological approach based on Menthonlogy[20], including some adaptations extracted from Uschold-King, Noy-McGuinnesandUPON[21]. The construction of these models has been organized according to a series of stages, each of them providing a specific (sub)model:

1) Specification model: it states the objectives and the scope of the semantic model by means of *competence questions*. These artefacts are questions expressed in natural language to which the system being developed should provide responses. Competence questions are extremely useful because they make the scope of the underlying ontology explicit, and provide an instrument to facilitate quality assessment in evaluation processes. In the framework of this research, competence questions like the ones collected in Table I have been defined.

2) Conceptualisation model: it identifies concepts and relations among concepts, as illustrated in Figure 1 representing the model produced along this project.

3) Formalisation model: instantiates the Conceptualisation Model through descriptive logic and Horn formalisms.

*4) Coding model:*translates the Formalisation Model into executable code. More specifically,ontology coding in our case was supported by the W3C-recommended language OWL and the Protégé software platform.

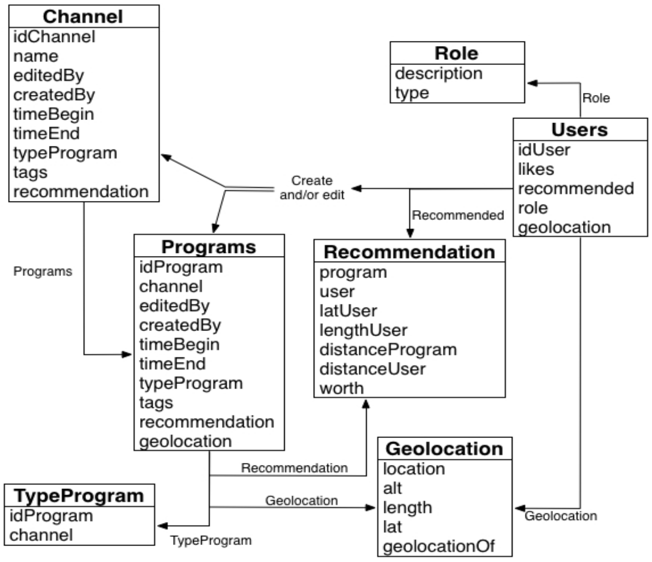


Figure 1. Excerpt of the conceptualisation phase's outcome

TABLE I. EVALUATION MODEL

	Evaluation Model – Competence questions	
	Natural language	***SPARQL sentences***
1	Which programmes (P) are relevant to user (U)?	*PREFIX channelTV:<http://193.146.210.125/canleTV/ontology.owl#> SELECT DISTINCT ?P WHERE {?UserchannelTV:idUser ?U . ?UserchannelTV:interests ?I . ?PchannelTV:tags ?e . Filter(regex(?U,"U")). Filter (regex(?I,?e))}*
2	Which relevant programmes (P) are broadcast in the vicinity of user (U)?	*PREFIX channelTV:<http://193.146.210.125/canleTV/ontology.owl#> SELECT ?P WHERE { ?User channelTV:idUser ?U. ?UserchannelTV:recommended?R. ?RchannelTV:distanceProgramme ?D. ?RchannelTV:distanceUser ?DU. ?RchannelTV:programme ?P Filter(regex(?U,"U")). Filter(?D < 5). Filter(?DU < 10)}*
3	Which programmes (P) are available to user (U)?	*PREFIX channelTV:<http://193.146.210.125/canleTV/ontology.owl#> SELECT ?P WHERE { ?User channelTV:idUser ?U. ?UserchannelTV:recommended ?R. ?RchannelTV:programme ?P. ?RchannelTV:value ?v. Filter(regex(?U,"U")). Filter(?v<10)}*

5) Evaluation model: it supports the verification of the final system prior to its deployment using the competence questions defined in the specification stage. In this phase, competence questions are expressed in a semantic query language, namely *SPARQL*, as illustrated in Table I.

IV. A MULTI-DEVICE SYSTEM TO SUPPORT TOURISM SERVICES

The final outcome of this project consists of a multi-device application to provide information to pilgrims walking the Way by means of a network of personalized and geolocated TV channels. These channels' programming grids are configured according to the pilgrims' profiles, that is, their mother tongue, their interests (e.g., gastronomy, nature, religion, etc.), and their specific location along the pilgrimage route. Users' interests are captured from the information provided by users upon registration, and this information is updated every time users actualize their profile or visit relevant landmarks along the route. In other words, users' interests are modelled as a blend of static and dynamic data organised as objects and classes according to the developed ontological model (cf. Sect. III).

Thus, the foundation of this proposal is a multi-platform video distribution and access system configured as a network of personalized channels able to provide smart recommendations using sematic algorithms.

This system provides a triple path to content:

- Anywhere, anytime, using a mobile device with 3G/3G+ connectivity (e.g., smartphone, tablet).
- At hostels along the way, using a mobile device (e.g., smartphone, tablet) with Wi-Fi connectivity.
- At hostels along the way, through standard TV sets where pilgrims may access, besides standard broadcasted channels, a specific TV channel customized to each accommodation facility.

This work involves several interacting agents that provide multiple services through the network. More specifically, the main actors involved in the platform are:

a) Pilgrims: they use their mobile devices to access the provided personalized TV channels.

b) Hostels: they may configure their own channel to provide information through standard TV sets.

*c) City councils:*they produce and broadcast thematic channels with information on services, places of interest, celebrations, etc. within their influence area.

d) Public administrations at regional and national level: they produce specific channels to promote and provide detailed information on routes and places along the Way.

e) Small/mediumentrepreneurs: they create commercials to promote their products and services (e.g., rural accommodation, souvenirs, local gastronomy, local produce, etc.).

A. Architecture

The deployment of this video distribution system requires the design, development and deployment of several sub-systems, each of them providing a specific functionality. This section describes the most relevant characteristics of the main logical subsystems in the system architecture depicted in Figure 2:

*1) Backend:*it is the central repository providing a storehouse for all relevant system's data. Information is kept about user profiles, programming grid status, programming schedule, active channels, service identification, multimedia content, metadata used for channel customisation, etc.

*2) Content repository (CR):*it stores program content and associated metadata needed for classification and location by external search engines. Content providers access this component to upload new content or to update existing content, and programmers fetch programmes to be included in programming grids. This repository has been implemented as a multimedia content management system (CMS) and ancillary subsystems. It supports several interfaces adapted to different content sources (e.g., TV-HTPC, mobile devices, etc.)

*3) Channel generator (CG):*it processes content descriptions stored in content repositories and creates personalized playlists according to users' profiles, including the most relevant content depending on user preferences and user (geo)location along the Way.

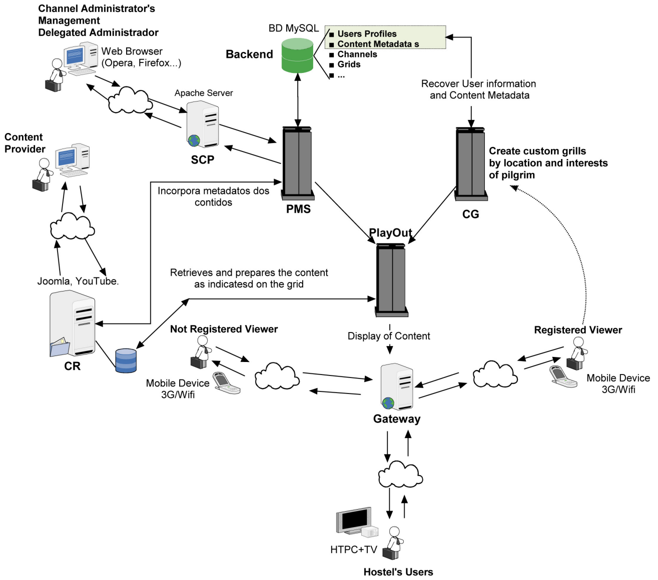


Figure 2. System's architecture

*4) Programming Management System (PMS):*it is used to manage the programming grid associated to each channel, providing concurrent access to programmers according to their profiles (e.g., role, authentication level, etc.). This subsystem also supports the implementation of programming scopes (i.e., the assignment of programming slots to specific programmers) and guarantees the coherence of the overall programming grid.

5) Programming Front-End (PFE): implements a Web-based user interface to the PMS. It provides programmers' authentication, the assignment of programming scopes to programmers, and concurrent access to the PMS.

*6) Play-Out:*this system feeds to video terminals the video content in each programming grid slot. This module also conditions the piece of content according to each specific user platform (i.e., TV-HTPC or mobile terminal) performing the required conversions (e.g., applicable codec, aspect ratio, video container, etc.) to guarantee the best possible user experience.

7) Access portal: it provides a front-end for users' video terminals. It embeds a streaming video viewer matching the user's terminal and receives content from the Play-Out module. It also displays information about the content being visualized (e.g. starting and ending time, duration, synopsis, etc.) and a simple electronic programme guide (EPG).

B. Services

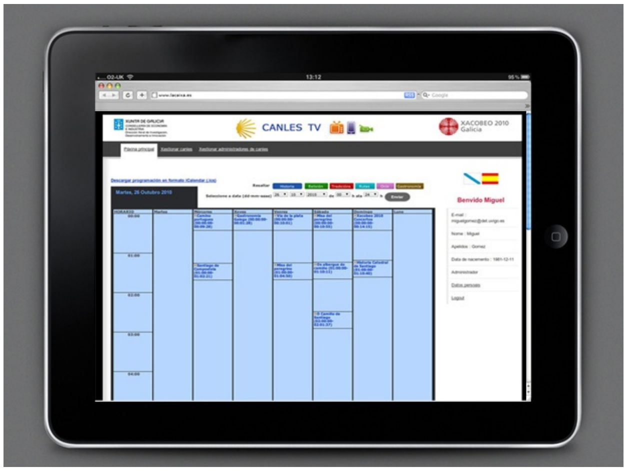
Users have access to a portfolio of multi-platform services. These services are classified according to the user platform:

1) PC interface: it is a Web portal where users may configure a programming grid for a given channel. Figure 3 depicts an example of a weekly programming grid. This grid may also be visualized according to specific time intervals (horizontal view) or specific days (vertical view).The main activities that can be performed through the PC interface are:

*a) Programming grid management:*As discussed above, users may populate with content a channel's programming grid.

*b) Channel management:*encompasses all tasks related to the channels' life cycle: creating, listing, searching, modifying, and deleting channels.

c) Channel Administrator's management: includes all activities related to the management of the individuals that may manage channels and their programming grids: create, modify and delete administrators; assign channels to administrators; enable and disable programming slots (scopes) to be managed by administrators, etc. Figure 4 illustrates the channel administration functionalities.


CANLES TV

XACOBEO 2010 Galicia

Benvido Miguel

Figure 3. Example I of the PC Interface

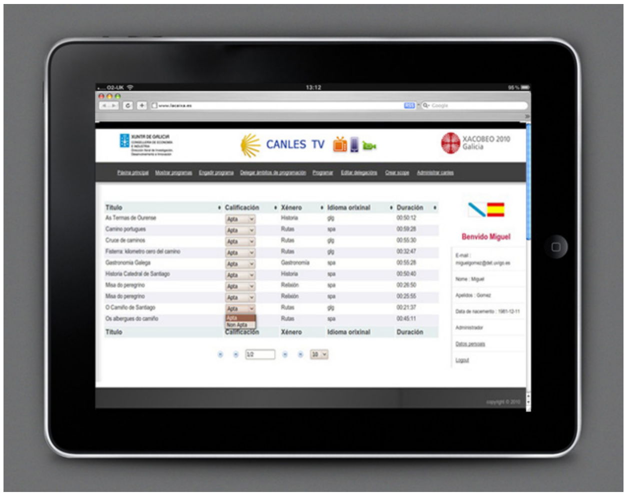
2) Mobile interface: to access this interface, users (typically pilgrims) should have an Android-based smartphone with Internet access. Figure 5 illustrates the mobile interface. Through this interface, users may access customized and geolocated services according to their profile along the Way. The most relevant services available through the mobile interface are:

a) Visualisation of multimedia channels availablein a given geographical range, according to users' location.

b) Geolocation services supported by an embedded GPS device or through Google Maps[1].

c) Multimedia content upload.Content submitted will be supervised by system administrators prior to being stored in the content repository.

d) Content geolocation. Users may indicate a range for the validity of the uploaded content along the Way of St. James.



Titulo	Calificación	Xénero	Idioma orixinal	Duración
As Termas de Ourense	Apta	Historia	glg	00:50:12
Camino portugues	Apta	Rutas	spa	00:59:28
Cruce de caminos	Apta	Rutas	glg	00:55:30
Fisterra: kilometro cero del camino	Apta	Rutas	glg	00:32:47
Gastronomia Galega	Apta	Gastronomia	spa	00:55:28
Historia Catedral de Santiago	Apta	Historia	spa	00:50:40
Misa do peregrino	Apta	Relixión	spa	00:26:50
Misa do peregrino	Apta	Relixión	spa	00:25:55
O Camiño de Santiago	Apta	Rutas	glg	00:21:37
Os albergues do camiño	Apta Non Apta	Rutas	spa	00:45:11
Titulo	Calificación	Xénero	Idioma orixinal	Duración

Benvido Miguel

E-mail : miguelgomez@det.uvigo.es

Nome : Miguel

Apelidos : Gomez

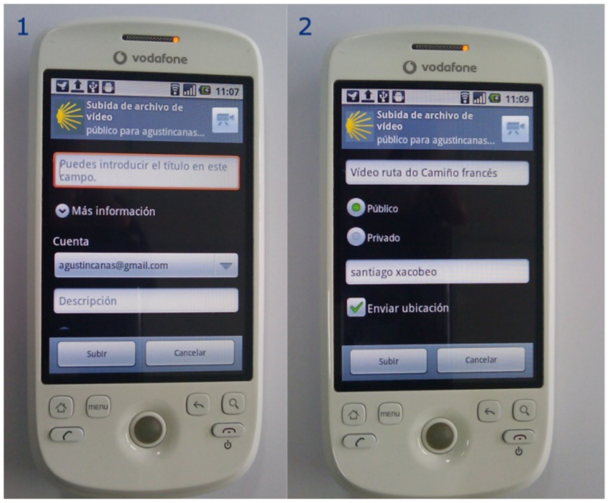
Data de nacemento : 1981-12-11

Administrador

Datos persoais

Logout

Figure 4. Example II of the PC Interface



Subida de archivo de vídeo público para agustincanas...

Puedes introducir el título en este campo.

Más información

Cuenta

agustincanas@gmail.com

Descripción

Subir Cancelar

Subida de archivo de vídeo público para agustincanas...

Vídeo ruta do Camiño francés

● Público

○ Privado

santiago xacobeo

☑ Enviar ubicación

Subir Cancelar

Figure 5. Example of the mobile interface

3) *TV-HTPC interface:* this access interface is targeted to users (i.e., pilgrims) staying in a hostel along the Way. Besides traditional TV channels available in the area, users may visualize specific channels providing multimedia content to prepare and enhance their experience. Figures 6, 7 and 8 provide some examples of this interface. The main services in this case are:

a) Searching of channels in the geographical vicinity of a given hostel, or relevant to it.

b) Multimedia content visualisation.

c) Access to specific content slots in a programming grid.

C. Validation

At the time of writing this paper the developed platform is in its validation stage, and more specifically in the second stagebelow. A collection of mechanisms has been designed to test the usability and evaluate the users' degree of satisfaction. Validation has been organized around two stages:

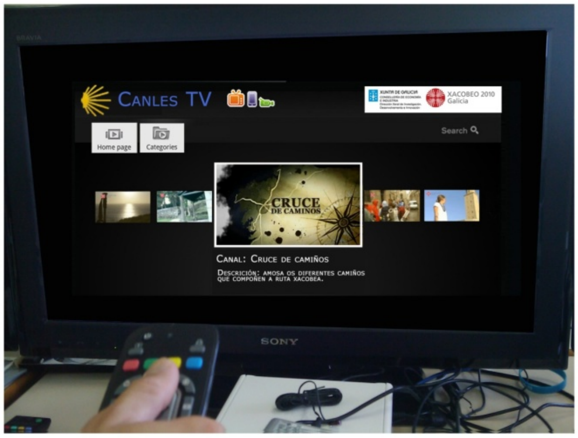


Figure 6. Example I of the TV-HTPC interface

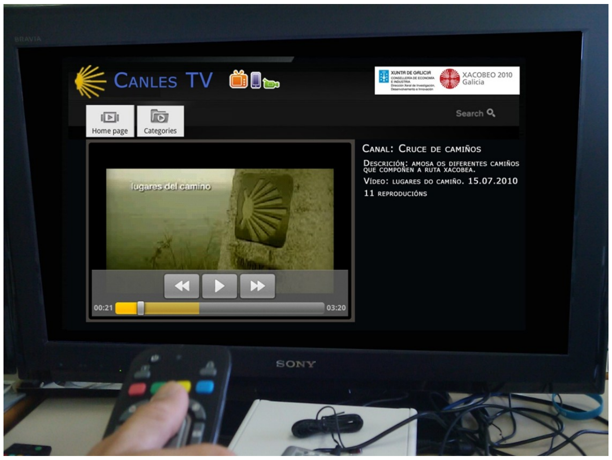
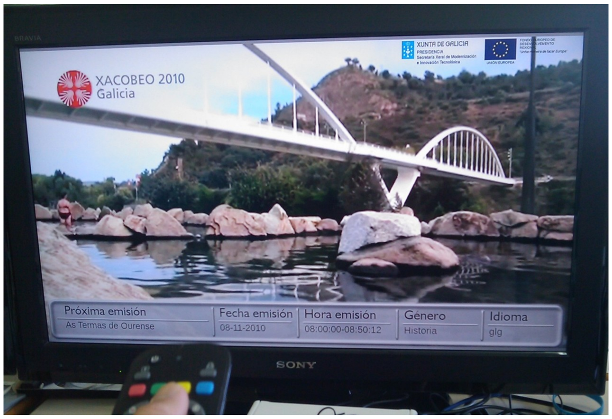


Figure 7. Example II of the TV-HTPC interface



Próxima emisión	Fecha emisión	Hora emisión	Genero	Idioma
As Termas de Ourense	08-11-2010	08:00:00-08:50:12	Historia	glg

Figure 8. Example II of the TV-HTPC interface

- The members of the research group performed the final testing phases of any software project. This testing also included usability assessment. This phase produced a fully functional beta release that was transferred to the second stage.

- Once the beta release has been produced, it was transferred to field test. For that, a representative user sample has been selected including all user profiles in the final system (i.e., pilgrims, administrators, content producers, and programmers). Instruments being used along this validation stage include users' perception surveys about usability and functionality; analysis of user interactions from collected system logs; direct observation of user interactions; and analysis of users' satisfaction in relation to pilgrims not using this system.

V. CONCLUSIONS

The focus of this work is the development of multi-technology, multi-platform multimedia online services. We may consider its outcome as a proof-of-concept that will be eventually deployed to enhance users' experience in the framework of one of the key elements of the Spanish tourism sector, namely the Way of St. James.

This proposal blends some of the most relevant state-of-the-art information and communication technologies to provide anovelonline service. For example, it introduces semantic technologies to characterize user profiles for advanced customisation, and to contribute smart planning algorithms for multimedia content broadcasting. It also integrates later developments in content adaptation and interfacing to configure a multi-device system to enable access to personalised multimedia content from smartphones, TV-HTPC sets and personal computers. The system also supports collaborative content tagging and geolocation.

Privacy of personal data is a major concern in any modern online application. In our case, data collected in the central server about users' interests is utilized to provide personalised recommendations. Apart from the validation of the security mechanisms implemented to protect this information, we are also defining a privacy policy to be eventually enforced when the final system is deployed. The perception surveys performed during the validation phase are also relevant to obtain feedback about users' privacy concerns.

Finally, and besides the technical contributions, this system dramatically improves the quality and relevance of the information available to pilgrims in the Way of St. James, which in turn contributes to the touristic promotion of this world-relevant pilgrimage way.

ACKNOWLEDGMENT

This work has been performed under the support of: (1) the European Regional Development Fund (ERDF) and the Galician Regional Government under projects "Customized and geolocated TV channels about The Way of Saint James" (10TIC050E) and "RedPlir" (CN2012/319); and (2) the Spanish Ministry of Science and Innovation under grant "Methodologies, Architectures and Standards for adaptive and accessible e-learning (Adapt2Learn)" (TIN2010-21735-C02-01).

REFERENCES

[1] Website of http://www.jacobeo.info/index.php. [Last Accessed in November 2012]

[2] Website of http://www.caminosantiago.com/. [Last Accessed in November 2012]

[3] Website of http://www.xacobeo.es/ .[Last Accessed in November 2012]

[4] Official website of http://www.team-mediaportal.com/ [Last Accessed in November 2012]

[5] Official website of Windows Media Center. http://www.microsoft.com/latam/windows/products/windowsvista/features/details/mediacenter.mspx [Last Accessed in November 2012]

[6] Official website of Moovida. http://www.moovida.com/ [Last Accessed in November 2012]

[7] Official website of XBMC. http://xbmc.org/ [Last Accessed in November 2012]

[8] Official website of Google Latitude. http://www.google.com/intl/en_us/latitude/intro.html[Last Accessed in November 2012]

[9] Official website of FireEagle. http://fireeagle.yahoo.net/ [Last Accessed in November 2012]

[10] Official website of Plazes. http://plazes.com/ [Last Accessed in November 2012]

[11] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web. Scientific American (May Issue) ", 2011.

[12] G. Klyne, and J. Carrol, "Resource Description Framework (RDF): Concepts and Abstract Syntax". W3C Recommendation, World Wide Web Consortium, 2004.

[13] D. McGuinness, and F. Harmelen, "OWL Web Ontology Language Overview". W3C Recommendation, World Wide Web Consortium. 2004.

[14] I. Harrocks, P. Patel, B. Harold, S. Tabet, B. Grosof, and M. Dean, "SWRL: A Semantic Web Rule Language Combining OWL and RuleML". W3C Member Submission. World Wide Web Consortium, 2004.

[15] D. Beckett and J. Broekstra, "SPARQL Query Results XML Format". W3C Recommendation, World Wide Web Consortium, 2008.

[16] P. McCarthy, "Introduction to Jena". IBM DeveloperWorks Report, 2004.

[17] V. Haarslev and R. Möller, "Racer: A Core Inference Engine for the Semantic Web". 2nd International Workshop on Evaluation of Ontology-based Tools (EON2003), Sanibel Island, Florida (EE. UU.), 2003, pp. 27-36.

[18] D. Tsarkov and I. Horrocks, "FaCT++ Description Logic Reasoner: System Description". Lecture Notes in Computer Science, 2006, pp. 292-297.

[19] E. Sirin, B. Parsia, B. Cuenca, A. Kalyanpur, and Y. Katz, "Pellet: A practical OWL-DL reasoner. Journal of Web Semantics", vol.5 (2), 2007.

[20] METHONTOLOGY: From Ontological Art Towards Ontological Engineering. http://oa.upm.es/5484/1/METHONTOLOGY_.pdf [Last Accessed in November 2012]

[21] A. De Nicola, M. Missikoff and R. Navigli, "A Proposal for a Unified Process for Ontology Building: UPON", 2005, pp. 655-664.

[22] Official website of Google Maps. http://www.google.com/maps [Last Accessed in November 2012

ISBN: 978-1-61208-256-1

User and Device Adaptable Service Managing Mechanism In Ubiquitous Computing Environment

Jung-Sik Sung , Jong-uk Lee, and Jaedoo Huh
Green Computing Department, IT Convergence Technology Laboratory, ETRI, Daejeon, Korea
jssung@etri.re.kr

Abstract*— Service mobility has become a new issue in the area of service convergence with the advent of versatile mobile devices. Hence, we propose a user and device adaptable service managing mechanism supporting service mobility. This mechanism is performed by adaptive data synchronization. The adaptive data synchronization service performs synchronization with part of data, not all of data, just used in frequent by user. Also, it manages data list in separate. So, it increases the performance of synchronization. The ubiquitous service framework presented in this paper suggests best available service for the user and the device when the user moves to other domains with other device. We implemented a prototype service framework to verify continuity and synchronization of service. Also, we showed a scenario for mobility of video conference service using the prototype service framework.

Keywords-data synchronization;service mobility; user adaptable service management; device adaptable

I. Introduction

In the past decade, many prototypes have been made for both ubiquitous computing and convergence of services, but it is not easy to service automatically according to users' and environment's context. Users want to be provided continuous and user adaptable service with multiple devices by moving the places. So, the design of a convergence service should address mobility, heterogeneity, and user-centric issues [1]-[3]. Service mobility is considered as maintaining a connection even when terminals or networks are changed due to user movement or personal preference [4]. There are several researches to support service mobility. In [5] it was suggested the method supporting service mobility by moving service components between devices in a serving network. But in this approach, information consistency is harder to achieve for a personal service because the data is scattered across several computers and some of them are disconnected to the network. In [6] it was proposed service mobility method based on Bluetooth but the user-centric mobility was not provided. Mobile agent based frameworks [7], [8] were proposed to provide personal mobility in accessing Internet services. In [7] it supports three Internet services, namely, Web, e-mail, and FTP using four assistants: user, HTTP, mail, and FTP assistants. Assistants operate at a proxy server close to the user. In [7] it supports a personalization scheme only. In order to provide true personal mobility that requires the integration of contact and personalization.

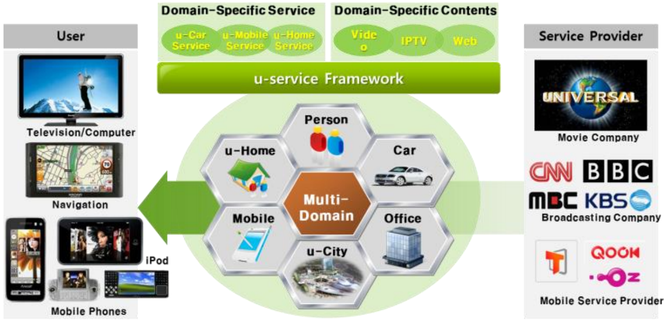


Figure 1. A conceptual model of u-service framework in ubiquitous computing environment

Focusing on service mobility under user movement and heterogeneous devices, a major problem with this service convergence is to build a platform that is applicable to services supported by heterogeneous service platforms and devices with their own platforms. A platform for convergence services plays the role of an infrastructure to execute content and application programs smoothly without obstacles, and provide interoperability between devices and services. We propose an user and device adaptable service managing mechanism in ubiquitous service(u-service) platform to provide the facilities stated above.

Fig. 1 shows a conceptual model of u-service framework in ubiquitous computing environment, where users can be provided continuous services although they move one place to another with different devices. There are several domains in the ubiquitous computing environments, where domains can be defined two terms, one is the device of users such as PC, TV, mobiles and smart phones, and the other is the location of the users, such as home, car, office, hot-spot, and so on. In existing environment, users could only utilize domain-specific services with specific devices. If users move to other domains with other devices, it is difficult to use the service properly, because the device or the network may not support the functions for the service. The u-service framework server presented in this paper suggests the best available service for the user and device, and then constructs and executes a service execution engine for the selected service on the connected device. As the user changes the device due to movement or personal preference, the service is provided continuously with transformed content suitable for the new device.

ISBN: 978-1-61208-256-1

II. u-SERVICE FRAMEWORK ARCHITECTURE

We propose a u-service framework as an open service framework to support convergence services. It is a technology to generalize a ubiquitous computing environment by providing an environment that eases execution and combination of domain-subordinated services and/or contents by organically integrating independent service domains. Therefore, the u-service framework is an optimized integrated service framework that provides continuous services that are not constrained by physical user environments. Fig. 2 shows an architecture of u-service framework proposed in this paper.

The u-service framework supports the registration of service and execution engines using service profiles, the registration of user and device using user and device profile when users login in u-server framework server. Users and service providers can make profiles and rules using a profile authoring tool which generates and edits profiles of device, execution engine and service, vocabularies, and rules. The functions of u-SF Middle in Fig. 2 are provided by the u-service framework manager. The u-service framework manager supports the registration and management of service and execution engines, the management of service profile, device, subscriber, and service category. It controls user accounts and u-service session. The prototype provides a Web- and proprietary GUI-based method for access of the framework so that devices with conventional Web browsers connect to the framework through a Web site, while devices without Web browsers use a proprietary GUI to access the framework. When services, devices or users are registered or modified to the u-service framework, the u-service framework manager updates service lists and service category lists that are suitable for the user. But service lists that are suitable for both the user and the targeted terminal device are computed in real-time when the targeted terminal device is connected to the u-service framework.

The proposed framework can recommend currently available services based on user preference and device characteristics. Fig. 3 describes a learning algorithm according to user preference gathered by examining the service usage history and a recommendation algorithm for services that can be run on the used device. When the user simply selects one of the recommended services, and the according execution engine is then downloaded automatically/dynamically forming an optimized service execution environment. The dynamic configuration of a service execution environment includes the following procedures: user device profiling, a search for user and device specific services, execution engine search procedure to find an engine that suites both the service selected by the user and the user device profile, transfer of selected execution engine to the user's device, and automatic installation/execution of downloaded engine. The downloaded engine is managed by the execution engine loader and updated automatically and periodically.

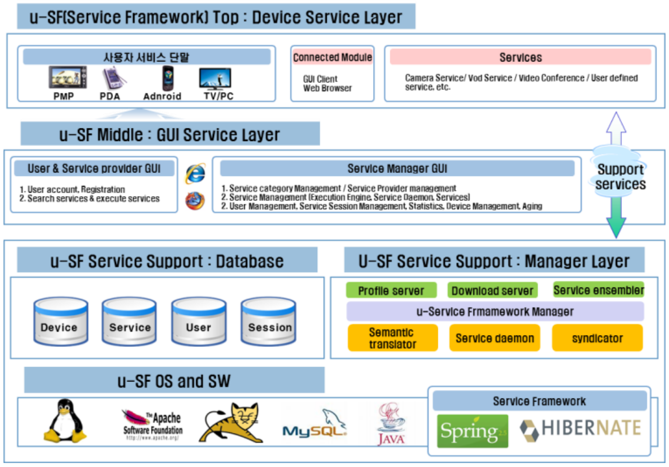

u-SF(Service Framework) Top : Device Service Layer
사용자 서비스 단말
PMP
PDA
Adnroid
TV/PC
Connected Module
GUI Client
Web Browser
Services
Camera Service/ Vod Service / Video Conference / User defined service, etc.
u-SF Middle : GUI Service Layer
User & Service provider GUI
1. User account, Registration
2. Search services & execute services
Service Manager GUI
1. Service category Management / Service Provider management
2. Service Management (Execution Engine, Service Daemon, Services)
2. User Management, Service Session Management, Statistics, Device Management, Aging
Support services
u-SF Service Support : Database
Device
Service
User
Session
U-SF Service Support : Manager Layer
Profile server
Download server
Service ensembler
u-Service Frmamework Manager
Semantic translator
Service daemon
syndicator
u-SF OS and SW
Apache Software Foundation
MySQL
JAVA
Service Framework
Spring
HIBERNATE

Figure 2. The Proposed u-Service Framework Architecture

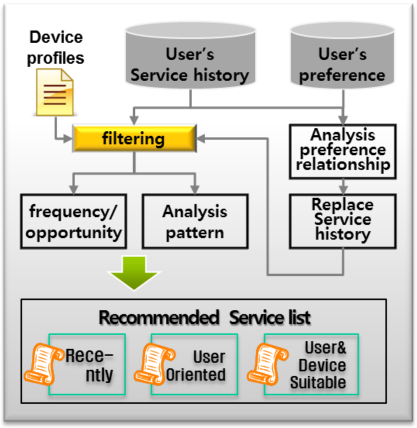


Figure 3. User & device available service recommendation

The framework performs automatic detection of a suitable service execution engine and also supports a search function for services that are suitable for the user and the targeted terminal device. Moreover, it supports real-time content adaptation for targeted terminals, semantic translation including a communication protocol translation, and seamless service continuity so that a user can continue using a service across different terminals. Fig 4 shows processing in series provided by u-service framework in order to support seamless service synchronization at last when a user connects to the u-service framework server.

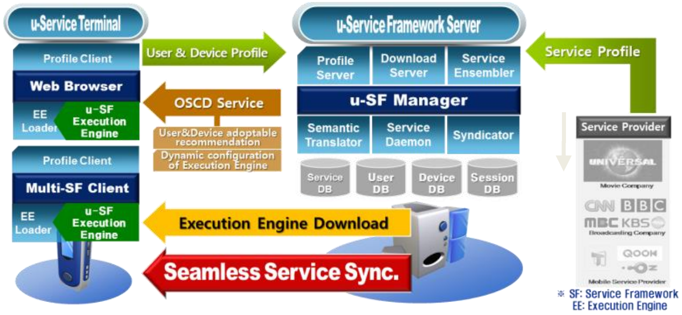


Figure 4. Seamless service synchronization in u-service framework

III. User & Device Adaptable Service Management

We suggest the use and device adaptable service managing mechanism via data synchronization. The syndicator is located on middleware and it provides seamless service in ubiquitous computer environment by supporting data synchronization between terminals and users. Existing data synchronization service has some inefficient problems such as data duplications on terminals and low performance of synchronization. We propose an adaptive data synchronization service mechanism in order to solve these problems. The adaptive data synchronization service does processing of synchronization with part of data used frequently by user not all of data and it manages data list in separate. So, it increases the performance of synchronization. Fig. 5 shows functional components of the syndicator, adaptive data synchronization block and a syndicator manager.

The Syndicator manager maintains information of users and devices, which are independent with a synchronization target. It also communicates with each service synchronization terminal.

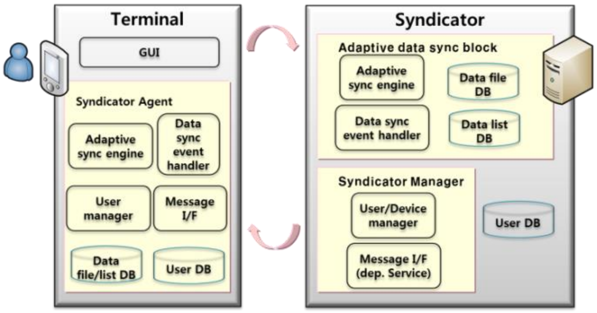


Figure 5. An architecture of the syndicator

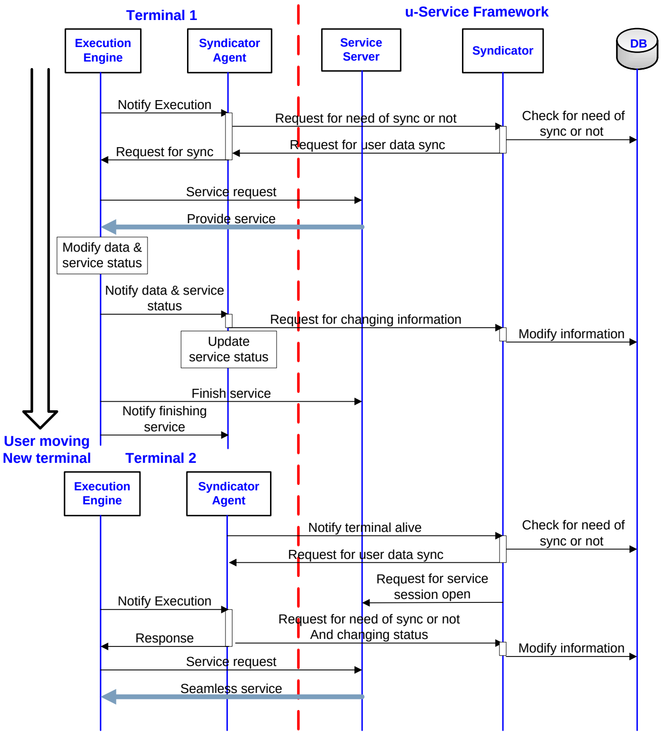


Figure 6. The procedure of seamless service synchronization

The adaptive data synchronization block supporting a data synchronization function composes of a data sync event handler and an adaptive sync engine. The data event handler provides data synchronization based on data update event, while the adaptive sync engine generates adaptive synchronization files based on terminal's characteristics. The adaptive sync engine supports data synchronization with data file DB, data catalogue DB, and user information DB. The followings are functions in order to provide data synchronization.

- Synchronization recent file and data list according to the events of a terminal
- Generation of an adaptive synchronization file list according to the characteristics of a terminal
- Remote synchronization with a terminal using service

First and the last functions are executed by the data sync event handler while second one is executed by the adaptive sync engine. Fig. 6 shows the procedure for supporting service synchronization seamlessly using the syndicator of a u-service framework server and the syndicator agent of a terminal.

IV. A Scenario for Supporting Service Synchronization

We composed a scenario as shown Fig. 7 for supporting service synchronization such as managing service status continuously and providing the data synchronization. The syndicator of the u-service framework is implemented using Java and GNU C++ developing languages. Three kinds of terminals like Android, Windows 7, and openSUSE11 were used to test service synchronization seamlessly by switching user's terminals. The scenario is as follows: User-1 is talking with video conference server using a smartphone on his way office. When he arrived at his office, he wanted to keep video conference using his laptop. So, he moves his call from the smartphone to the laptop, the session between the two devices (the smartphone and the video conference server) is closed.

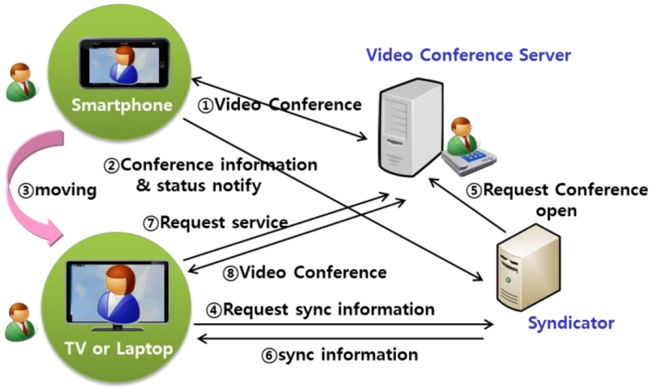
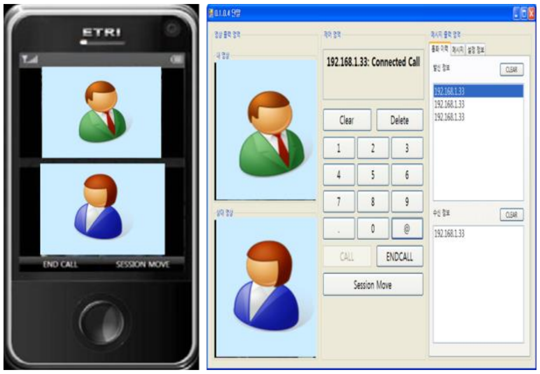


Figure 7. A video conference service scenario for supporting service synchronization

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-256-1



192.168.1.33: Connected Call

Figure 8. Video conference program capture screens of the smartphone & the laptop

A new link is established between User-1's new device, the laptop and the video conference server. After completing the new connection process between two devices, User-1 can seamlessly talk with the video conference with his new device, the laptop. This service synchronization was provided by the syndicator. The syndicator of the u-service framework supports data synchronization between the smartphone and the laptop and manages video conference information and status continuously. It also asks the video conference server to open the conference with the new device and notifies the synchronization information to the new device. Fig. 8 shows the actual demonstration environment for the scenario in Fig. 8. The video conference execution engine program can detect the device's camera and display the owner in the upper side of the displayer and the other in the lower side as shown in Fig. 8. If the device has no displayer, then only voice can be transferred between users. The left side of Fig. 8 illustrates the program for mobile devices such as smartphone, PMP, PDA, and so on, while the right one is the execution engine program on windows 7. Using the "Session Move" button in Fig. 8 users can change their devices during a call. If they press the "Session Move" button, the agent of syndicator in device notifies the conference information to the syndicator. So, the user moves one domain to another, he can be on the continuous phone with the other using his device that he has at that time. Using the proposed video conference service, users can use video conference services anywhere, any devices and any network.

V. Conclusion

In ubiquitous computing environment, users move from one domain to another with many kinds of devices usable in each domain. In that case, users want to use services seamlessly irrespective of their location and devices they have. The technique necessary in this case is called service mobility. This paper proposed a ubiquitous service framework that supports convergence services including heterogeneous service platforms and devices with their own independent platforms. It plays the role of an infrastructure to execute content and application programs with a dynamic configuration using mechanisms such as user preference learning, service and execution engine profiling, and real-time device profiling. It also supports service mobility to provide continuity and service synchronization when terminals or networks are changed due to user's movement or a change of personal preference. Existing data synchronization service has some inefficient problems such as data duplications on terminals and low performance of synchronization. But our proposal adaptive data synchronization service mechanism solved these problems. It processed synchronization with part of data used frequently by user, not all of data and it manages data list in separate. So, it increases the performance of synchronization

We implemented a prototype service framework to verify continuity and synchronization of service. We describe a scenario for mobility of video conference service in ubiquitous computing environment where a user moves from one domain to another with being provided seamless service through his device. Also, we design and implement the model for video telephony mobility and its clients. Using our scheme, users can use video telephony anywhere, any devices and any network. We showed not only service mobility of one user's migration but also of multiple users' sharing one single session among them. In order to minimize the delay time for seamless service mobility [9], a further study such as a reliable prediction about the movable target terminal through the analysis of user's context information is needed.

Acknowledgment

This work was supported by the IT R&D Standardization program of MKE/KATS, [2011-PM10-02, Development of Smart Utility based Green WPAN Standardization]

Reference

[1] E. Lavinal, N. Simoni, M. Song, and B. Mathieu, "A Next-Generation Service Overlay Architecure," Annals of Telecommunications, vol. 64, no. 3-4, Apr. 2009, pp. 175-185.

[2] P. Maniatis, et al., "The Mobile People Architecture," Mobile Computing and Communications Review, vol. 1, no. 2, July 1999, pp. 36–42.

[3] H. J. Wang, et al., "ICEBERG: An Internet-core Network Architecture for Integrated Communication," IEEE Personal Communications, Aug. 2000, pp. 10–19.

[4] H. Si, Y. Wang, J. Yuan, and X. Shan, "A Framework and Prototype for Service Mobility," 2009 World Congress on Computer Science and Information Engineering, Mar. 2009, pp. 315-319.

[5] Z. Chen, C. Lin, and X. Wei, "Enabling On-Demand Internet Video Streaming Services to Multi-terminal Users in large scale," IEEE Transactions on Consumer Electronics, vol. 55, no. 4, Nov. 2009, pp. 1988-1996.

[6] M. Hasegawa, H. Morikawa, M. Inoue, U. Bandare, H. Murakami, and K. Mahmud, "Cross-device handover using the service mobility proxy," in Proceedings of 6th International Symposium on Wireless Personal Multimedia Comunications (WPMC2003), vol. 2, Yokosuka, Japan, October 2003, pp. 357–361.

[7] A. D. Stefano and C. Santoro, "NetChaser: Agent Support for Personal Mobility," IEEE Internet Computing, Mar. 2000, pp. 74–79.

[8] B. Thai, R. Wan, and A. Seneviratne, "Integrated Personal Mobility Architecture: a Complete Personal Mobility Solution," ACM Mobile Networks and Applications, Feb. 2003, pp. 27–36.

[9] Y. Cui, K. Nahrstedt, and D. Xu, "Seamless User-Level Handoff in Ubiquitous Multimedia Service Deliver," Multimedia Tools and Applications, vol. 22, no. 2, 2004, pp. 137–170.

Automated Audio-visual Dialogs over Internet to Assist Dependant People

Thierry Simonnet, Samuel Ben Hamou
R&D department
ESIEE-Paris
Noisy le Grand, France
{t.simonnet, s.benhamou}@esiee.fr

***Abstract*—With today's advancements in medical treatments and care fields, an increasing number of people need help or assistance at home. Concerned categories are mostly elderly, isolated or disabled persons even though persons with mild cognitive impairment are also considered. One of the main issues these categories are facing is the lack of constant communication to help maintaining some kind of social link, with families, friends and eventually caregivers. Moreover, the fact that current hardware accessories and technologies are not always suited to allow for such functionalities, in a generic situation, tend to increase that particular gap. Many studies suggest the use of Automatic Speech Recognition (ASR) to have a global control over devices and communication means. The resulting architecture allows for a pseudo butler to be put in place, serving as the main entry point to manage daily common communication tasks as well as giving access to different web services via a voice controlled environment. Once put in application, it could stand as a complete and integrated solution for remote communication and monitoring of the designated population target.**

Keywords-ASR; Voice over IP (VoIP).

I. INTRODUCTION

Our ongoing research projects are focusing on creating a unified solution/channel to be used for daily tasks management but also audio/video communications between people. The main reason for this unified solution comes from an increasing demand for maintaining dependent people at home [5][9]. In [20], the World Health Organization assessed the restructuration of hospital services, with an increased role for substitution between different levels of care, strengthening primary health care services, increasing patient choice and participation in health services and improving outcomes through technology assessment and quality development initiatives. According to these recommendations, the number of telesurveillance implementations and pilot experimentations has been growing in Europe, especially within the Northern countries.

To reduce financial costs, hospitals load but also improve the patients quality of life, it has been recently considered to keep them at home, thus a need for suitable communication and telecare technologies has arisen. For such a specific panel, we also often need medical assistance. This has a direct implication on the quality of services as we need reliable communication tools and a really easy learning curve for the end users. We are relying on IP technologies as Asymetric Digital subscriber Lines (ADSL [28]) are available all across Europe for affordable prices. It has a small drawback as upload bandwidth is generally limited and thus has to be taken into account for data/video communications. Also, considering that mobile networks are relying on the same technology for data exchange, IP solutions represent the obvious choice as they will probably be suited for future network evolutions. At the moment, in order to offer a good service, we have to focus on video and audio quality, which are directly linked to the available bandwidth and compression algorithms.

We will present the environment with a platform overview and an explanation of our choices in Section 2, while Section 3 will focus on technical descriptions. Integration will be covered in Section 4 and results in Section 5. The conclusion will focus on identifying current and remaining issues, but also put this in perspective.

II. REMOTE AUTOMATIC SPEECH RECOGNITION INTERFACE

A. Speech recognition

The most natural and obvious way for humans to interact and communicate nowadays is speech. Considering our end-users, who may not always be acquainted with traditional computer interfaces, it makes sense to focus on Automatic Speech Recognition technologies as the primary way of interaction with the system. It allows for vocal commands to be passed, but is also able to eventually identify mood states or for example detect particular/distress situation.

B. Usage scenarios

The system has to pose as a virtual butler with access to a centralized and collective memory database, with either audio and or visual representations. Some ways to interact with it are as follow:

- Find his way: the butler, as a service, can be used on a gps enabled smartphone allowing some guidance.
- Manage a diary, appointments, bills payment...
- Answer the phone, messaging, mail...
- Find information on the web.
- Detect abnormal situations, behaviors through a wearable vital/actimetric device [2].

ISBN: 978-1-61208-256-1

- Provide recipes; keep history of menus prepared for friends/family.
- Remember faces/names/information through the phone camera.

Some of these features are already available on smartphones, others are being developed such as the Microsoft MyLifeBits project [7].

III. VOIP ARCHITECTURE AND SERVICES

A. Existing Platform

As part of different projects (CompanionAble, vAssist) the current platform can take many shapes even though some areas are commonly shared:

- Asterisk Internet Protocol Private Branch eXchange (IPBX) for all video/audio communications.
- Julius ASR server.

At the users' homes:

- At least one platform featuring a camera, display and VoIP client (computer/phone/tablet/tvbox).
- Sensors for person monitoring.

B. A Unified and standardized communication solution

As a result of the devices heterogeneity, we needed to be able to handle different kinds of media. The VoIP solution allows us not only to take care of that aspect, but gives us the possibility to extend functionalities via plugin developments. Moreover, this infrastructure has the ability to be inserted into a public VoIP network for cross domains/technologies communications. Current advantages are:

- Support for various Internet infrastructures (e.g., public/private IP, ADSL box).
- Interoperability with public and private telecommunication networks (e.g., google talk, skype).
- Low latency (less than 100ms with H263 video)
- Automatic bandwidth adjustment for Quality Of Service.
- Support for various clients (e.g., softphones -phone softwares-, IP phones, mobile phones).
- Large choice of audio and video codecs.
- Ability to set up centralized services (low cost of deployment) as IVR (Interactive Voice Responce), ASR, multi-conferencing, voice and video messaging.
- The user is linked to a unique identifier (a phone number).
- Centralization of data (voice, video).
- Out of the box internationalization.

C. Communication infrastructure

As VoIP solutions imply the use of a PBX, we decided to use Asterisk from DIGIUM. It has standard configurations for regular calls but allows us to tweak it extensively for our purposes. Regarding the fact that patient networks will use private IP addresses, we initially believed a local PBX was needed not only for local communications, but also for call forwardings/translation from the public to private domains. After some tests, we would probably need this in some restricted cases for PSTN translations, but for the average user, it will probably not be necessary (depending on the id-number allocation).

When a call is started, a SIP [23] request is sent to the PBX, which transmits it to the end-client. When this signalling communication is done, a direct tunnel is established using RTP (Real Time protocol) [24]; (see Fig. 1). This protocol, over UDP [25], keeps the packet order and drops old ones. Fig. 2 shows how different components are set on ISO layers. Depending on the service definition, it might be necessary to use trunking services to allow all communications through PBXs (see Fig. 3).

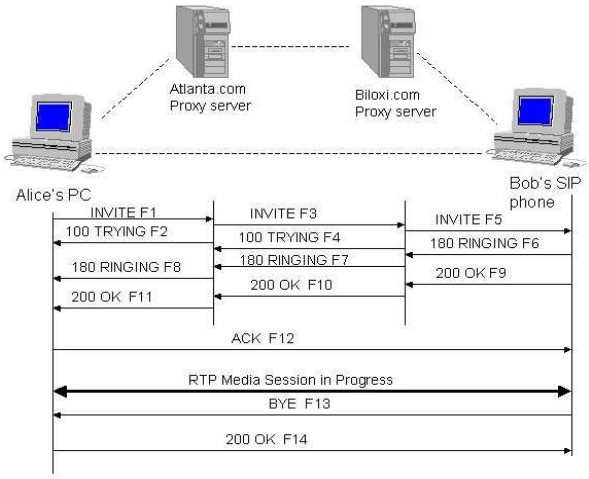


Figure 1. Call Dialog.

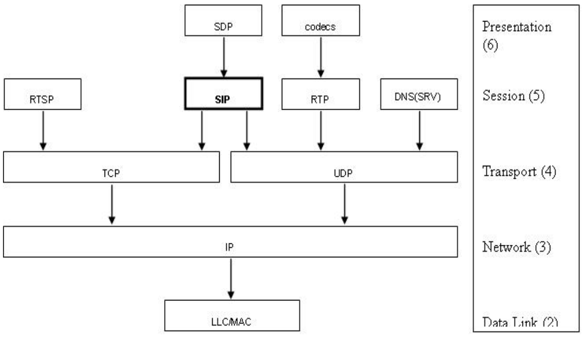


Figure 2. SIP and OSI.

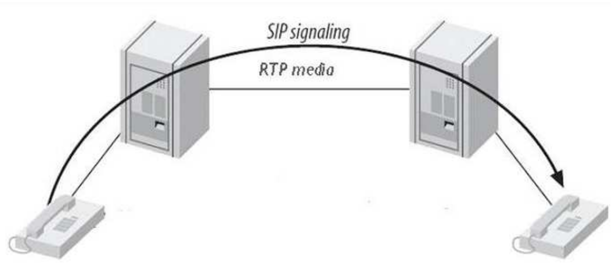


Figure 3. SIP trunking architecture.

1) A codec, is a module that can COde and DECode an analog or a digital signal. For VoIP communications, many codecs are available. As PBXs are not designed for stream translation, we initially needed to make sure both clients used the same normalized codecs. Later on we worked with a transcode plugin to eventually reduce this impact.

Asterisk can handle at least:

- Voice: ulaw, alaw, gsm, ilbc, speex [10][22], g726, adpcm, lpc10, g729, g723;
- Video: h261 [11], h263 [12], h263+, h264 [13][19], MPEG-4, VP8.

For our systems, we decided to use: µlaw, alaw, speex for the voice encoding and H261, H263 and H264 for the video part. The key point is finding a good fit between "compression", "delay" and "video quality" as increasing the compression rate increases the delay due to buffer use and a higher processing load per time unit.

2) Alarms: There are multiple ways to handle and transmit alarm signals he goal of the vAssist project is to provide specific voice controlled Home Care and Communication Services for two target groups of older persons: Seniors suffering from chronic diseases and persons suffering from (fine) motor skills impairments. The main goal is the development of simplified and adapted interface variants for tele-medical and communication applications using multilingual natural speech and voice interaction (and supportive graphical user interfaces where necessary).and all could be implemented in parallel. The first one is to use the SIP MESSAGE method. (see Table 1 for SIP Methods). As Asterisk does not handle it, it is necessary to implement RFC 3428 [26] for SIP channel. We also could use T.140 (RFC 4103 [27]) method for Instant Messaging/Alarms communications. Last would be to simply automate calls/messages/mails to emergency services when a specific signal has been sent from the monitoring device.

TABLE I. SIP METHODS

SIP Method	*Description*	*RFC*
ACK	Acknowledge final response to Invite	3261
BYE	Terminate a session	3261
CANCEL	Cancel a previous call	3261
INFO	Mid-session signaling	2976
INVITE	Initiate a session	3261
MESSAGE	Allows the transfer of IMs	3428
NOTIFY	Event notification	3265
OPTIONS	Query to find the capabilities	3261
PRACK	Acknowledgement for Provisional responses	3262
PUBLISH	Publish event state	3903
REFER	Transfer user to a 3rd party	3515
REGISTER	Register with a SIP network	3261
SUBSCRIBE	Request asynchronous event notification	3265
UPDATE	Update parameters of a session	3311

D. Performances

Two different VoIP clients are currently used for performances and codecs compatibility: ekiga [29] for the PC platform and linphone [30] for either PC, Android, iOS platforms. These clients are customized for HD and low delays communication. We currently use wideband Speex audio codec and H263 or H264 video codecs depending on the platform with a specific bandwidth adaptation module. It makes sure instant messaging and voice delays are being kept as low as possible by reducing video resolution in case of congestion. This ensures low delays communication over internet (less than 100ms for a PC to PC communication over internet, less than 200ms for a PC to smartphone communication using WiFi). Tests with other standard VoIP clients and skype gave delays between 200ms and 500ms for long term communication (more than 3 hours long). All these tests were done between two private networks with their own Asterisk IPBX.

IV. VOICE-BASED SYSTEM INTERFACE

A. ASR and VoIP

The main advantage of such a centralized platform is that services and tools can be accessed with all connected devices. Regarding ASR, there is no embedded ASR tool into Asterisk. Julius, an Open source project, offered all the services we needed and has the ability to redirect either input and/or output streams to any ip socket. Its speed and real time speech recognition for large vocabulary made it a perfect candidate for this purpose. We found the app_julius module [15], developed by Danijel Korzinek and Dikshit Thapar, which allowed us to make a direct connection between Asterisk and Julius which follows this flow :

- Create an ASR object using SpeechCreate()
- Activate your grammars using SpeechActivateGrammar(Grammar Name)
- Call SpeechStart() to indicate you are going to do recognize speech immediately
- Play back your audio and wait for recognition using SpeechBackground(Sound File|Timeout)
- Check the results and do things based on them
- Deactivate your grammars using SpeechDeactivateGrammar(Grammar Name)
- Destroy your speech recognition object using SpeechDestroy()

A simple macro is used in the dialplan to confirm word recognition. ARG1 is equal to the file to play back after "I heard..." is played.

```
   [macro-speech-confirm]
exten => s,1,SpeechActivateGrammar(yes_no)
exten => s,2,Set(OLDTEXT0= ${SPEECH_TEXT(0)})
exten => s,3,Playback(heard)
exten => s,4,Playback(${ARG1})
exten => s,5,SpeechStart()
exten => s,6,SpeechBackground(correct)
exten => s,7,Set(CONFIRM=${SPEECH_TEXT(0)})
exten => s,8,GotoIf($["${SPEECH_TEXT(0)}" = "1"]?9:10)
exten => s,9,Set(CONFIRM=yes)
exten => s,10,Set(CONFIRMED=${OLDTEXT0})
exten => s,11,SpeechDeactivateGrammar(yes_no)
```

The voice-based MMI (Maximum Mutual Information) functionality uses a voice recognition module based on Julius and HTK (Hidden Markov Toolkit) softwares (Julius for recognition, HTK for training) with adaptation facilities to customize the system to our speakers' constraints and needs..

B. Julius, HTK

The voice recognition module is based on the use of conventional Hidden Markov Models (HMM) to model statistically the acoustic models of phonemes and / or words in the vocabulary. We use software tools such as HTK [21] and Julius [16]. Language models (linguistic probabilities, which are complementary to acoustic probabilities) are implicitly addressed in the use of such models to make robust word recognition in a given sentence (use of statistical N-grams and rules of grammar).

V. CURRENT PROJECTS

The following projects came to life after realizing that no particular solution was fitted to offer a unified solution for this matter. Of course we could find some products purely specialized in the video or audio communication or more recently some virtual “butlers” like SIRI started to appear though with limited functionalities of the medical context, but regarding the unified user experience, the simplicity of use, a lot of work could have been done.

A. CompanionAble

The main idea behind the CompanionAble project was to provide a synergy between Robotics and Ambient Intelligence technologies and integrate them into a fully assistive environment. A robotic companion was working collaboratively with a smart home environment. CompanionAble addressed the issues of social inclusion and homecare of persons suffering from chronic cognitive disabilities prevalent among the increasing European older population. This is obviously a subsection of a more generic group of persons with elevated requirements and constraints. ASR has been used for service managements and SIP technologies have been put in place for audio/video communications, integrated into a standardized GUI. Yet, the two technologies were not linked and ran concurrently. Also, SIP had been proposed to handle commands for the robot movements via the messaging service. Usage of the robot proved to be quite accepted but costs and infrastructure requirements (a smarthome environment) unfortunately reserved it to a very few: It would be unable to fit in a small flat as the ones found in the big european cities and at the same time would not be able to handle stairs in the case of a house with multiple floors.

B. vAssist

As for CompanionAble, the goal of the vAssist project is to provide specific voice controlled Home Care and Communication Services for two target groups of older persons: Seniors suffering from chronic diseases and persons suffering from (fine) motor skills impairments. The main idea is the development of simplified and adapted interface variants for tele-medical and communication applications. The main difference with the previous project stands in the two following points:

- To target a wider audience, standard equipment is preferred. It reduces the development costs and existing hardwares and interfaces in the home of the users can be used such as PC, TV, mobile phone or tablets.
- Every service of the system must be defined not only to use graphical user interfaces but also multilingual natural speech and voice interaction. In a sense, existing services can be adapted/enhanced to support these aspects.

In this aspect, Asterisk and Julius represent the first accesspoints to such a service.

VI. CONCLUSION AND FUTURE WORK

The infrastructure for testing a physical or virtual butler is in place. Open source software components were selected for telecommunications (PBX - Asterisk) and for automatic processing of speech (Julius). Experimental results were obtained during the CompanionAble project. Under vAssist, more common devices (smartphones, tablets...) will be preferred [1]. The Asterisk server is ready for testing services related to usage scenarios listed in Section 2.

So far, telephony signals have been sampled at 8kHz but our experimentations showed that we would probably need to work with higher-rates codecs (e.g., Speex 16kHz), better acoustic models and then finally to improve the platform from Narrowband to Wideband.

It is definitely interesting to achieve such a flexible level of communication using open source softwares. Although we would need to work more on the modelization of more robust acoustic models for ASR (in order that it is capable to handle to increase the recognition rates), all the needed infrastructure is ready to be used and to make progress towards multiple kinds of applications, in many contexts (e.g., telemedicine, security, vocal commands, etc).

Acknowledgment

Some parts of this research leading to these results have received funding from the European Commission Seventh Framework Program (FP7/2007-2013) under grant agreement number 216487.

References

[1] N. Armstrong, C. Nugent, G. Moore, and D. Finlay, “Using smartphones to address the needs of persons with Alzheimer's disease,” Annales des Télécommunications, vol. 65, pp. 485-495, 2010.

[2] J.L. Baldinger, et al., “Tele-surveillance System for Patient at Home: the MEDIVILLE system,” 9th International Conference, ICCHP 2004, Paris France, Series : Lecture Notes in Computer Science, Ed. Springer, 2006.

[3] R. Bayeh, “Reconnaissance de la Parole Multilingue: Adaptation de Modeles Acoustiques Multilingues vers une langue cible,” Thèse (Doctorat) TELECOM Paristech, 2009.

[4] D.R.S Caon, et al., “Experiments on acoustic model supervised adaptation and evaluation by k-fold cross validation technique,” In: ISIVC. 5th International Symposium on I/V Communications and Mobile Networks. Rabat, Morocco: IEEE, 2010.

[5] N. Clement, C. Tennant, and C. Muwanga, “Polytrauma in the elderly: predictors of the cause and time of death,” Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine, v. 18, n. 1, p. 26, 2010. ISSN 1757-7241, http://www.sjtrem.com/content/18/1/26, [retrieved: February, 2013]

[6] A. Constantinescu and G. Chollet, “On cross-language experiments and data-driven units for alisp (automatic language independent speech processing),” In: IEEE Workshop on Automatic Speech Recognition and Understanding. Santa Barbara, CA, USA, 1997, pp. 606-613.

[7] Digium, The Open Source PBX & Telephony Platform, http://www.asterisk.org/, [retrieved: February, 2013].

[8] J. Gemmell, G. Bell, and R. Lueder, “MyLifeBits: a personal database for everything,” Communications of the ACM, vol. 49, Issue 1, pp. 88-95, 2006. http://research.microsoft.com/en-us/projects/mylifebits/, [retrieved: February, 2013].

[9] L.N. Gitlin and T. Vause Earland, “Améliorer la qualité de vie des personnes atteintes de démence: le rôle de l'approche non pharmacologique en réadaptation,” J.H. Stone, M. Blouin, editors. International Encyclopedia of Rehabilitation, 2011. http://cirrie.buffalo.edu/encyclopedia/fr/article/28/, [retrieved: February, 2013].

[10] G. Herlein, J. Valin, A. Heggestad, and A. Moizard, “RTP Payload Format for the Speex Codec,” draft-ietf-avt-rtp-speex-07, http://tools.ietf.org/html/draft-ietf-avt-rtp-speex-07 , 2009, [retrieved: February, 2013].

[11] International Telecommunication Union, “H.261: Video codec for audiovisual services at p x 64 kbit/s,” Line Transmission of Non-Telephone Signals, 1993.

[12] International Telecommunication Union, “H263: Video coding for low bit rate communication,” SERIES H: Audiovisual and Multimedia Systems Infrastructure of audiovisual services, Coding of moving Video, 2005.

[13] International Telecommunication Union, “H264: Advanced video coding for generic audiovisual services”, SERIES H: Audiovisual and Multimedia Systems Infrastructure of audiovisual services, Coding of moving Video, 2003.

[14] Julius ASR, http://julius.sourceforge.jp/en_index.php, [retrieved: February, 2013].

[15] D. Korzinek, module app_julius, http://forge.asterisk.org/gf/project/julius/, [retrieved: May, 2012].

[16] A. Lee, T. Kawahara, and K. Shikano, “Julius - an open source real-time large vocabulary recognition engine,” EUROSPEECH, pp. 1691-1694, 2001.

[17] A.S. Rigaud, et al., “Un exemple d'aide informatisé á domicile pour l'accompagnement de la maladie d'Alzheimer : le projet TANDEM”, NPG Neurologie - Psychiatrie - Gériatrie. N°6, Vol.10, ISSN :1627-4830, LDAM édition/Elsevier, ScienceDirect, April 2010, pp 71-76.

[18] T. Schultz and K. Katrin, “Multilingual Speech Processing,” Elsevier, 2006.

[19] T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra, “Overview of the H.264/AVC Video Coding Standard,” IEEE Transactions on Circuits and Systems for Video Technology, 2003.

[20] World Health Organization, The European Health Report, European Series, #97, 2002.

[21] S. Young, et al., “The HTK Book”, version 3.4. Cambridge, UK: Cambridge University Engineering Department, 2006.

[22] Xiph.Org Foundation, “Speex: A Free Codec For Free Speech”, http://speex.org/, [retrieved: February, 2013].

[23] SIP protocol, RFC 3261, http://www.ietf.org/rfc/rfc3261.txt, [retrieved: February, 2013].

[24] RTP , RFC 3550, http://www.ietf.org/rfc/rfc3550.txt, [retrieved: February, 2013].

[25] UDP, RFC 0768, http://www.ietf.org/rfc/rfc0768.txt, [retrieved: February, 2013].

[26] SIP Message Extension, RFC 3428, http://www.ietf.org/rfc/rfc3428.txt, [retrieved: February, 2013].

[27] RTP Payload for Text Conversation, RFC 4103, http://www.ietf.org/rfc/rfc4103.txt, [retrieved: February, 2013].

[28] http://en.wikipedia.org/wiki/Asymmetric_Digital_Subscriber_Line, [retrieved: February, 2013].

[29] http://www.ekiga.org, [retrieved: February, 2013].

[30] http://www.linphone.org, [retrieved: February, 2013].

Tiered Interior Gateway Routing Protocol

Yoshhiro Nozaki, Parth Bakshi, and Nirmala Shenoy
College of Computing and Information Science
Rochester Institute of Technology
Rochester, NY, USA
{yxn4279, pab8754, nxsvks}@rit.edu

***Abstract*— Most ISPs and Autonomous Systems on the Internet today use Open Shortest Path First (OSPF) or Intermediate-System-to-Intermediate-System (IS-IS) as the Interior Gateway Protocol (IGP). Both protocols are Link-State routing protocols and require distribution of link state information to all routers. Topological changes require redistributing updates and refreshing routing tables, resulting in high convergence times. Routing table sizes grow linearly with network size, indicating scalability issues. Future Internet initiatives provide new venues to address the routing problem. In this article, a Tiered Routing Protocol (TRP) is presented as a candidate protocol for intra-AS routing. TRP is supported by a tiered addressing scheme. TRP replaces both IP and the routing protocol. TRP's performance is compared with OSPF using Emulab test-beds.**

Keywords-Intra-domain Routing; Network Convergence; Internetworking Architectures; Tiered architectures; Routing Table sizes.

I. Introduction

In IP networks, routers use routing protocols to discover and maintain routes to other networks. Routing table sizes maintained by current routing protocols increase linearly with increase in network size and is indicative of scalability issues which can manifest as performance deterioration. Also, the time taken for the network to adapt to topological changes increases with network size resulting in higher convergence times during which routing is unreliable. Patch and evolutionary solutions address the problem both at inter and intra domain level [1, 2].

Interior Gateway Protocols (IGP) such as Routing Information Protocol and OSPF were designed to work with IP. Large ISP networks use Link-State (LS) IGPs such as IS-IS or OSPF which uses the area concept to segment networks into manageable size. LS routing protocols require periodic updates and redistribution of updates to all routers in the network on link state changes. Each router running the LS routing protocol executes the Dijkstra's algorithm on the link state information to populate routing tables. Dissemination of network-wide (or area-wide) link state information adversely impacts scalability and convergence time when using OSPF.

A primary contribution in this work is the decoupling of the routing table sizes from the network size. A major goal was to investigate a solution that is acceptable to the service provider community. Thus, the proposed internetworking model derives from the structures used by ISPs to define their business relationships namely the *tiers*. The routing protocol proposed under this internetworking model is called the *tiered routing protocol* (TRP). A new tiered addressing scheme was introduced. The tiered address inherits attributes of the tiered structures. To decouple dependencies between connected entities, a nesting concept is introduced [3].

TRP replaces both IP and routing protocol. In this article, TRP operation as an IGP is described and evaluated. The tiered structure within an AS is identified and used for the purpose. Its performance is compared with OSPF using Emulab [4] test-beds. The performance metrics evaluated were: initial convergence times, convergence times after link failures, routing tables sizes, and control overhead during initial convergence and convergence after link failure.

Section II describes some related work in reduction of convergence times in IGPs. Section III describes the two routing protocols under study. Section IV provides details of the emulations tests and Section V analyses the results of the tests. Section VI provides the conclusions.

II. Related Work

Significant research effort has been directed towards reduction and optimization in IGP convergence time to link state changes in the network. The work can be broadly categorized into: (a) reducing failure detection time and (b) reducing routing information update time.

A. Reduction in Failure Detection Time

Layer-2 notification is used to achieve sub-second link / node failure detection. This relies on types of network interfaces and does not apply to switched Ethernet [5].

Hello protocol is used to identify link/node failure in many routing protocols and is called layer-3 failure detection. OSPF sends *hello* packets to adjacent routers at regular intervals. On missing four *hello* packets consecutively, OSPF routers recognize an adjacency failure with a neighboring router. Reducing *hello* packet interval time to sub-seconds can significantly reduce the failure detection time at the expense of increased bandwidth use.

B. Reduction in Link State Propagation Time

Although link/node failure detection time can be reduced to sub-seconds, propagating the link status to all routers in the network takes time and is dependent on the network size.

To reduce such delays, several pre-computed back up routing path approaches have been proposed. Pan et al. [6] proposed the Multi Protocol Label Switching (MPLS) based on a back up path to reroute around failures. However, having all possible MPLS back up paths in a network is not efficient. Multiple Routing Configurations (MRC) [7] uses a small set of backup routing paths to allow immediate packet forwarding on failure detection. A router in MRC maintains additional routing information on alternative paths. MRC guarantees recovery from only single failures. Liu at el. [8] proposed use of pre-computed rerouting paths if resolved locally. Otherwise multi-hop rerouting path had to be set up by signaling to a minimal number of upstream routers.

While the above two delays are of significance, SPF recalculation time can also be almost a second in large networks [5]. As packet loss/delay or routing loops occur during convergence, it is important to reduce this time.

III. Routing Protocols and Operations

In this section, we describe the operations of the two protocols that are OSPF and TRP. In the case of OSPF, only a few basic operations necessary to explain the performance metrics are presented. Details are available in [1]. TRP operations include implementing tiered structures within an AS, tiered address allocation to devices in the tiers, routing table maintenance with TRP, and the packet forwarding algorithm and failure handling.

A. Open Shortest Path First (OSPF)

Basic operations of OSPF include: (a) establishing adjacencies with neighbors and electing a Designated Router (DR) and a Backup DR (BDR); (b) maintaining Link State Database (LSDB) and; (c) executing Dijkstra's algorithm. The operations are invoked during startup and also in response to link state changes. Convergence in each case is impacted differently and described separately below.

1) Initial Convergence in OSPF

a) Establishing Adjacencies: OSPF establishes adjacencies with direct neighbors using the *Hello* protocol. Once *hello* packets are exchanged, each router recognizes the adjacent routers and elect the DR and BDR.

b) Maintaining Link State Databases: On link state establishment as nodes come up, distribution of adjacency information to all routers is initiated by flooding Link State Advertisements (LSA). Each router maintains the flooded link state information in LSDBs.

c) Populating Routing Tables: Using the topology information in the LSDB, each router computes shortest paths from itself to all other routers in the network (area), using the Shortest Path First (SPF) algorithm to populate the routing tables or Forwarding Information Bases (FIB).

2) Convergence After Link / Node Failures

a) Failure Detection: Missing 4 *hello* consecutive packets from a neighbor indicates link or router failure on that link and hence is one mechanism for failure detection.

b) LSA Propagation: After failure detection, a router generates new LSAs to be propagated to all routers in the network (area). The time for generating new LSAs for a single failure is between 4ms and 12ms [8] and OSPF specifies that LSAs cannot be created within 5 seconds from the last LSA generation time to provide sufficient time to update the LSDB from the last event. LSA propagation time also depends on the number of hops between the routers in the network and the processing delay at each router/hop.

c) SPF Recalculation Time: New LSAs update the LSDB and trigger new SPF calculations to update the FIB. Two parameters delay SPF calculations; a *delay timer*, which is 5 seconds and a *hold timer*, which is 10 seconds by default. *Delay timer* is the time between the *new LSA arrival time* and *start of SPF calculation time. Hold timer* limits the interval between two SPF calculations.

B. Tiered Routing Protocol (TRP)

Identifying the tiered structure is described first. In large ISP and AS networks, backbone routers provide connectivity between distribution routers, which, in turn, connect to access routers or sub-networks. In the proposed tiered architecture, the set of backbone routers are designated as tier 1 routers, distribution routers as tier 2 routers and the access routers and sub-networks that they connect as tier 3. This is adopted in the presented studies. A Tiered Routing Addresses (TRA) is required [9] for the purpose. Some features of TRA and resulting impacts on TRP are described below.

1) TRA Allocation: TRA depends on the tier level in a network and carries the tier value as the first field. The tier levels were assigned as stated above. Basically, nodes near backbone or default gateway have lower tier value and nodes near network edge have higher tier value. TRA can be allocated to a *network cloud* (that comprises of a set of routers used for a specific purpose, such as backbone, distributions and so on) or a node. It is not allocated to network interface, which will be identified by port numbers. TRA assignment is made to the node. However, a node can have multiple TRAs based on its connection to the upper tier nodes or networks to support multi homing.

2) TRA Guarantees Loop-Free Routing: TRA allocation starts from a lower value tier to higher value tiers. The parent's address (without the tier value) precedes a child's address. As TRAs determine the packet forwarding paths, this attribute avoids packet looping. However, the dependency can be decoupled at any level through *nesting*.

3) Nested TRA: TRAs can be assigned to network cloud. A new TRA can be started for entities within a network cloud, allowing nesting of TRAs. If a network administrator wishes to incorporate clouds in a cloud, nested TRAs can be used where TRA of an inner cloud does not depend on the TRA of the outer cloud. This decoupling provides a high level of scalability and flexibility in the internetworking.

4) Inherent Routing Information: TRA carries the path information between a lower tier entity and an upper tier entity due to the inheriting the parent's TRA in the child TRA (without tier values). Thus, a route between two communicating nodes can be identified by comparing the nodes' TRAs. If a node has multiple TRAs, a sender node can select a communication path based on criteria such as a shorter path or better resources.

5) TRP Convergence Time: TRP does not require distribution of routing information due to the inherent route information carried by the TRA. Network convergence in TRP is the time required for direct neighbors to recognize the topology change in the neighborhood. This will be several magnitudes less than that required by current routing protocols. The extent of information dissemination can be controlled for optimization.

6) TRP Routing Table Size: The packet forwarding decision in TRP is based on next-hop tier level in the direction of packet forwarding, and has only three choices: same tier level, upper tier level, and lower tier level. Thus, the routing table has to be minimally populated with the

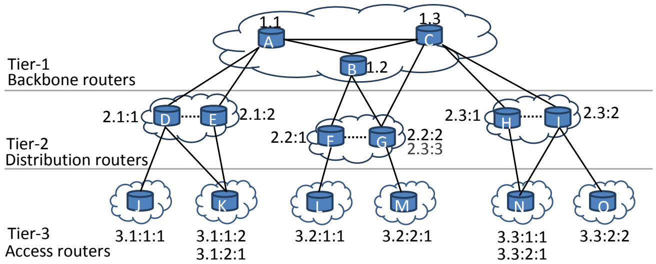


Figure 1. Example Tiered Topology and TRA

directly connected neighbor networks /routers. Optimization is possible by including the two-hop or three-hop neighbors.

C. TRP Operation

TRP address allocation, packet forwarding, link/node failure detection/recovery, address re-assignment, and addition/deletion of nodes are explained in this section.

1) Address Allocation Process: TRA allows automatic address allocation by a direct upper tier cloud / node. Once tier 1 nodes acquire their TRAs, tier 2 nodes will get their TRA from the serving tier 1 node.

a) TRA Allocation: The process starts from the top tier (tier 1). A tier 1 node advertises its TRA to all direct neighbors. A node, which receives an advertisement, sends an address request and is allocated an address. For example in Fig 1, Router A with TRA 1.1 sends Advertisement (AD) packets to Routers B, C, D, and E. Routers D and E send Join Request (JR) to Router A because they do not have TRA yet. Router B and C do not request address to Router A because they are at the same tier level. In Fig. 2, Router A allocates new address (2.1:1) to Router D using a Join Acceptance (JA) packet. Another new address (2.1:2) is allocated to Router E in Fig. 1. The last digit of the new address is maintained by the parent router - Router A. Once Router D registers its TRA, it starts sending AD packets to all its direct neighbors and address assignment continues to the edge routers.

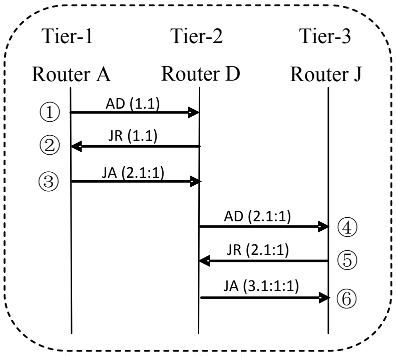


Figure 2. TRA allocation process

TABLE I. ROUTING TABLES OF ROUTER F AND G FROM FIGURE 1

Router F {2.2:1}						**Router G {2.2:2, 2.3:3}**					
Uplink		Down		Trunk		Uplink		Down		Trunk	
Port	Dest	Port	Dest	Port	Dest	Port	Dest	Port	Dest	Port	Dest
1	1.2	3	3.2:1:1	2	2.2:2, 2.3:3	1	1.2	3	3.2:2:1	4	2.2:1
Dest – directly connected neighbor						2	1.3				

b) Mutli-Addressing: If a router has multiple parents, like Router G in Fig.1, it can get multiple addresses. A router with multiple addresses may decide to use one address as its primary address to allocate addresses to its children routers. This implementation was adopted in this work.

2) Routing Tables: TRP maintains three routing tables based on the type of link it shares with its neighbors. In a tiered structure, links between routers are categorized into three different types: up-link which connects to an upper tier router; down-link which connects to a lower tier router; and trunk-link which connects to routers in the same tier level.

A router can identify the type of link from which the AD packet arrives by comparing its tier value with the tier value in the received packet.

Router F has three different types of links to Routers B, G, and L on port numbers 1, 2, and 3 respectively. Advertisement from Router B is received at port 1 and compared with the tier level of Router B (which is 1) and its own tier level (which is 2). Since tier level of Router B is less than tier level of Router F, the link connected on port number 1 is recognized as up-link and the information is stored in the up-link table. Likewise information about Router G is stored in the trunk-link table, and information about Router L is stored in the down-link table.

In Table 1, the 'port' column shows port number of router and 'dest' column shows TRA of direct neighbor obtained from the advertisements. There are multiple entries against a single port in trunk-link table of Router F because Router G has two TRAs. The routing table for Router G is also provided.

```
if( R.TV == P.TV ) then
   if( R.TA.last_digit == P.TA.1st_digit ) then
      if( port_num = find ( P.TA.2nd_digit, down-link table ) ) then
         remove( P.TA.1st_digit );
         P.TV++;
         forward( P, port_num );
         return();
      end if
   else if( R.TV == 1 ) then  //at Tier-1
      if( port_num = find ( P.TA.1st_digit, up-link table ) ) then
         forward( P, port_num ),
         return();
     end if
else if( R.TV – P.TV == 1 && R.TA.parent_digit == P.TA.1st_digit ) then
   if( port_num = find ( P.TA.2nd_digit, trunk-link table ) ) then
      remove( P.TA.1st_digit );
      P.TV++;
      forward( P, port_num );
      return();
   end if
else if( R.TV < P.TV ) then
   discard( P ); //wrong packet
   return();
end if
if( port_num = find (up-link table ) ) then
   forward( P, port_num );
   return();
end if
discard( P );  //no entry in routing tables
return();
```

Algorithm 1. Packet forwarding at router *R* and incoming packet *P*.

The TRA carries the shortest path information inherently. Hence, initial convergence time in TRP is significantly lower than OSPF because, with one advertisement packet from each direct neighbor, the routing tables converge. This also results in less number of control packets and traffic.

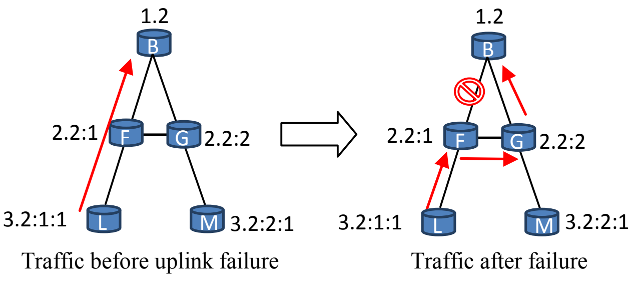
In the network in Fig. 1, three tier levels have been identified, and the TRA for the routers in this network are noted beside them. The TRA is made up of *TV. TA*, where *TV* is the tier value to identify the tier level and *TA* is the address of the router. A '.' notation in the tiered address separates a TV and the Tree Addresses (TA). Thus, the TRA starts with a TV followed by ':' separated addresses which are the TA's. Thus, TRA 3.1:1:1 has TV=3 and TA= 1:1:1.

3) Packet Forwarding in TRP: Packet forwarding in routers running TRP is done as follows. The source router compares the source and destination TRAs to determine TV of a common parent (grandparent) router between them. Assume source is Router L and destination is Router M in Fig. 1. Source Router L compares TA in its TRA namely ***2***:1:1 with the TA of the destination router's TRA namely ***2***:2:1 from left to right to find the common digit in these addresses. In this case, it happens to be the **1**st digit 2 (shown bold italic character) in the *first place*. This provides the information that a common parent (grandparent) between the two routers resides at *tier 1*. The TV in the forwarding address is thus set to 1. To this TV is then appended the TA of the destination router to provide the forwarding address 1.2:2:1. Another example, for a forwarding address between source Router J ***1:1***:1 and the destination Router K ***1:1***:2 will be ***2***.1:2 because a common parent is identified at *tier 2*. The pseudo code for the forwarding decisions at a TRP router is provided in Algorithm. 1 and it is self-explanatory.

D. Failure Detection and Handling

Failure detection in TRP is *hello* packet based, i.e. typical of layer 3 notification proposed for use with current routing protocols. In TRP, 4 missing AD packets is recognized as link/node failure. A TRP router tracks all neighbors AD packets interval and if ADs from a neighbor is missing 4 consecutive times, the TRP router updates its routing table accordingly.

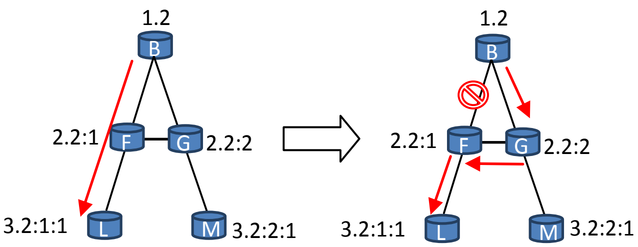
However, in TRP packet forwarding on link/node failure does not have to wait for the 4 missing AD packets. An alternate path, if it exists, can be used on detecting a single missed AD packet irrespective of the routing table update. With the current high speed and reliable technologies, it is highly improbable to miss AD packets and redirecting packets on missing one AD packet is justified.



Traffic before uplink failure

Traffic after failure

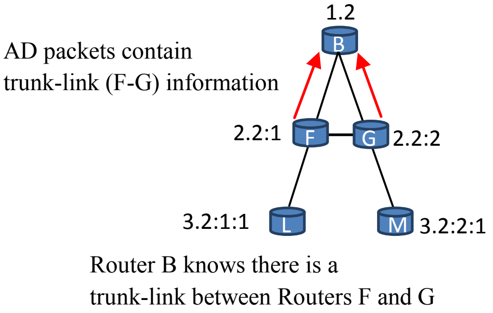
Figure 3. Failure handling with uplink



Traffic before downlink failure

Traffic after failure

Figure 4. Failure handling with downlink



AD packets contain trunk-link (F-G) information

Router B knows there is a trunk-link between Routers F and G

Figure 5. Trunk-link information sharing by the parent router

1) Uplink failure: If a node detects an uplink failure and has a trunk link, it can use the trunk link, because trunk link exists between routers that have the same parent router, or if a router has another uplink, it can use it. In Fig. 3, sibling router connected to Router F derives its address from the same parent. So, Router F knows that the uplink router on Router G will be its parent Router B.

2) Down link failure: Let link failure occur between Routers B and F in Fig. 4. To detour around the link failure, down link traffic between Router B and F needs to take a path Router B-G-F. To achieve this, Router B needs to know if there exists a trunk link between Router F and G. A parent router must know all trunk links between its children routers. The trunk link information can be set in AD packets to help a parent router maintain all trunk link information as described in Fig. 5. Due to the inheritances, routers can assume responsibilities to forward for their directly connected neighbors as the TRAs carry relationship information.

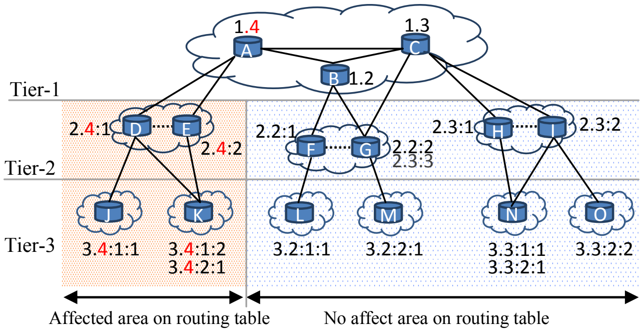
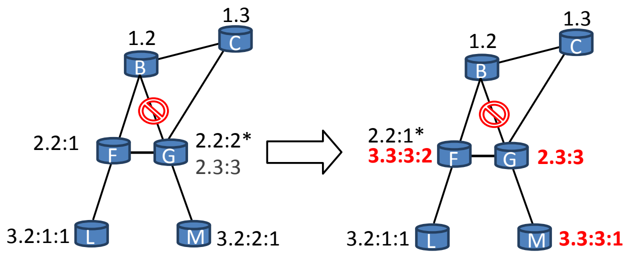


Figure 6. Address changes in TRP



Addresses before failure Addresses changed after failure

Figure 7. Primary address change

3) Address Changes: Address changes can happen because of node failure, topology change, or administrative decisions. In TRP, address changes affect limited area and incur very low latency as no updates have to be propagated.

For example, if Router A changed its TRA from 1.1 to 1.4 in Fig. 6, all neighbor Routers B, C, D, and E notice the change from the AD packet sent by Router A. Router D and E will change their TRAs without notifying Router A. Therefore, children of Router A can change their addresses rapidly. The same procedure continues to Routers J and K by the next AD packet from Routers D and E. The pruning operation is triggered on change detection.

4) Primary Address Change: If a node has multiple addresses and a link to a primary address failed, the node changes one of its secondary address to primary address and advertises the same. The child of the node also changes its address in the same manner as described in the case above and keeps the last digit. For example, Router G has two addresses and let 2.2:2 be the primary address in Fig. 7. When a failure occurs between Routers B and G, Router G changes its primary address to 2.3:3 and then advertises it. As the result, Router M changes its address to 3.3:3:1.

IV. EMULATIONS

A. Emulab Test Setup

A TRP router was implemented on Linux machines in Emulab. Emulab is an experimentation facility which allows creation of networks with different topologies to provide a fully controllable and repeatable experimental environment. Emulab uses different types of equipment for this purpose. Two different types of machines were used during the course of this experiment, as allocated by the Emulab team.

Quagga 0.99.17 [11], a software routing suite for configuring OSPF was used for the comparison studies. Iperf [10] was used to generate traffic.

A 21 node topology is shown in Fig. 8. The configuration details are provided in Table II. In the 45 nodes topology, the additional 24 nodes were added to the outer circle of routers utilizing a topological connection similar to that of the outer routers in Fig. 8. The IP addresses were allocated from address space 10.1.x.x/24 to the segments as shown for OSPF. The TRAs for TRP were allocated using the scheme described in section III-B.

B. Assumptions

1) More complex or meshed topologies could not be created due to the limitations on the number of interfaces on the Emulab machines.

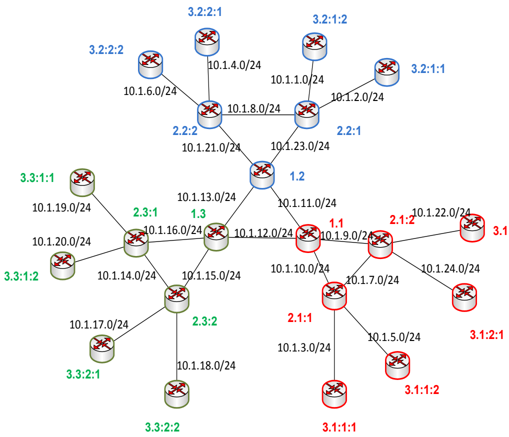


Figure 8. Testbed Topology with IP and Tiered Addresses

TABLE II. EMULAB TESTBED CONFIGURATIONS

Topology	**21 Nodes**	**45 Nodes**
Type of processor	Pentium III	Quad Core Xeon Processor
Number of links	24	54
Link shaping nodes	12	20
Connection speed	100 Mbps	100 Mbps

2) TRP code operates on Linux user space and hence timings and dependent variables such as packet loss during convergence project a higher value than if the code were run in kernel space. Quagga OSPF code runs in kernel space.

3) To provide a random environment for the tests, they were conducted in two different sets of networks and the experiments repeated five times in each case.

4) To emulate link failures, Emulab uses link shaping nodes that can be placed on the segments.

5) For OSPF evaluations, only one area was defined, as the intention is to demonstrate the performance impacts to increase in the number of routers in a networks or an area.

B. Tiered Routing Protocol Code

TRP runs above layer 2, *bypassing all layers* between layer 2 and the application layer. It replaces both IP and its routing protocols. To run applications on TRP, SIPerf, a modified clone of Iperf which allows bandwidth and link quality measurement in terms of packet loss, was used.

C. Performance Statistics on Initial Convergence

1) Convergence Times: In OSPF, initial convergence takes place after the FIB update is run on all routers. To improve the veracity of collected data, the timestamps when SPF was run as well as the time when the routing table was updated was logged. For TRP, the timestamp for a new entry in the routing tables is logged and if the routing table at the routers remains unchanged for the next three *hello* intervals then the network was deemed to have converged.

2) Routing Table Size: In OSPF, this value was logged using the built-in commands provided by Quagga. In TRP, this information was logged in a file and sent to the server.

3) Control Overhead: Tshark [12] which is a command line tool similar to Wireshark [12] was utilized for the purpose. Bytes in the packets exchanged during convergence were summed to determine the control overhead at each node and then sent to the server. In TRP, a utility to record the number of control packets exchanged during initial convergence time was built in.

D. Performance Statistics on Link Failures

Convergence time after link failure has two components.

1) Link failure detection time: This is the same for OSPF and TRP as they detect a link failure on missing 4 *hello* messages. With a *hello* interval of 10 seconds, this was recorded to be 30 seconds with an additive time - time between the first missing packet and the time when the link was actually brought down.

2) Time to update routing tables: This time is different for OSPF and TRP and are explained using Figs 9 and 10.

3) TRP Response to Link Failures: In Fig. 9, the time t_1 when the link failed is noted along with time t_3 it took to remove the link from the routing table. Total time for convergence T_c is then given by

$$T_c = T_{ru} - T_{fd} \quad (1)$$

where T_{fd} is the failure detection time given by

$$T_{fd} = t_2 - t_1 \quad (2)$$

and T_{ru} is the routing table update time given by

$$T_{ru} = t_3 - t_2 \quad (3)$$

Thus,

$$T_c = t_3 - t_1 \quad (4)$$

T_{fd} will be the same for OSPF, but T_{ru} is negligible in the case of TRP as this is the time to for the TRP code to access the routing tables and update its contents. In Figs 9 and 10, these times are identified based on the operations of TRP and OSPF respectively.

4) OSPF Response to Link Failure: OSPF uses several timers on link failures, to rerun SPF algorithm and a few other hold times to avoid toggling. They are *Hold_Time*, which is the seperation time in ms between consecutive SPF calculations. An *Initial_hold_time* and *Max_hold_time* is also specified. SPF starts with the *Initial_hold_time*. If a new event occurs within the *hold_time* of any previous SPF calculation then the new SPF calculation is increased by *initial_hold_time* up to a maximum of *max_hold_time*.

Let T_{LSA} be the LSA propagation delay, T_{SPF} be the time to run SPF on subsequent LSA messages and T_{TU} be the table update delay, then Tru of OSPF is given by

$$T_{ru} = T_{LSA} + T_{SPF} + T_{TU} \quad (5)$$

T_{SPF}, *initial_hold_time* and *max_hold_time* were set to 200ms, 400ms, and 5000ms respectively for the test. Fig. 10 captures the relationship between the delays for OSPF.

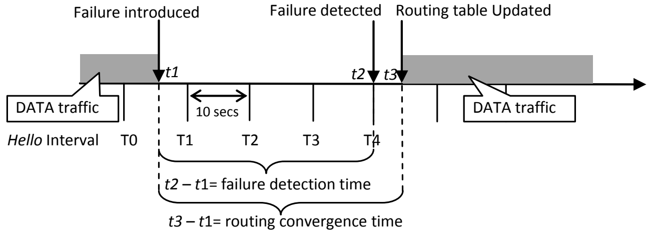


Figure 9. TRP Routing Convergence Time

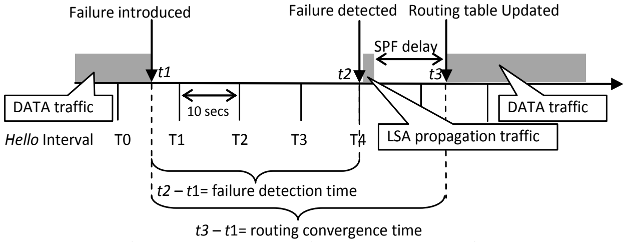


Figure 10. OSPF Routing Convergence Time

V. Performance Analysis

The performance of OSPF and TRP, during the initial convergence phase and their response to subsequent link failures are presented in this section. In the histograms, data collected for the two test sites are provided separately, to show the closeness of the two data sets under different environments to reflect the reliability of the experiments.

1) Initial Convergence Times

Fig. 11 records the average initial convergence times in seconds collected from the two test sites and for the two different topologies, one with 45 routers and the other with 21 routers. While the convergence times recorded for OSPF range from 55 secs in the case of the 21 router network to over 60 secs in the case of the 45 router network, the convergence times for the network running TRP was around 1 sec. While convergence times are stable irrespective of the number of routers running TRP, in the case of OSPF, the convergence times showed an increase by 5 to 6 secs, indicating dependency of convergence times to the network size. TRP has 50-60 times improvement compared to OSPF.

2) Control Overhead During Initial Convergence

Fig. 12 shows the plot of control overhead in Kbytes for OSPF and TRP. Control overhead in the case of OSPF varies from 250 Kbytes for the 21 router network to around 750 to 800 Kbytes for the 45 router network. Increase in overhead almost triples as network size doubles. Control overhead for TRP was 2.6 Kbytes for 21-router network and around 6 Kbytes for 45-router network. The improvement achieved with TRP 100 times in the case of the 21-router network and 130 times in the case of the 45-router network.

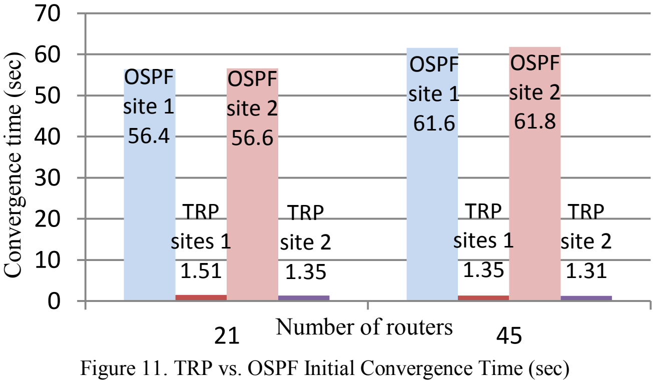


Figure 11. TRP vs. OSPF Initial Convergence Time (sec)

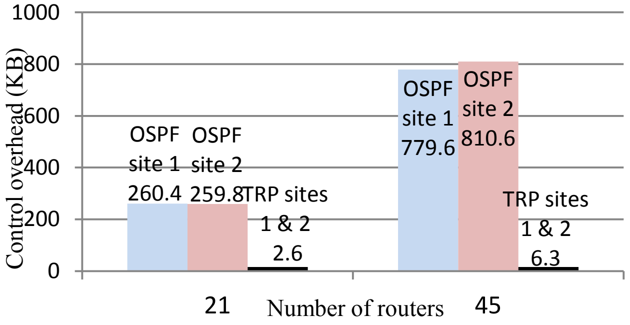


Figure 12. TRP vs. OSPF Routing Control Overhead Size (KB)

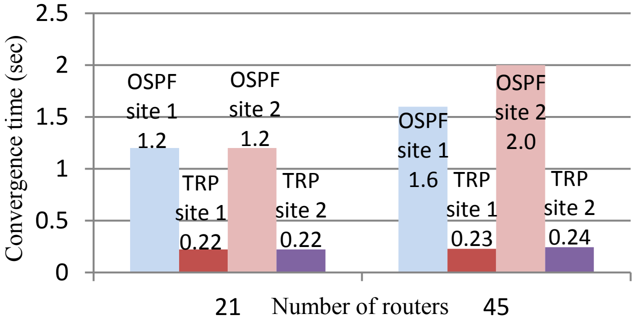


Figure 14. TRP vs. OSPF Convergence Time after Failure (sec)

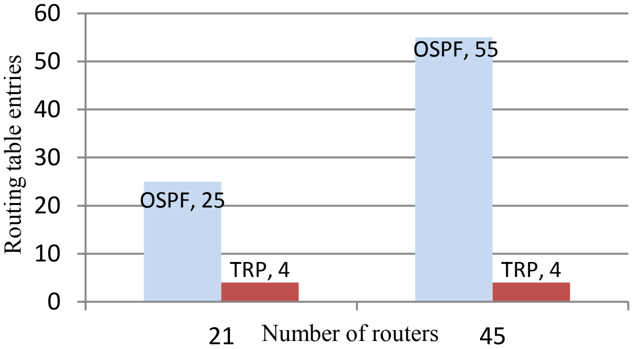


Figure 13. TRP vs. OSPF Routing Table Entry Size

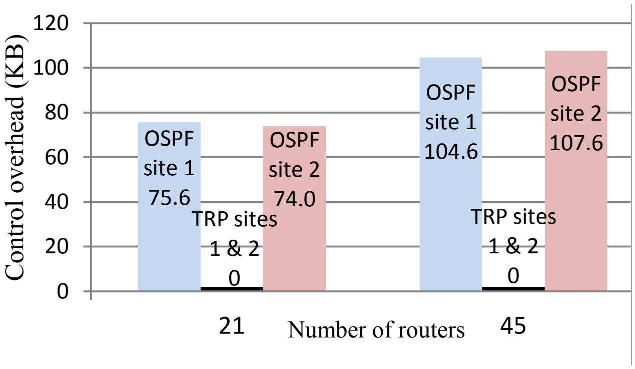


Figure 15. TRP vs. OSPF Control Packet Size after Failure (KB)

3) *Routing Table Sizes*

In Fig. 13, the routing table sizes collected were the same in the case of OSPF and TRP for the two test sites and hence one graph with maximum routing table entries is provided. In the case of OSPF, this value is 25 for the 21-router network (as there are 25 segments) and in the case of 45-router network this value was 55. In the case of TRP, the routing table entries reflects number of directly connected neighbors, so in both cases, the maximum routing table entry was 4 – there is no dependency on the network size.

4) *Convergence Time After Link Failure*

Fig. 14 has the routing table update time in seconds subsequent to link failure detection. While OSPF shows an update time of 1.5 to 2 secs for the 45-router network and around a second for the 21-router network, TRP update times were 200 to 240 milliseconds; a magnitude of 6 improvement for the smaller network and a magnitude of 8 improvement for the larger network. Routing table update time is invariant to the network size in the case of TRP.

5) *Control Overhead After Link Failure*

Control overhead for TRP and OSPF collected during the convergence times, includes the time to detect a failure and also time to update routing tables. For the given topologies no control overhead was incurred with TRP. In Fig. 15, OSPF required around 100 Kbytes and 70 Kbytes of control packets for the 45-router and 21-router networks respectively. For complex topologies, in TRP change information may have to be propagated to downstream networks. Similarly, upstream router may also have to be informed when a downstream link fails. These features were not tested.

6) *Data Packets lost*

The packets lost during failure detection will be the same for both protocols as the failure detection time is 4 missing *hello* packets. The time to update routing tables was recorded to be around 0.2 sec for TRP and 1.2 to 2.0 sec for OSPF. Thus the packets lost during routing table update time was a maximum of 1 packet for TRP and a maximum of 10 packets with OSPF at a data rate of 5 packets per second.

VI. Conclusions and Future Work

A Tiered Routing protocol was developed under a new tiered Internet architecture. The tiered addresses in this architecture are used by TRP for packet forwarding. In this article, TRP is evaluated as an IGP using Emulab test facility. Initial convergence time and control overhead with networks running TRP is very low as the protocol does not require message flooding or any calculations subsequent to a link status change. Due to the inherent routing information in the tiered addresses, the routing table sizes in TRP are significantly low. Stability in the routing entries and their invariance to network size also indicates the strengths of such new approaches. Comparison with OSPF validates this.

There are several possible directions for future work. OSPF supports area concept for large network, so apply the area concept for larger network to compare with TRP. Validating TRP for inter-domain routing is another direction. Since tier levels in Autonomous System (AS) level topology can also be identified, based on their business relationships such as provider-customer and peer-peer relationship, TRP can be applied for inter-domain routing. Thus, Border

Gateway Protocol (BGP) and TRP are compared to validate TRP as inter-domain routing protocol.

ACKNOWLEDGMENT

This work was sponsored by NSF under grant number 0832008.

REFERENCES

[1] J. Moy, “RFC 1245 - OSPF Protocol Analysis,” RFC Editor, 1991.

[2] M. Yannuzzi, X. Masip-Bruin, and O. Bonaventure, “Open issues in interdomain routing: a survey,” *Network, IEEE* , vol.19, no.6, pp. 49- 56, Nov.-Dec. 2005

[3] Y. Nozaki, H. Tuncer, and N. Shenoy, “A Tiered Addressing Scheme Based on Floating Cloud Internetworking Model,” Distributed Computing and Networking, Lecture Notes in Computer Science, Vol 6522/2011, pp. 382-393. 2011.

[4] “Emulab: Network Emulation Testbed,” http://www.emulab.net. (accessed March 2013)

[5] C. Alaettinoglu, V. Jacobson, and H. Yu, “Towards Milli-Second IGP Convergence,” Internet Draft draft-alaettinoglu-isisconvergence-00.txt, IETF, November 2000.

[6] P. Pan, G. Swallow, and A.Atlas, “RFC-4090, Fast Reroute Extensions to RSVP-TE for LSP Tunnels.” May 2005.

[7] A. Kvalbein, A.F. Hansen, T. Ciˇci´c, S. Gjessing, and O. Lysne, “Multiple routing configurations for fast IP network recovery,” IEEE/ACM Trans. Netw. 17, 2, pp. 473-486, 2009

[8] Y. Liu, and A.L.N. Reddy, “A fast rerouting scheme for OSPF/IS-IS networks,” In Proceedings of ICCCN 2004, pp. 47- 52, 11-13 Oct. 2004.

[9] N. Shenoy, M. Yuksel, A. Gupta, K. Kar, V. Perotti, and M. Karir, “RAIDER: Responsive Architecture for Inter-Domain Economics and Routing," GLOBECOM Workshops (GC Wkshps), 2010 IEEE , pages 321-326, 6-10 Dec. 2010.

[10] “Iperf: The TCP/UDP Bandwidth Measurment Tool,” http://www.iperf.sourceforge.net. (accessed March 2013)

[11] “Quagga Software Routing Suit,” http://www.quagga.net. (accessed March 2013)

[12] “Tshark and Wireshark,” http://www.wireshark.org. (accessed March 2013)

A Reconfiguration Trial on the Platform of Allied Information for Wireless Converged Networks

Jie Zeng, Xin Su, Yuan You, Lili Liu, and Xibin Xu
Tsinghua National Laboratory for Information Science and Technology
Research Institute of Information Technology, Tsinghua University
Beijing, China
e-mail: zengjie@tsinghua.edu.cn

Abstract—**Platform of Allied Information (PAI) promotes persistent innovation and application of the wireless technologies among different institutes, based on an open, safe and controllable network architecture. The platform architecture, unified interface, security mechanism, and reconfiguration are designed to achieve the convergence of various wireless experimental resources. In this paper, the reconfiguration mechanism is designed and a reconfiguration trial is implemented based on the Platform of Allied Information, to verify the ability of integrating experimental resources in the related units.**

Keywords-wireless communication; reconfiguration; convergence; PAI.

I. Introduction

In recent years, wireless communications have had a rapid development with the emerging of new wireless technologies, including cognitive radio [1], network coding [2], cooperation and coordinated transmission [3][4], self organizing network [5], and so on. But, most of them are still in the stage of theoretical research or small-scale validation without large-scale trials verified which restricts their further development. At the same time, wireless experimental resources in different forms, such as the wireless devices, simulation environment, experimental instruments, software modules, scatter in different places with the lack of unified usage. In order to meet the needs of resource sharing, these different resources should be integrated to work for the innovation and application research of wireless technology. Thus, with the support of national Science and Technology major project, Tsinghua University initiates and constructs an environment, named Platform of Allied Information, as shown in Fig. 1, for development and verification of various new wireless technology and new services with related universities and companies.

One of the basic tasks of PAI is to build an open, safe and controllable transmission platform. With PAI, the cross-region and inter-unit remote collaborative researches and developments of wireless technology can be achieved, and the distributed experimental resources can be fully integrated and utilized for experimental verifications researches of new wireless technology and new services by more than one allied unit.

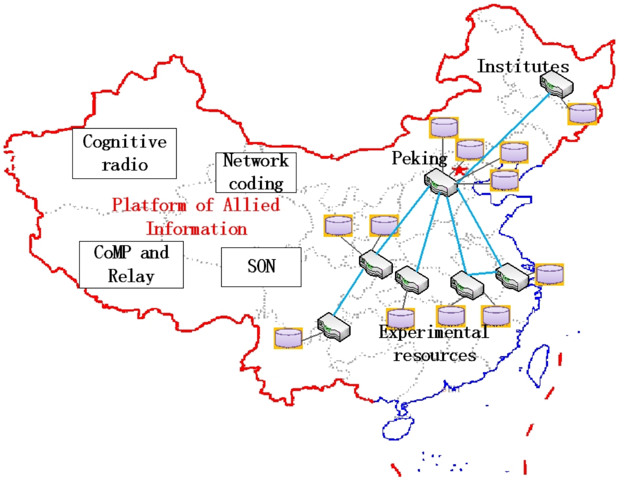


Figure 1. Resources sharing and collaborative development of PAI.

In this paper, the reconfiguration mechanism is designed and a reconfiguration trial is implemented based on PAI. Through appropriate resource configuration and experiment configuration, wireless innovative experiment could be tested to prove the feasibility of the proposed reconfiguration mechanism.

The rest of this paper is organized as follows. We will introduce the major problems solved by PAI in Section II, and then describe the design and implementation of PAI in Section III. In Section IV, the reconfiguration mechanism is described. In Section V, we show the running of real-time reconfiguration, then give a conclusion in the last section.

II. Major Problems Solved by PAI

Considering the extensive flexibility, the open platform architecture is adopted. We use the China Education and Research NETwork (CERNET) [6] as the wired transmission backbone as well as dedicated gateway devices and unified local interface as the access of the various experimental resources. To ensure real-time control with high-speed data transmission, there will be separate paths for data and control signal. Resource management, user management, experiment operation and other aspects are in-depth considered and optimally designed.

A. *Unifying Specifications of the Interface*

The specifications of the interface between the platform and experimental resources are clearly defined. A scalable universal interface is designed to support the access of various wireless devices, simulation environments, experimental instruments, and software modules. Some typical experimental resources are given as follows.

- Wireless devices, such as Universal Software Radio Peripheral (USRP) [7], WARP [8].
- Simulation environments, such as simulation platforms based on MATLAB, OPNET and NS2.
- Experimental instruments, such as signal generators, spectrum analyzers, and channel simulators that are available to Internet.
- Software modules, such as software developed by OSSIE [9] and GNU Radio [10].

B. *Centralizing the Scheduling and Management*

Since the experimental resources are highly distributed, the unified management mechanism should be studied and developed to support different experimental resources, end-users and experimental procedures. A centralized management system is established to achieve the following functions.

- Convert the end-users' instructions into local or remote operations.
- Manage the data exchange between different experimental resources.
- Submit results to end-users.

C. *Improving the Security Mechanisms*

The security mechanism is designed to ensure safety during the access of hardware and software and implementation of experimental verification. The communications procedures are completed by a collaboration of the hardware and software modules, the security of the verification platform aim at software security and hardware security.

D. *Keeping the Compatibility*

The new wireless technologies are emerging continuously. The high requirements for the verification platform should be satisfied, such as rate matching, adaptive access and experimental approach selection. The core factors, parameters and experimental needs of the mainstream wireless technologies at present are analyzed and summarized. The basic interface specifications, operating procedures guidelines and feasibility assessment methods are established to ensure regular development of the experiments.

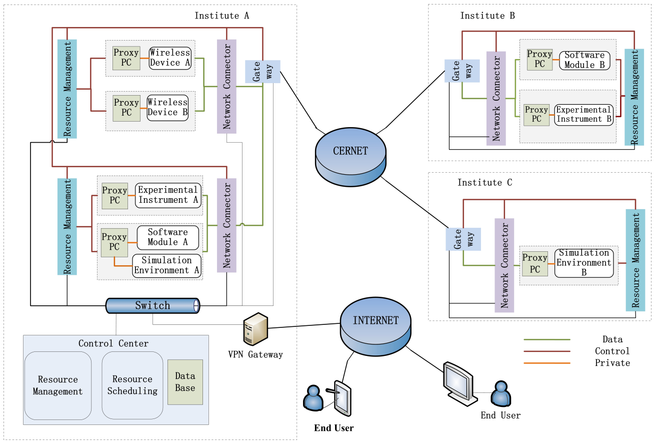


Figure 2. Basic architecture of PAI.

E. *Building the Experimental Resources Library*

The experimental resources library is built with a variety of commonly used hardware devices, unit components, software modules, simulation functions, and system components combined with the hardware and software. New components developed by different users must meet certain specifications before adding to the library. They must go through self testing, system testing and multi-parameter testing, to ensure proper functioning.

F. *Brief Summary*

In order to construct the platform, the major problems mentioned above should be addressed based on the technology accumulation of the cooperative institutes. Because this paper focuses on the scheduling and reconfiguration of wireless communication system, the interface, security and library technologies would not be described in detail.

III. Design and Implementation of PAI

A. *Basic Architecture*

PAI consists of four major parts: Experimental Resources, End-Users, Control Center, and Backbone Networks, as shown in Fig. 2.

All accessed Experimental Resources need to be reprogrammed in accordance with the unified interface specifications. End-Users are the operators of the experiments. The Control Center works with reconfiguration mechanism to configure and manage all Experimental Resources; it provides further management functions such as authentication and security management. Backbone networks is the basis of the entire platform, using special network equipments to provide a high-speed, stable and secure network access.

B. *Interface Specifications*

Interface specifications guide the Experimental Resources transforming the private Application Program Interface (API) into the public interface with the help of Web services.

Special access mechanism such as manually upload data by End-User would be provided for the Experimental Resources without suitable interface adaptation. Some typical Experimental Resources are adapted as follows.

- USRP is a widely used software defined radio test bed, which provides RF, ADC/DAC and IF and customizes the baseband processes. USRP is connected to the proxy PC via the USB port; GNU Radio provides a major modules library to operate USRP peripheral devices. The common features are not difficult to show through the Web service like the basic file send / receive applications.
- MATLAB is the most popular simulation tool; many institutes have accumulated lots of MATLAB wireless communications simulation environments for many years, which could be released to authorized end-users through Web services. MATLAB demonstrations could be adapted by the MATLAB engine. MATLAB engine is a stand-alone C / C++ program, which can call the MATLAB functions through COM objects, send commands to the MATLAB process, transmit parameters and receive results.
- The interfaces of Agilent instruments are GPIO, USB, LAN, etc. A proxy PC is established to operate the instrument through an adaptive process that is developed based on the Agilent VISA and SICL libraries. The proxy PC needs to send the parameters assigned to the reserved instrument, and capture the measurement report.

C. *Configuration and Management*

The Control Center showed in Fig. 3 allows End-Users to reserve the Experimental Resources with permission, after that various services provided by the Experimental Resources could be called. Some experiments are carried out with the interconnection of the data from different Experimental Resources; others are carried out with the cooperation of several End-Users. The experiments could be operated not only by remote end-user but also at the appropriate given time.

IV. Reconfiguration Mechanism

Reconfiguration can be realized by software reconfiguration, research of software reconfiguration is embodied in software architecture description language and its support system [11]. Related methods are software reconfiguration based on mapping rules, software reconfiguration based on hierarchical message bus architecture [12], software reconfiguration based on the C2 style software architecture [13], etc. These software reconfiguration mechanisms separate the whole software system into modules, and architecture description language is used to describe the connection between modules and connectors. The modeled software system increases system flexibility, so the system became enabled to refresh in order to meet new environments and demands [14].

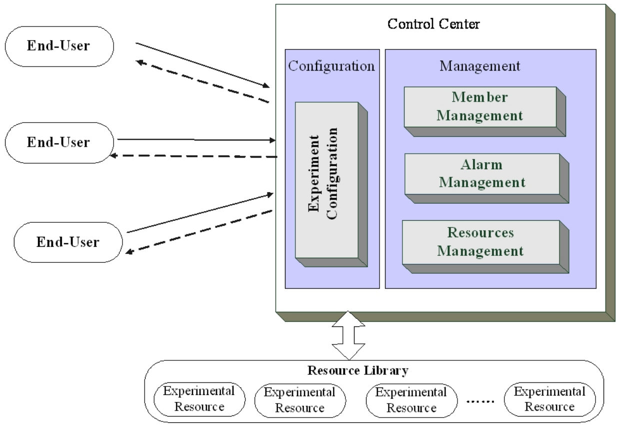


Figure 3. Centralized configuration and management of the Control Center.

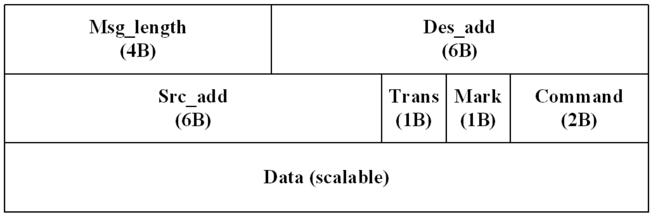
With the development of network, the reconfiguration mechanisms discussed above are not suited to the complex distribution environment and flexible application model. This paper extends the application of reconfiguration mechanism into the real converged network environment, giving a method of reconfiguration on PAI. With the proposed reconfiguration mechanism, PAI can be configured into appropriate applications according to different experimental needs.

The reconfiguration mechanism is realized by Resource Management and Experiment Configuration module in Control Center. Resource Management mainly solve the issue of ”resources’accessibility”, while the Experiment Configuration is mainly responsible for the scheduling of resources and interaction between resources. Through appropriate Resource Management and Experiment Configuration multiple wireless innovations have integrated effectively, which promotes the loach of wireless innovative experiment.

A. *Resource Management*

Resource Management starts from resource certification. In PAI, resources with a characteristic of wide distribution are belong to different network nodes and connected through the CERNET. These resources can be specific equipment, software modules, virtual instruments, etc. They are highly heterogeneous; typically do not have a unified interface to the configuration and management module. So the prime problem to be solved is how to promote the sharing of resources, providing a unified standard for resource description.

Web services [15] were defined as the abstract form of accessed resources and unified XML resource description format as a resource certification and identification rule. Through the adaptation, the private interface is released into the public way for the Resources Management module API



<table>
<tr><td colspan="2">Msg_length
(4B)</td><td colspan="4">Des_add
(6B)</td></tr>
<tr><td colspan="2">Src_add
(6B)</td><td>Trans
(1B)</td><td>Mark
(1B)</td><td colspan="2">Command
(2B)</td></tr>
<tr><td colspan="6">Data (scalable)</td></tr>
</table>

Figure 4. Resource commnication packet format.

calls. XML description files include some basic properties and operation commands of resource, such as resource name, functional description, the command name, command description, the command parameters, return values and so on.

In PAI, the main function of Resource Management is carried out by the Resource Management module, in which the XML file was analyzed to get related information, identify resource function, and then provide users with the ability of operating resources.

B. Experiment Configuration

Experiment Configuration schedules experiments and ensures the interaction between resources, which are completed by the cooperation of Experiment Scheduling and Resource Communication sub-module.

Experiment scheduling sub-module is responsible for the initial configuration. When users log in, they can create a project on demand visually, finishing resources selection, topological link, command parameters configuration and other configuration operations. The initial configuration information is programmed by Experiment Scheduling sub-module and saved to the database for rechecking.

Resource Communication sub-module defines the communication protocol between resources. It completes the data interaction between resources and realizes the mapping of resources to the actual function based on the initial configuration information. The protocol format is shown in fig.4. *Msg_length* specifies the length of the entire packet, 4Byte; *Des_add* and *Src_add* show the destination and source address of the packet, each 6Byte; *Trans* identifies the process of communication between resources, starting from 0, 1 to end, 1Byte; *Mark* identifies whether the packet has segments, 0 for no segment, others for the sequence number of the segments, 1Byte; *Command* is used to store a specific order of resources, 2Byte; *Data* holds the data associated command; its length is variable according to the corresponding order. A larger amount of data can be transmitted by serial segments.

V. Running of Real-time Reconfiguration

In this section, a trial is given to test the reconfiguration mechanism. As shown in Fig. 5, this experiment includes two sets of USRP equipments (tagged as USRP1, USRP2), convolution coding module and Viterbi decoding module (deployed in PC1), source coding and decoding modules (deployed in PC2). All resources are deployed in network and connected by special equipments. Through USER PC, users get access to the Control Center (server) to configure the whole experiment.

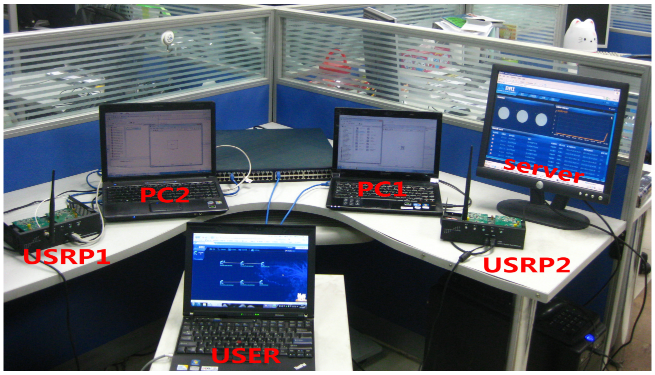
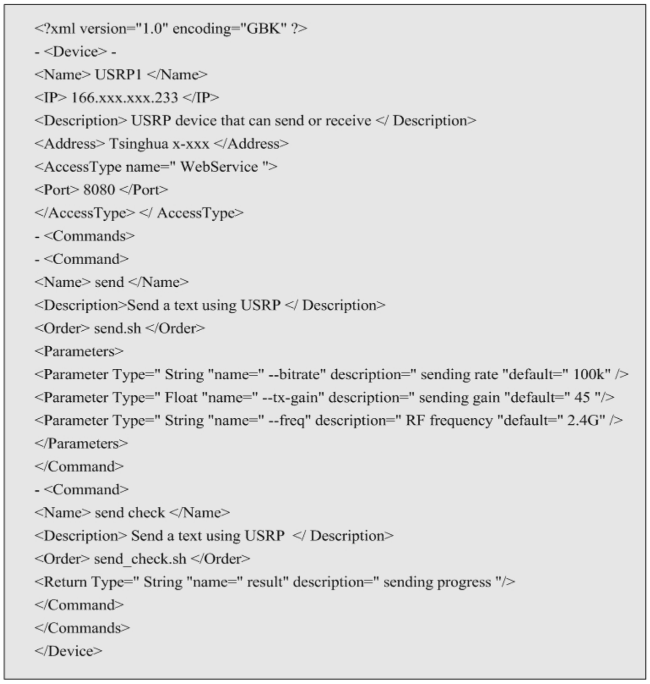


Figure 5. Scene graph of the reconfiguration trial.



```
<?xml version="1.0" encoding="GBK" ?>
- <Device> -
<Name> USRP1 </Name>
<IP> 166.xxx.xxx.233 </IP>
<Description> USRP device that can send or receive </ Description>
<Address> Tsinghua x-xxx </Address>
<AccessType name=" WebService ">
<Port> 8080 </Port>
</AccessType> </ AccessType>
- <Commands>
- <Command>
<Name> send </Name>
<Description>Send a text using USRP </ Description>
<Order> send.sh </Order>
<Parameters>
<Parameter Type=" String "name=" --bitrate" description=" sending rate "default=" 100k" />
<Parameter Type=" Float "name=" --tx-gain" description=" sending gain "default=" 45 "/>
<Parameter Type=" String "name=" --freq" description=" RF frequency "default=" 2.4G" />
</Parameters>
</Command>
- <Command>
<Name> send check </Name>
<Description> Send a text using USRP  </ Description>
<Order> send_check.sh </Order>
<Return Type=" String "name=" result" description=" sending progress "/>
</Command>
</Commands>
</Device>
```

Figure 6. USRP1 XML description file.

All the related resources access to PAI need to be reconfigured as defined in section IV, using Web services as a public API to the Control Center and loading XML files to describe the information of the resources. Take USRP1 as an example, the XML description file is shown in Fig. 6. In this case, the reconfigured parameters includes transmission rate, transmit power gain and RF frequency.

Resource Management module verifies the legitimacy of XML documents, and parses out the relevant information,

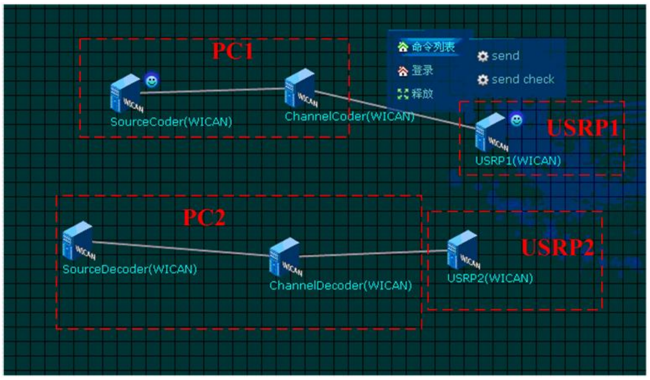


Figure 7. Initial configuration interface.

then feedback a list of commands. When users log in, select the relevant experiment resources, render the link topology, and configure the appropriate experimental parameters, as shown in Fig. 7. We place the SourceCoder and ChannelCoder modules in PC1 and send the coded bitstream to USRP1, convert it to radio frequency and then transmit, through the airinterface, USRP2 receive the radio frequency signal and convert it to baseband signal, the baseband signal is transported to PC2 and be processed with the ChannelDecoder and SourceDecoder modules to get the original bitstream.

After the initial configuration is completed, Experiment Scheduling sub-module generates configuration file according to the operations above and save it to the database. Resource Communication sub-module follows the communication protocol defined in Section IV to package the data. Experimental data transmits via source coding and convolution coding module to USRP1, where GMSK modulation and digital up conversion are completed, and then the experimental data are transmitted. Correspondingly, in the USRP2, data is received and digital down conversion and demodulation will be done, the processed data will be passed on to Viterbi channel decoding module and the source decoding module to reconstruct the source data. The reconstructive data as well as the source data is passed back to the USER. All the input/output data streams mentioned above are retransmitted by the Control Center; thus all the intermediate results can be collected and observed.

VI. Conclusion

We have designed and developed PAI to meet the convergence of various innovative wireless technologies. A novel method of reconfiguration has been proposed and a reconfiguration trial of wireless transmission has been done based on the overall configurability of PAI. The distributed Experimental Resources in related research units can be fully integrated and utilized for experimental verification of new wireless technologies, such as cognitive radio, network coding, cooperation and coordinated transmission, self organizing network and so on. PAI supports the research for future broadband wireless technology effectively. More work about security and user interface of PAI could be in-depth studied to improve the safety and applied range.

VII. Acknowledgment

This work has been supported by National S&T Major Project 2011ZX03003-002 and Beijing Natural Science Foundation 4110001.

References

[1] J. Mitola, “Cognitive radio: making software radios more personal,” Personal Commun., IEEE, vol. 6, no. 4, 1999, pp. 13-18.

[2] S. Li, R. Yeung, and C. Ning, “Linear network coding,” IEEE Trans. Inf. Theory, vol. 49, no. 2, 2003, pp. 371-381.

[3] J. Laneman, D. Tse, and G. Wornell, “Cooperative diversity in wireless networks: efficient protocols and outage behavior,” IEEE Trans. Inf. Theory, vol. 50, no. 12, 2004, pp. 3062-3080.

[4] 3GPP, “Aspects of coordinated multi-point transmission for advanced E-UTRA,” 3GPP TSG RAN WG1 Meeting#54, R1-083530, Texas Instruments, Sept. 2008.

[5] J. L. van den Berg et al., “Self-organization in future mobile communications networks,” ICT-Mobile Summit, Stockholm, Sweden, 2008.

[6] JP. Wu, “Current state and future of China education and research network,” New Technology of Library and Information Service, 1997, pp. 9-11.

[7] Universal software radio peripheral: the foundation for complete software radio systems. [Online]. Available: http://www.ettus.com/downloads/ettus-ds-usrp

[8] K. Amiri, Y. Sun, P. Murphy, C. Hunter, J. R. Cavallaro, and A. Sabharwal, “WARP, a unified wireless network test bed for education and research,” IEEE Int. Conf. Microelectronic Systems Education, San Diego, CA, USA, 2007, pp. 53-54.

[9] C. R. A. Gonzalez et al., “Open-source SCA-based core framework and rapid development tools enable software-defined radio education and research,” IEEE Commun. Mag., vol. 47, no. 10, Oct. 2009, pp. 48-55.

[10] Free Software Foundation, Inc. (2009). GNU Radio - the GNU software radio. [Online]. Available: http://www.gnu.org/software/gnuradio

[11] N. Medvidovic and D. S. Rosenblum, “Domains of concern in software architectures and architecture description languages,” in Proc. of USENIX conf. on DSL, Santa Barbara, CA, USA, oct. 1997, pp. 199-212.

[12] C. Yu and S. Huang, “Real-time software reconfiguration based on software architecture,” Computer Engineering and Applications, vol. 36, no. 3, 2000, pp. 47-54.

[13] S. Huang, Y. Fan, and Y. Zhao, “Research on generic adaptive software architecture style,” Journal of Software, vol. 17, no. 6, 2006, pp. 1338-1348.

[14] Z. Wang and X. Xie, “Software reconfiguration based on component-oriented architecture,” Computer Development, vol. 14, no. 7, 2004, pp. 8-15.

[15] C. Hu, J. Huai, and H. Sun, “Web service-based grid architecture and its supporting environment,” Journal of Software, vol. 15, no. 7, 2004, pp. 1064-1073.

Optimization of Overlay QoS Constrained Routing and Mapping Algorithm for Virtual Content Aware networks

Radu Dinel Miruta, Eugen Borcoci
Telecommunication Dept.
University POLITEHNICA of Bucharest
Bucharest, Romania
radu.miruta@elcom.pub.ro, eugen.borcoci@elcom.pub.ro

Abstract—Multimedia Services including video distribution are increasingly required by the current market and will be also a target of the Future Internet. One method to customize the multi-domain guaranteed transport with several QoS classes of services is to create Virtual Content Aware Networks (VCAN) constructed as overlays over IP networks. The mapping of VCANs onto real multi-domain topologies is needed. This paper develops new optimizations to increase the performances of a previously proposed combined hierarchical multi-domain algorithm performing VCAN mapping with QoS constraints.

Keywords-Virtual Content-Aware Networking; Network Aware Applications; Multi-domain; QoS; Management; Constrained routing; Future Internet.

I. Introduction

The transport of media streams over heterogeneous IP networks is a need of the current and also Future Internet. However, assuring the quality of services (QoS) and other special needs of the high level services, (security, reliability, etc.) are still not answered satisfactorily by the current public networks, except the "wall gardened" networks, fully owned and controlled by operators (e.g., IPTV distribution networks).

One new solution, content-oriented, is to transport media flows over some previously created (on demand) Virtual Content Aware Networks (VCANs). They are usually constructed as overlays on top of IP level [1][4], based on (light) virtualisation techniques. In a multi-domain network and several operators context, the VCANs can be multiple-domain spanning, therefore several Network Providers (e.g., ISP) might cooperate towards this goal. Given that a VCAN is an overlay and the fact that NP/ISPs are independent entities, it is useful to define new business role, i.e., a new provider level called CAN Provider (CANP) [4][5][13]. This is capable to aggregate network resources offered by several NP/ISPs and to create VCANs on top of them. The VCANs are offered by CANPs to by High Level Services Providers (SP) which deploy media services for communities of users, or they are asked by the SPs to CANPs. Each VCAN can be associated to a given QoS class.

A VCAN solution to media flow dedicated transport is proposed in ALICANTE European FP7 ICT research project, "Media Ecosystem Deployment Through Ubiquitous Content-Aware Network Environments" [4]. The VCANs are realized as parallel data planes [10] and are content-type recognition capable under control of a single management and control – M&C plane. The solution, while not fully content oriented as in [2][3], is attractive because it can offer a possibility of seamless deployment and put much less processing tasks on the content aware routers than Information/Content Centric Networking (CCN/CCN) approach.

The network contains several Core Network Domains (CND) and access networks (ANs). The ANs are out of scope of ALICANTE and to VCANs; access network resource control is considered as a separate problem. The CNDs belong to NPs and can be Autonomous Systems (AS). The CAN layer Management and Control (M&C) is partially distributed: one *CAN Manager* (CANMgr) belonging to CANP exists for each IP domain, performing VCAN planning, provisioning, advertisement, offering, negotiation installation and exploitation. Each domain has an *Intra-domain Network Resource Manager* (Intra-NRM), which configures the network nodes. The EU terminals are connected to the network through Home Boxes (HB). The novel CAN routers are called *Media-Aware Network Elements (MANE)* to emphasize their additional capabilities: content and context – awareness. The CAN layer cooperates with HB and SE by offering them CAN services. In the CNDs DiffServ and/or MPLS technologies can support splitting the sets of flows in QoS classes (QC), with a mapping between the VCANs and the QCs with several levels of QoS granularities, [4][5]. The QoS behavior of each VCAN (seen as one of the parallel Internet planes) is established by the SP-CANP.

The VCANs asked from a CANP by an SP should be mapped onto real multi-domain network topology, while respecting some QoS constraints. This provisioning is done through negotiations performed between CAN Managers associated to each network domain. One CANMgr is the initiator of VCAN construction, at request of an SP. If necessary the initiator communicates with other CANMgrs, to finally agree a reservation and then a real allocation (i.e., installation in the network routers) of network resources necessary for a VCAN. A CAN Planning module inside each CANMGr is the entity which runs a *combined algorithm doing QoS constrained routing, VCAN mapping and resource logical reservation*. In this set of actions it is supposed that the initiator CANMgr knows the inter-domain topology at an overlay level and also a summary of each network domain topology, in terms of abstract trunks (e.g., {ingress, egress, bandwidth, QoS class, ..}). This knowledge is delivered by an additional discovery service and is out of scope of this article. Previous papers, of the same authors [5][13], developed and implemented the combined VCAN mapping algorithm. *This article continues the previous work by proposing several techniques for performance and scalability improvement.*

The paper organization is described below. Section 2 makes a short overview on samples of related work. Section 3 summarizes the original VCANs planning and mapping algorithm. Sections 4 and 5 contain the main contribution of this paper. Section 4 develops the optimization techniques, and Section 5 presents some performance analysis results. Section 6 contains conclusions and future work orientation.

II. Related Work

The basic algorithm proposed in [5] and [13] by these authors has as goal to map customized QoS capable VCANs over several network domains, independently managed, to efficiently transport of real-time and media traffic. This paper proposes some optimizations of that basic algorithm. Therefore some related works presented previously are only summarized here. Given that generally such mapping problems are NP-hard [12], a convenient solution has been selected to fit the ALICANTE architecture specific needs. In particular, the CAN Managers and Intra-domain Network Resources Managers– have knowledge on the status of their resources. After paths finding, a negotiation protocol is run, [4][5], between domain managers, to establish inter-domains SLAs. If no QoS constraints are used during routing there are significant chances that the SLA negotiation will fail. A better solutions is to first search for QoS enabled paths, as in [5][6][7][13], followed by SLS conclusions.

The Service Overlay Networks are discussed in [14] which are partially similar to our VCANs. The following assumptions have been considered - part of them similar to our VCAN case: pre-determined location of the overlay nodes; the overlay link metric is the delay; the overlay path between a pair of overlay nodes is selected by using the Dijkstra algorithm; each overlay path is composed of IP-layer links. At IP layer, the cost of each link is 1/Bandwidth, and the shortest path between a pair of IP nodes is computed by using the Dijkstra algorithm. Several overlay topologies have been studied: Full-Mesh (FMsh), K-Minimum Spanning Tree (KMST), Mesh-Tree (MT), Adjacent Connections (ACON), K-Shortest Path Tree (KSPT), Pruned Adjacent Connection (PAC) and Demand –aware adjacent connection (DAC). The overall optimization cost function is a weighted sum of delays on different overlay links weighted with the traffic demands between pair of overlay nodes. The “best” overlay topology was considered if, on equal terms of accepted traffic and performance, has the lowest overhead (minimum number of interfaces per/node), due to the overlay network maintenance traffic. This is not a primary criterion of our solution.

The algorithm in [15] considers both link capacity, and overlay servers capacities. However, this last parameter is out of the scope of our proposal.

The ALICANTE solution is similar to the K-Shortest Path Tree (KSPT), [14] in terms of topology. However, ALICANTE includes in the algorithm for VCAN mapping not only QoS constrained routing based on modified Dijkstra algorithm, but also resource reservation – thus supporting the QoS assurance.

Our solution assumes that an inter-domain overlay QoS peering and routing [13][14], has been solved in the sense that a topology discovery protocol and service exists, capable to make the CAN Mangers aware of topology (at overlay level) and capacities aware.

Overlay networks having QoS capabilities are described in several papers, [6][7][8][9][14]. The solution proposed in [5][13] has a new characteristic that it tries to combine in the same algorithm QoS enabled (constrained) routing, admission control, mapping and resource reservation for VCANs.

III. Basic VCAN mapping Algorithm

This section summarizes the initial algorithm, [5][13], run by the CAN Manager/Intra-NRM in order to map VCAN QoS requirements onto physical network resources onto one or more core network domains (CND). The main input data will be: the multi-domain network graph (topology, capacities) - collected by the topology discovery service; Traffic Demand Matrix (TDM) - asked by SP to the initiator CAN Manager. Note that actually the SP request can contain more parameters in an SLA template including several aspects of the VCAN life and operation. We only considered here the relevant parameters for the VCAN mapping algorithm. The output of the algorithm will be the mapping of TDM on real paths after admission control is done to check respecting the minimum bandwidth constraints and also optimize the network resource usage.

The CANMgr/Intra-NRM runs a combined constrained routing, mapping and admission control and resource reservation algorithm. The metric proposed for a link, is selected as to lead to selection of the widest path.

The cost of a intra-domain link (i,j) in the overlay graph is defined as additive metric *C(i,j) = Breq/Bij= Breq/Bavail*, where Bij is the available bandwidth on this link and *Breq* is the bandwidth requested for that link, [5][13]. The ratio also is seen as link utilization factor; that is the alternative notations will be used: C(i,j) = $U_{link\_ij.}$. The constraint is: *(Breq/Bij < 1)*. Therefore in each action of path search the branches not satisfying this constraint should be not considered. The metric is additive, so one can apply modified Dijkstra algorithm to compute the *Shortest Path Trees (SPT),* one tree for each ingress node where the traffic flows will enter. Note that *Breq/Bij* can be only computed if we know the mapping TT - link (i.e., we know *Breq* for a given link), which is not yet our case. The mapping is to be done jointly with the routing process. So in the first approximation we consider 1/Bij as an additive link metric. Other more sophisticated metrics could be considered, e.g., including the delay, provided that this can be estimated/measured by a monitoring system.

The solution presented here is *valid for both unicast and multicast VCANs*; a multicast TDM is actually a particular case of a unicast TDM matrix. In unicast case the TDM entries are tuples including information like *(ingress, egress, bandwidth, ..)* where each egress may have a different bandwidth request. In multicast case the whole TDM is representing a tree or a set of trees where the bandwidth of a tree is the same for all egress points associated with a root (i.e., ingress of the TDM).

IV. VCAN MAPPING ALGORITHM OPTIMIZATION

One problem discussed in this section is how to reduce the complexity of calculus given that we may have large graphs in a multi domain topology. The Dijkstra's original algorithm runs in $O(|V|^2)$ complexity, or in the best case if the implementation is based on a min-priority queue implemented by a Fibonacci heap, then one has $O(|E| + |V| \log |V|)$ (Fredman & Tarjan 1984). In our case, a TDM may have *n* ingress points (lines), so the complexity is n* O(Dijkstra). For each computation (out of a total of n) the algorithm will determine a constrained Shortest Path Tree (SPT), and then will map the TDM hoses (each TDM line corresponds with a hose) on this SPT. The reservation is done by subtracting the requested capacities from the initial ones per each branch of the graph.

However, the order in which the hoses (i.e., requests) are analysed (and subsequent subtraction) may change the final result. Then if the CAN Manager wants the best VCAN mapping and least overall utilization, then it should check all combinations of computation. The most trivial solution is to recompute the step 2 of the algorithm for other order of inputs given by the bijective function $f(GR_1, ..GR_n) \rightarrow \{GR_{k1}, GR_{k2}, ..GR_{kn}\}$ which creates actually permutations of the set $\{GR_1, ..GR_n\}$, where each GR_k represents a group of requests (i.e., a hose) associated to an ingress point of traffic. The final mapping solution will be the one having the least overall utilization. The overall complexity will be *n*n!* O(Dijkstra)* which has not so good scalability, [13]. Acceptance of such a solution could exist however, given that VCANs are constructed for medium-long term and the frequency of SP requests for VCANs are rather low (non hard-real time computation).

A. Service Provider Driven Priorities

The order of analysis can be more deterministic and the number of computation reduced if the SP assigns a priority order to its requests; then less or even no permutations are needed. Note that a group of requests is represented by a tuple (ingress, egress1, egress2, …). The network available capacities are priority reserved in order, first for the most important requests. In ALICANTE context the SP is the appropriate business actor to know which traffic pipes of the TDM are more important.

In other contexts, the NP could create some particular rules for establish an honoring list. One possible rule could be that the request with the higher requested capacity to be solved the first one. However, in ALICANTE and not only, not always the higher capacity value signifies the most important request.

Two cases are for analysis: a. *strict monotonic row of groups_of_requests priorities;* b. *monotonic row of group_of_requests priorities* (i.e., some of them may be equal). In case *a.* the complexity will be reduced drastically, i.e., we have *complexity = n* O(Dijkstra),* given that the order is strictly determined. In case *b.* one has a structure: $\{(GR_{1,1}, GR_{1,2}, \dots GR_{1,n1}), (GR_{2,1}, GR_{2,2}, \dots GR_{2,n2}), ... (GR_{k1}, GR_{k2}, \dots GR_{k,nk})\}$, where, inside each set of a group (…), all requests have the same priority. Additionally we suppose that the priorities for groups are in strict decreasing order. We also have $n1 + n2 + ... + nk = n$, i.e., the total number of requests. Still in this case one gets a serious reduction in number of computations, given that n1! + n2! + … is much less than *n!.*

As a simple example we suppose that n1= n2 = ..nk =n/k. In this case the total number of permutations will be k [(n/k)!]. Using Stirling approximation formula $n! \sim (2\pi n)^{1/2}(n/e)^n$, we get a reduction factor equal to

$$n!/[k^* (n/k)!] \sim (k)^{n-1/2} \qquad (1)$$

For instance if we have n=10, k=2 we have a reduction factor in number of computation of ~714 and this increases rapidly with *n.* Therefore the solution is much more scalable for large network graphs.

B. Priority Specification Model

The proposed model in this section can be used for both VCAN mapping solutions (in one or two steps) presented in [5][13]. All requests from the received set are grouped based on the source node and *group priority* is defined (lower value means higher priority). In the case of several groups with the same priority, as shown in the sub-section above, the algorithm will permute the processing order obtaining the best cost. Note that the algorithm details have been already described in [5].

In the proposed algorithm, once a group is chosen for analysis, all the requests from that group are processed- but in which order? To offer a maximum flexibility solution w.r.t. SP interests, one should admit that SP can specify a priority for each individual request. So, the model will allow two levels of priorities: *per group* and *per request* inside the group. The choice here is that *group priority has precedence on the individual request one.* However in practice this is not always true. In such cases the solution is to define distinct groups for some requests, for which we want given priorities, despite that the ingress point is the same.

Fig. 1 shows an example on how a TDM can have a split of request in groups assigned to a given ingress point, where the individual requests may have different priorities inside each group. The values P represent the priorities of an individual request.

The future VCAN satisfying this TDM is represented as an outer circle. The actual network may have several interconnected Core Network Domains (three in our example).

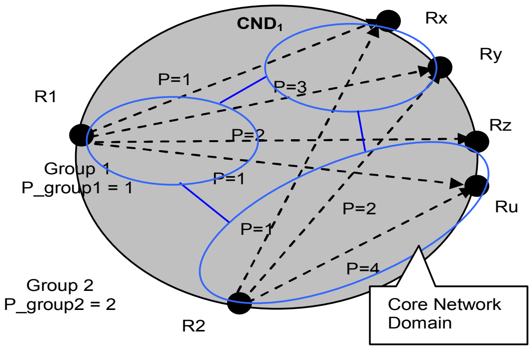
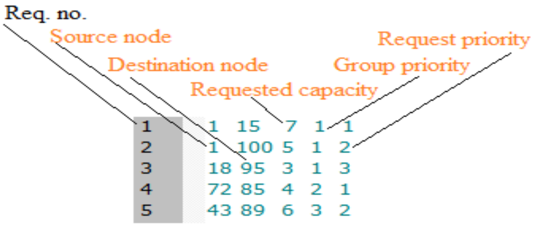


Figure 1. Example of prioritized requests for resources

The sequence of solving the requests in the above example is: 1. {R1-Rx, R1-Ru, R1-Rz, R1-Ry }; 2. {R2-Rx, R2-Ry, R2-Ru}.


Req. no.
Source node
Request priority
Destination node
Group priority
Requested capacity

1	1	15	7	1	1
2	1	100	5	1	2
3	18	95	3	1	3
4	72	85	4	2	1
5	43	89	6	3	2

Figure 2. Traffic Demand Matrix Example

Fig. 2 shows a simplified example of a TDM, containing several requests. This TDM is produced by SP and delivered to the CAN Manger initiating the VCAN construction. Each line of the matrix specifies an individual request as *{source node, destination node, requested capacity, group priority, individual request priority}*. The groups are associated with source nodes being {1, 18, 72, 43}. One can see that the groups {1, 18} have equal same priority =1. This will determine two permutations when analyzing the requests. Each individual request priority inside a group has only local significance.

V. Performance Analysis

This section will present simulation results. The basic algorithm implementation proposed in [5] has been upgraded to leverage priorities. The network has been simulated, by generating the topology using specialized tools.

A. Simulation settings

The tools have been a Network Analysis and Routing eValuation – NARVAL module 2.0.1-1 [16] from Scilab 5.4.0 [17] to generate complex multiple-domains network topologies. Fig. 3 illustrates a two level hierarchy where the bottom part represents the inter-domain network graph and the top one signifies the intra-domain one.

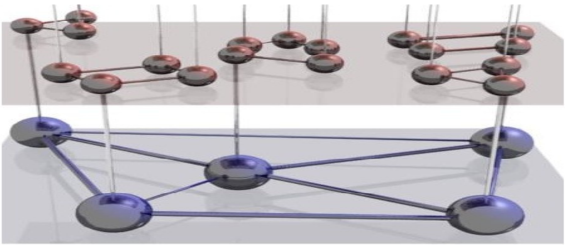


Figure 3. Two level hierarchy topology [16]

Creation of some scripts using the NARVAL module allowed constructing a large hierarchical network with backbone of inter-domains links; each node at inter-domain level represents an abstraction of an intra-domain topology. Each segment for both inter and intra-domain areas has an associated bandwidth generated in respect to a Gaussian distribution centered in 70.

Some default functions have been modified in order to obtain a two levels hierarchical topology (by default there are five levels). The network backbone of size *n* is assumed to be created based on the Waxman model [18], with parameters *a* and *b*. The largest connex subnetwork was extracted. As a matter of use case for the algorithm the backbone needs to be fully connected (through fully connected we understand that there are not isolated nodes, not that the topology is connected in a full mesh fashion). Thereafter the second layer was added according to the Waxman algorithm, too (the same parameters *a* and *b* are used for each network layer). New nodes are added by small groups of size randomly selected into the range [1, 2, 3, …, *nl*]; *cv* is a *s*-length vector, where *s* is the no of layers, that contains the colors used to display each layer [17]. The nodes of the first layer have a diameter (diameter of the circle from the figure representing a node) equal to *db*. The nodes diameter is constant for layer, but we reduce this value when we move to the next layer with a rate of *dd*. The network generated has with 27 backbone nodes and 116 intra-domain nodes; (total is 143 nodes). The topology obtained is presented in Figure 4.

ISBN: 978-1-61208-256-1

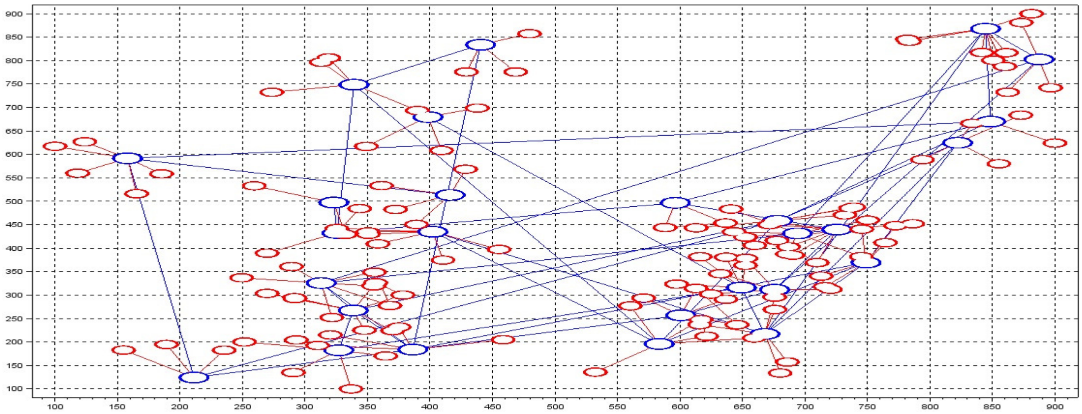


Figure 4. Two level hierarchical network topology generated with Narval tool

The topology was generated using the following parameters values:

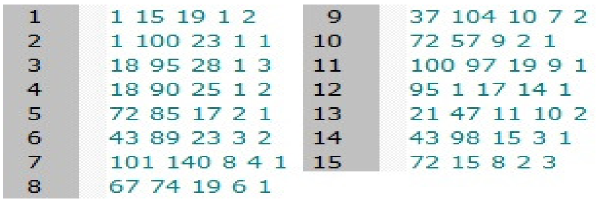
a=0.3;//first parameter of the Waxman model;
b=0.4;//second parameter of the Waxman model;
n=27;//network backbone size; l=1000;//network squared area side;
nl=7;//maximal quantity of nodes per subnetwork;
db=20;//original diameter of nodes;
dd=5;//diameter difference between successive network layers;
cv=[2 5];//color of each network layer.

Using some scripts the adjacency matrix has been extracted with values of 0 and 1 (0 means no link between nodes and 1 means the presence of a link). In order to obtain an adjacency matrix where the presence of a link is represented with the available bandwidth value instead of simply 1, as a last part of our simulations settings we assigned an element of a previous created weight vector (using a Gaussian distribution centered in 70) to all of the adjacency matrix elements different from 0. The weight vector contains elements to be assigned as bandwidth value to each existing link. This assignation process of the corresponding bandwidth value for each link can be seen below:

```
[l c]=size(AM);
ind=find(AM==1);//presence of a link
AMW=AM;//matrix with weight initialized with AM
for i=1:length(ind)
   il=modulo(ind(i),l);//line index
   if (il==0) then
      il=l;
   end
   ic=ceil(ind(i)/c);//column index
   AMW(il,ic)=g.edge_length(NARVAL_G_Nodes2Edge(g,il,ic));
end
```

B. Simulation Results

The TDM proposed contains a set of 15 requests divided into 9 groups with different priorities and different individual priorities as in Fig. 5.



1	1 15 19 1 2	9	37 104 10 7 2
2	1 100 23 1 1	10	72 57 9 2 1
3	18 95 28 1 3	11	100 97 19 9 1
4	18 90 25 1 2	12	95 1 17 14 1
5	72 85 17 2 1	13	21 47 11 10 2
6	43 89 23 3 2	14	43 98 15 3 1
7	101 140 8 4 1	15	72 15 8 2 3
8	67 74 19 6 1		

Figure 5. Example of a set of requests with priorities

Running the algorithm, it produces below results (for only *two permutations* in case of groups with the same priorities):

```
==============================
Input file Scilab1.in:
==============================
Request 1->100, load 23: 1 6 4 16 100
Request 1->15, load 19: 1 14 15
Request 18->90, load 25: 18 4 16 13 90
Request 18->95, load 28 unsatisfied on 18->4,avail.cap. 11. path traveled: 18 4 14 95
Request blind 18->95, load 28, cost 2.860116: 18 8 19 23 14 95
….
Request 37->104, load 10 unsatisfied. Node unreachable.

Cost: 28.76079 of which blind: 9.19815 Satisfied req: 13 / 15
------------------------------------------
Request 18->90, load 25: 18 4 16 13 90
Request 18->95, load 28 unsatisfied on 18->4,avail.cap. 11. path traveled: 18 4 14 95
Request blind 18->95, load 28, cost 3.795075: 18 8 19 6 1 14 95
Request 1->100, load 23: 1 6 4 16 100
Request 1->15, load 19: 1 6 19 23 14 15
…
Request 43->89, load 23 unsatisfied on 4->14,avail.cap. 14. path traveled: 4 14 15 24 25 13 89
Request blind 43->89, load 23, cost 6.202275: 43 4 6 19 23 14 15 24 25 13 89
…
Cost: 31.10171 of which blind: 9.99735 Satisfied req:  13 / 15
------------------------------------------
Best cost: 28.760790
Satisfied Requests: 13 / 15
Total time: 0.018000
```

As it can be seen, only 13 requests from 15 are solved and a better cost is associated to the first order (excepting the situation of node unreachable). Two requests could not be solved using the modified Djikstra algorithm and in this special case the blind search found an alternative path. Only this blind search adds an extra cost because of the longer found path compared to the Djikstra one. All requests are honored according to the group and individual priorities. As an alternative choice, in the case of many groups with the same priority, one can be specified how many permutations are desired. We used only 2 in this example.

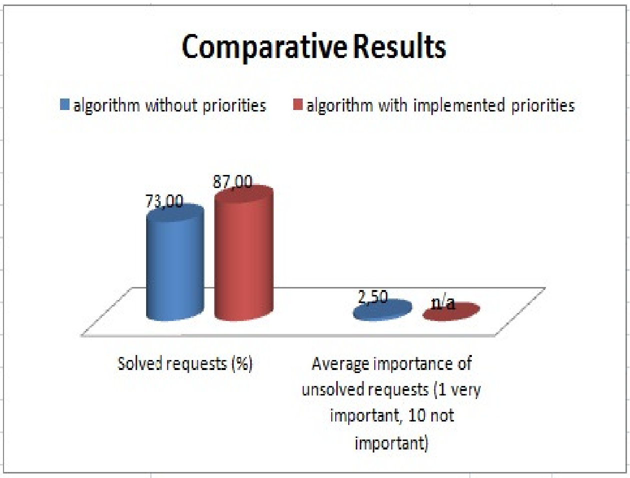


Figure 6. Comparative results

Running the basic algorithm without any priorities, but for the same input file, *only 11/15 requests are solved.* The most important thing is that some important requests (with priorities 2 and 3 in this new context) have not been solved, while and some less important requests have been solved. Fig. 6 shows comparative results for initial algorithm and that one having priorities.

For both cases, there were not taken into consideration the cases of unreachable nodes. Even if the network graph is constructed as a connex one, because of some optimization techniques used during the implementation (removing from the existing graph all segments which do not respect the condition: *available bandwidth >= minimum request* bandwidth value from the group) some nodes could become unreachable. For the case with prioritized requests the two unsolved requests are because of the unreachable node, so in this comparison we consider unsolved requests as n/a.

VI. CONCLUSIONS

This paper proposed optimization methods to increase the performances of a previously developed combined algorithm, having the goal to map Virtual Content Aware Networks on top of multi-domain IP topologies, while respecting QoS constraints. It is shown how introduction of priorities in the Traffic Demand Matrices asked by the Service Provider can greatly reduce the number of computations while increasing the number of solved requests, in comparison with the basic algorithm. Future work will extend the evaluation on several types of topologies (sparse, dense) and allocate resources for several types of QoS classes. Currently, the algorithm is developed inside the CAN Manger of the ALICANTE FP7 project.

ACKNOWLEDGMENTS

This work was supported partially by the EC in the context of the ALICANTE project (FP7-ICT-248652) and partially by the project POSDRU/88/1.5/S/61178.

REFERENCES

[1] T. Anderson, L. Peterson, S. Shenker, and J. Turner, “Overcoming the Internet Impasse through Virtualization”, Computer, vol. 38, no. 4, Apr. 2005, pp. 34–41.

[2] J. Choi, J. Han, E. Cho, T. Kwon, and Y. Choi, A Survey on Content-Oriented Networking for Efficient Content Delivery, IEEE Comm. Magazine, March 2011.

[3] V. Jacobson et al., “Networking Named Content,” CoNEXT ’09, New York, NY, 2009, pp. 1–12.

[4] FP7 ICT project, “MediA Ecosystem Deployment Through Ubiquitous Content-Aware Network Environments”, No248652, “D2.1: ALICANTE Overall System and Components Definition and Specifications”, http://www.ict-alicante.eu/ (last access March 2013)

[5] R. Miruta, E. Borcoci and E. Palis „Planning and Provisioning of Virtual Content Aware Networks over IP Infrastructures”, International Conference on Telecommunication and Multimedia, July 2012, pp. 118-123

[6] Z. Li, P. Mohapatra, “QRON: QoS-Aware Routing in Overlay Networks”, IEEE Journal on Selected Areas in Communications, Vol. 22, No. 1, January 2004, pp.29-39.

[7] J. Galán-Jiménez and A. Gazo-Cervero, “Overview and Challenges of Overlay Networks”, International Journal of Computer Science & Engineering Survey (IJCSES) Vol.2, No.1, Feb 2011, DOI : 10.5121/ijcses.2011.2102 19

[8] Z. Li, P. Mohapatra, and C. Chuah, Virtual Multi-Homing: On the Feasibility of Combining Overlay Routing with BGP Routing, University of California at DavisTechnical Report: CSE-2005-2, 2005.

[9] L.F. Verdi and F. Magalhaes “Using Virtualization to Provide Interdomain QoS-enabled Routing”, Journal of Networks, April 2007, pp. 23-32.

[10] M. Boucadair, et al., “A Framework for End-to-End Service Differentiation: Network Planes and Parallel Internets”, IEEE Communications. Magazine, Sept. 2007, pp. 134-143.

[11] Z. Wang and J. Crowcroft, “Quality-of-service routing for supporting multimedia applications”, IEEE Journal on Selected Areas in Communications, vol. 14, no. 7, 1996, pp. 1228—1234.

[12] A. Haider, et. al., , “Challenges in Resource Allocation in Network Virtualization”, 20th ITC Specialist Seminar, 18.-20. May 2009, Hoi An, Vietnam, http://www.itcspecialistseminar.com/ paper/itcss09Haider.pdf (last access Feb 2013)

[13] E. Borcoci, R. Miruta and S. Obreja, “Multi-domain Virtual Content-Aware Networks Mapping on Network Resources” , 20th European Signal Processing Conference, August 2012, pp. 2223-2227

[14] D. Adami, et. al., Design and Performance Evaluation of Service Overlay Networks Topologies, Journal of Networks, Vol. 6, No. 4, April 2011.

[15] A. Karamoozian, M. Erfani and A. H. Abdullah, "QoS-Satisfied Dynamic Routing based on Overlay Service Network", Second International Conference on Communication Software and Networks, ICCSN 2010, Feb 2010, pp. 441-445

[16] http://atoms.scilab.org/toolboxes/NARVAL (last access March 2013)

[17] http://www.scilab.org/ (last access March 2013)

[18] http://www2.math.uu.se/research/telecom/software/stgraphs.html (last access March 2013)

Electric-Vehicle-based Ad Hoc Networking and Surveillance for Disaster Recovery

Proposal of Three-Dimensional Mobile Surveillance Using Electric Helicopters

Kenichi Mase
Graduate School of Science and Technology, Niigata University
Niigata-shi, Japan
mase@ie.niigata-u.ac.jp

Takuya Saito
Research Institute for Natural Hazards & Disaster Recovery, Niigata University
Niigata-shi, Japan
takuya@net.ie.niigata-u.ac.jp

Abstract—**In this paper, we argue that small electric vehicles (mini-EVs) may become increasingly prevalent in the future, leading to the realization of the so-called ubiquitous EV society. EVs have the potential to be a great resource for recovery from large-scale disasters. Specifically, each EV can be equipped with wireless communication devices, and so EVs in a disaster area can link together to form an EV-based mobile ad hoc network (EVANET). Two use cases of EVANET are emergency networks and disaster area surveillance. Focusing on the latter application, we present the concept of three-dimensional mobile surveillance (3DMS). In 3DMS, EVs on the ground cooperate with small lightweight unmanned electric helicopters (mini-EHs) to cooperatively collect information about the disaster area. Each EV and mini-EH pair is equipped with cameras and other sensors to monitor disaster damage. We present an effective method of solving the problem of limited continuous flying time of EHs. We investigate the requirements for a mini-EH in 3DMS. In particular, we demonstrate that autonomous piloting is necessary. To support autonomous piloting, we propose that EH position information be obtained using GPS and a low-power wireless transmission system between the mini-EH and the corresponding EV. We call this system the EH positioning system. Based on real-world experiments, we show that a prototype of the EH positioning system has adequate capabilities in terms of both transmission range and energy consumption.**

Keywords-ad hoc network; disaster; electric vehicle; helicopter; GPS

I. Introduction

The Great East Japan Earthquake and the resulting tsunami on March 11, 2011, caused severe disruption in both the telecommunications and power supply networks [1]. In the fixed telephone system (NTT-East), 18 telecom buildings were destroyed and 23 were flooded, 90 transmission routes were disconnected, 1,000 telecom buildings were powered off and 300 ceased to operate owing to a shortage of fuel for generating electric power, and 1.5 million subscriber lines became unusable. In the cell phone networks (NTT DoCoMo, KDDI, and Softbanks), 12,000 base stations ceased to operate (mostly because of commercial power failures and consequent battery power shortage), and the traffic regulation ratios were 70–95% and 30% for voice and mail, respectively. After the occurrence of the earthquakes and the tsunami over a wide area of East Japan, telecom companies made remarkably continual efforts to repair the disrupted buildings, facilities, and equipment in order to restore their services under the harsh conditions. However, the resources of the telecom companies were very limited in comparison with the scale of the damage caused by the disaster. Depending on their location, the population of the affected areas had either no communication services or disrupted services for days, weeks, or even months after the earthquake.

Today, environmental destruction and global warming have become serious concerns. Thus, there is an urgent worldwide need for the evolution of low-carbon sustainable societies. Automobile exhausts have been, and continue to be, a major cause of air pollution, but the markets for emission-reducing hybrid cars and electric vehicles (EVs) have experienced significant growth recently. EVs are an ideal exhaust-free means of minimizing air pollution in large cities, and they are expected to become increasingly popular as their associated costs continue to decrease. Further, in both developing and advanced nations, small one- or two-seater EVs (mini-EVs) are attracting attention as economical vehicles for use in communities, and it is expected that they will become increasingly prevalent in the near future. A further advantage of such vehicles is that they can provide mobility to the elderly, many of whom find regular walking difficult, so mini-EVs are an attractive proposition for aging societies.

When a large-scale disaster occurs, the first priority is to recognize the nature and scope of the disaster damage over the affected area in order to most effectively begin rescue and disaster recovery activities. The efficiency of activities such as surveying the damage, discovering survivors, and saving lives, however, is reduced by the lack of information on the disaster area owing to the destruction of and damage in the communications infrastructure and prolonged confusion in telecommunications services, as mentioned earlier. There is an urgent need to establish a more effective way of quickly providing a temporary communications and surveillance network over a large disaster area.

In this study, which assumes a growing market penetration by EVs in the near future, a novel approach for efficient networking and surveillance over a wide disaster area is presented. In our proposal, EV-based mobile ad hoc networking is used to rapidly deploy a communications network in the disaster area. A key player in our proposed system is the unmanned electric helicopter (EH). EVs and EHs are cooperatively employed to perform surveillance activities within the disaster area, allowing three-dimensional mobile surveillance (3DMS). There are many studies on applications of unmanned aerial vehicles (UAVs) including helicopters [2]-[14]. Some of these are targeted at disaster surveillance. However, the potential and benefits of EHs for use in wide-area disaster surveillance have not been fully explored.

Our main contributions are as follows: 1. The novel idea of cooperatively using EVs and unmanned EHs for wide-area disaster surveillance. 2. An effective method of solving the significant problem of limited continuous flying time of EHs. 3. The prototype of a feasible EH positioning system for autonomous piloting. The rest of this paper is organized as follows: Section II describes the usefulness of ad hoc networks using EVs to provide emergency communication networking in disaster areas. Section III presents the concept of 3DMS using EVs and EHs, with an analysis of the availability of EHs for airborne surveillance. Section IV describes not only the EH positioning system to support autonomous piloting but also the prototype development and initial experimental results. Finally, Section V offers our conclusions.

II. Ad Hoc Networks Using Electric Vehicles Deployed in Disaster Areas

A. Background

The electric vehicle (EV) is considered an attractive alternative to the gasoline-powered car as a means of reducing air pollution caused by automobile exhaust gases. With the objective of replacing gasoline-powered cars with electric ones, major automobile manufacturers have been developing EVs whose production costs and continuous driving ranges are comparable with those of conventional automobiles powered by gasoline engines. Although these goals have not yet been fully achieved, the EV market has recently experienced significant growth.

A very small EV with one or two seats, which we call a mini-EV (see the specifications of mini-EVs in [15]), is another promising development, which in the near future may be popular in communities regardless of issues of air pollution from CO_2 emissions. Mini-EVs are cheaper to purchase or maintain and require smaller parking spaces, allowing households to possess them as their first car or as additional cars. They are also ideal for B2B and B2C delivery services in communities. They have further uses, especially in an aging society. It has been reported that cars driven by elderly drivers typically carry only one or two passengers, and the distance driven each day is relatively short, conditions that are well matched by the limited capabilities of current EVs. Unlike a gasoline-powered vehicle, an EV has no mechanical engine and requires little maintenance, which is convenient for the elderly. Gasoline-powered vehicles must also be refueled at gas stations, which can be a burden on an elderly driver. In contrast, a mini-EV can be powered at home or solely by a solar battery mounted on the roof, thereby eliminating the time needed to attend a gas station. These features significantly reduce the burden of vehicle maintenance on the elderly.

A serious problem in many communities is the inconvenience caused by changes in public transportation services, such as where the frequency of a bus service is reduced or cancelled on routes that are unprofitable. In this context, the availability of mini-EVs in a community becomes important. A personal EV may be conveniently used to support personal mobility for daily activities such as shopping and visiting others in the neighborhood, particularly for the aged who may have difficulty walking. Thus, personal EVs can improve the quality of life (QOL) for the elderly and contribute to the realization of a vibrant community. Mini-EVs will potentially create a new EV market in the near future, achieving significantly greater penetration in the community to fashion what in this paper is referred to as the ubiquitous EV society. The battery capacity of an EV, even a mini-EV, is not comparable to that of the battery in a gasoline-powered car. An improvement in the energy density of batteries is further expected. The ubiquitous EV society will be one of ubiquitous large-capacity batteries, which will additionally prove to be useful during any prolonged blackout following on from a large-scale disaster.

B. EV-based Ad Hoc Networks

Communications after large-scale disasters and between cars have been considered major applications in the research and development of mobile ad hoc networks (MANETs) [16]. With regard to car-to-car communication, the vehicular ad hoc network (VANET) [17] has been studied particularly within the framework of intelligent transport systems (ITS). A vehicle is equipped with a communication device, allowing the vehicle to act as a communication station to form a MANET with other vehicles. The presence of gasoline-powered vehicles is implicitly assumed in VANET research, and the applications of a VANET while the vehicle is driven are of major interest. With its increasing penetration, however, the mini-EV can become a principal player in the formation of a VANET. An EV-based ad hoc network is called an EVANET [15]. It is noteworthy that an EV can operate as a communication station for a long time by using its large-capacity battery, regardless of whether it is moving or stationary, whereas a gasoline-powered vehicle cannot provide the same utility when its engine is switched off. Consequently, EVANET applications need not be limited to situations where the vehicle is being driven, but may be enhanced to provide new services when the vehicle is parked. For example, suppose that mini-EVs in individual household parking spaces at night form an EVANET within a community. In such a scenario, the EVANET might act as a sensor network to detect or prevent crimes such as burglary.

C. Use Cases for Disaster Recovery

In the ubiquitous EV society, a tremendous number of EVs may be in use by a community. When a disaster occurs, many of these may be utilized for disaster recovery. Mini-EVs owned by public offices and those volunteered by individuals can be deployed to form an EVANET. In order to

TABLE I. ALTERNATIVES FOR SURVEILLANCE FROM THE AIR IN COOPERATION WITH MINI-EVS

	Mobility	**Hovering**	**Wind resistance**	**Operation time**	**On-EV charging**	**Logistics**
Tethered balloon	No	Yes	Limited	Fair	Feasible	Full backup needed
Airship	Low	Yes	Limited	Fair	Difficult	Full backup needed
Helicopter	Low	Yes	Fair	Limited	Feasible	Light support needed (electric helicopter)
Airplane	High	No	Fair	Limited	Difficult	Full backup needed

reach a position suitable for networking, a mini-EV can take advantage of its small body to pass through narrow routes while avoiding any obstacles created by the disaster. Two major applications of EVANETs are presented below.

1) Emergency network: When telecommunications services are degraded or disrupted by traffic congestion or damage to network facilities, an EVANET can function as a secondary telecommunications infrastructure (emergency network) throughout the affected area. Forming an EVANET can drastically shorten the network construction time compared to conventional methods. The surplus battery power available can be remotely monitored using the EVANET itself, and mini-EVs whose batteries have been exhausted can easily be replaced with others whose batteries are fully charged. Such emergency networks can be used for rescue purposes, damage surveying, and communicating with people in shelters [18]. The higher their antennas are located, the greater the probability that a line-of-sight (LOS) exists between two communication stations. A longer transmission range is also expected because of the decrease in radio reflection from the ground. For reasons of wind resistance and driving safety, a 10-m-high antenna cannot be used on an EV while it is being driven. However, a high antenna can be used when the EV is parked. Specifically, an elastic pole fixed vertically to the car body can be used to support an antenna mounted on its top. This pole can be retracted and carried by the EV while it is driven. The experimental results show that a higher antenna can extend the transmission range owing to the increase in the line-of-sight range and the decrease in radio reflection from the ground. The throughput of a 9-m-high antenna is significantly greater and more stable than that of a 2-m-high antenna [19].

2) Disaster area surveillance: An EVANET can function as a surveillance infrastructure in a disaster area. In addition to its communication device, each EV is equipped with cameras and sensors for collecting information on the affected area. The information collected by each EV is delivered through the EVANET to the data collection station, where an Internet gateway is available. It is thus possible for a disaster recovery headquarters to efficiently and quickly obtain basic information over a wide disaster area in order to acquire an overview of the damage to allow for an effective allocation of rescue and recovery resources.

III. THREE-DIMENSIONAL MOBILE SURVEILLANCE

A. Surveillance from the Air

By deploying a number of EVs in the disaster area, it is possible to achieve mobile surveillance that is both extensive and efficient. In addition, airborne surveillance is considered an effective way to efficiently obtain a complete view of a disaster area [11]-[14]. By combining the surveillance results from the air and ground, improvements in the coverage, accuracy, and efficiency of the surveillance can be expected.

The various means of airborne surveillance are listed and qualitatively evaluated in Table I. The airship, helicopter, and airplane can be either manned or unmanned for piloting. The former is costly and the availability of human pilots is limited. In this paper, the latter is investigated. The tethered balloon and airship have relatively little wind resistance, since both use bladders inflated with helium gas for lift. In the case of the tethered balloon, it is necessary to bring helium gas bottles to the site for inflating or refilling the balloon, for which the balloon needs to be brought to the ground and then released again. The surveillance coverage is limited, but a power source is not required, allowing for relatively lengthy surveillance [20]. The airship allows extensive flying and hovering over the disaster area, but fuel is necessary for flying and continuous operation time is limited by both the decrease in gas pressure and the fuel consumption. The helicopter has characteristics in common with the airship, but fuel consumption is greater and continuous operation time is further limited. Due to its flight characteristics, the airplane has difficulty both in continuously monitoring the same area and in adapting its route according to the situation. Like the helicopter, its continuous operation time is limited by its fuel consumption. As the power source for flying, an engine using a fuel such as gasoline or a motor with a battery may be considered, but the appropriate choice is necessarily based on the application requirements. A motor can be selected if the helicopter or airplane is small and light and a relatively short continuous operation time is acceptable. The need for a human pilot on the ground should be avoided by means of autonomous piloting, which allows for a longer period of continuous operation. Since a battery is necessary to provide the energy source for mounted cameras and communication devices, its capacity is also a limitation on continuous operation time. Such constraints need to be carefully balanced at the design stage with respect to continuous operation time.

Each method requires support services such as the supplementing of helium gas or engine fuel and battery charging for a motor or communication device. In mobile surveillance, where a large number of EVs are deployed within the disaster area, an EV on the ground can be used as a base station for battery charging (on-EV charging). On-EV charging can easily be used for a tethered balloon or a helicopter. In the case of a very small EH, the roof of an EV can be used as the heliport and the battery of the EH can then be charged from the battery of the EV. An airship is considerably bigger than a tethered balloon and requires a relatively large space for landing, and an airplane needs a runway for takeoff and landing, while a tethered balloon needs periodic topping-up of its helium gas. These supports are unnecessary for an EH owing to on-EV charging.

B. *System Concepts*

Based on the qualitative evaluation of the possibilities of airborne surveillance in Section III.A, we have chosen the EH for our system. The proposed surveillance system is thus composed of EVs, including mini-EVs, and tiny unmanned EHs (mini-EHs). In the proposed system, each EV also functions as the carrier of a mini-EH. That is, each EV is equipped with a mini-EH on its roof for the deployment, takeoff, and landing of the mini-EH. Moreover, the EV provides charging from its battery to that of the mini-EH. The EVs and mini-EHs are equipped with cameras and sensors used for surveillance. These unique features differentiate the proposed system from existing approaches [11]-[14]. This system is referred to as a three-dimensional mobile surveillance (3DMS) system. The smallest configuration of the 3DMS system is a single pairing of an EV and an EH. Many EVs and EHs, however, are required to participate in cooperation for the surveillance of a large disaster area. In principle, the pairing between EVs and EHs is fixed, but more flexible pairing relations can also be considered. For example, some EVs may perform surveillance without their partner EHs, or may cooperate with and provide charging to EHs other than their partner.

To minimize the total weight of the mini-EH, the weight of the battery for its motor is strictly limited, which means that the battery capacity is also limited, allowing only for a relatively short continuous flight. However, a mini-EH can engage in airborne surveillance repeatedly, owing to on-EV charging. A carrier EV may be equipped with a fast charger to shorten the charging time. In addition, a carrier EV can be equipped with spare EH batteries. By substituting a spare EH battery for the spent battery of the landed mini-EH, the waiting time of the mini-EH on the heliport of the EV is shortened. The spent EH battery from the mini-EH is recharged on the EV in preparation for reuse. The driver of the EV is typically someone who cannot manually pilot the corresponding mini-EH. Autonomous piloting of the mini-EH should thus be supported in principle.

It is possible to view the entire disaster area from the air. A mini-EH can monitor from the air locations that an EV cannot access owing to a lack of roads or the presence of obstacles, and from there can then take pictures and search for survivors. EVs and mini-EHs form a mobile ad hoc network and share the information collected by each EV and mini-EH, delivering such information to a data collection station. When two EVs are out of transmission range with each other, a mini-EH in the air can be used as their relay node or message carrier [21].

TABLE II. REQUIREMENTS FOR MINI-EH IN 3DMS

Mini-EH Features	Required Values
Size (Length and width)	Less than 1 m
Mass	1–3 kg
Maximum altitude	100–200 m
Maximum speed	50–60 km/h
Payload	500–1000 g
Continuous flying time	No less than 15 min
Wind resistance	6 m/s
Rain resistance	10 mm/h
Remote-control range	200–1000 m

C. *Requirements*

Based on the system concepts given in Section III.B, the requirements for a mini-EH in 3DMS are summarized in Table II.

A mini-EH can be remotely piloted through the use of commands (simple piloting). Specifically, the following functions should be supported:

- Takeoff and ascent to the designated altitude
- Hovering at the designated altitude and position
- Landing from the current position
- Progressing or retreating, sliding right or left, and ascending or descending at the designated speed and time
- Turning right or left with the designated angle
- Circuitous flight around the designated position with the designated speed

In addition, autonomous piloting should be supported for a mini-EH. Some examples of autonomous piloting are as follows:

- Landing on the heliport on the roof of the corresponding carrier EV
- Flying at the same altitude, direction, and distance from the carrier EV
- Returning to the takeoff location and landing when communication with the carrier EV is lost
- Returning to the carrier EV and landing before the battery is spent

The development of a piloting system that meets the above requirements is beyond the scope of this paper.

D. *Availability of EH for Airborne Surveillance*

As mentioned in Section III.B, a mini-EH can fly continuously for a relatively short time due to its limited battery capacity, and the on-EV charging time may be considerably longer than the continuous flying time, resulting in a low availability, which is defined as the ratio of the total flying time to the total surveillance period. By using spare batteries, the availability can be improved. Without loss of generality, we assume that the mini-EH consumes a single battery for each flight. Let *n* be the total number of

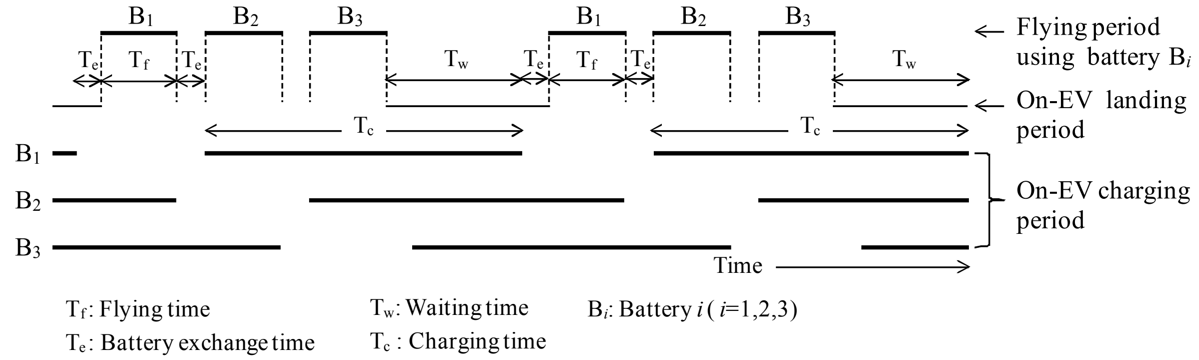


Figure 1. Rotation of charging and use of batteries for EH flying.

batteries, including spare batteries. In the example of Fig. 1, three batteries are used in rotation, where the ith battery B_i ($i = 1, 2, 3$) is used during the flying periods marked by B_i in the upper part of the figure, and is charged on the EV during the periods marked by B_i in the lower part of the figure. The following relation is easily obtained from Fig. 1:

$$\begin{aligned} T_w &= T_c - (n-1)T_f - (n-2)T_e \qquad && (n < \frac{T_c + T_f + 2T_e}{T_f + T_e}) \\ &= 0\ . && (Others) \end{aligned} \tag{1}$$

The availability, A, is then given by the following equation:

$$\begin{aligned} A &= \frac{nT_f}{T_c + T_f + 2T_e} \qquad && (T_w > 0) \\ &= \frac{T_f}{T_f + T_e}\ . && (T_w = 0) \end{aligned} \tag{2}$$

A numerical example is given in Fig. 2, where the availability versus the total number of EH batteries n are shown for three different charging times T_c: 15, 30, and 60 min, assuming that T_e = 1 (min) and T_f = 15 (min). The availability increases with increasing n and becomes

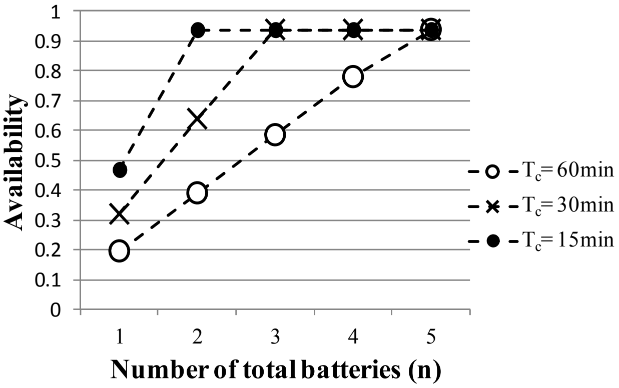


Figure 2. Number of total batteries.

saturated when n reaches and exceeds its minimum required value; at this point, the mini-EH does not need to wait for the on-EV charging to be completed, and maximum availability is obtained. It is shown that the waiting time and the total number of batteries are in the trade-off relation, and relatively fewer EH batteries are sufficient to obtain the maximum availability under practical assumptions.

IV. EH Positioning System to Support Autonomous Piloting

A. Basic Approach

Most EH products in the market support simple operator-controlled piloting and some partly support the autonomous piloting mentioned in Section III.C. There are two approaches to autonomous piloting for 3DMS. In the first approach, autonomous piloting is developed on the basis of existing products. In the second approach, autonomous piloting is developed independently of the existing products. We adopt the second approach because we want to develop an autonomous piloting system that can be applied to any EH product, allowing the user to select an appropriate EH product for 3DMS from many candidates. In the proposed method, the basis of the autonomous piloting is to recognize the current positions of the mini-EH and the partner (carrier) EV, which is responsible for piloting the mini-EH (and is thus referred to as the piloting EV). To this end, both the mini-EH and the EV are equipped with GPS receivers. The GPS measurement data are periodically transmitted from the mini-EH to the partner EV. A similar method was also used for managing the formation flight of unmanned aerial vehicles (UAVs) in [8]. However, our system and prototype are both optimized to our application in terms of transmission range and the maximum speed and payload of the mini-EH. The mini-EH controller on the EV compares the GPS data history of the mini-EH with that of the EV and sends appropriate commands to guide the mini-EH toward the desired position. In order to realize such autonomous piloting, the EV needs to have functions to periodically collect GPS data from the mini-EH. We call this system the EH positioning system.

The transmission range of the radio system used for the

EH positioning system should be designed sufficiently greater than that of the one used for remotely piloting the mini-EH. When the mini-EH happens to wander beyond its remote-control range, the carrier EV can still receive position data from the mini-EH. According to the requirements given in Section III.C, the mini-EH should be designed to land on the site when it has lost control from the piloting EV. The mini-EH continues to periodically transmit location messages so that the piloting EV can easily search for the lost mini-EH. In this sense, the EH positioning system is independent of the remote-piloting mechanism of the mini-EH, in accordance with the aforementioned second approach, which is more reliable than the first approach.

B. System Components

The EH positioning system is composed of a tag mounted on the mini-EH and a base station in the partner EV. An EH-tag is composed of a CPU, GPS receiver, barometer, radio transceiver, antenna, memory, and battery (see Fig. 3). The GPS receiver in the EH-tag periodically performs a measurement and the GPS data are transmitted to the base station on the EV. To reduce the battery consumption of the EH-tag, a license-free low-power transmission system with a maximum transmission power of 10 mW is used for data transmission from the EH-tag to the base station on the EV. We expect that when the EH is in flight, the GPS measurements by the EH-tag are always successful, line-of-sight is maintained between the EH-tag and the base station, and the measured GPS data can be successfully transmitted to the base station.

A base station includes a communication module, micro server, memory, and router. The communication module (see Fig. 4) is used to receive GPS data sent from the EH-tag. The base station is connected to the mini-EH controller on the EV (see Fig. 5). It is also equipped with various sensors, including a camera, weather sensor, and barometer to monitor the surroundings of the EV.

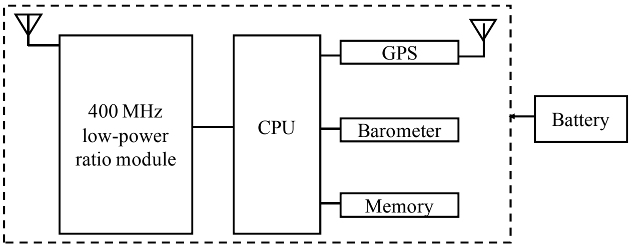


Figure 3. Functional blocks of an EH-tag.

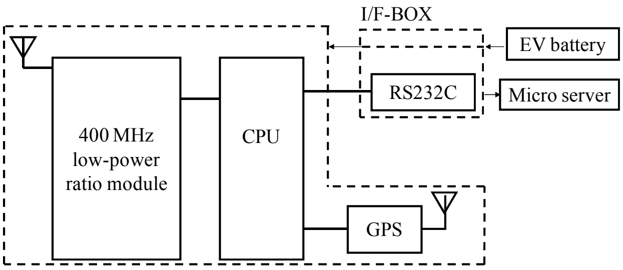


Figure 4. Functional blocks of the communication module of a base station on an EV.

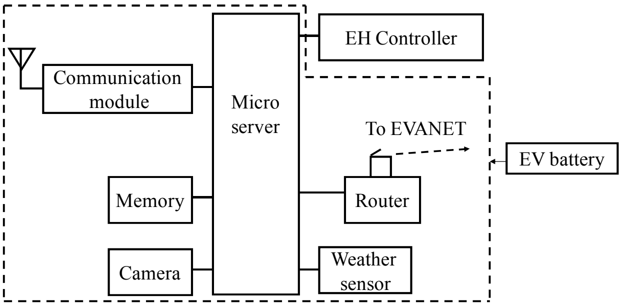


Figure 5. Functional blocks of a base station on an EV.

C. Power-Saving Mechanism

The mini-EH controller on the EV periodically sends status information, such as flying/landing status and flying speed, of the EH to the base station. The base station in turn determines the frequency of GPS measurements and sends it to the EH-tag. For energy conservation, GPS measurements in the EH-tag are periodically performed only when in the flying state. The measured GPS data and the barometer readings are subsequently sent to the base station. The base station periodically updates the period of GPS measurements. If a pre-determined number of consecutive updates are not received, the pre-determined minimum period is configured in the EH-tag.

The frequency of the GPS measurements during the flight of the mini-EH should be configured to detect a change of distance between the mini-EH and the piloting EV of the order of 10 m flown at the maximum relative speed. When it is 10 m/s, the frequency should be set to around once per second. This frequency may be dynamically configured to decrease, reflecting the normally lower speed of the mini-EH, so that energy consumption may be minimized.

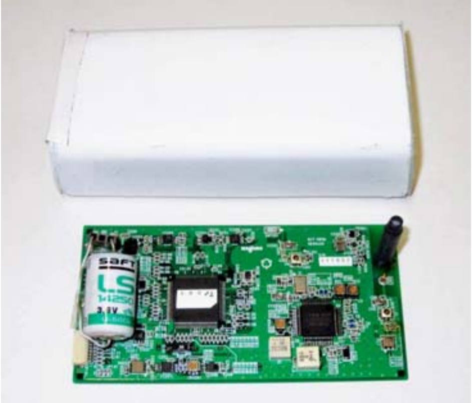


Figure 6. The case and IC motherboard developed for an EH-tag.

TABLE III. SPECIFICATIONS OF COMPONENTS FOR EH-TAG

Components	Size (mm)	Mass (g)
IC motherboard	96 × 50 × 0.8	10.5
Case	110 × 53 × 49	33
Lithium-thionyl chloride battery	*Φ14.5 × 24.5*	9
GPS antenna	3.2 × 1.6	0.9
Radio antenna	*Φ5.5*	0.9
GPS module	15 × 12.5	1.1
Microcomputer	17×17	0.5
Resistor, condenser	1.0×0.5	Not measurable

D. *Prototype Development*

We developed prototypes of the EH-tag and base station mentioned in Section IV.B. The main challenge was the development of a small lightweight EH-tag that can be mounted on the mini-EH. A picture of the developed EH-tag is shown in Fig. 6, and the specifications of the components selected for the EH-tag are listed in Table III. The total mass is about 60 g, which meets the mini-EH payload limit of the requirements in Table II. We used only existing products for the components of the developed EH-tag. Further reductions in size and weight can be expected by using components designed specifically for the EH-tag.

E. *EH-tag Battery Lifetime*

A small lightweight battery that is capable of operating at 3 V and of the order of 10 mA is necessary as the energy source of the EH-tag. To meet this requirement, a lithium-thionyl chloride battery with a capacity of 1,100 mAh was selected. To estimate the lifetime of the battery used for the EH-tag, we conducted experiments in which GPS measurements were performed every minute and the measurement data were transmitted to the base station each time the measurement succeeded. Based on the results of the experiments, the battery lifetime is about 31 h. The current draw is highest, at 45 mA, during GPS measurements. Assuming that a GPS measurement is conducted every second in the worst case, as mentioned in Section IV.C, and that the current draw is constant at 45 mA, the calculated lifetime of the EH-tag battery is about 24 h. It is also expected that the GPS measurements will stop or slow down, as mentioned in Section IV.C. Considering these factors, which reduce the energy consumption, the EH-tag can be expected to work continuously for at least one whole day without the need for recharging or replacement, which is well suited to a 3DMS application.

F. *Real-World Data Delivery Experiments*

We conducted experiments to verify the transmission range of the EH positioning system by using the base stations on the Niigata University campus and in Sawata on Sado Island (see Fig. 7). GPS measurements were performed by the EH-tag every 1 min for 10 min at each distance from the base station and at each height from 20 m to 100 m above the ground in 20 m steps. Unlike our objective of using EH-tag on the EH, we used a balloon to hold the EH-tag in the air at each height in this experiment, because a prototype EH was not ready at that time. It is worth noting that a similar data delivery performance would be obtained when the EH-tag is actually mounted on a mini-EH. The measured GPS data were sent to the base station. As expected, there were no failures in GPS measurements. The average data transmission success ratios are shown in Fig. 8. It is observed that a data transmission success ratio of almost 100% was achieved at both sites for distances up to 6 km. However, there were examples of the case in which the transmission is unsuccessful even though the distance is relatively short. We surmise that these failures occur when the EH-tag posture is changed due to the balloon swinging in the wind.

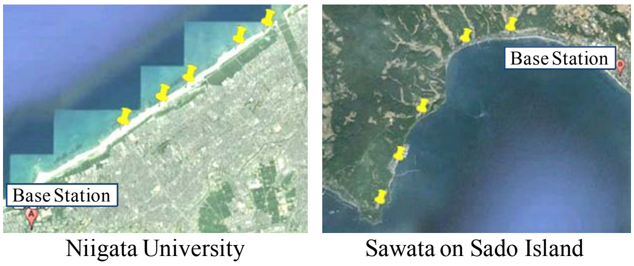


Figure 7. Experimental sites.

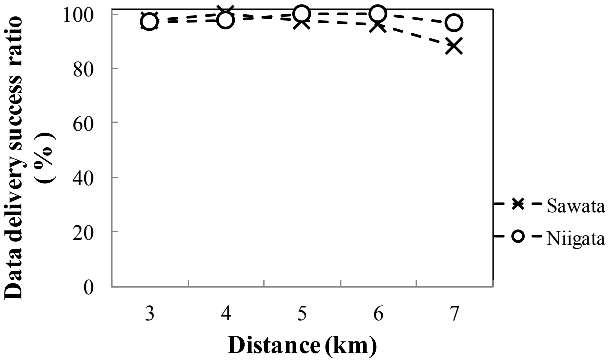


Figure 8. Data delivery success ratios for the EH positioning system.

Thus, it has been confirmed that the EH positioning system has sufficient data transmission range, significantly exceeding the remote piloting range of the mini-EH, and can be applied to 3DMS.

V. CONCLUSION AND FUTURE WORKS

In this paper, we argued that small electric vehicles (mini-EVs) may be increasingly common in the future, leading to the realization of the so-called ubiquitous EV society. In the future, these EVs can be a great resource for recovery from large-scale disasters. Specifically, where each EV is equipped with wireless communication devices, a number of EVs available in a disaster area can form EV-based mobile ad hoc networks (EVANETs). Two use cases of EVANET are emergency networks and disaster area surveillance. Focusing on the latter application, we presented the concept of three-dimensional mobile surveillance (3DMS). In 3DMS, the EVs on the ground cooperate with small lightweight unmanned electric helicopters (mini-EHs) to collect information on the disaster area. Each EV and mini-EH pair is equipped with cameras and other sensors to monitor damage arising from the disaster. We gave an effective method of solving the problem of the limited continuous flying time of EHs. We investigated the requirements for a mini-EH in 3DMS. In particular, we pointed out that autonomous piloting is necessary. To support autonomous piloting, we proposed that EH position information be obtained using the GPS and a low-power wireless transmission system between the mini-EH and the corresponding EV. We call this system the EH positioning system. Based on real-world experiments, we showed that a prototype of the EH positioning system has adequate capabilities in terms of transmission range and energy consumption.

Future works include resolving the EV positioning problem in a large disaster area, autonomous creation of EH flight routes based on the position of a partner EV, the assignment of surveillance work among EVs and their partner EHs, performance evaluation of surveillance data collection, and the development of a prototype of the autonomous piloting system for 3DMS in order to demonstrate the feasibility of the 3DMS on the basis of real-world experiments and performance evaluations.

ACKNOWLEDGMENT

This work was supported by JSPS Grants-in-Aid for Scientific Research (KAKENHI) Grant Number 24246068.

REFERENCES

[1] K. Mase, "Communication Service Continuity under a Large-Scale Disaster-A Practice in the Case of the Great East Japan Earthquake," IEEE ICC 2012 Workshop, June 2012.

[2] D. Kingston, R. Beard, T. McLain, M. Larsen, and W. Ren, "Autonomous Vehicle Technologies for Small Fixed Wing UAVs," In AIAA 2nd Unmanned Unlimited Systems, Technologies, and Operations-Aerospace, Land, and Sea Conference and Workshop & Exhibit, San Diego, CA. Paper no. AIAA-2003-6559, pp. 1-10, 2003.

[3] R. Beard, D. Kingston, M. Quigley, D. Snyder, R. Christiansen, W. Johnson, T. McLain, and M. A. Goodrich, "Autonomous Vehicle Technologies for Small Fixed-Wing UAVs," AIAA Journal of Aerospace Computing, Information, and Communication, Vol. 2, pp. 92-108 January 2005.

[4] T. J. Koo, D. H. Shim, O. Shakernia, B. Sinopoli, Y. Ma, F. Hoffman, and S. Sastry, "Hierarchical Hybrid System Design on Berkeley UAV," International Aerial Robotics Competition, 1998.

[5] F. N. Webber and R. E. Hiromoto, "Assessing the Communication Issues Involved in Implementing High-Level Behaviors in Unmanned Aerial Vehicles," IEEE Military Communication Conference (MILCOM), pp. 1-7, 2006.

[6] Z. Han, A. L. Swindlehurst, and K. J. R. Liu, "Smart Deployment/Movement of Unmanned Air Vehicle to Improve Connectivity in MANET," In Proceedings of WCNC 2006, Vol. 1, pp. 252-257, 2006.

[7] P. Zhan, K. Yu, and A. L. Swindlehurst, "Wireless Relay Communications with Unmanned Aerial Vehicles: Performance and Optimization," IEEE Trans. Aerosp. Electron. Syst., Vol. 47, No. 3, pp. 2068-2085, 2011.

[8] S. Gil, M. Schwager, B. Julian, and D. Rus, "Optimizing Communication in Air-Ground robot networks Using Decentralized Control," pp. 1964-1971, 2010.

[9] O. Burdakov, P. Doherty, K. Holmberg, J. Kvarnstrom, and P-M. Olsson, "Positioning Unmanned Aerial Vehicles as Communication Relays for Surveillance Tasks," In Proceedings of the 5th Robotics: Science and Systems Conference (RSS), 2009.

[10] C. Dixon and E. W. Frew, "Optimizing Cascaded Chains of Unmanned Aircraft Acting as Communication Relays," IEEE Journal on Selected Areas in Communications, Vol. 30, No. 5, pp. 883-898, June 2012.

[11] G. M. Saggiani and B. Teodorani, "Rotary wing UAV potential applications: An Analytical Study through a Matrix Method," Aircraft Engineering and Aerospace Technology, Vol. 76, No. 1, pp. 6-14, 2004.

[12] D. W. Casbeer, D. B. Kingston, R. W. Beard, T. W. McLain, S. Li, and R. Mehra, "Cooperative Forest Fire Surveillance Using a Team of Small Unmanned Air Vehicles," International Journal of Systems Science, Vol. 37, No. 6, pp. 351-360, 2005.

[13] K. Alexis, G. Nikolakopoulos, A. Tzes, and L. Dritsas, "Coordination of Helicopter UAVs for Aerial Forest-Fire Surveillance," Applications of Intelligent Control to Engineering Systems, pp. 169-193, 2009.

[14] R. W. Beard, T. W. McLain, D. B. Nelson, D. Kingston, and D. Johanson, "Decentralized Cooperative Aerial Surveillance Using Fixed-Wing Miniature UAVs," Proceedings of the IEEE, vol. 94, No. 7, pp. 1306-1324, July 2006.

[15] K. Mase, "Information and Communication Technology and Electric Vehicles – Paving the Way towards a Smart Community," IEICE Trans. Commun., Vol. E95-B, No. 6, pp. 1902-1910, June 2012.

[16] J. Hoebeke, I. Moerman, B. Dhoedt, and P. Demeester, "An Overview of Mobile Ad Hoc Networks: Applications and Challenges," Journal of the Communications Network, Vol. 3, pp. 60-66, 2004.

[17] H. Hartenstein and K. P. Laberteaux, "A Tutorial Survey on Vehicular Ad Hoc Networks," IEEE Communications Magazine, Vol. 46, No. 6, pp. 164-171, 2008.

[18] K. Mase, "How to Deliver Your Message from/to Disaster Area," IEEE Communications Magazine, Vol. 50, No. 1, pp. 52-57, 2011.

[19] K. Mase and J. Gao, "Electric Vehicle-based Ad-hoc Networking for Large-Scale Disasters- Design Principles and Prototype Development," The 5th Ad Hoc, Sensor and P2P Networks Workshop (AHSP 2013), March 2013. (To be presented).

[20] H. Oka, H. Okada, and K. Mase, "Experimental Evaluation of SKYMESH Using Terrestrial Nodes," Proceedings of 16th Asia-Pacific Conference on Communications, 2010.

[21] W. Zhao, M. Ammar, and E. Zegura, "A Message Ferrying Approach for Data Delivery in Sparse Mobile Ad Hoc Networks," Proceedings of the 5th ACM International Symposium on Mobile Ad Hoc Networking and Computing, pp. 187-198, 2004.

Resilience of a Network Service System: Its Definition and Measurement

Junwei Wang and Kazuo Furuta
Department of Systems Innovation, the University of Tokyo
Tokyo, Japan
juw623@mail.usask.ca, furuta@sys.t.u-tokyo.ac.jp

Wenjun Zhang
Department of Mechanical Engineering, University of Saskatchewan
Saskatoon, Canada
Chris.Zhang@Usask.Ca

***Abstract*— Modern human society relies on different critical service systems. One important feature of these systems is that they work in a network manner. Thus, they could also be called networked service systems. Resilience is a necessary property of systems; in particular, the networked service systems should meet customers' demands facing various uncertainties. However, the understanding of resilience, especially in the context of service system, is still not very clear; the concept of resilience was confused with other similar safety concepts, such as reliability and robustness in the current literature. In this paper, we present a definition of resilience in the context of networked service system. Furthermore, some criteria of resilience measurement following this definition are also proposed. Two particular measurement models focusing on the different measurement criteria are presented with two simple examples to show how the proposed models work.**

***Keywords-resilience; measurement; networked service system; rebalance*.**

I. Introduction

Modern human society relies on different critical service systems, such as transportation system, power grid system, communication system and so on. One important feature of these system is that they work in a networked manner. Thus, they could also be called networked service systems. The networked service systems are complex socio-tech systems. It is unfortunate that today's networked service systems face serious safety issues. Here are two very recent examples. The breakdown of the power supply system in the north India happened in July 2012 led one-half of the country into serious trouble, which further caused the failure of other critical service systems, such as transportation system, financial system, water supply system, and hospital system [1]. Hurricane Sandy in late October 2012 killed at least 199 people in seven countries and half of New York lost the functions of almost all the service systems [2]. To address such safety challenges, the concept of resilience was introduced into safety engineering and used to describe the system's safety property [3].

From the safety engineering perspective, the resilience concept has been investigated by different researchers [3]-[6]; particularly, there are three categories of understanding of resilience. The first category mixed up resilience with other safety concepts, such as reliability and robustness [7], [8]. The second category focused on the system's recovery ability from partially damage [5]. The third category viewed resilience as in intrinsic ability to adjust its functioning prior to, during, or following changes and disturbances under both expected and unexpected conditions [12]. The three categories of definitions can not well reflect the features of the service systems. Therefore, the objectives of this paper are to (1) clarify the concept of resilience by giving a definition of resilience in the context of networked service systems, and (2) propose resilience measures based on the definition.

The organization of the remainder of this paper is as follows. Section II proposes a new definition of resilience concept. Section III presents the criteria of the resilience measurement and two particular measurement models, as well as simple examples. Section IV discusses the conclusions and future work.

II. Resilience Definition for Service Systems

A. *Resilience Definitions in the Literature*

Resilience is a popular term in material science [9], medicine [10] and ecology [11]. This concept was introduced into engineering field to understand safety as the ability to succeed under varying conditions [3], [12]. Two typical definitions of resilience from engineering perspective are given here. Zhang defined resilience as the ability of a system to recover to meet the demand from a partial damage [5]. Another well-known definition viewed resilience as the intrinsic ability of a system to adjust its functioning prior to, during, or following changes and disturbances so that the system can sustain required operations under both expected and unexpected conditions [12]. This definition actually covers the traditional concepts of safety, reliability and robustness, and makes further extension; in particular, resilience concept deals with the damaged situation – a point of resilience stressed by Zhang [5]. However, these definitions and understandings do not well describe the safety features of the networked service systems.

B. *Definition of Service System*

To facilitate the further discussion, the concepts of service and service system we proposed elsewhere [13] are given below.

"A service is a function that is achieved by an interaction between a human and an entity under a protocol [13]."

"A service system consists of three subsystems: (i) an infrastructure, (ii) a substance, and (iii) a management to directly meet demands of humans who are defined as consumers. The infrastructure is of network, and substance “flows” over the infrastructure. The management plays the roles such as coordinating, leading, planning and controlling, which are applied to both the infrastructure and substance systems [13]."

The substance in a service system may refer to: material, human or animal, energy, data or signal [13]. It is noted that a service system, in this paper, is also called a networked service system, as it has networked feature as mentioned in Introduction.

Following the definitions above, we may have two important understandings: (1) a service system has different functions to meet human's different demands; in particular, a service system with only one function could be viewed as a special case, and (2) relationships between the multiple demands and the multiple supplies determine the safety performance of a service system; furthermore, the relationships are dynamic.

C. Definitions of Resilience

From the perspective of functions, the output of a service system are different functions, which are defined as supply functions, which are represented as $F_i^S, 1 \le i \le m$ where *m* is the number of supply functions. The demands from the customers are also expressed as functions, which are defined as demand functions, which are represented as $F_i^D, 1 \le i \le m$. The relationships between the supply functions and demand functions, balance and imbalance, are defined as follows. Balance between the supply functions and the demand functions is defined as the relationship between the supply functions and demand functions which satisfies the conditions: $F_i^S = F_i^D, 1 \le i \le m$. **Imbalance** between the supply functions and the demand functions is defined as the relationship between the supply functions and demand functions which satisfies the conditions: $\exists i, F_i^S < F_i^D, 1 \le i \le m$. The situations of balance and imbalance may be transferable; a service system may be rebalanced. **Rebalance** is defined a process that a service system transfers from an imbalance situation to a balance situation. Thus, **safety** of a service system is defined as the dynamic balance between the multiple supply functions and the multiple demand functions. Such a definition of safety is quite different from traditional definition of safety. The traditional understanding of safety implies that if a system is not safe, there must be something wrong. However, the proposed definition above indicates that even there is nothing wrong in the system, it may be still not safe, as it may not meet the customers' demands. For example, a large athletic meeting leads to the demand of wireless communication in a particular area much larger than the supply provided by the regular wireless communication system. There are no damages in such a situation; however, the wireless communication system is not safe, as it could not meet the customers' demand.

The definition of safety above has shown that an unsafe service system is not necessarily damaged. Thus, the resilience definitions discussed in Section II do not well reflect the features of a service system, as they are related with partial damage situations. Our definition is given below.

The **resilience** of a service system is defined as a property that allows the system rebalance the supply functions and demand functions from imbalance situations. Four remarks are given below for further explanation on this definition.

Remark 1: According to the definition of safety above, imbalance situation means unsafe situation.

Remark 2: A service could be examined through different perspectives, such as state, structure, functions and so on. The proposed definition of resilience implies that the resilience property should be examined from the perspective of function.

Remark 3: Resilience does not aim at returning to the original states, or structure or functions of the system; it aims at making the supply functions meet the demand functions.

Remark 4: The imbalance situation implies that the supply functions are less than the demand functions. Therefore, the key ability of resilience is to respond to the imbalance situation and to improve the supply functions to meet the demand functions.

The proposed definition of resilience is different from other three categories of definitions of resilience introduced in Section I. According to the new definition of safety in this paper, category I is related to the balance situation of a service system. Category II could be viewed as a special case of the new definition, as it focuses on the imbalance situations of partially damage. Category III is an all-inclusive definition and covers the scope of the new definition.

III. Resilience Measurement for Service System

A. Criteria of Resilience Measurement

Following the proposed definition in Section II, the resilience can be measured through the maximization of imbalance situation, which can be rebalanced with the bounded time and cost (or resources). Four important corollaries could be derived from the definition as the criteria of resilience measurement.

Criterion 1: The resilience of a system is only measured in terms of particular imbalance situations.

Criterion 2: A system which can rebalance the supply functions and demand functions from larger imbalance situation is more resilient.

Criterion 3: A system which can rebalance the supply functions and demand functions from the imbalance situation with less time is more resilient.

Criterion 4: A system which can rebalance the supply functions and demand functions from the imbalance situation with less cost (or resources) is more resilient.

Based on these measurement criteria, it is obvious that there are three important factors affecting resilience performance: (1) rebalance solutions, (2) rebalance time, and

(3) rebalance cost. For a networked service system under a particular imbalance situation, there may be different rebalance solutions and the best solution will determine the resilience performance of the system. The rebalance solutions certainly depend on the available rebalance time and resources. The criteria actually imply that given different conditions, there may be different measurements. For example, given rebalance time and cost, the measurement is maximization of imbalance situation. Given imbalance situation and available rebalance cost, the measurement is minimization of rebalance time. Next, two measures following Criterion 2 and Criterion 3 are expressed respectively.

B. Resilience measurement following Criterion 2

Criterion 2 is the main concern of resilience measurement. Based on Criterion 2, the following model is proposed to measure the resilience of a service system.

1) Variable Definition

m : the total number of functions;

n : the category number of resources;

w_i : weight of function i, $i = 1, 2, \cdots, m$, $\sum_{i=1}^{m} w_i = 1$;

$r_{i,j}^t$: the number of resource j needed by function i at time t, $i = 1, 2, \cdots, m$, $j = 1, 2, \cdots, n$;

R_j : the total number of resource j, $j = 1, 2, \cdots, n$;

$D_i^t(x)$: the demand function i at time t with rebalance solution x, $i = 1, 2, \cdots, m$;

$S_i^t(x)$: the supply function i at time t with rebalance solution x, $i = 1, 2, \cdots, m$;

T_i : the demand time for the rebalance of function i, $i = 1, 2, \cdots, m$;

x : rebalance solution

2) Objective Function and Constraints

$$\text{Max} \ \sum_{i=1}^{m} w_i \frac{D_i^0(x) - S_i^0(x)}{D_i^0(x)} \qquad (1)$$

$$\text{s.t.} \ \ S_i^t(x) \ge D_i^t(x), t \ge T_i \qquad (2)$$

$$\sum_{i=1}^{m} r_{i,j}^t \le R_j, \ j = 1, 2, \cdots, n, t \le \max\left\{T_i \left| i = 1, 2, \cdots, m\right.\right\} \qquad (3)$$

In the above, formula (1) represents the imbalance situation at beginning; formula (2) represents the constraint of rebalance time for different functions; formula (3) represents the resource constraints.

3)A Simple Example of Transportation System: A very simple transportation system example is given to show how the model work. In this example, we only consider the maximum imbalance situation that the system can rebalance and the rebalance time and cost are ignored. Fig. 1 is an original transportation system with only two nodes. Two edges link the two nodes. The travel time and edge capacity are as shown in the figure. Suppose that the unit of travel time is minute. In this example, we consider the imbalance situation is that the transportation demand from A to B is increased to 15 per 2 minutes due to some reason. The rebalance solution is contraflow approach, namely reverse and edge from B to A, as shown in Fig. 2.

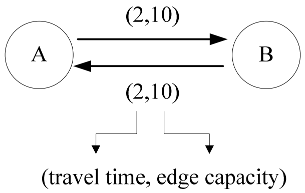


Figure 1. Original transportation system with only two nodes.

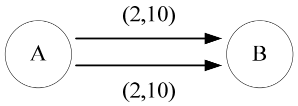


Figure 2. Rebalance solution.

With the rebalance solution in Fig.2, the maximum transportation ability from A to B is 20 per minutes. Therefore, the largest imbalance situation that can be rebalanced is that the transportation demand from A to B increased to 20 and the imbalance degree is (20-10)/20=50%. Thus, the resilience of this transportation system facing the particular imbalance situation of increased transportation demand from A to B is 50%.

C. Resilience measurement following Criterion 3

Criterion 3 implies that given the same imbalance situation that could be rebalanced, the minimization of the rebalance time could be used to measure the resilience of a service system.

1) Variable Definition:

The variable definition is given below.

m : the total number of functions;

n : the category number of resources;

w_i : weight of function i, $i = 1, 2, \cdots, m$, $\sum_{i=1}^{m} w_i = 1$;

$r_{i,j}^t$: the number of resource j needed by function i at time t, $i = 1, 2, \cdots, m$, $j = 1, 2, \cdots, n$;

R_j : the total number of resource j, $j = 1, 2, \cdots, n$;

$c_i(x)$: the completion time for the rebalance of function i, $i = 1, 2, \cdots, m$;

T_i : the demand time for the rebalance of function i, $i = 1, 2, \cdots, m$;

2) Objective Function and Constraints

The objective function and constraints are given below.

$$\text{Max} \ \sum_{i=1}^{m} w_i \frac{c_i(x)}{T_i} \qquad (4)$$

$$\text{s.t.} \ \ \sum_{i=1}^{m} r_{i,j}^t \le R_j, \ j = 1, 2, \cdots, n, t \le \max\left\{T_i \left| i = 1, 2, \cdots, m\right.\right\} \qquad (5)$$

$$c_{(i)}(x) \le T_i, i = 1, 2, \cdots, m \qquad (6)$$

In the above, formula (4) represents the resilience of the system; formula (5) represents the resource constraints; formula (6) represents the constraint of rebalance time.

3)An Example of Enterprise Information System: A simple example of enterprise information system is employed to show how the model works. An enterprise information system is a very special service system in that such a system usually has backup and could be rebalanced from even 100% lost of functions. We consider a scenario that an enterprise information system is fully damaged. There are two functions in this system, which are totally lost. There are two categories of resources. The total number of resource 1 is 2; the total number of resource 2 is 4. All the functions are treated with the same importance and the weight for each function is 0.5. Due to the limitation of resources, the two functions can not be rebalanced synchronously. This implies that a rebalance solution needs to choose a particular order among a set of rebalance tasks.

TABLE I. REBALANCE PARAMETERS

Function	Resource 1 $r_{i,1}$	Resource 2 $r_{i,2}$	p_i (min)	T_i (min)
i=1	2	3	2	2
i=2	2	4	5	8

The information of rebalance solutions is given in Table 1. p_i is the process time for the recovery of function *i*. Obviously, the optimum solution for this problem is with the order of [1,2]. The fitness value is 0.857, which implies that the system could be rebalanced a little earlier than the demand recovery time.

IV. CONCLUSION AND FUTURE WORK

This paper proposes a new definition of resilience for the networked service system by considering its important safety features. Furthermore, the measurement criteria of resilience are discussed. Following these criteria, two particular measures are presented with two simple examples to show how the measures work.

This work describes only very preliminary results with the two models. As future work, realistic examples of networked service systems will be adopted to validate the proposed measures.

REFERENCES

[1] India today online, http://indiatoday.intoday.in/story/massive-power-outage-hits-delhi-again-halts-metro-services/1/211127.html. [retrived: Jan 2013].

[2] The New York Times, http://www.nytimes.com/2012/10/31/us/hurricane-sandy-barrels-region-leaving-battered-path.html?pagewanted=all&_r=0 [retrived: Jan 2013].

[3] E. Hollnagel, D. D. Woods, and N. Leveson, Resilience engineering concepts and precepts. Burlington, VT: Ashgate, 2006.

[4] E. S. Patternson, D. D. Woods, R.I. Cook, and M.L. Render, "Collaborative cross-checking to enhance resilience," Cogn. Tech. Work, vol. 9, Aug. 2007, pp. 155-162, doi 10.1007/s10111-006-0054-8.

[5] W. J. Zhang, "Resilience engineering - a new paradigm and technology for systems?" Presentation at the Hong Kong Polytechnic University, 2008. http://homepage.usask.ca/~wjz485/PPT%20download/Resilience%20engineering%20-%20HKPolyU%202008.ppt [retrived: Jan 2013].

[6] T. Kanno, T. Fujii, R. Watari, and K. Furuta, "Modeling and Simulation of a Service System in a Disaster to Assess Its Resilience," Proc. 4th. Symp. Resilience Engineering, Presses des mines, Jun. 2011, pp. 128-134.

[7] Victoria Transport Policy Institute, http://www.vtpi.org/tdm/tdm88.htm [retrived: Jan 2013].

[8] J. Bongard, V. Zykov, and H. Lipson, "Resilient Machines Through Continuous Self-Modeling," Science, vol. 314, pp. 1118-1121, doi: 10.1126/science.1133687.

[9] K. Nagdi, "Rubber As Engineering Material: Guideline For Users," Munich : Hanser Publisher, 1993.

[10] S. S. Luthar, D. Cicchetti, and B. Becker, "The construct of resilience: A critical evaluation and guidelines for future work," Child Development, vol.71, May/Jun. 2000, pp. 543-562, doi: 10.1111/1467-8624.00164.

[11] L. H. Gunderson, "Ecological Resilience- in Theory and Application," Annual review of ecological and systematics, vol. 31, 2000, pp. 425-439.

[12] E. Hollnagel, J. Paries, D. D. Woods, and J. Wreathall, Resilience engineering in practice: a guide book. Burlington, VT: Ashgate,, 2010.

[13] J. W. Wang, H. F. Wang, W. J. Zhang, W.H. IP, and K. Furuta, "On a Unified Definition of the Service System: What is its Identity?" IEEE Systems Journal, in press.

Survey on Survivable Virtual Network Embedding Problem and Solutions

Sandra Herker, Ashiq Khan, Xueli An
DOCOMO Communications Laboratories Europe GmbH
Munich, Germany
{herker, khan, an_de_luca}@docomolab-euro.com

Abstract—**Survivability in networks has always been an important issue and lately becomes for network virtualization. Network virtualization provides to run multiple virtual networks on a shared physical network. Since a failure in the physical network can affect several virtual resources, therefore, the survivability has to be considered in the embedding of the virtual resources. In this paper, we present a survey on the survivable virtual network embedding problem and different approaches to solve this problem. The different approaches and algorithms are evaluated on their type of survivability.**

Keywords-survivability; virtualization; virtual network embedding; embedding algorithms

I. Introduction

Network virtualization is receiving more and more attention lately. It is the sharing of physical resources by subdividing a physical node or link into many virtual nodes or virtual links. Network virtualization is a technology which allows a service specific (virtual) network to be embedded onto a substrate network in a dynamic way. Using end-to-end virtualization it will be possible to create various service specific networks within one operator's network. The network can be tailored to the specific needs of a service with respect to topology, routing or QoS.

Multiple configurations of the virtual networks maybe created over the same physical setup. Some configurations may be more efficient than others in terms of different requirements such as, optimal use of physical resources, maximizing the revenue and/or minimizing the power consumption. The calculation of the effective allocation of the physical resources among the virtual network requests is known as the virtual network (VN) embedding problem. Since multiple virtual networks can share the physical resources of the underlying substrate, even a single failure in the substrate can affect a large number of VNs and the services they offer. Thus, the problem of efficiently mapping a VN to a substrate while guaranteeing the VNs survivability in the event of failures in the substrate becomes important. Many different basic solutions for embedding VNs are existing [1][2][3][4], however, the survivability issue in the VN embedding is not considered in these works. These algorithms are assuming that the substrate network after the embedding is operational at all times and ignoring the possibility of substrate link/node failures.

Link failure survivability problems and survivable routing have already been investigated for optical [5] and multi-protocol label switched (MPLS) networks [6]. However, the problems studied there are an offline version or assume the traffic demand matrix has been available in advance which is not the case in virtual network embedding.

In this paper, the focus is on survivable Virtual Network Embedding problem and solutions. The remainder of this paper is organized as follows. We first describe general and survivable Virtual Network Embedding problem in Section II. In Section III recent algorithms for solving the survivable Virtual Network Embedding problem are evaluated. Section IV and V gives a discussion on the algorithms and a conclusion.

II. The Survivable Virtual Network Embedding Problem

A. The Virtual Network Embedding Problem

The virtual network (VN) embedding problem deals with finding a mapping of a virtual network request onto the substrate network/physical network. When an operator wants a Virtual Network (VN) to offer a specific service to his customers and he sends a VN request to a Virtual Network Provider (VNP). The VNP requests resources which meet the requirements of the VN request from the Physical Infrastructure Provider (PIP), who owns the substrate network/physical network, for the VN creation.

The substrate network/physical network is presented as a graph $G^S = (N^S, E^S)$ where vertices N^S represent the substrate nodes and edges E^S represent the links between nodes in the network. Both substrate nodes and links have constraints. Node constraints can be CPU, RAM, geographical location, etc. Link constraints can be bandwidth, delay, etc.

The virtual network request consists of virtual nodes and virtual links, which is also described by a graph $G^V = (N^V, E^V)$ with constraints that describes the requirements of the virtual nodes and links. The mapping of virtual nodes and links onto the substrate network is realized by an embedding algorithm.

The objective of the VN embedding is to find an effective and efficient embedding algorithm for the VN request. Embedding has been proven to belong to the NP-hard category of problems in [1][7]. Three approaches are commonly used to solve a heuristic for the embedding problem: backtracking [4], simulated annealing [8] and approximation algorithms [9].

The VN embedding problem can be divided into two separate problems:

ISBN: 978-1-61208-256-1

a) Node mapping:

$$N^V \mapsto N^S \tag{1}$$

One virtual node needs to be mapped to exactly one substrate node, which satisfies the resource requirements of the virtual node (equation (1)). The node mapping problem is still a NP-hard problem, similar to the multi-way separator problem [1][7]. For node mapping, greedy methods [1][2] are often used.

b) Link mapping:

$$E^V \mapsto P^S \tag{2}$$

P^S is denoted as the set of all loop-free paths of substrate network. A virtual link between two virtual nodes can be mapped on a substrate path, which could consists one or multiple substrate links (equation (2)). For this problem, (k-) shortest path [2] or multi-commodity flow algorithms [10] are used.

B. The Survivable Virtual Network Embedding

1) Types and characteristics of failures: Survivable virtual network embedding deals with failures in the substrate and virtual network. The challenges to be considered are link and node failures, which have to be backed up before the failure or recovered after failure. Failures can occur at different layers in the network. For example at the physical layer, a fiber cut may cause a physical dis-connectivity. In [11], it is shown that 20 % of all failures in an IP backbone are resulting from maintenance activities. About 53 % of the unplanned link failures are due to router-related [11]. In a network, single and also multiple failures can occur. The single failure case happens more often than multiple simultaneous failures. The study [11] states that about 70 % of the unplanned link failures are single link failures. A study [12] about network-related failures in data centers found out that link failures happen about ten times more than node failures per day. Usually node failures are due to maintenance [12].

2) Survivable failure methods: There are two main survivability methods: protection and restoration [5]. Failure protection is done in a proactive way to reserve the backup resources before any failure happens. Reactive mechanisms, which are called restoration mechanisms, react after the failure occurs and start the backup restoring mechanism. However, some data loss is possible in the reactive case. There exist two kinds of backups for the protection scheme: dedicated backup or shared backup. In shared backup, the resources for the backup may be shared with other backups. In the dedicated case the backup resources are not shared for other backups.

Failures in the virtual network can be repaired through re-instantiation of the failed virtual network element (link or node) on the same substrate elements or some other suitable substrate elements. Failures in the substrate network require more effort to be restored or backed up, since sharing the substrate can affect several virtual resources. For substrate node failure, the virtual node or nodes has to be migrated to some other substrate nodes. For substrate link failure, a backup path over different substrate links has to be found, which can be done with a link or path based method. Link based methods means that each primary link is backed up by a pre-configured bypass path. In the path based methods, each end-to-end primary path is backed up by a disjoint path from the source node to the destination node.

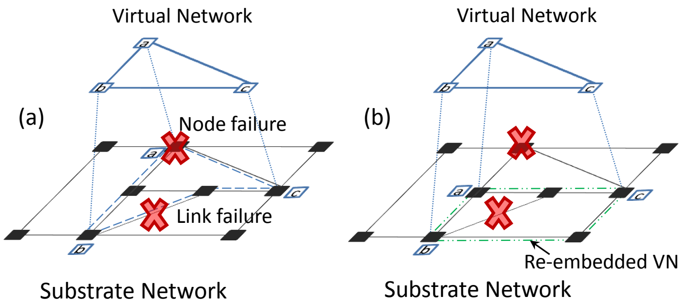


Figure 1. Survivable virtual network (VN) embedding

The task is to embed a virtual network that can deal with virtual and substrate network failures in a way, that after the failure, the virtual network is still operating. The failure and the fixing/recovery should be transparent to the users of the virtual network.

One possibility can be to extend the virtual network graph with backup nodes N_B and backup links E_B (equations (3)) and embed the extended graph G_B^V. The backup links E_B are links between backup nodes and working nodes.

$$G_B^V = (N^V \cup N_B, E^V \cup E_B) \tag{3}$$

In the survivable mapping, virtual nodes of one virtual network should not be mapped on the same substrate node. Due to the fact, that a possible failure of this substrate node could affect several virtual nodes. For links, different virtual links should use distinct paths in the substrate network.

Figure 1 (a) shows a mapping (dashed lines) of a virtual network (upper graph) onto a substrate network (lower graph). After embedding, a substrate node and link failure (represented by crosses) occur. The failed node has mapped the virtual node *a*, which need to be remapped. The substrate link failure is on the substrate path for the virtual nodes *b* and *c*. A possible re-embedding of the virtual network on the substrate after the failure is drawn in Figure 1 (b), where virtual node is migrated to a new substrate node and the links are re-embedded for the migrated node and the failed substrate link.

III. Algorithms for the Survivable Virtual Network Embedding

This section discuss existing algorithms and methods for survivable/resilient virtual network embedding for link or node failures.

A. Survivable VN Embedding against Link Failures

The following algorithms embed VN against links failures in the substrate network.

1) Link restoration and protection methods: In [13], a reactive backup mechanism to protect against a single substrate link failure for VN embedding is proposed. The idea is a fast rerouting of the links and to reserve bandwidth for backups

ISBN: 978-1-61208-256-1

on each physical link. The polynomial time heuristic consists of three parts. Before any VN request arrives, backup paths for each substrate link are calculated with a path selection algorithm. Then node and link embedding is done for the arriving request with an existing embedding algorithm. When a substrate link failure occurs, the calculated backup paths are used to reroute the bandwidth of the affected link using their reactive online optimization mechanism. The optimization goal is to maximize revenue for the PIP. This backup mechanism is a restore approach, therefore after a failure it cannot guarantee 100 % recovery. In cases that the bandwidth resources are used for new VN requests, there may be not enough resources left for the recovery. With increase in traffic load, a failure can cause a big amount of data loss and the backup mechanism may not restore the VN.

Authors in [14] also investigate the problem of shared backup network provision for a single substrate link failure for VN embedding. In their solution, a link based backup approach is used to protect against the link failure similar to [13]. Two schemes are proposed: In Shared On-Demand approach, bandwidth resources are allocated to the primary flows and to restoration/backup flows when a new VN request arrives. Bandwidth sharing is possible for the restoration flows, however, not for the primary flows. After every VN embedding, the residual resource information needs to be updated. In Shared Pre-Allocation approach, backup bandwidth for each substrate link is pre-allocated during the configuration phase before any VN request arrives. Since the bandwidth pre-allocation only needs to be done once and not for every VN request, there is less computing done during the VN embedding phase. The overall optimization is to maximize the revenue for the Infrastructure Provider through accepting most VN requests. Advantage to the previous algorithm [13] is that the backup bandwidth is already allocated before the failure happens and not after the failure. Disadvantage of the Shared Pre-Allocation approach is that backup bandwidth is reserved independent of the VN requests and may not be used at any time if few VN requests arrive.

2) Path protection methods with node migration: In [15], the problem of survivability for link failure is tried to solve with optimizing the networking and computing resources to tolerate link failures through a node migration technique. Instead of backing up the each primary link like in [13] and [14], each end-to-end primary substrate path is protected by a backup path. Their approach, migratory shared protection, migrates and maps a VN node to another substrate node to increase the resource efficiency when a failure occurs. The relocated node should need less backup path length to the destination node than before the migration and save resources. All VN links connected with the migrated VN node have to be remapped, and the backup links must be link-disjoint to the primary links. The re-established paths from the new migrated node form a tree: the migratory backup tree. The survivable mapping solution with migratory protection includes: an one-to-one node mapping from the VN nodes to the substrate nodes, a mapping of each VN link to a primary path from the original source node to the original destination node and a mapping of each VN link to a link-disjoint backup path or migratory backup tree. For this protection method, intra-share can be applied, that means sharing resource among the migratory backup tree and the corresponding migrated primary paths. Also inter-share is possible that means sharing of backup resources between different backup paths. Migratory shared backup tree is only calculated to improve performance of the traditional backup protection or if no traditional link-disjoint backup path can be found. The optimization goal is to minimize the sum of the computing and bandwidth resource. However, the cost of less bandwidth resources cannot be compared to the cost of a node migration, since node migration costs are considered higher. Compared to the traditional backup protection where only one path needs to be migrated in their approach several links and at least one node need to be migrated.

3) Path protection methods with QoS: A mechanism, named QoSMap, attempting to consider both quality of service (QoS) and resiliency in constructing VNs over a substrate network is presented in [16]. Its aim is to map a QoS-specified overlay onto the substrate network using direct paths between nodes that are pre-selected possible candidates. Nodes with higher quality are selected first. Node quality depends on the average backup paths that a substrate node can provide. Path resiliency is provided by constructing alternate backup paths via one intermediary node that could be additional underlying nodes or selected hosting overlay nodes. However, the substrate topology is not considered when selecting backup paths. It could be possible that disjoint overlay paths share common substrate links or nodes. There is also high degree of overlap for working and backup paths in the mapped solution. They might fail together if they share common point of failures. It may not always be possible to find direct backup paths. Since QoSMap uses direct paths, back-tracking in the algorithm is required to find these (backup) paths. This may take exponential time and affect scalability of the algorithm.

The authors in [17] formulate and solve the previous QoS and resilience mapping problem [16] with an Integer Linear Program (ILP). Since the heuristic QoSMap solution [16] cannot guarantee the best QoS performance, due to sequential and heuristically node selection, a mathematical formulation is used to achieve a optimal solution. A simplified topology, that contains the candidate nodes connected for the mapping of the request, is constructed from the substrate network. This logical topology enables the mapping of the request with reduce in computational complexity. In the logical topology, the links between the candidate nodes are calculated using the shortest path first routing and considering the overlay delay requirements. The object to optimize is to minimize the delay and the number of additional substrate nodes used for backup path mapping for the overlay links. The ILP considers the substrate topology and assure working and backup paths avoiding link overlaps in the substrate network. Therefore, multiple overlay link failures caused by a single substrate link failure should be reduced.

In [18] a heuristic is developed for the previous ILP [17]. This heuristic improves the QoSMap heuristic [16] by considering the substrate topology information in the mapping procedure.

B. Survivable VN Embedding against Node Failures

The following different approaches try to embed VNs with backup for virtual nodes and protections against node failures in the substrate network.

1) Two-step approaches: In [19], a two-step paradigm to fully recover a VN from facility node failures is presented. The first step is to construct a graph of the VN request with backup virtual nodes and links, and then this enhanced VN request has to be mapped onto the substrate network. Two solutions are proposed: the 1-redundant scheme and the K-redundant scheme. A 1-redundant solution is a reliable VN graph with one redundant virtual node (backup node) and redundant connections, which is then mapped onto the substrate network. Assuming only single failure, the backup node of a certain virtual node can also be used as backup of some other virtual node for resource sharing. For the mapping, it can share the physical link resources when mapping them onto the substrate network (backup share) and also share the bandwidth link resources between the original working path and its associated backup path (cross share). In the K-redundant solution, a K-redundant reliable VN graph is designed, in which each critical node is permitted to have a corresponding backup node. The optimization objective is to minimize network resource costs. However, this approach may fail to provide a joint optimization for the allocation of both the active and backup resources. In worst case, there need to reserve a backup node for every critical node and links to every neighbor node.

Another two-step method is presented in [20] for surviving single facility node failures. This approach designs the enhanced VN with a failure-dependent strategy, instead of a failure-independent strategy like in the previous one [19]. It manages to further reduce the needed virtual resources and, therefore, less allocated backup resources compared to failure-independent strategy. The idea is that, when node *i* fails, the role of node *i* may be replaced by any other nodes after a rearrangement of all the nodes (including the backup node(s)) using graph transformation/decomposition and bipartite graph matching. The disadvantage of this approach is that the large amount of possible migrations of working nodes after a failure makes the approach less applicable in large networks.

2) Node protection for regional failures: In [21], an approach for solving the problem of survivable VN mapping for single regional failures in a federated computing and networking system is presented. In a federated computing and networking system, facility nodes from a data center are interconnected. These facility nodes need to be backed up to achieve a survivable VN mapping. Their approach is based on the assumption, that the number of distinct regional failures is finite in a specified geographical area and that a regional failure refers to a set of substrate nodes and links, which is in the same shared risk group. The proposed approach first solves the non-survivable VN mapping problem with a heuristic and extends this heuristic to handle the survivable VN mapping problem. Two failure dependent survivable VN mapping algorithms are developed. The Separate Optimization with Unconstrained Mapping (SOUM) decompose problem into separate non survivable problems for each possible regional failure plus one for the initial working mapping of the VN request. Each problem is mapped in a way that the costs of the used resources are minimized. The other approach, Incremental Optimization with Constrained Mapping, maps first the initial working mapping and then handles each regional failure after another. Compared to the SOUM, the additional computing and networking resources, that are needed to handle the failure, are tired to minimize. With this strategy, the mapping of unaffected virtual nodes is not changed. The disadvantage of SOUM is the re-calculating virtual mapping of unaffected nodes, which results in more costs and more time to be calculated.

3) Node protection with location-constraint: The Location-constrained Survivable Network Embedding (LSNE) problem to protect against any single facility node failure is investigated in [22]. The location constraint of a virtual node is considered for its backup node. The goal is to map the VN with minimum resources while satisfying the bandwidth constraints for the links and capacity constraints for the nodes including meeting the location constraints for the primary and protection node. The idea is to construct a graph with the virtual and substrate graph in one graph. Thereby, each virtual node is connected to some candidate substrate nodes, which satisfy the location and capacity constraints. This problem is formulated as an ILP framework and for large scale a heuristic algorithm is developed. The heuristic algorithm (sequential survivable embedding algorithm) is based on the decomposition of the LSNE problem. First the VN request is mapped with an existing embedding algorithm and then the backup request is mapped.

4) Backup node sharing with reliability: Authors in [23] tried to recovery from both substrate node and link failures while minimizing backup resources through pooling. Further a relationship between reliability and the amount of redundant resources is tried to be found. Redundant (backup) virtual servers are created dynamically and are pooled together to be shared between VNs to assure the requested reliability level. The higher the reliability level, the higher number of backup nodes needed. It is possible to share the backup nodes such that the total number of backup nodes is lower than each VN separately has their own backup nodes. Every backup node can be a standby node for all other critical nodes. With the Opportunistic Redundancy Pooling (ORP) mechanism, backup nodes can be shared between VNs as long as the reliability of every network is satisfied. The ORP shares these redundancies for both independent and cascading types of failures. Therefore, VNs with different reliability guarantees can be pooled together and it is flexible in adding or removing VNs to the exciting ones.

5) Node protection in data centers: An optimization framework for the survivable virtual infrastructure mapping in virtualized data centers is presented in [24]. Multiple correlated Virtual Machines (VMs) and their backups are grouped together to form a Survivable Virtual Infrastructure (SVI) for a service or a tenant. The aim is to minimize the backup resources (number of active servers and needed bandwidth) while guaranteeing no-disruption no-degradation fail-over. This problem is similar to the VN embedding problem, however, multiple VMs are allowed to be placed on a common server to minimize the number of active servers. An additional goal is to minimize the total reserved bandwidth. This problem of a SVI can be divided in the VM Placement (VMP) and Virtual Link Mapping (VLM) Problem that can be solved separately. For the Virtual Machine Placement subproblem, an efficient heuristic algorithm (back tracking) based on Depth First search is designed. For each VMP solution, the Virtual Link Mapping subproblem is calculated using a Linear Program (LP)-based algorithm (LP-VLM). Further an algorithm to jointly solve the two subproblems at the same time is developed. This joint mapping algorithm determines a server pair for each virtual link and allocates the bandwidth between with a LP and after that solves the LP-VLM. For the VMP problem, quite a large number of possible solutions are calculated, even when it is restricted, and again for all the possible VMP solutions, the VLM must be calculated. This results into high computing overhead for large networks and not guaranteeing to get always closed to the optimum.

C. Distributed Survivable VN Embedding

In all the previous approaches, the survivable VN embedding is done by centralized entity. In [25], an adaptive VN embedding framework is proposed for a distributed survivable VN mapping algorithm, without a centralized controller. The proposed system is distributed and based on *agents*, which monitor physical elements. *Agents* detect failures and change the VN allocation to maintain the constraints of each VN. The fault-tolerant embedding algorithm can handle three resource failures: virtual node, substrate node and link failure. When a virtual node failure is detected by a substrate agent, a new virtual node has to be created on the same substrate node or on another substrate node. When a substrate node fails, alternative nodes have to be found and the affected virtual nodes and links have to be migrated. For link failure, the agent substrate nodes, which are connected to this link, try to find an alternative link or path. The embedding algorithm also monitors the bandwidth in the substrate nodes, therefore, can recognize congestion or overload in the substrate links. When a failure occurs the distributed embedding algorithms works following: If a substrate node agent detects a node failure, it sends a failure notification message to all substrate agents in the same cluster. All agents receiving this message check if they can host the node. Each agent calculates dissimilarity metric to compare their similarity to requested node. The substrate node, which metric is minimal, will be used. The last step is to map the virtual links to the substrate paths between substrate nodes using a distributed shortest-path algorithm [26]. For link failures only the last step need to be done.

IV. Discussion

In summary, Table I presents a comparison of the embedding algorithms presented in this paper.

A. Limitation of Previous Work

The approaches are mostly protection methods for link or node failures, which reserve/backup before any failure happens. Restoration methods like in [13] may need less reserved bandwidth compared to protection methods, however, it cannot provide against a possibility of data loss during the failure. Most works focus on single substrate failure. Types of failures are single link, single facility node and single regional failures in the network. They assume that the network failures are independent from each other and only one failure happens at a time. In [21], a single regional failure which destroys more than one facility node is addressed. Several approaches [15], [16] uses path protection against link failures which could provide bandwidth saving over link protection. However, path protection is more vulnerable to multiple link failures than link protection. Shared protection for the backup links or nodes is also part of some approaches [14], [15], [19] which saves resources over dedicated protection, however, it is more vulnerable to multiple link failures. Also none of them deal with node and link failure occurring at the same time. They only focus on link backup or node backup with the concerned links. However, combining node and link failure for survivability in the network is also important.

The approaches focus on solving the survivable embedding problem in a single PIP environment. At least in [21] a federated computing and networking system is considered.

The main objective for optimization of the presented approaches is maximizing the revenue while minimizing the total cost through minimizing the redundant resources. Each substrate resources like bandwidth or computing resource has a unit cost. The total cost is the sum of all resource costs of the used substrate resources.

B. Open Research Issues

Multiple node or link failures occur at the same time in the network and the correlations between node/link failures are not addressed in any approach. Further work could be done to extend the existing heuristics/algorithms to deal with multiple link or node failures and to combine link and node protection or migration methods.

Survivability in a multi-domain VN environment could have new challenges for inter and inter domain link failures. Multiple simultaneous inter-domain and intra-domain failures could require to develop new mechanism than for single domain environment.

V. Conclusion

This paper presented a survey on existing survivable VN embedding algorithms. Several approaches to solve the problem are examined. The redundancy and the survivability issue in networks has always been an important aspect of network operators and especially for mobile operators. Due

TABLE I
SUMMARY OF THE SURVIVABLE EMBEDDING ALGORITHMS

References	Survivability	Type of failure	Optimization Objective	Survivable failure mechanism
Survivable virtual network embedding [13]	Link	Single substrate link failure	Maximize revenue for Infrastructure Provide	reactive, after failure (Restoration)
Shared backup network provision for virtual network embedding [14]	Link	Single substrate link failure	Maximize revenue/accepting VN requests	proactive, before failure (Protection)
Migration based protection for virtual infrastructure survivability for link failure [15]	Link	Single substrate link failure	Minimize sum of costs	before failure
QoSMap: Achieving Quality and Resilience through Overlay Construction [16]	Link	Single substrate link failure	Minimize delay and additional resources for backup	before failure
An overlay mapping model for achieving enhanced QoS and resilience performance [17] / An overlay mapping model for achieving enhanced QoS and resilience performance [18]	Link	Single substrate link failure	Minimize delay and additional resources for backup	before failure
Survivable virtual infrastructure mapping in a federated computing and networking system under single regional failures [21]	Node	Single regional failure	Minimize sum of cost	before failure
Cost efficient design of survivable virtual infrastructure to recover from facility node failures [19]	Node	Single facility node failure	Minimize sum of cost	before failure
A novel two-step approach to surviving facility failures [20]	Node	Single facility node failure	Minimize resources/total cost	before failure
Location-constrained survivable network virtualization [22]	Node	Single facility node failure	Minimize resources	before failure
Designing and embedding reliable virtual infrastructures [23]	Node	Single substrate node failure	Minimize amount of resources used	before failure
Survivable virtual infrastructure mapping in virtualized data centers [24]	Node	single server failure	Minimize operational cost	before failure
Adaptive virtual network provisioning [25]	Node or Link	single node failure or single link failure	-	after failure

to revenue reduction of the operators, the redundancy has to be optimized against cost. Most operators have very large nationwide networks that fast approximation algorithms for the survivability embedding problem have to be found. The ILP and MILP are limited in scaling and not applicable to larger networks. Therefore, approximations and heuristic algorithms are necessary which can cope with multiple failures in the network. Future directions could be to investigate the survivability VN embedding issue in a multi-domain NV environment or to extend to handle multiple link and node failures at the same time.

REFERENCES

[1] M. Yu, Y. Yi, J. Rexford, and M. Chiang, “Rethinking virtual network embedding: substrate support for path splitting and migration,” *SIGCOMM Comput. Commun. Rev.*, vol. 38, pp. 17–29, March 2008.

[2] Y. Zhu and M. Ammar, “Algorithms for assigning substrate network resources to virtual network components,” in *IEEE INFOCOM 2006. Proceedings*, April 2006, pp. 1 –12.

[3] J. Lu and J. Turner, “Efficient Mapping of Virtual Networks onto a Shared Substrate,” Washington University in St. Louis, Tech. Rep., 2006. [Online]. Available: http://www.arl.wustl.edu/˜jl1/research/tech_report_2006.pdf

[4] J. Lischka and H. Karl, “A virtual network mapping algorithm based on subgraph isomorphism detection,” in *Proceedings of the 1st ACM workshop VISA*, ser. VISA ’09. ACM, 2009, pp. 81–88.

[5] S. Ramamurthy, L. Sahasrabuddhe, and B. Mukherjee, “Survivable wdm mesh networks,” *Journal of Lightwave Technology*, vol. 21, no. 4, pp. 870 – 883, April 2003.

[6] Y. Liu, D. Tipper, and P. Siripongwutikorn, “Approximating optimal spare capacity allocation by successive survivable routing,” *IEEE/ACM Transactions on Networking*, vol. 13, no. 1, pp. 198 – 211, Feb. 2005.

[7] D. G. Andersen, “Theoretical approaches to node assignment,” Dec. 2002, unpublished Manuscript.

[8] R. Ricci, C. Alfeld, and J. Lepreau, “A solver for the network testbed mapping problem,” *SIGCOMM Comput. Commun. Rev.*, vol. 33, pp. 65–81, April 2003.

[9] N. Chowdhury, M. Rahman, and R. Boutaba, “Virtual network embedding with coordinated node and link mapping,” in *INFOCOM 2009, IEEE*, April 2009, pp. 783 –791.

[10] W. Szeto, Y. Iraqi, and R. Boutaba, “A multi-commodity flow based approach to virtual network resource allocation,” in *GLOBECOM ’03. IEEE*, vol. 6, Dec. 2003, pp. 3004–3008.

[11] A. Markopoulou, G. Iannaccone, S. Bhattacharyya, C.-N. Chuah, and C. Diot, “Characterization of failures in an ip backbone,” in *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, vol. 4, march 2004, pp. 2307 – 2317.

[12] P. Gill, N. Jain, and N. Nagappan, “Understanding network failures in data centers: measurement, analysis, and implications,” in *Proceedings of the ACM SIGCOMM 2011 conference*. ACM, 2011, pp. 350–361.

[13] M. R. Rahman, I. Aib, and R. Boutaba, “Survivable virtual network embedding,” in *Proceedings of the 9th IFIP TC 6 international conference on Networking*, ser. NETWORKING’10, 2010.

[14] T. Guo, N. Wang, K. Moessner, and R. Tafazolli, “Shared backup network provision for virtual network embedding,” in *IEEE International Conference on Communications (ICC)*, June 2011, pp. 1 –5.

[15] H. Yu, V. Anand, C. Qiao, and H. Di, “Migration based protection for virtual infrastructure survivability for link failure,” in *Optical Fiber Communication Conference and Exposition (OFC/NFOEC), 2011 and the National Fiber Optic Engineers Conference*, March 2011, pp. 1 –3.

[16] J. Shamsi and M. Brockmeyer, “Qosmap: Achieving quality and resilience through overlay construction,” in *4th International Conference on Internet and Web Applications and Services, ICIW ’09*, May 2009.

[17] X. Zhang, C. Phillips, and X. Chen, “An overlay mapping model for achieving enhanced qos and resilience performance,” in *3rd International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2011*, Oct. 2011, pp. 1 –7.

[18] X. Zhang and C. Phillips, “A novel heuristic for overlay mapping with enhanced resilience and qos,” in *IET International Conference on Communication Technology and Application (ICCTA 2011)*, Oct. 2011.

[19] H. Yu, V. Anand, C. Qiao, and G. Sun, “Cost efficient design of survivable virtual infrastructure to recover from facility node failures,” in *IEEE International Conference on Communications (ICC)*, June 2011.

[20] C. Qiao, B. Guo, S. Huang, J. Wang, T. Wang, and W. Gu, “A novel two-step approach to surviving facility failures,” in *Optical Fiber Communication Conference and Exposition (OFC/NFOEC), 2011 and the National Fiber Optic Engineers Conference*, March 2011, pp. 1 –3.

[21] H. Yu, C. Qiao, V. Anand, X. Liu, H. Di, and G. Sun, “Survivable virtual infrastructure mapping in a federated computing and networking system under single regional failures,” in *GLOBECOM 2010, IEEE*, 2010.

[22] Q. Hu, Y. Wang, and X. Cao, “Location-constrained survivable network virtualization,” in *Sarnoff Symposium (SARNOFF), 2012 IEEE*, May 2012, pp. 1 –5.

[23] W.-L. Yeow, C. Westphal, and U. Kozat, “Designing and embedding reliable virtual infrastructures,” in *Proceedings of the second ACM SIGCOMM workshop VISA*, ser. VISA ’10. ACM, 2010, pp. 33–40.

[24] J. Xu, J. Tang, K. Kwiat, W. Zhang, and G. Xue, “Survivable virtual infrastructure mapping in virtualized data centers,” in *IEEE 5th International Conference on Cloud Computing (CLOUD), 2012*, June 2012.

[25] I. Houidi, W. Louati, D. Zeghlache, P. Papadimitriou, and L. Mathy, “Adaptive virtual network provisioning,” in *Proceedings of the second ACM SIGCOMM workshop VISA*. ACM, 2010, pp. 41–48.

[26] I. Houidi, W. Louati, and D. Zeghlache, “A distributed virtual network mapping algorithm,” in *ICC ’08. IEEE International Conference on Communications*, May 2008, pp. 5634 –5640.

The Impact Of Extra Traffic On The Control Channel Over The Performance Of CCA Applications

Iallen Gábio de Sousa Santos, Felipe Eduardo N. Mazullo *and* André Castelo Branco Soares
Departamento de Computação
Universidade Federal do Piauí - UFPI
Teresina-PI, Brazil
iallen16@gmail.com, felipemazullo@gmail.com, andre.soares@ufpi.edu.br

***Abstract*— This paper presents an analysis of the impact of competing traffic on the control channel on the performance of applications like Chain Collision Avoidance (CCA). These vehicular network's applications are addressed to traffic safety aiming chain collision avoidance. This work shows the impact of signaling messages of other applications (in the control channel) in the CCA application's performance. The results were carried out considering since an ideal scenario (without concurrent traffic) until a scenario with high utilization level of control channel. Besides, the impact of the transmission power in the CCA applications is also evaluated.**

Keywords-vehicular; network; CCA; safety; traffic.

I. Introduction

With growing number of automobiles traveling through the roads, the probability of accidents increases. Such accidents endanger drivers and passengers, while still causing financial loses.

There are studies being conducted with the purpose of developing new mechanisms to reduce traffic accidents. Among these mechanisms lies the development of Intelligent Transportation System (ITS). Vehicular Ad Hoc Networks (VANETs) are networks formed by automobiles and/or fixed equipments usually localized at the side of the roads. These networks support the ITS through the communication between vehicles and operate to avoid accidents [1] [8]. One of the main motivators for the development of vehicular networks is the development of applications that aim to increase traffic safety [1]. One problem often found in traffic is chain collision. In this case, the braking of one vehicle could lead to collisions among the vehicles that come behind the first one. VANETs are able to support applications with the purpose to avoid chain collisions, they are called Chain Collision Avoidance (CCA) applications [7].

This paper presents a study on the extra traffic impact over the control channel on the performance of manual and automatic CCA applications. The CCA applications are evaluated on 4 scenarios with different levels of traffic competition on the control channel. Their performance is analyzed in terms of vehicle collisions and percentage of successful delivery of emergency messages.

This paper is organized as follows: Section 2 presents the fundamentals of vehicular networks and the Wireless Access in Vehicular Environments (WAVE) architecture. Section 3 describes the related work and contributions of this article. Section 4 presents the problem of chain collisions and the CCA applications. Finally, in sections 5 and 6 are the results and conclusions, respectively.

II. VANETs and the WAVE Architecture

The primary goal of vehicular networks is to establish the conditions to allow for communication between vehicles. Vehicles can communicate directly (Vehicle to Vehicle Communication – V2V) or by making use of an infra-structure located at the side of the road (vehicle to roadside infrastructure – V2R) [5].

The applications for VANETs can be classified into: i) entertainment applications, ii) driver assistance applications, and iii) traffic safety applications.

The entertainment applications include file sharing, text messaging and Internet access. The driver assistance applications can, for example, provide information about the traffic in a particular region, on the existence of free parking, tourist spots etc. The traffic safety applications aim to prevent accidents. In this class are the applications that warn about the possibility of collisions at intersections, warn the driver about speeding and, in general, anticipate the reaction of the driver in order to decrease the risk of traffic accidents [1].

VANETs have some peculiarities in relation to traditional mobile networks, such as: nodes moving at high speed, short contact time, highly dynamic network, the fact that nodes have their movement restricted by roads, the need to provide high scalability, among others. Those characteristics make the existing protocols for mobile networks not suitable for use in a vehicular environment [2].

An architecture called Wireless Access in the Vehicular Environment (WAVE) is being developed to support vehicular networks [8]. In the experiments performed for this paper, we used the WAVE architecture implemented in the NCTUns 6.0 simulator, including the medium access protocol IEEE 802.11p based on the IEEE 802.11 but modified to work on multiple channels [3].

The WAVE architecture works with multiple channels, a control channel (Control Channel - CCH) and several service channels (Channels Service - SCHS). The control channel can be used to send frames containing Wave Short

Message (WSM) packets, sent by critical applications. Service channels can be used to send both frames with WSMs and frames containing IPv6 packets [8].

The Wave Basic Service Set (WBSS) is a set of WAVE stations that can communicate using service channels. To join a WBSS a node must first receive a beacon frame that is sent through the control channel. This frame is sent by the WBSS provider node and contains all information necessary for the association of the receiving node. In this paper, WBSS provider nodes were used to transmit beacon frames generating extra traffic on the control channel.

III. RELATED WORK AND CONTRIBUTIONS

The development of vehicular networks has been widely promoted in order to permit the use of applications to enhance traffic safety [4], [5], [7], [6] and [9]. Among these there are the CCA applications that aim avoid chain collisions through the exchange of emergency messages aimed at preventing sequential collisions of vehicles [7], [9].

Tomas-Gabarron et al. [7] made an evaluation of CCA applications using the IEEE 802.11p, varying parameters such as the power of the signal transmission and the vehicle speed (alternating between 108 km/h and 144 km/h). The authors evaluated the performance of the CCA application in cases where only some of the vehicles involved supported the application. CCA applications presented a deficiency in these scenarios.

S. Xu et al. [9], evaluate a multi-hop broadcast protocol for the transmission of emergency messages in a highway scenario. This work primarily investigated the successful message delivery while considering the end-to-end delay, without assessing the amount of accidents.

This main contribution of this paper is a detailed study of the impact of concurrent transmissions on the control channel over the performance of CCA applications. This paper evaluates four scenarios with different levels of competition on the control channel. The level of competition on the control channel is influenced by actual characteristics of the roads, such as two-way traffic and roads with more than one traffic lane. These characteristics influence the density of cars and therefore cause greater competition in the control channel, which directly impacts the performance of CCA applications. In addition, we also evaluated the impact of various transmission power levels in the success of CCA applications.

IV. CHAIN COLLISIONS AND CCA APPLICATIONS

Chain collisions are a frequent problem in traffic, they are caused by the sudden braking of a vehicle. Typically, such braking occurs by technical or human error, leading to a series of sequential collisions behind the vehicle that first braked.

Fig. 1 illustrates a chain collision scenario. Vehicle 1 sees the obstacle in front of him and brakes. The second vehicle only realizes what happened after vehicle 1 starts reacting. The braking only occurs after the reaction time of vehicle 2’s driver. This reaction time varies from driver to driver and is influenced by the driver’s level of attention.

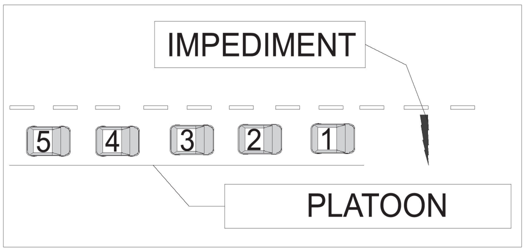


Figure 1. Example of a chain collision scenario.

In this article, we call perception instant the moment the driver perceives the obstacle or the car braking immediately in front him. The reaction time is the time it takes the driver to react after the moment of perception.

J.-B. Tomas-Gabarron et al. [7] mentioned that the reaction time is normally between 0.5 and 1 second. The reaction instant is the moment when the driver begins its braking. Therefore, the reaction instant can be obtained by adding the reaction time of the driver to the perception instant.

Given two vehicles, v1 and v2, that travel in the same direction and in the same traffic lane. Vehicle v2 is right behind v1. When v1 brakes, it lets v2 know through its brake lights. In general, the perception instant of v2’s driver is very close to v1’s driver reaction time. However, v2’s driver will take some time to react. This time depends on the level of attention of v2’s driver.

The reaction instant of each driver depends on his own reaction time and the reaction time of all drivers in front him. Thus, the problem of chain collision worsens the more vehicles there are.

Here are some definitions considering that the first car in a platoon of *n* cars brakes sharply.

RI(n) is the reaction instant of the vehicle at position *n*. The time to detect the abrupt braking (*Tdab*) is the time required by the CCA application in the first vehicle to brake. This is interpreted by the CCA application as a high probability of chain collision involving the vehicles that are coming right behind the first one. After that time interval the CCA application sends the emergency message to the cars that are right behind the first one. *TPT(n)* is the transmission and propagation time of the message until it reaches vehicle *n*. *MPT* is the message processing time. *DRT(n)* is the time that the driver spends to react and start braking process (non-automatic) after the instant he identifies the need for braking. The automatic reaction time involves the time the CCA application needs to start the reaction upon the arrival of an emergency message.

The reaction instant of the vehicle at position *n* using the automatic CCA application is defined by *RI (n)* = min {*RI (n-1)* + *DRT (n)*, *RI* (1) + *Tdab* + *TPT (n)* + *MPT*}. In this case, the difference between the reaction instants is essentially the difference between the transmission and propagation times to each vehicle in the platoon.

The reaction instant of the vehicle at position *n* using the manual CCA application is defined by *RI (n)* = min {*RI (n-1)* + *DRT (n)*, *I* (1) + *Tdab* + *TPT (i)* + *MPT* + *DRT (i)*}. In this case the difference between the reaction instants is influenced by the manual reaction time of each driver.

The reaction instant of the vehicle at position *n* without using the CCA application is defined by *RI (n)* = *RI (n-1)* + *DRT (i)*. The difference between the reaction instants of each driver is influenced by the manual reaction time of the driver at position *n* and the reaction time of all drivers in front of him.

The difference between the reaction instants of each vehicle in the platoon influences the number of collisions. One solution to minimize this problem is to approximate the reaction instants of the vehicles in the platoon. This can be achieved through the exchange of messages between the vehicle which caused the accident and the other vehicles in the platoon. This is the foundation of CCA applications.

CCA applications start their procedures upon detecting an abrupt braking. The application then sends an emergency message to vehicles behind the vehicle that braked. On the receiving end, upon receiving the emergency message the CCA application can start am automatic braking or simply issue a warning to the driver so that he can start braking.

CCA applications that trigger an automatic reaction on the vehicle are called automatic CCA applications. If the CCA application only sends an alert to the driver then it is called manual CCA application [7].

Fig. 2 illustrates how two vehicles (v1 and v2) get closer due to an emergency braking made by vehicle v1.

At the i1 instant, v1 and v2 are separated by a distance d0. At i2, v1's driver starts braking abruptly (for example, by suddenly realizing there is an obstacle in front of him). Therefore, *RI* (v1) = i2. Still, at i2 v2's driver observes v1's brake lights. It is worth noting that only in the instant i3, i3 = i2 + *DRT* (v2), v2 starts its braking process. At i3, the distance between v1 and v2 is equal to d1, which is smaller than d0.

From the instant i3 onwards, the two vehicles are already slowing down, but v1's speed remains smaller than v2's. Therefore, at the time of stopping (i4) the distance between the two vehicles is d2, which is smaller than d1. The value of d2 might be zero, characterizing the collision of vehicles.

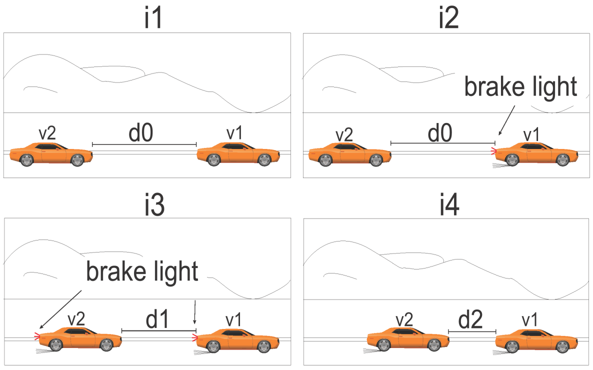


Figure 2. Two vehicles during a braking situation.

In a scenario with more than two vehicles, the approach between two vehicles happens exactly as showed in Fig. 2 and from time i3 onwards cycle restarts for the next pair of vehicles.

CCA applications are driven towards traffic safety and aim to prevent chain collisions. The CCA application functioning is based on the exchange of emergency messages between vehicles. If the vehicle is using an application automatic CCA application, the vehicle reacts automatically. If the vehicle uses a manual application an alert is sent to the driver, and he is responsible for the braking.

V. Performance Evaluation

The performance evaluations were performed with the help of the NCTUns 6.0 simulation tool. This simulator was chosen because of several characteristics: i) it has integrated traffic and network simulators, ii) it has the WAVE protocol stack, therefore directly supports vehicular network simulations and iii) it allows for microscopic modeling, i.e. simulations with traffic parameters and network events defined for each node. All of this is favorable for modeling chain collision situations.

The WAVE architecture, which is implemented in the NCTUns 6.0 simulator, was used for inter-vehicle communications. The WSM protocol was used for sending emergency messages.

Table I shows the traffic parameters and Table II shows the parameters concerning the CCA applications used in the simulations.

The traffic parameters characterize a scenario with high risk of collision. While this it may not occur constantly, this is the kind of scenario where chain collisions usually happen. This situation can be found mostly in medium and big sized cities. Studies have been

TABLE I. Traffic Parameters

Traffic parameters	
Average speed	16 m/s
Number of vehicles on the platoon	30
Average distance between vehicles	10 m
Max deceleration	10m/s²
Vehicle length	3m

ISBN: 978-1-61208-256-1

TABLE II. CCA APLICATIONS PARAMETERS

CCA aplications parameters	
Transmission power	21 dBm, 28 dBm, 35 dBm
Transmission rate	6 Mbps
Transmission channel	178 (control channel)
Emergency deceleration detection time	0.4s
Emergency message processing time	0.2s
Driver reaction time	0.5s - 1s

conducted addressing scenarios with different average car speeds to evaluate the influence of speed on the number of collisions.

The number of vehicles generating extra traffic on the control channel was varied to observe the impact of extra traffic on the control channel over the performance of CCA applications.

In addition, we evaluated the impact the signal power used to transmit emergency messages has over the delivery rate of these messages and the number of vehicle collisions.

The header of a WSM contains information on which channel to be used, power and transmission rate associated with each packet enabling the control of these parameters by applying them to each package individually.

To assess the impact of extra traffic on the control channel over the performance of CCA applications we considered four scenarios. Scenario 1 presents ideal conditions, there is no competition on the control channel. In a single traffic lane, there are 30 vehicles with 10 meters of distance between vehicles. Experiments were carried out to evaluate the performance of CCA applications from the moment the first driver in line performs an abrupt braking. The metric of interest is the number of collisions between vehicles.

In Scenario 2, besides Scenario 1 characteristics, we also consider concurrent transmissions on the control channel. On the single traffic lane there are also vehicles providing on average a WBSS every 200 meters. This means that the WBSS provider vehicle periodically sends beacon frames on the control channel. These beacon frames compete with emergency WSM messages from CCA application.

Scenario 3 also considers only one direction of traffic, but has three traffic lanes with vehicles using the control channel. Scenario 4 has vehicles traveling in 6 traffic lanes, 3 lanes in each direction. In scenarios 3 and 4, for each lane at each 200 m there is vehicle providing a WBSS. In all scenarios considered, the platoon of vehicles 30 (discussed in terms of vehicle collision) is present in only one of the lanes. In all scenarios we assume an average speed of 16m/s. Initially, we assume a transmission power of 28dBm. All results are presented with a confidence interval with a confidence level of 95%.

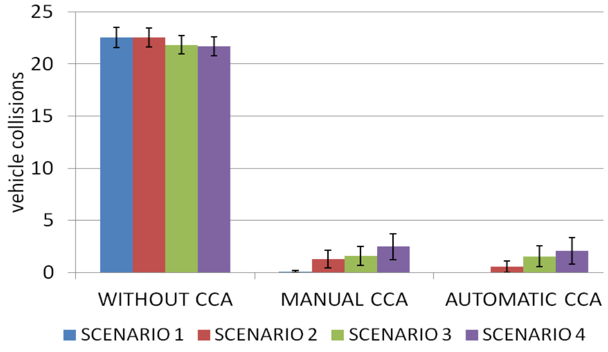


Figure 3. Performance Comparison of the CCA applications in the studied scenarios.

It is noteworthy that the maximum number of collisions in a platoon of n vehicles is n -1, because in this work we assume that the first vehicle does not collide with any obstacle, therefore the maximum number of collisions in the studies presented is 29.

Fig. 3 shows the performance without the CCA application, and the performance with the automatic and manual application for each of the scenarios.

In general, there is a significant decrease in the number of collisions when the CCA application is used. Without the application over 20 collisions occurred. When the application was put to use at the worst case there were less than 4 collisions.

Fig. 4 shows the manual and automatic CCA performance concerning the number of collisions for each of the 4 scenarios previously presented.

By increasing the number of traffic lanes the number of beacon transmissions on the control channel also increases. The beacon transmissions compete with the emergency messages from the CCA application. The increase of this concurrent traffic on the control channel negatively impacts the performance of CCA applications. This is shown through the increasing number of vehicle collisions as a function of increasing the number of traffic

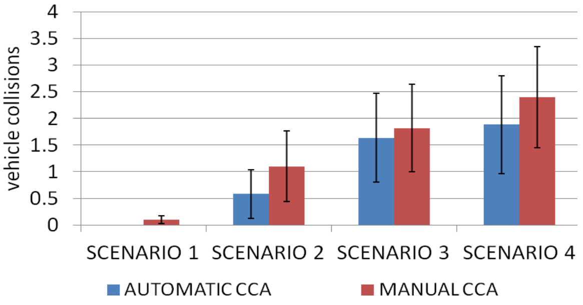


Figure 4. Manual and automatic CCA performance in each scenario studied.

lanes. This behavior occurs in both manual and automatic CCA application. This increased competition in the control channel increases the odds of collisions involving emergency messages frames.

As a direct consequence from the increased competition, there is a decrease in the successful delivery rate of messages from the CCA application, causing more vehicle collisions.

Table III shows the successful delivery rate of emergency messages and its impact over the average number of vehicle collisions in each scenario.

In Table IV, there is a decrease in the successful delivery rate of CCA messages as competition increases in the control channel. For example, in the scenario 4 with 6 traffic lanes, the successful delivery rate is 87.16%, which means a loss of 12.84% of emergency messages. This explains the decrease in the CCA performance. It is noteworthy that this behavior was also observed in Scenarios 2, 3, but with less intensity.

Afterwards, a study was conducted to assess the impact of the WSM messages transmission power over the performance of CCA applications. In this study we considered a scenario similar to the one in scenario 4.

Fig. 5 compares the performance of manual and automatic CCA as a function of the transmission power of emergency messages. Table IV shows successfully delivery rates for each transmission power level used.

Fig. 5 shows that the smallest number of vehicle collisions presented itself with the highest transmission power, 35 dbm.

In Fig. 5, there is a small average number of collisions (below 0.5) with a transmission power of 35 dBm. It's important to point out that 35 dBm represents a longer range and obtained a 91.12% emergency message delivery rate. The CCA application performances while transmitting with 21dbm to 28 dBm were very close, enough to cause the confidence intervals to overlap.

TABLE IV. SUCCESSFUL DELIVERY RATE OF EMERGENCY MESSAGES AND AVERAGE NUMBER OF COLLISIONS IN EACH SCENARIO

Scenarios	Delivery rate	Average of vehicle collisions with automatic CCA aplication	Average of vehicle crashes with manual CCA aplication
Scenario 1	100.00%	0	0.1
Scenario 2	92.78%	0.583333333	1.1
Scenario 3	88.40%	1.633333333	1.816666667
Scenario 4	87.16%	1.883333333	2.4

TABLE III. SUCCESSFUL DELIVERY RATE OF EMERGENCY MESSAGES.

DELIVERY RATE	
21 dBm	90.93%
28 dBm	87.16%
35 dBm	91.12%

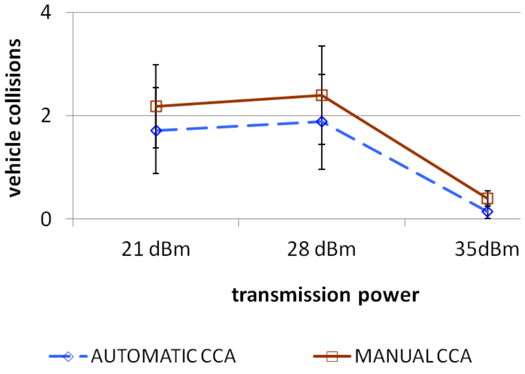


Figure 5. CCA application performance while varying the transmission power.

Through experiments it was found that the extra traffic on the control channel reduces the performance of CCA applications, since the number of collisions increases. This is associated with a decrease in the successful delivery rate of emergency messages. We believe that this increase in the number of emergency messages collisions is strongly related to the hidden terminal problem.

By increasing the transmission power the range also increases. Although it also increases the collision domain on the control channel (which may increase the frame collision probability), the increased transmission power decreases the occurrence of hidden terminals.

The increased transmission power has increased the delivery success rate of emergency messages and consequently improved the CCA application. This behavior, observed at least in the scenarios studied in this article, can be used to point the hidden terminal problem as the main cause of decreased performance of CCA applications in scenarios with concurrent traffic on the control channel.

Fig. 6 illustrates a simplified way how the hidden terminal phenomenon may cause a decrease in the successful delivery rate of emergency messages.

In Fig. 6, vehicle A transmits an emergency message to vehicles coming after it. However, the message collides with beacons transmitted by the vehicle B. As A and B do not know about each other, the CSMA-CA protocol is not very successful in controlling medium access. In scenarios with the higher density of vehicles providing WBSSs the

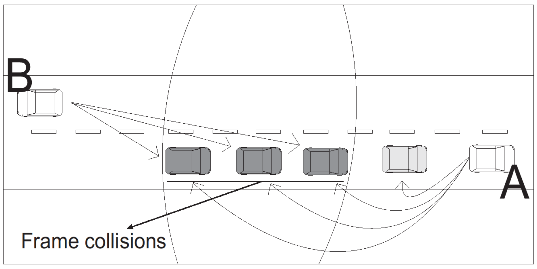


Figure 6: Example scenario with the hidden terminal problem.

occurrence of this phenomenon is more likely. This decreases the performance of CCA applications.

VI. CONCLUSION AND FUTURE WORK

This paper presented a performance evaluation of manual and automatic CCA applications considering four scenarios with different levels of concurrent traffic on the control channel. Generally, at least under the conditions considered in this article, the applications proved efficient, significantly reducing the amount of vehicle collisions.

We identified that with by increasing the concurrent traffic on the control channel the CCA application performance worsens. This effect was caused due to frame collisions, which prevent vehicles from receiving emergency messages sent in WSM packets through the control channel.

The performance of CCA applications as a function of the transmission power was also evaluated. In the studies carried out for this paper, increasing the transmission power decreased the occurrence of the hidden terminal problem and improved the application's performance.

The scenarios where chain collisions can occur are highly diversified. Factors such as the number of vehicles, speed, one way traffic or two-way traffic impact the vehicle density and consequently the level of competition for access control to the control channel.

Studies are being conducted with the goal of minimizing the loss of performance of CCA applications in scenarios with competition on the control channel. These future studies aim to identify further issues with the use of safety applications in VANETs and to propose ways to minimize these deficiencies.

REFERENCES

[1] K.A. Hafeez , L. Zhao, L. Zaiyi, and B.N. Ma, "Impact of Mobility on VANETs' Safety Applications" Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE, December 2010, pp. 1-5.

[2] J. Toutouh and E. Alba, "Performance analysis of optimized VANET protocols in real world tests", Wireless Communications and Mobile Computing Conference (IWCMC), 2011 7th International, July 2011, pp.1244-1249.

[3] D. Jiang and L. Delgrossi, "IEEE 802.11p: Towards an international standard for wireless access in vehicular environments". In Vehicular Technology Conference, 2008. VTC Spring 2008. IEEE, May 2008, pp.2036-2040.

[4] M. Koubek, S. Rea, and D. Pesch, "Reliable broadcasting for active safety applications in vehicular highway networks". In IEEE 71st Vehicular Technology Conference (VTC 2010-Spring), May 2010, pp.1-5.

[5] B.M. Mughal, A.A. Wagan, and H. Hasbullah, "Efficient congestion control in vanet for safety messaging". In International Symposium in Information Technology (ITSim), June 2010, pp.654-659.

[6] T. Taleb, A. Benslimane, and K. Ben Letaief, "Toward an effective risk-conscious and collaborative vehicular collision avoidance system". In IEEE Transactions on Vehicular Technology, March 2010, pp.1474-1486.

[7] J.-B. Tomas-Gabarron, E. Egea-Lopez, J. Garcia-Har, and R. Murcia-Hernandez, "Performance evaluation of a cca application for vanets using ieee 802.11p". In IEEE International Conference on Communications Workshops (ICC), March 2010, pp.1-5.

[8] R. Uzcategui and G. Acosta-Marum, "Wave: A tutorial". Communications Magazine, IEEE, 47(5), May 2009, pp.126-133.

[9] S. Xu, H. Zhou, C. Li, and Y. Zhao, "A multi-hop v2v broadcast protocol for chain collision avoidance on highways". In IEEE International Conference on Communications Technology and Applications, October 2009, pp.110-114.

Different Scenarios of Concatenation at Aggregate Scheduling of Multiple Nodes

Ulrich Klehmet Kai-Steffen Hielscher
Computer Networks and Communication Systems
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
Email: {klehmet, hielscher}@informatik.uni-erlangen.de

***Abstract*—Network Calculus (NC) offers powerful tools for performance evaluation in queueing systems. It has been proven as an important mathematical methodology for worst-case analysis of communication networks. One of its main application fields is the determination of QoS guarantees in packet switched communication systems. One issue of nowadays' research is the applicability of NC concerning the performance evaluation of aggregate multiplexing flows either at one node or at multiple nodes. Then, we have to differ whether the FIFO property at merging single flows can be assumed or not as in case of so-called *blind multiplexing*. In this paper, we are dealing with problems of computing the service curve for the single individual flow at demultiplexing in connection with aggregate scheduling of both – a singular service system (node) or of multiple nodes, at least two. These service curves are relevant for worst-case delay computation. In particular we define important application scenarios and compare their resulting single flow service curves. These are of practical benefit in many applications and can not be found in literature.**

***Index Terms*—Network Calculus; FIFO Multiplexing; Blind Multiplexing; Concatenation of nodes; Pay Multiplexing Only Once**

I. Introduction

In the framework of NC, the modelling elements *arrival curve* and *service curve* play an important role. They are the basis for the computation of maximal deterministic boundary values like backlog bounds and delay bounds found in [1], [2].

Definition 1 (Arrival curve): Given a system S with input flow $x(t)$. Let $\alpha(t)$ be a non-negative, non-decreasing function. $x(t)$ is constrained by or has arrival curve $\alpha(t)$ iff
$x(t) - x(s) \leq \alpha(t-s)$ for all $t \geq s \geq 0$.
Another speech is: flow F is α-smooth.

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

$\alpha_{r,b}(t) = b + rt$ for $t > 0$ and zero otherwise.

As one can see in Fig. 1 this arrival curve forms an upper limit for traffic flows $x(t)$ with (average) rate r and instantaneous burst b . That means $x(t) - x(s) \leq \alpha_{r,b}(t-s) = b + r \cdot (t-s)$. For $\Delta t := t - s$ and $\Delta t \to 0$ it holds

$$\lim_{t \to s}\{x(t) - x(s)\} \leq \lim_{\Delta t \to 0}\{r \cdot \Delta t + b\} = b$$

An important definition of NC is the following one:

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 by

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

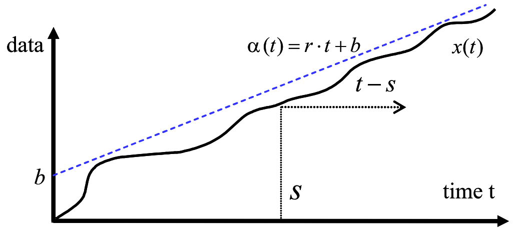


Fig. 1. Token Bucket Arrival Curve

Applying Definition 2 we can characterize the arrival curve $\alpha(t)$ with respect to $x(t)$ as:

$$x(t) \leq (x \otimes \alpha)(t)$$

The concept of arrival curves describes an upper bound to an input stream of a system processing some type of data. Concerning the output of this system we are interested in some service guarantees, i.e. is there a guaranteed minimum of output $y(t)$ – the amount of data leaving system S? The modeling element *service curve* deals with this problem.

Definition 3 (Service curve): Given a system S with input flow $x(t)$ and output flow $y(t)$. The system offers a (minimum) service curve $\beta(t)$ to the flow iff $\beta(t)$ is a non-negative, non-decreasing function with $\beta(0) = 0$ and $y(t)$ is lower bounded by the convolution of $x(t)$ and $\beta(t)$:

$$y(t) \geq (x \otimes \beta)(t).$$

Fig. 2 demonstrates $(x \otimes \beta)(t)$ as an example for the lower bound of the output $y(t)$ and any given input $x(t)$.

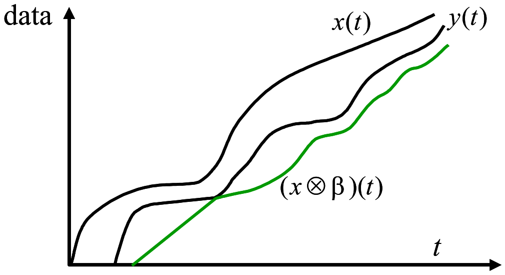


Fig. 2. Convolution as a Lower Output Bound

Example 2: One commonly used service curve is the rate-latency function: $\beta(t) = \beta_{R,T}(t) =$
$R \cdot [t-T]^+ := R \cdot \max\{0; t-T\}$. The rate-latency function reflects a service element which offers a minimum service of rate R after a worst-case latency of T. Having in mind a worst case performance analysis, it is possible to abstract away

from complex (queuing) systems with different scheduling strategies.

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

Theorem 1 (Backlog bound and output bound): Consider a system S with input flow $x(t)$ and output flow $y(t)$. Be $x(t)$ α-smooth and S offers a service curve $\beta(t)$. The backlog v at time t, $v(t) = x(t) - y(t)$, is bounded by the supremum of the vertical deviation of arrival curve and service curve:

$$x(t) - y(t) \leq \sup_{s \geq 0}\{\alpha(s) - \beta(s)\}$$

and output $y(t)$ is constrained by the arrival curve

$$\alpha^*(t) = \alpha \oslash \beta = \sup_{s \geq 0}\{\alpha(t+s) - \beta(s)\}.$$

The complete backlog $v(t) = x(t) - y(t)$ at time t within a system is sometimes denoted as *buffer(t)*.

If the node or system serves the incoming data of a flow in FIFO order (First In First Out), the following bound is computable:

Theorem 2 (Delay bound): Assume a flow constrained by arrival curve $\alpha(t)$ passing a system with service curve $\beta(t)$. The maximal virtual delay d is given as the supremum of all possible virtual delays of data, i.e. is defined as the supremum of the horizontal deviation between arrival curve and service curve:

$$d \leq \sup_{s \geq 0}\{\inf\{\tau : \alpha(s) \leq \beta(s+\tau)\}\}.$$

Fig. 3 depicts both theorems.

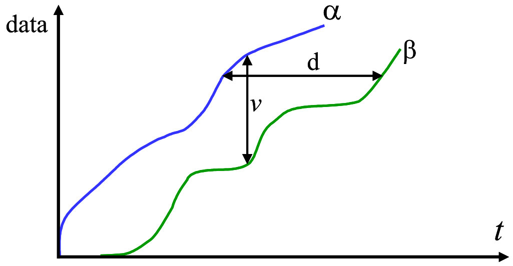


Fig. 3. Backlog and delay bound

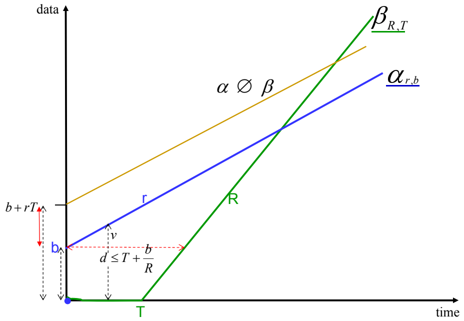


Fig. 4. Example for the bounds

Example 3: Suppose there is a system with input according to a token bucket, thus $x(t) - x(s) \leq \alpha_{r,b}(t-s)$ and rate-latency output:

$$y(t) \geq \inf_{s \leq t}\{x(s) + \beta_{R,T}(t-s)\}$$

Based on the above theorems we get the delay bound $d \leq b/R + T$, the output bound $\alpha^*(t) = r(t+T) + b$, and the backlog is bounded by $v = b + rT$. Fig. 4 shows the results.

Remark: Always in case of token-bucket like input and rate-latency output the worst-case delay d_{max} is computable by

$$d_{max} = \frac{burst}{servicerate} + latency.$$

II. AGGREGATE SCHEDULING

Until now, only per (single) flow-based scheduling have been considered. But in real systems, *aggregate scheduling* arises in many cases. Always, if there are more than one separat input flows entering some kind of data processing/transferring system and then dealt as a whole stream of data – we speak of aggregate scheduling. Important examples are aggregate based networks such as Differentiated Service domains (DS) of the Internet [3]. In order to address such class-based networks, we have to look for rules of multiplexing and aggregate scheduling. Assume that m flows enter a system (network) or system node and are scheduled by aggregation. According to [4] the aggregate input flow and arrival curve are given as follows.

Theorem 3 (Multiplexing): An aggregation, or multiplexing of m flows can be expressed by addition of the input functions respective arrival curves. W.l.o.g. be $m = 2$, 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.

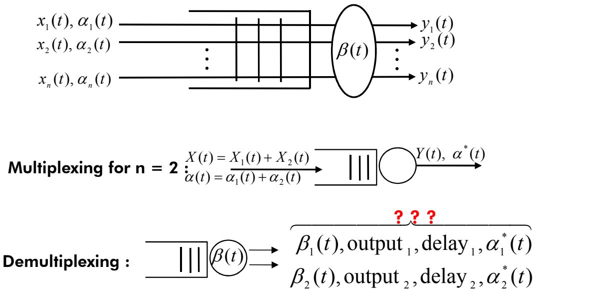


Fig. 5. Multiplexing of input x_i, output y_i with arrival & service curve α_i, $\beta = \beta aggr$

Now, as is shown in Fig. 5 important questions arise: Is it possible to apply the same analysis e.g. of buffer bounds and maximal delay of Theorems 1 and 2 to the single flows x_i? Does there exists a service curve β_i for the individual flow x_i, sometimes denoted as left-over service curve? What is the maximal delay, say of flow x_1, after servicing the aggregate and subsequently demultiplexing? The answers are based on the type of multiplexing in each case, i.e. in which manner the aggregate scheduling is done: **FIFO**, priority-scheduling, or multiplexing by complete unknown arbitration between the flows, which is the definition of **Blind** scheduling [5]. Together with the particular scheduling type one has to take into consideration the service curve of the aggregate flow. From a practical point of view we will discuss here the two important scheduling disciplines: FIFO and Blind. Regarding aggregate flow servers the next both theorems given by [1] are important.

Theorem 4 (FIFO Service curves): Consider a node serving the flows x_1 and x_2 in FIFO order. Assume first that the node guarantees a service curve β to the aggregate of the flows and secondly, flow x_2 is $\alpha_2 - smooth$. Define the family of functions $\beta^1_\theta(t) := [\beta(t) - \alpha_2(t-\theta)]^+ \; if \; t > \theta$ otherwise $\beta^1_\theta(t) := 0$. Then for any $\theta \geq 0$ it holds $y_1 \geq x_1 \otimes \beta^1_\theta$, where y_1 is the output of flow x_1. If β^1_θ is a non-negative, non-decreasing function, flow x_1 has the service curve β^1_θ.

Note $[x]^+ = x$ if $x \geq 0$ otherwise 0.

If no knowledge is given about the choice of service between the flows, i.e. in case of *blind multiplexing* one has to differ between *strict* or *non-strict* aggregate service curves [1].

Theorem 5 (Blind Multiplexing): Consider a node serving the flows x_1 and x_2, with some unknown arbitration between the two flows. Assume the node guarantees a strict service curve β to the aggregate of the two flows and that flow x_2 is $\alpha_2 - smooth$. Define $\beta_1(t) := [\beta(t) - \alpha_2(t)]^+$. If β_1 is wide-sense increasing, then it is a service curve for flow x_1.

But what does it mean, a service curve is *strict* ?

Definition 4 (Strict service curve): A system S offers a strict service curve β to a flow if during any backlogged period $[s,t]$ of duration $u = t - s$ the output y of the flow is at least equal to $\beta(u)$, i.e. $y(t) - y(s) \geq \beta(t-s)$, or equivalently $y(z) \geq \beta(z) \; \forall z \in [s,t]$.
Of course, any strict service curve is a service curve in terms of definition 3, but not vice versa - see for instance [1] or [6].

Example 4: The constant rate server in Fig. 6 with input flow x and output y has a strict service curve $\beta(t) = ct$. Let s be the start of a busy period, that means $y(s) = x(s)$, then $y(t) - y(s) = c(t-s)$, and so $y(t) - x(s) \geq \beta(t-s)$.

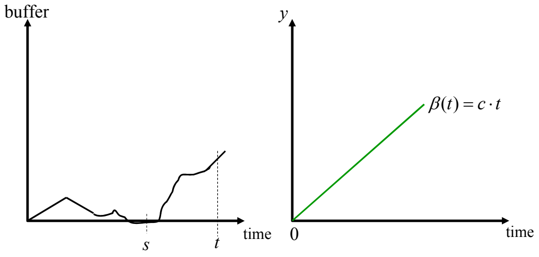


Fig. 6. Constant rate server

Our main objective in this paper is the consideration of *typical application scenarios* concerning multiplexed flows in FIFO or blind schedule situations: Based on Theorems 1 and 2 we want to apply the same analysis, e.g., for getting buffer bounds and maximal delay-values of for instance the single flows x_i after being demultiplexed. For that, what is the 'best' service curve β_i for the individual flow x_i, respectively?

Concerning most practical applications we focus in particular on input flows with token bucket like arrival constraints $\alpha_{r,b}$ and rate-latency service curves $\beta_{R,T}$.

A. Determination of the best service curve at FIFO scheduling:

First, let us come back to Theorem 4 for FIFO schedule in case of two flows x_1 and x_2. The main statement is that for any θ with $0 \leq \theta < t$ the expression $\beta^1_\theta(t) := [\beta(t) - \alpha_2(t-\theta)]^+$ is a service curve for flow x_1. Because that is valid for each $t = t_0$ – we may ask for which especial θ we get the 'best' service curve, i.e. the least pessimistic – or in other words the greatest β^1_θ, (so guaranteeing the least worst case delay etc.) Of course, since α_2 is a wide-sense increasing function – formula $\beta^1_\theta(t) := [\beta(t) - \alpha_2(t-\theta)]^+$ in general will get the largest value if θ is converging to t from left: $\theta \to t$ with $\theta < t$, for that we use the notation $\theta \to \; t_-$.
As we said before concerning practical applications, the arrival and service curves are often a token bucket-type $\alpha_{r,b}(t) = b + rt$ and rate-latency function $\beta_{R,T}(t) = R \cdot [t-T]^+$, respectively. Therefore, in order to demonstrate the search for a 'best service' we will take these both types of curves. That is to say get the supremum of
$\beta_{1,\theta} = [\beta_{R,T}(t) - \alpha_2(t-\theta)]^+$ with $\alpha_2(t) = r_2 t + b_2 \Rightarrow$ $\sup_{0 \leq \theta < t}\{R \cdot (t-T)^+ - [r_2 \cdot (t-\theta) + b_2]\} = \sup_{0 \leq \theta < t}\{Rt - RT - r_2 t + r_2\theta - b_2\}$ which outcomes to $\theta = \theta_{opt} := T + \frac{b_2}{R}$. Thus, the 'best' rate-latency service curve is $\beta_{1,\theta} = \beta(t) - \alpha_2(t - (T + \frac{b_2}{R}))$ with $\beta(t) \;\; (= \beta_{aggr})$ as service curve to the aggregate.

If we now compare the service curves $\beta_{1,\theta}$ of both the FIFO ($\theta = \theta_{opt} = T + \frac{b_2}{R}$) and blind Multiplexing ($\theta = 0$) – of the same multiplexed server – one could expect in case of FIFO the service curve of single flow x_1 is larger, and consequently the better one w.r.t. the worst-case delay d_{max}. Let's denote this as $\beta_{1,FIFO}(t) > \beta_{1,Blind}(t)$. Our following computation will conform to this.

Blind Multiplexing:

$\beta_{1,\theta=0}(t) = \beta_{1,Blind}(t) = \beta(t) - \alpha_2(t-0) = R(t-T)^+ - (r_2 t + b_2) = \cdots = (R - r_2)[t - \frac{RT+b_2}{R-r_2}]^+$
The result is (again) a rate-latency service curve:
$\beta_{1,Blind}(t) = \beta_{R',T'}(t)$ with rate $R' = R - r_2$ and latency $T' = \frac{RT+b_2}{R-r_2}$.

FIFO Multiplexing:

$\beta_{1,\theta}(t) = \beta_{1,FIFO}(t) = \beta(t) - \alpha_2(t - \theta_{opt}) = R(t-T)^+ - (r_2 \cdot (t - (T + \frac{b_2}{R})) + b_2) = \cdots = (R - r_2)(t - [T + \frac{b_2}{R}]^+)$
The result, again a rate-latency service curve:
$\beta_{1,FIFO}(t) = \beta_{R',T'}(t)$ with rate $R' = R - r_2$ and latency $T' = T + \frac{b_2}{R}$.
It is easy to see:

$\beta_{1,FIFO} = (R - r_2)(t - [T + \frac{b_2}{R}])^+ > (R - r_2)[t - \frac{RT+b_2}{R-r_2}]^+ = \beta_{1,Blind}$.

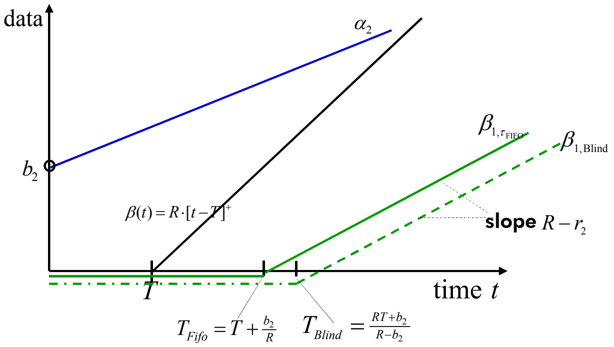


Fig. 7. Service curve FIFO vs. Blind of single flow x_1

In summary is to state:

For both we get the same service rate $R' = R - r_2$, however – as we expected – the latency increases from FIFO to Blind multiplexing. And because a service curve by definition 3 defines a lower output limit $\beta_{1,FIFO}$ specifies a greater lower limit to a single flow x_1 than $\beta_{1,Blind}$. Fig. 7 shows this issue.

Given now a service system multiplexing two flows x_1 and x_2. Theorems 4 or 5 provide a service curve β_i e.g. β_1 for the single flow x_1.

If x_1 is α_1-smooth and by Theorems 1 and 2 – the maximum backlog bound for the demultiplexed single x_1 is given by

$$x_1(t) - y_1(t) \le \sup_{s\ge 0}\{\alpha_1(s) - \beta_1(s)\}$$

and the important worst case end-to-end-delay parameter of x_1 by

$$d \le \sup_{t\ge 0}\{\inf\{\tau : \alpha_1(t) \le \beta_1(t+\tau)\}\}$$

at which expression $d_\tau(t) = \inf\{\tau \ge 0 : \alpha_1(t) \le \beta_1(t+\tau)\}$, the so-called *virtual delay*, is needed: If an input x at time t has arrived it is assured that not later than $d_\tau(t)$ it has left the service facility. This is guaranteed for FIFO scheduling but not for blind Multiplexing. However, we may presume FIFO per single flow x_i within the aggregate and thus apply all bounding theorems without any restrictions.

III. Different aggregate scheduling scenarios

So far, we have considered elementary service nodes (network elements). We now want to discuss the concatenation of aggregate network elements. First of all let's give the important theorem given in[1]:

Theorem 6 (Concatenation of nodes): Assume a flow traverses systems S_1 and S_2 in sequence and β_i is a service curve of S_i, i = 1, 2. Then the concatenation of the two systems offers a service curve of $\beta = \beta_1 \otimes \beta_2$ to the flow, like in Fig. 8.

Using this service curve $\beta = \beta_1 \otimes \beta_2$ we mention the important property [1] **Pay Burst Only Once (POO)**: Applying delay bound Theorem 2 one gets tighter end-to-end delay bounds if the delay computation is based on the concatenated end-to-end service curve β: $D_\otimes \le D_1 + D_2$ with: $D_1 \le \frac{b}{R_1} + T_1$, $D_2 \le \frac{b+rT_1}{R_2} + T_2$ and $D_\otimes \le \frac{b}{\min(R_1,R_2)} + (T_1 + T_2)$, again token-bucket and rate-latency curves supposed. (The burst b affects the sum $(D_1 + D_2)$ twice whereas $D_\otimes$ only once.)

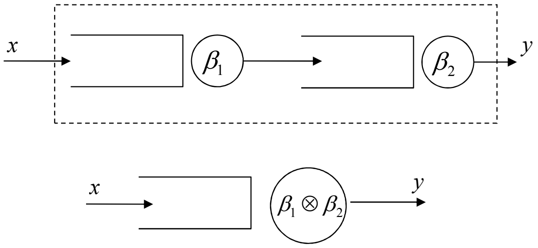


Fig. 8. Service curve of concatenated nodes

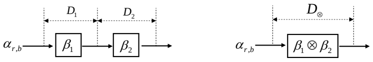


Fig. 9. Pay Burst Only Once-Principle

A. *Concatenation of aggregated nodes*

Now we will regard the concatenation of aggregate nodes, exemplarily for an input of two flows x_1, x_2 and a concatenated two-node system as shown in Fig. 10.

What is the end-to-end service curve of let's say flow x_1 ? By Theorem 6 and the (aggregation-) Theorems 4 (or 5) with $\beta^1_\tau(t) = [\beta(t) - \alpha_2(t-\tau)]^+$ we get:

$\beta_1^{tot}(t) = (\beta^I_{1,\tau} \otimes \beta^{II}_{1,\vartheta})(t)$

$= [\beta^I(t) - \alpha_2^I(t-\tau)]^+ 1_{t>\tau} \otimes [\beta^{II}(t) - \alpha_2^{II}(t-\vartheta)]^+ 1_{t>\vartheta}$,

where β^I, β^{II} are service curves of the aggregated flows of node I or node II, and α_2^I and α_2^{II} the arrival curves of the individual flow x_2 at the corresponding nodes.

(The term $1_{t>\theta}$ is zero for $t \le \theta$. In the following, formulas, for the sake of clarity we will omit this term frequently). As given in Fig.10 flow x_1 is aggregated with x_2 only at node I, i.e. multiplexing happens only once. These thoughts lead to the PMOO-principle (Pay Multiplexing Only Once) [5]: First do the concatenation $\otimes$ of both nodes w.r.t. service curve β and afterwards apply Theorem 4 (or 5 in case of Blind):

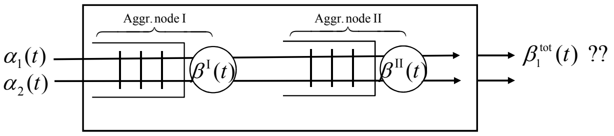


Fig. 10. Concatenation of aggregate nodes

$$\beta^{tot}_{1,PMOO}(t) = [(\beta_1^I \otimes \beta_1^{II})(t) - \alpha_2^I(t-\kappa)]^+.$$

Question: Is $\beta^{tot}_{1,PMOO}$ better than β_1^{tot} or in other words $\beta^{tot}_{1,PMOO}(t) \ge \beta_1^{tot}(t)$?

Again, suppose: rate-latency service curves $\beta^I(t) = R^I \cdot [t - T^I]^+$, $\beta^{II}(t) = R^{II} \cdot [t - T^{II}]^+$ and token bucket arrival curves $\alpha_2^I(t) = r_2 \cdot t + b_2^I$, $\alpha_2^{II}(t) = r_2 \cdot t + b_2^{II}$.

At this point, we have to differ between FIFO and blind multiplexing, that means in formula $\beta^1_\tau(t) = [\beta(t) - \alpha_2(t-\tau)]^+$ we define $\tau = \tau_{opt} = T + \frac{b_2}{R}$

or $\tau = 0$, respectively.

*1) **Case FIFO**:* Let be $\tau = T^I + \frac{b_2^I}{R^I}$ of node I, $\vartheta = T^{II} + \frac{b_2^{II}}{R^{II}}$ of node II It follows:

- $\beta_1^{tot}(t)$= $\min(R^I - r_2, R^{II} - r_2) \cdot [t - T^I - T^{II} - \frac{b_2^I}{R^I} - \frac{b_2^{II}}{R^{II}}]^+$

And according to **PMOO** with $\tau = T^I + \frac{b_2^I}{R^I}$, $\vartheta = T^{II} + \frac{b_2^{II}}{R^{II}}$, $\kappa = (T^I + T^{II}) + \frac{b_2^I}{\min(R^I, R^{II})}$ we get

- $\beta_{1,PMOO}^{tot}(t) = [\min(R^I, R^{II}) - r_2] \cdot [t - T^I - T^{II} - \frac{b_2^I}{\min(R^I, R^{II})}]^+$

The results are two service curves β_1^{tot} or $\beta_{1,PMOO}^{tot}$ of flow x_1 again of type rate-latency.

Since $b_2^I \leq b_2^{II}$ it follows: $\beta_{1,PMOO}^{tot}(t) \geq \beta_1^{tot}(t)$.

Computing the worst-case delay D by $D = \frac{burst}{servicerate} + latency$, for $D = D_1$ or $D = D_{1,PMOO}$ and using β_1^{tot}, respectively $\beta_{1,PMOO}^{tot}$ we get:

- $D_1(t) = \frac{b_1}{\min(R^I - r_2, R^{II} - r_2)} + [T^I + T^{II} + \frac{b_2^I}{R^I} + \frac{b_2^{II}}{R^{II}}]$
- $D_{1,PMOO}(t) = \frac{b_1}{\min(R^I - r_2, R^{II} - r_2)} + [T^I + T^{II} + \frac{b_2^I}{\min(R^I, R^{II})}]$,

thus $D_{1,PMOO}(t) < D_1(t)$.

Result: $\beta_{1,PMOO}^{tot}$ is better than β_1^{tot}, since it produces a shorter worst case delay D.

*2) **Case Blind**:* $\tau = 0$, $\vartheta = 0$, $\kappa = 0$ (after Theorem 5) It follows for the end-to-end service curve $\beta_1^{tot}(t)$ of x_1:

- $\beta_1^{tot}(t) = [\min(R^I - r_2, R^{II} - r_2)] \cdot [t - (\frac{R^I T^I + b_2^I}{R^I - r_2} + \frac{R^{II} T^{II} + b_2^{II}}{R^{II} - r_2})]^+$

And applying the PMOO-principle here again, we get:

- $\beta_{1,PMOO}^{tot}(t) = [\min(R^I - r_2, R^{II} - r_2)] \cdot [t - \frac{\min(R^I, R^{II}) \cdot (T^I + T^{II}) + b_2^I}{\min(R^I, R^{II}) - r_2}]^+$

Unfortunately, now it is not always true: $\beta_{1,PMOO}^{tot} \geq \beta_1^{tot}$: $\beta_{1,PMOO}^{tot}$ per se does not causes less delay than β_1^{tot}. We get

$$\beta_{1,PMOO}^{tot} \geq \beta_1^{tot} \Leftrightarrow \begin{cases} (*) & b_2^{II} \geq \frac{r_2 T^{II} (R^{II} - R^I)}{R^I - r_2} \\ (**) & b_2^I \geq \frac{r_2 T^I (R^I - R^{II})}{R^{II} - r_2} \end{cases}$$

if (*) $\min(R^I, R^{II}) = R^I$ or (**) $\min(R^I, R^{II}) = R^{II}$. That means: $D_{1,PMOO}(t) < D_1(t)$ for condition (*) or (**).

B. More general concatenation settings

For practical application and comparisons we complete these scenarios and introduce the following definitions.

Definitions – FIFO:

$\beta_1^{FIFO} := [\beta^I(t) - \alpha_2^I(t - \tau)]^+ \otimes [\beta^{II}(t) - \alpha_2^{II}(t - \vartheta)]^+$

$\beta_{1,PMOO}^{FIFO} := [(\beta^I \otimes \beta^{II})(t) - \alpha_2^I(t - \kappa)]^+$

$\tilde{\beta}_{1,PMOO}^{FIFO} := [\beta^I(t) - \alpha_2^I(t - \tau)]^+ \otimes \beta^{II}(t)$ or

$\tilde{\beta}_{1,PMOO}^{FIFO} := \beta^I(t) \otimes [\beta^{II}(t) - \alpha_2^{II}(t - \tau)]^+$ with $\tau = T^I + \frac{b_2^I}{R^I}$, $\vartheta = T^{II} + \frac{b_2^{II}}{R^{II}}$ and $\kappa = (T^I + T^{II}) + \frac{b_2^I}{\min(R^I, R^{II})}$.

Definitions – Blind:

$\beta_1^{Blind} := [\beta^I(t) - \alpha_2^I(t - 0]^+ \otimes [\beta^{II}(t) - \alpha_2^{II}(t - 0]^+$

$\beta_{1,PMOO}^{Blind} := [(\beta^I \otimes \beta^{II})(t) - \alpha_2^I(t - 0]^+$

$\tilde{\beta}_{1,PMOO}^{Blind} := [\beta^I(t) - \alpha_2^I(t - 0]^+ \otimes \beta^{II}(t)$ or

$\tilde{\beta}_{1,PMOO}^{Blind} := \beta^I(t) \otimes [\beta^{II}(t) - \alpha_2^{II}(t - 0]^+$ here $\tau = \vartheta = \kappa = 0$.

But what does it mean for instance
$(i) : [\beta^I(t) - \alpha_2^I(t - 0]^+ \otimes \beta^{II}(t)$ or
$(ii) : \beta^I(t) \otimes [\beta^{II}(t) - \alpha_2^{II}(t - 0]^+$?

Fig.11 explains in (i) and (ii) the semantic equivalent of first or second expression. In picture (i) the single flow x_2 with arrival curve α_2 leaves the system after served in node I, and in (ii) flow x_2 enters the system being served by node II only.

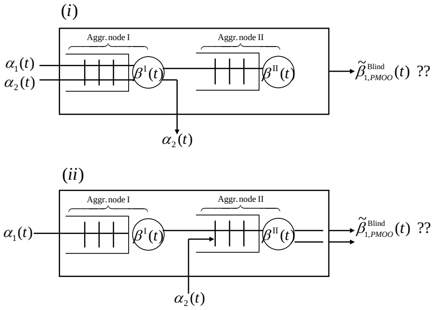


Fig. 11. Service curves of flow x_1 in (i) and (ii)

Using these definitions we come to the following results of the

Different scenarios:

Within FIFO:
$\tilde{\beta}_{1,PMOO}^{FIFO} \geq \beta_{1,PMOO}^{FIFO} \geq \beta_1^{FIFO}$

Within Blind:

- $\beta_{1,PMOO}^{Blind} \geq \beta_1^{Blind}$ if above condition (*) or (**) is given

ISBN: 978-1-61208-256-1

- $\tilde{\beta}^{Blind}_{1,PMOO} \geq \beta^{Blind}_{1,PMOO}$ if $t \to \infty$
- $\tilde{\beta}^{Blind}_{1,PMOO} \geq \beta^{Blind}_{1,PMOO} \geq \beta^{Blind}_{1}$ if $t \to \infty$ and at condition (*) or (**)

Figure 12 shows the relations of left-over service curves between FIFO- and Blind-scheduling. Herein, the relations between the service curves symbolized by '?' are to be computed from case to case. Depending on the parameters $R^I, R^{II}, \tau, \vartheta$ and the concrete value of parameter t – both inequations are possible, either '$\geq$' or '$\leq$' respectively. Of course, one has to ask how to deal with more complex scenarios, e.g. in case of more than two aggregated flows or more than two service nodes. In principle we can apply an approach resulting from aggregation Theorems 3, 4 and 5 together with the concatenation Theorem 6. However one has to check whether the solutions are of practical benefit, may be the service curves $\beta_i(t)$ of single service x_i are too pessimistic which means they create to large worst-case delay bounds based on Theorem 2.

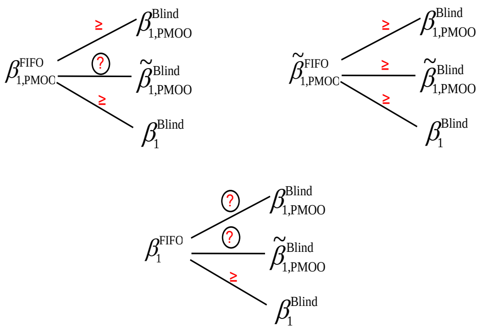


Fig. 12. Comparison between FIFO and Blind

An example setting from [4] for FIFO scheduling with 3 flows and 3 server nodes is given here, where in Fig. 13 (*i*) flow 3 enters node II and after service is given out immediately. The other both flows are served by all 3 nodes. According to the aggregation and concatenation theorem we get:

$\beta_1(t) = [\beta^I(t) \otimes [(\beta^{II}(t) - \alpha_3^{II}(t-\tau))]^+ \otimes \beta^{III}(t) - \alpha_2^I(t-\vartheta)]^+$.

As before taking rate-latency service curves and token bucket arrival curves, $\tau = T^{II} + \frac{b_3^{II}}{R^{II}}$, and $\vartheta = T^I + T^{II} + T^{III} + \frac{b_2^I}{\min(R^I, R^{II}-r_3, R^{III})}$ thereupon resulting in $\beta_1(t) = [\min(R^I, R^{II} - r_3, R^{III}) - r_2] \cdot [t - T^I - T^{II} - T^{III} - \frac{b_3^{II}}{R^{II}} - \frac{b_2^I}{\min(R^I, R^{II}-r_3, R^{III})}]^+$.

The scenario in Fig. 13 (*ii*) w.r.t. left-over service of flow 2 leads to the end-to-end service curve

$\beta_2(t) = \min(R^I - r_1, R^{II} - r_3 - r_1, R^{III} - r_3) \cdot [t - T^I - T^{II} - T^{III} - \frac{b_1^I}{\min(R^I, R^{II}-r_3)} - \frac{b_3^{III}}{\min(R^{II}-r_1, R^{III})}]^+$.

Hereby flow 1 get service by node I and node II and leaves the system whereas flow 3 is served by node II and node III before leaving the server system. Only flow 2 get service by all 3 nodes.

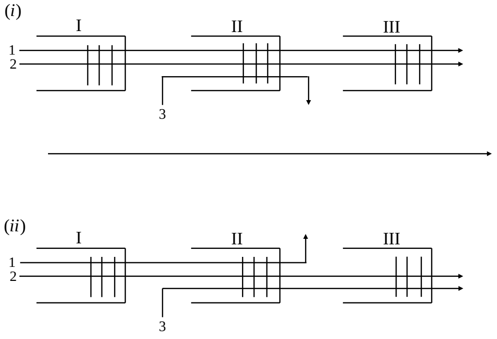


Fig. 13. Server networks with more flows and nodes

IV. Conclusion

In this paper, we considered the subject of service curves in connection with aggregate scheduling mechanisms. Based on these service curves the maximum end-to-end delays of single flows x_i (left-over flow) after being demultiplexed are computable. In particular we discussed different scenarios of multiple aggregated nodes - which are typical for practical applications: Token bucket input flows and rate-latency service curves together with the main scheduling principles FIFO and Blind multiplexing. In a sense of case study we computed corresponding formulas and compared the results w.r.t. 'best service curves', i.e. the largest one and such producing the shortest worst-case end-to-end delays, which have great practical benefit for many hard real-time server systems.

In conclusion, for FIFO and Blind-scheduling of concatenated aggregation systems we computed service curves of demultiplexed single flows and compared them in different practice-relevant scenarios, which so far in the literature are not given. With our formulas and comparisons for single end-to-end service curves we move a step closer to allowing the design of complex systems.

References

[1] J.-Y. Le Boudec and P. Thiran, *Network Calculus*. Springer Verlag LNCS 2050, 2001.

[2] R. Cruz, "A calculus for network delay, part i: Network elements in isolation," *IEEE Trans. Inform. Theory*, vol. 37-1, pp. 114–131, 1991.

[3] A. Charny and J.-Y. Le Boudec, *Delay Bounds in a Network with Aggregate Scheduling*. Springer Verlag LNCS 1922, 2000.

[4] M. Fidler and V. Sander, "A parameter based admission control for differentiated services networks," *Computer Networks*, vol. 44, pp. 463–479, 2004.

[5] 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.

[6] U. Klehmet and K.-S. Hielscher, "Strictness of Rate-Latency Service Curves," in *Data Communication Networking(3rd DCNETS Conference), DCNET/ICE-B/OPTICS, page 75-78. SciTePress, Rome*, July 2012.

Full Dedicated Optical Path Protection in the WDM Mesh Networks without Wavelength Conversion

Stefanos T. Mylonakis
University of Athens
Faculty of Information and Telecommunication
e-mail: smylo@otenet.gr

***Abstract*-In this paper, a WDM mesh network is planned and designed by two methods so that to satisfy all its demands and each connection has to be protected with the dedicated way by both nodes. In the first method, each connection uses the free available wavelength after the maximum busy wavelength (higher index) of each optical link from full complementary working and protection lightpaths. In the second method, each connection uses the first free (lowest index) available wavelength of each optical link from full complementary working and protection lightpaths.**

Keywords -WDM networks; dedicated protection.

I. INTRODUCTION

Optical networks using Wavelength Division Multiplex (WDM) make use of the enormous bandwidth of an optical fiber. WDM divides the tremendous bandwidth (~50THz) of a single mode optical fiber in to many non overlapping wavelengths (or wavelength channels with bandwidth 1-10 Gbps or more) which can operate simultaneously, with the fundamental requirement that each of these channels operate at different wavelengths. WDM basically is frequency division multiplexing in the optical range where the carrier frequencies are referred as wavelengths. These high capacity WDM optical mesh networks that based on optical technologies, provide routing, grooming and restoration at the wavelength level as well as wavelength based services.

In this paper, all network parameters are known and the WDM mesh network is planned and designed so that to satisfy all its demands using the shortest path algorithm and each connection has to be protected with the dedicated way and based on the spare capacity which is allocated as a “dedicated” resource for sole use of the connection. The assignment of the suitable wavelengths for each connection is done using a) the free available wavelength after the maximum busy (higher index) wavelength of each optical link from full complementary working and protection lightpaths, b) the first free **(**lower index**)** available wavelength of each optical link from full complementary working and protection lightpaths and c) the benefits of the second method versus the first one also shown when wavelength conversion is not used as well as the performance improvement when the wavelength conversion is used. The demand tables are large and they are not showed.

This paper is broken down in the following sections. Section II describes the related works. Section III describes the problem and provides a solution, the method synoptic description, an example and the discussion and proposals. Section IV draws conclusions and finally ends with the references.

II. RELATED WORKS

Research has been done [1]-[13] in relation to the methods and the problems associated with planning, protection and restoration of optical networks. A modelling and analysis was performed by H. Kobayashi [1]. Advanced software engineering course is showed by F. L. Bauer et al. [2]. There are several approaches to ensure fibre network survivability, as described by T. Wu [3] and A. Bononi [4]. V. E. Benes [5] analyzes Mathematical Theory of Connecting Networks and Telephone Traffic. In [6], B. Ramamurthy et al. write about Wavelength Conversion in WDM Networking. In [7], J. Emirghani et al. offer an overview of the enabling technologies and extend the treatment to the network application of the wavelength converters. In [8], C. Xiaowen et al. demonstrate that their paper network architecture can significantly save the number of wavelength converters, yet achieving excellent blocking performance. In [9], M. O’Mahony et al. begin with an overview on the future of optical networking. A historical look at the emergence of optical networking is first taken, followed by a discussion on the drivers

ISBN: 978-1-61208-256-1

pushing for a new and pervasive network, which is based on photonics and can satisfy the needs of a broadening base of residential, business and scientific users. J. Zhang et al. [10] present that, for fault management in optical WDM mesh networks end to end path protection, is an attractive scheme to serve customers' connections. In [11], the modelling methods and simulation tools are described and used for the analysis of a new integrated restoration scheme operating at multilayer networks. In [12], T. Ingham et al. deal with the modelling and simulation effectively help and validate the design of various components constituting the service delivery platform. In [13], J. Burbank deals with the modelling and simulation and gives practical advices for network designers and developers.

III. THE PROBLEM AND ITS SOLUTION

A. The problem

The network topology and other parameters are known as WDM and optical fibre capacity, the number of node pairs and the node pairs that the demands (requests for connection) must be satisfied and the m:N (1:7) WDM and optical fibre shared protection protocol. So, this network is characterized as multifibre network by working and protection fibres per link (multifibre link) and edges of two opposite direction links. The WDM mesh network is planned and designed so that to satisfy all its demands using the shortest path algorithm and each connection has to be protected with the dedicated way and based on the spare capacity which is allocated as a “dedicated” resource for sole use of the connection. The assignment of the suitable wavelengths for each connection is done using a) the free available wavelength after the maximum busy (higher index) wavelength of each optical link from full complementary working and protection lightpaths, b) the first free (lower index) available wavelength of each optical link from full complementary working and protection lightpaths and c) the benefits of the second method versus the first one also showed when wavelength conversion is not used as well as the performance improvement when the wavelength conversion is used. The demand tables are large and they are not showed.Table I with symbols is showed below. The network has identical nodes. Each node can be assumed to have two functionalities: first, a lightpath or connection request generation/termination capability and, second, a wavelength routing capability. This essentially means that a node can either act as the source/destination node of a lightpath or as wavelength routing node. On the network nodes are installed the Optical Cross Connects (OXCs). The Wavelength Division Multiplex- Optical Cross Connect (WDM - OXC) has multiplex and demultiplex systems that convert the aggregated optical signal to simple optical signals and vice versa. A lightpath is an optical channel from source to destination to provide a connection between these nodes and using a same free wavelength on all of the fiber links in the path.

TABLE I. THE SYMBOLS

Symbol	Comments
q	The node set element number
p	The edge set element number
G(V,E)	The network graph
V(G)	The network node set
E(G)	The network edge set
2p	The number of links
n	The number of source – destination nodes pairs of the network
(S_n,D_n)	The order pairs of the node pairs
Xn	A column matrix (nx1) with elements the connection group size of the corresponding source-destination node pairs and corresponds to the successful requests for connection.
n(i)	The total number of the connection groups that passes through the fiber (i) and means that each fiber has different number of connection groups pass through it
k	The number of the wavelengths channels on each fiber that is the WDM system capacity
Yw	The column matrixes (2px1) with the working wavelengths of network links.
Aw	Matrix (2p x n) which shows the network active links that pass working lightpaths
$aw_{i,j}$	Element of the matrix Aw and takes the value one if the node pair (j) passes all its primary connections from the fiber (i) and zero (0) if no passes
Adp	Matrix (2p x n) which shows the network active links that pass protection lightpaths
$adp_{i,j}$	Element of the matrix Adp and takes the value one if the node pair (j) passes all its backup connections from the fiber (i) and zero (0) if no passes
Ydp	The column matrixes (2px1) with the dedicated protection wavelengths of network links
A	Matrix (2p x n) which shows the network active links that pass lightpaths
$a_{i,j}$	Element of the matrix A and takes the value one if the node pair (j) passes all its connections from the fiber (i) and zero (0) if no passes
λmax,i	The maximum (highest index) busy wavelength of each optical link
m	The shared protection WDM and optical fiber systems of each link
N	The WDM and optical fiber systems of each link
Tw	The total working WDM and fiber systems
Tp	The total protection WDM and fiber systems

This is called wavelength continuity constraint. An optical channel passing through a cross-connect node may be routed from an input fiber to an output fiber on the same wavelength. It is assumed that no different wavelengths are assigned on all links along the route if nodes have not wavelength conversion capabilities. It means that the initial wavelength which carries the traffic does not shift to other wavelengths by intermediate nodes of the lightpath. The networks with this capability are called “Networks without wavelength conversion”. If nodes have wavelength conversion different wavelengths can assign on all links along the route for each lightpath. It means that the initial wavelength which carries the traffic can shift to other wavelengths by intermediate nodes of the lightpath. The networks with this capability are called “Networks with wavelength conversion”. For this example, the connection group size of each node pair is set to the number two (2).

B. The formulation

To make the problem computationally feasible, the problem is generally divided into subproblems, the working lightpath assignment and the backup lightpath assignment. But the lightpath assignment is different for each method. At the first method, the free available wavelength after the maximum busy (higher index) one on all of the links is assigned and at the second one, the first free (lower index) available wavelength on all of the links is assigned (if wavelength conversion is not used these wavelengths must keep along each lightpath but if wavelength conversion is used it is not valid). So each different assignment method produces different needs for optical fibers. So the maximum (higher index) occupied wavelength of the first method for each fiber is greater than the second one (lower index). So the critical factor is the number of wavelengths required to satisfy the network demands.

The solution of the planning and designing problem is based on the following equations.

$$Yw = Aw * Xn \qquad (1)$$

Aw is a matrix that shows the active optical fiber network links (*2p*) from which the (*n)* working connection groups pass so its dimension is (2p x n), Y_w the column matrix (2p x 1) which has elements the working busy capacity of each optical fiber network link and Xn the column matrix (n x 1) which has elements the connection group size of each node pair. The total working wavelengths for all links (*TYw*) are given below but the total working wavelengths of each link is the term in the bracket.

$$TYw = \sum_{i=1}^{2p} \left[\sum_{j=1}^{n} Aw\, i,j * Xj \right] \qquad (2)$$

The knowledge of each node pair demands which are its requests for connection and their shortest full disjoint dedicated protection lightpaths create the necessary wavelengths for their satisfaction for each link.

$$Ydp = Adp * Xn \qquad (3)$$

Adp is a matrix that shows the active optical fiber network links (*2p*) from which the (*n)* protection connection groups pass so its dimension is (2p x n), Y_{dp} the column matrix (2p x 1) which has elements the protection busy capacity of each optical fiber network link. The total dedicated protection wavelengths for all links (*TYdp*) are given below but the total dedicated protection wavelengths of each link is the term in the bracket.

$$TYdp = \sum_{i=1}^{2p} \left[\sum_{j=1}^{n} Adp\, i,j * Xj \right] \qquad (4)$$

$$A = Aw U Adp \qquad (5)$$

The unity of the matrixes *Aw* and *Adp* gives the matrix A in which there are common active links for working and protection lightpaths.

The total wavelengths are the following

$$TY = TYw + TYdp \qquad (6)$$

The maximum (higher index) busy wavelength of each optical link is λmax,i. The maximum (higher index) busy wavelength of all optical links is λmax and λmax= maximum (λmax,1, λmax,2, …, λmax,2p).

The total working WDM and fiber systems given by

$$Tw = \sum_{i=1}^{2p} \left[\frac{\lambda max,i}{k} \right] \qquad (7)$$

The total protection WDM and fiber systems given by

$$Tp = \sum_{i=1}^{2p} \frac{m}{N} \left[\frac{\lambda max,i}{k} \right] \qquad (8)$$

The parentheses mean the rounding. When the protection network is not used the term which multiplied with (m/N) is neglected and the available resources are less. The equation (8) means that multiplying the number of the necessary working WDM and optical fiber systems of each link with the m:N ratio creates the necessary protection WDM and optical fiber systems. The m:N=1:7 shared protection WDM and optical fiber systems of each link means that the maximum number of working WDM and optical fiber systems that sharing a protection WDM and optical fiber system is seven. It is a practical way to reduce the cost of the protection network. The roundup is always done for the larger integer. If there are not protections WDM and fiber systems the equation (8) is zero.

The total WDM and optical fiber systems are

$$T = Tw + Tp \qquad (9)$$

The wavelength protection ratio for dedicated protection is written below

$$PRd = \frac{TYdp}{TYw} \qquad (10)$$

C. Synoptic description of the methods

Two methods are used in this paper. These methods have two parts, the first part or the planning and designing part that means network without failure and the second part or network with failure. The algorithm of allocation path and routing uses a more traditional approach which is the shortest path algorithm. The synoptic description of these methods is showed in the table II. On the failure free network phase, the third step (wavelength allocation step) is the more critical and it is different to each method. The network has not nodes with wavelength conversion. The working connection starts from the source node and progresses through the network occupying a wavelength on each optical fibre and switch to another fibre on the *same* wavelength by OXC, according to its shortest working optical path up to arrive at the destination node. Simultaneously, the protection lightpath of the connection starts from the source node and progresses through the network occupying a wavelength on each optical fibre and switch to another fibre on the *same* wavelength by OXC, so another full disjoint protection optical path is obtained. So the full dedicated protection for this connection is obtained. The difference between “without wavelength conversion” and “with wavelength conversion” is the occupied wavelength of the link after switching that is the “same” with the wavelength if input, for the case of “without” and “anyone” for the case “with”.

The assignment problem of both algorithms can be used by both cases (nodes without or with wavelength conversion) suitably modified. The assignment problem of each algorithm is working as follows to keep the capability of nodes without conversion. For the first algorithm, each node pair uses the free available wavelength after the maximum busy (higher index) one which is the same for all optical links of the working path of the connection and the same is done for the protection lightpath of the connection. The first algorithm is adapted in the problem of nodes without wavelength conversion. In this scheme, all wavelengths are numbered. The searching for available wavelengths is done after the maximum busy wavelength and a higher numbered wavelength is considered. The first available wavelength is then selected. This scheme is requires no global information. The assignment problem for the second algorithm is as follows. For the second algorithm, for each node pair the *first* same (lower index) free available wavelength of all optical links of the working path of the connection is assigned and the same is done for the protection lightpath of the connection for the same node pair. If there is any such wavelength for all optical links, the connection is done by them. If no, the first algorithm is done. The second algorithm is the First Fit one adapted in the problem of nodes without wavelength conversion. In this scheme, all wavelengths are numbered. When searching for available wavelengths, a lower numbered wavelength is considered before a higher numbered wavelength. The first available wavelength is then selected. This scheme is requires no global information. The second algorithm is used for the problem of nodes with wavelength conversion.

When a failure occurs and an optical link cut, the working and protection WDM and fiber systems of this link are also cut and the network topology changes and protection lightpaths pass the traffic. Table II presents the synoptic description of the methods. Its worst case time complexity of each method depends of the network topology and the total number of connections. *It is O(t*q^2) where t the total number of the connections*. The second method needs about 11 100th of the second (0.11 seconds) time to consume but the corresponded first protection method needs only 5 100th of the second (0.05 seconds). The time difference of these protection methods is small but the first method is faster than the second one.

In graph theory, the shortest path algorithm finds the shortest path between two given vertices in an undirected graph G= (V, E).The shortest path connects the two vertices and its length is minimum.

TABLE II. THE SYNOPTIC PRESENTATION
OF THE METHODS

FIRST PART
First step. Network parameters reading (q, p, V(G), E(G), G(V,E), 2,2p, k)
Second step. Connection selections (n, (S_n,D_n), Xn,)
Failure-free Network Phase
Third step. Wavelength allocation (Routing and wavelength assignment method)
Forth step. Results (Yw, Ydp, (λmax,i) ,(Tw,i,) Tw, (Tp,i), Tp)
SECOND PART
With failure Network Phase
Fifth step. Network parameter modifications (cut link, q, p, V(G’), E’(G’), G’(V,E’), 2,2p-1, k)
Sixth step. Traffic is passing by Protection lightpath (Protection method)
Seventh step. New Results (Y’w, Y’dp, (λ’max,i), (Tw,i), T’w, (Tp,i), T’p)

D. Example

It is assumed that the topology of the network is presented by the graph G(V,E). This mesh topology is used because it is a simple, palpable and it is easy to expand to any mesh topology. The vertex set has q=8 elements which are V= {v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8} and the edge set has p=12 elements which are E = {e_1, e_2, e_3,…, e_8, e_9, e_{10}, e_{11}, e_{12}}. Each edge has two optical links of opposite directions with their fibers for each direction. The connections of each node pair form connection groups according to its shortest path and transverse the network. Figure 1 presents the mesh topology.

Network planning and designing is flexible to meet the needs of the network. The capacity of WDM optical fiber system takes values of 8, 16 and 32 OCh (wavelengths). The number of node pairs is n=8*(8-1) =56. The source destination demands are not showed because their tables are large. The m:N WDM and optical fiber system shared protection is 1:7, which is a practical way to reduce the protection network. No further calculations and results are presented for the network planning and designing because their tables are very large and their meaning to solve the protection problem is small. The symbolism V_1,V_2,V_3 is usually used as follows on optical path layer (OCh layer). If the nodes have no wavelength conversion capability, it means that the wavelength of the input port wavelength is used at the output port. So a connection wavelength of the optical link <V_1, V_2> routes at the same wavelength one to the optical link <V_2, V_3> by WDM-OXC of node V_2 and thus, it is written V_1, V_2, V_3. If the nodes have wavelength conversion capability means that the wavelength of the input port wavelength is the same or different at the output port. The dedicated protection is done during the planning and designing steps allocating to each connection two links full disjoint lightpaths between the source and the destination node, one for the working lightpath and the other for the protection lightpath. If a failure occurs, the infrastructure will change and the protection lightpath passes the traffic. The capacity of the WDM system is 8,16 and 32 optical channels and each node pair has two connections. The total working wavelengths are 200 and the total protection wavelengths are 304, for each case of table 4. So the total busy wavelengths are 504. The wavelength protection ratio is 1.52. The <V_1, V_2> is assumed as a cut link. The table III shows the node pairs connections that pass through the cut link. The bold lightpaths are cut.

For the assignment (nodes without wavelength conversion), the tables which show the procedures and the results of each algorithm are large, so I don't show them. The connection of each node pair forms a connection group and by the both working and protection lightpaths proceed from source node to its destination node and by allocating a common free wavelength on all of the fiber links in the working lightpath and one corresponded in the protection lightpath. The entire bandwidth available on each lightpath (working or protection) is allocated to this connection during its holding time and the corresponded wavelengths cannot be allocated to any other connection. When a connection is terminated, the associated lightpaths are torn down and the wavelengths become free once again on all links along the routes. All connection group sizes are equal to two (2). The second algorithm has $\lambda_{max,i}$ wavelength for each link which is lower or equal than the corresponded of the first one and it means that equation (9) gives smaller or equal number WDM systems for the second algorithm versus first one. The difference of the performance is ought to the maximum wavelength index difference of two algorithms. The lightpath length is an important factor that effects on performance. The table IV shows the performance of each case as number of WDM and fiber systems. The performance of first algorithm is showed in the table IV(x). The performance of second algorithm if can find lower index free wavelengths for lightpaths with only one hop length, is showed in the table IV(1). The performance of second algorithm if can find lower index free wavelengths for lightpaths with only one and two hop length, is showed in the table IV(2). The performance of second algorithm if can find lower index free wavelengths for lightpaths with only one, two and three hop length, is showed in the table IV(3). The performance of second algorithm if can find lower index free wavelengths for lightpaths with only one, two, three and four hop length, is showed in the table IV(4). The performance of second algorithm if can find lower index free wavelengths for lightpaths with only one, two, three, four and five hop length, is showed in the table IV(5).Table IV(y) shows the improvement of performance if the second algorithm is used *with wavelength conversion*.

TABLE III. THE NODE PAIR CONNECTIONS THAT PASS THROUGH THE CUT LINK

A /A	**Node Pair**	**Working Lightpath**	**Protection Lightpath**
1	[V_1,V_2]	**V_1,V_2**	V_1,V_3,V_2
2	[V_1,V_3]	V_1,V_3	**V_1,V_2,V_3**
3	[V_1,V_5]	**V_1,V_2,V_5**	V_1,V_3,V_6 ,V_5
4	[V_1,V_6]	V_1,V_3,V_6	**V_1,V_2,V_5,V_6**
5	[V_1,V_7]	**V_1,V_2,V_5 ,V_7**	V_1,V_3,V_6,V_7
6	[V_1,V_8]	V_1,V_3,V_6,V_8	**V_1,V_2,V_5,V_7,V_8**
7	[V_3,V_2]	V_3,V_2	**V_3,V_1,V_2**
8	[V_4,V_2]	**V_4,V_1,V_2**	V_4,V_3,V_2
9	[V_4,V_5]	**V_4,V_1,V_2,V_5**	V_4,V_3,V_6,V_5
10	[V_4,V_6]	V_4,V_3,V_6	**V_4,V_1,V_2 ,V_5,V_6**
11	[V_4,V_7]	V_4,V_3,V_6,V_7	**V_4,V_1,V_2 ,V_5,V_7**
12	[V_4,V_8]	V_4,V_3,V_6,V_8	**V_4,V_1,V_2,V_5,V_7, V_8**

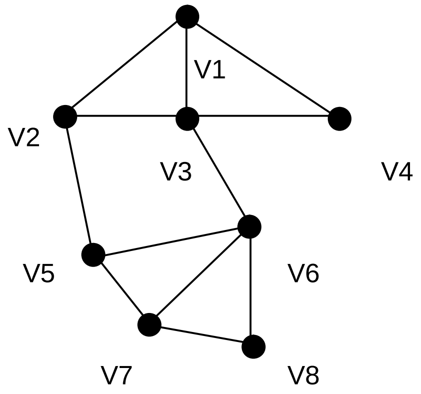


Figure 1.The mesh topology of the network

E. Discussion and Proposals

Protection strategies are critical for optical mesh networks. Although dedicated path protection mechanisms are simple and fast, they use 100% or more redundant capacity. Protection is the primary mechanism to deal with a failure and it is faster but no flexible. Spare wavelengths on the routes are dedicated or shared across working path demands that have no spans in common. It is not need to know the exact location of the failure because it is capable of protecting against multiple simultaneous failures on suitable working lightpaths. This protection method can give to the network surveillance after single failure as optical link cuts, node failures, wavelength scratches and multiple failures on working connection lightpaths. It is also uses the coherent protection of the mesh network. If the network is planned and designed to satisfy all its needs with protection paths, it means that no blocking probabilities should be calculated. The algorithm of the assignment is critical because it could be saved significant number of WDM and fiber systems and to do the network design less expensive. The second assignment algorithm has better performance than the first one (table IV, case (5) versus case (x)) because for each optical link its λmax is less or equal than the corresponded of the first one and produces smaller number of WDM and fibre systems. The other cases (1),(2),(3) and (4) give the transition to case (5). The wavelength conversion improves more the performance because reduced number of WDM and fiber systems is required, table IV(y). Its disadvantage is the usage of wavelength converters that are expensive. In a mesh network with wavelength conversion capability, each output port of the optical switch is associated with a dedicated wavelength converter. So it is able to convert all the input wavelengths to any other wavelengths without any limitation. This wavelength conversion method is called complete one. The number of converters is equal to the number of the fiber links multiplied by the number of wavelengths per fiber. So the number of converters will be large and the cost of such architecture can be high. The converters can be incorporated in the OXCs. The second algorithm of assignment is a variety of the First Fit. First Fit chooses the available wavelength with the lowest index. This is adjusted for both problems (nodes without and with wavelength conversion). When the capacity of WDM systems increases the number of these systems may be same, table IV, cases IV(4), IV(5) and IV(y) but their occupation percentage is different.

TABLE IV. THE COMPARISON OF THE PERFORMANCE

Up to Lightpath length	**WDM systems Capacity, 8λ**	**WDM systems Capacity, 16λ**	**WDM systems Capacity, 32λ**
x	224	124	83
1	208	115	75
2	143	85	61
3	124	78	54
4	118	72	48
5	118	72	48
y	92	64	48

IV. CONCLUSIONS

In this paper, the problem of using the shortest path algorithm to plan and design a WDM mesh network with dedicated protection path is showed. The wavelength assignment problem is studied by two algorithms. The first one is looking for available wavelengths out of the busy wavelength space and the second one in the busy wavelength space. The second algorithm has significant benefits versus the first one because it uses fewer resources than the other one when a dedicated protection lightpath is used with the shortest path algorithm for both lightpaths. The first algorithm is faster than the second one. But the conclusion is that both approaches are suitable for the telecommunication traffic protection for fault management. The selection of the suitable method is based on wavelength availability of the failure site. The wavelength conversion method has also better performance than without wavelength conversion method but it is more expensive and the optimization is obtained if this problem is studied.

REFERENCES

[1] H. Kobayashi, Modeling and Analysis, Addison -Wesley, 1981.

[2] F. L. Bauer et al., Software Engineering An Advanced Course, Springer –Verlag, 1973.

[3] T. Wu, Fibre Network Service Survivability, Artech House, 1992.

[4] A. Bononi, Optical Networking, Part 2, Springer, 1992.

[5] V. E. Benes, Mathematical Theory of Connecting Networks and Telephone Traffic, Academic Press, 1965.

[6] B. Ramamurthy and B.Mukherjee, “Wavelength Conversion in WDM Networking”, JSAC, Vol 16, No 7, pp. 1061-1073, September 1998.

[7] J. Emirghani and H. Mouftah, “All-Optical Wavelength Conversion: Technologies and Applications in DWDM Networks”, IEEE Comms Magazine, Vol 38, No 3, pp. 86-92, March 2000.

[8] C. Xiaowen, L. Jiangchuan and Z. Zhensheng, “Analysis of Sparse –Partial Wavelength Conversion in Wavelength-Routed WDM Networks”, Infocom 2004, Vol 2, pp. 1363-1371, Hong Kong, March 7-11, 2004.

[9] M. O’Mahony, C. Politi, D. Klonidis, R. Nejabati and D. Simeonidou, “Future Optical Networks”, IEEE Journal of LightWave Technology, Vol 24 , No 12, pp. 4684-4696, December 2006.

[10] J. Zhang, K. Zhu, L. Sahasrabuddhe, S. J. B. Yoo and B. Mukherjee, “On the study of routing and wavelength assignment approaches for survivable wavelength routed WDM mesh networks”, Optic Networks Magazine, pp. 16-27, November/ December 2003.

[11] G. Tsirakakis, and T. Clarkson, “Simulation tools for multilayer fault restoration”, IEEE Comms Magazine, Vol 47, No 3, pp. 1128-134, March 2009.

[12] T. Ingham, S. Rajhans, D. K. Sinha, K. Sastry and S. Koumar, “Design validation of service delivery platform using modeling”, *IEEE*, Comms Magazine, Vol 47, No 3, pp. 135-141, March 2009.

[13] J. Burbank, “Modeling and Simulation: A practical guide for network designers and developers”, IEEE Comms Magazine, Vol 47, No 3, pp. 118, March 2009.

Security Attack based on Control Packet Vulnerability in Cooperative Wireless Networks

Ki Hong Kim
The Attached Institute of ETRI
Daejeon, Korea
e-mail: hong0612@ensec.re.kr

***Abstract*—Cooperative wireless communication has been proposed as a way to improve channel capacity, robustness, reliability, delay, and coverage. Multiple research works have been done to support cooperative communication in the medium access control (MAC) layer. Synergy MAC is one of the MAC protocols that support cooperative communication using cooperative relay nodes. In this paper, some security attacks against control packets of Synergy MAC are identified and the potential security issues that arise in Synergy MAC due to these attacks are also discussed.**

Keywords–Synergy MAC; cooperative communication; control packet; security attack; security issue.

I. Introduction

Within the last ten years, cooperation communication in wireless networks has received significant attention. Cooperative wireless communication is an innovative communication scheme that takes advantage of the open broadcast nature of the wireless medium and the spatial diversity to achieve performance gain. It is also known to be essential for making ubiquitous communication connectivity a reality. In the cooperative wireless networks, when the source node transmits data to the destination node, some nodes that are close to source node and destination node can serve as relay nodes by forwarding replicas of the source's data. The destination node receives multiple data from the source node and the relay nodes and then combines them to achieve performance and quality improvement [1][2][3].

There are three major schemes employed by the relay node to forward data to the destination node: amplify-and-forward (AF), decode-and-forward (DF), and compress-and-forward (CF). In AF scheme, relay node receives a noisy version of the transmitted original data and then amplifies and retransmits this noisy data to the destination node. On the other hand, in DF scheme, relay node decodes data transmitted by the source node and then retransmits the decoded data to the destination node. Finally, CF scheme works by forwarding incremental redundancy of original data by the relay node to the destination node [1][4].

Several protocols in the MAC layer have been proposed to utilize the concept of cooperative transmission. A typical example is Synergy MAC protocol [5][6]. Synergy MAC is an IEEE 802.11b [7] based cooperative MAC protocol for mobile ad hoc networks. Synergy MAC was proposed to take advantage of cooperation, while remaining backward compatible with legacy IEEE 802.11b. This protocol is able to alleviate the ill effects of signal fading by realizing spatial diversity and transmit data at rates higher than otherwise possible by allowing nodes with low signal-to-noise ratio (SNR) to destination utilize intermediate relay nodes. It also outperforms standard IEEE 802.11b and mitigate some of the fairness problems caused by multiple modulation schemes.

Security is a principal issue that must be resolved in order for the potential of cooperative wireless networks to be fully exploited. However, security issues related to the design of cooperative wireless networks have largely not been considered. In this paper, a comprehensive study of security attack based on control packet vulnerability in Synergy MAC is presented. Security issues at each handshaking procedure while attacking the control packets such as request-to-send (RTS) and clear-to-send (CTS) is analyzed and discussed. This work differs from previous works in that it concentrates on one significant aspect of a security issue in the Synergy MAC, namely security issue of Synergy MAC caused by attack against the control packets at handshaking mechanism.

The remainder of this paper is organized as follows. In Section II, I present some related works and security issues on cooperative wireless networks. In Section III, I give a brief description of the Synergy MAC protocol. In Section IV, I identify some possible security attacks against control packets of Synergy MAC and then discuss the security issues caused by these attacks. Finally, in Section V, I conclude the paper and present plans for future work.

II. Related Works

Due to the rapidly increasing popularity of cooperative wireless networks, there have been multiple research works regarding cooperative communication protocols and security issues for cooperative wireless networks. The work in [1] described cooperative wireless communication that enables single antenna mobiles to share their antennas. The [2] proposed and analyzed opportunistic relaying as a practical scheme that forms a cooperative diversity. The [3] introduced

ISBN: 978-1-61208-256-1

an adaptive relay selection on demand with early retreat scheme to reduce the overall energy consumption significantly.

Some MAC protocols have been suggested to support cooperative transmissions in wireless networks. In [8], a new MAC protocol for the IEEE 802.11 [9], namely CoopMAC, was proposed and its performance was also analyzed. The Synergy MAC, an IEEE 802.11b [7] based cooperative MAC protocol for mobile ad hoc networks was studied in [5]. Also, COSMIC, a carrier sense multiple access/collision avoidance (CSMA/CA) based cooperative MAC protocol for wireless sensor network (WSN) with minimal control messages was proposed in [10], and cooperative MAC protocol of alleviating the problem from a pure MAC centric perspective, called CMAC was introduced to provide immediate improvements to the IEEE 802.11e [11] efficiency [4]. The [12] suggested a distributed MAC protocol, which uses an automatic relay selection with embedded relay collision avoidance and three-way handshaking to minimize signaling overhead.

Cooperative wireless communications are vulnerable to security attacks due to the open broadcast nature of the wireless communication channel and the cooperative transmissions with multiple transmitters. Several research groups have studied security issues including attacks, vulnerabilities, and mechanisms in cooperative wireless networks. The [13] formulated cooperative mechanisms for wireless networks with cooperative relays which help to give provable unconditional secrecy guarantees, while the [14] developed a framework for evaluating the trade-off between using cooperative transmissions or non-cooperative transmissions in sensor networks with a mix of malicious and non-malicious nodes. The [15] presented the distributed trust-assisted cooperative transmission mechanism handling relay's misbehavior as well as channel estimation error. The [16] described a security framework for leveraging the security in cognitive radio cooperative networks. The security vulnerabilities found in traffic adaptive cooperative wireless sensor-MAC (CWS-MAC), a flow specific medium access scheme were identified and analyzed in [17]. The work in [18] studied the coordinated denial of service (DoS) attacks against data packets using the concept of cooperative game theory on IEEE 802.22 [19] from the malicious nodes' perspective. The [20] proposed a detection technique of misbehaving nodes either based on the uniform most powerful (UMP) test or on the sequential probability ratio test (SPRT) in networks using CoopMAC and automatic repeat request (ARQ) protocols. The security concerns on data packets that a Synergy MAC introduces due to its reliance on a third party relay were discussed in [6]. Similarly, the potential security issues and vulnerabilities that arise in CoopMAC were addressed in [21][22].

In spite all the above mentioned researches, there is still no work that analyzes the security issues caused by the security attacks against control packets in the Synergy MAC. Most of the previous works are focused on efficient and reliable cooperative transmission scheme using the relay node and identification of general security issues caused by the malicious relay node. In this paper, I discuss the potential security issues that arise in Synergy MAC due to security attacks against control packets. This work is the reasonable attempt to analyze and compare security issues from possible security attacks based on control packets vulnerabilities in Synergy MAC.

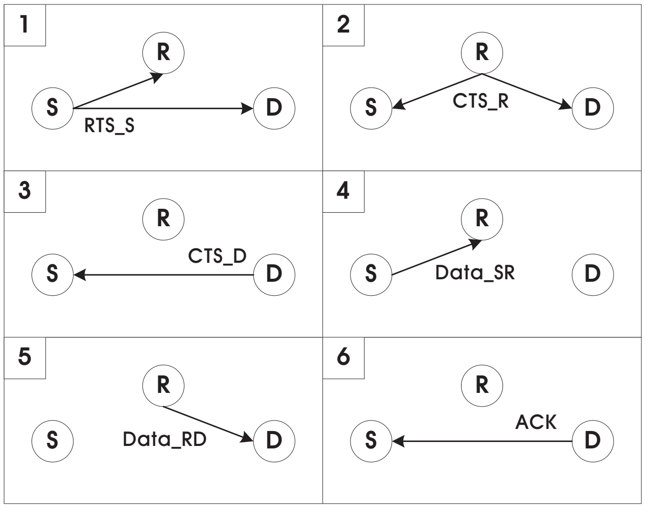


Figure 1. Handshaking mechanism followed by control packets exchange in Synergy MAC protocol.

III. SYNERGY MAC

Synergy MAC is a MAC protocol based on the IEEE 802.11b's distributed coordination function (DCF) mechanism to realize cooperative transmission at the physical layer. It employs control packets like RTS and CTS for sensing the wireless medium to determine if it is free. The Synergy MAC is completely compatible with IEEE 802.11b and can be easily extended to suit other version of the legacy IEEE 802.11. It achieves higher rates of data transmission than IEEE 802.11b despite leveraging on the multi-rate capability of IEEE 802.11b.

The three-way handshaking procedure for Synergy MAC is depicted in Fig. 1. When a source node (S) wants to send data packets to destination node (D), it first senses the wireless channel condition, busy or idle. If the channel is idle, S sends the RTS packet (RTS_S) to the D, reserving the channel for network allocation vector (NAV) duration needed to transmit data packets. If not, S should wait the channel is idle and then send the RTS_S. When a relay node (R) overhears RTS_S transmission and decodes it successfully, it broadcasts a self addressed CTS packet (CTS_R). When the D receives a CTS_R from R soon after receiving a RTS_S from S, it sends CTS packet (CTS_D) to the S. This CTS_D is used to reserve the

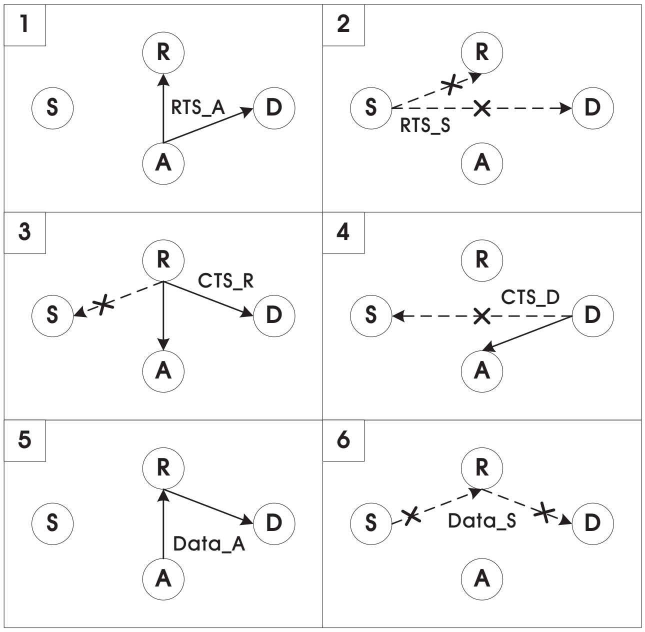


Figure 2. Source attack: false RTS transmission to relay and destination.

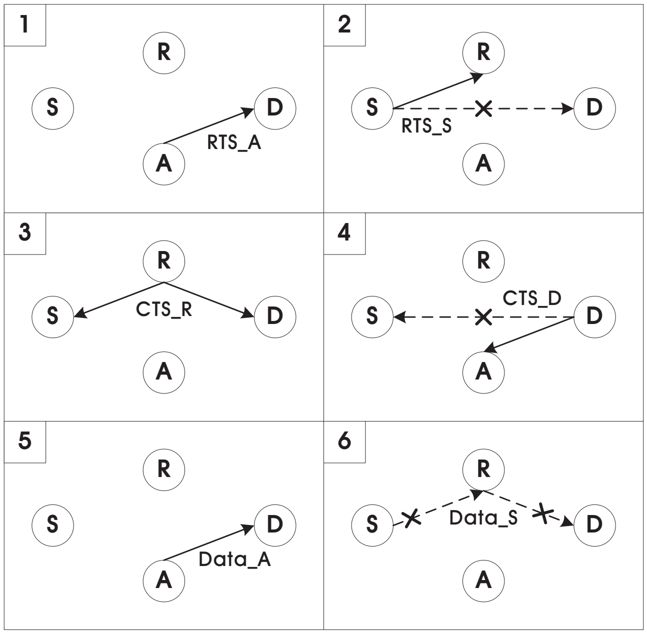


Figure 3. Source attack: false RTS transmission to destination.

channel for cooperative communication via the R. Once S receives the CTS_R from the R and the CTS_D from the D respectively, it starts transmitting its data packets ($Data\_SR$) to R. R then forwards the data packets ($Data\_SR = Data\_RD$) received from S to D. After D successfully receives $Data\_RD$ from R, it sends an acknowledgement packet (ACK) to S. Otherwise, D sends a negative acknowledgement packet ($NACK$), notifying S of the failure of cooperative transmission between S and D via the R. In addition, if S receives no response from D within a specific timeout period, it will also notice the failure of transmission to D. Data transmission cycle in Synergy MAC is complete when the S receives the ACK from the D. More details on Synergy MAC may be found in [5][6].

IV. Security Attack and Issue in Synergy MAC

Due to broadcast nature of the wireless transmission and cooperative transmission, Synergy MAC suffers from various attacks. For example, in Fig. 1, let's assume attacker node is closer to S than D or it is between the S and the D. In this environment, attacker node can disguise itself as D and respond with its CTS packet to S. There is no suitable countermeasure to prevent this attack and solution to authenticate D. Therefore, an attacker node close to the victim nodes can respond with a CTS packet to them thus it results in disruption of the normal cooperative transmission between nodes. The attackers' goal is focused on the network's performance, that means they want to disturb the communication between source node and destination node. They would exploit the weakness in cooperative procedure, especially in the control packets exchange, and disguise themselves as legitimate relay nodes to disturb the network's operation and to degrade the communication quality.

Security attacks based on the control packets resulting from attacker nodes can be classified into two categories: (1) false RTS attack and (2) false CTS attack. The former generates a false RTS packet in order to create the virtual jamming, while the latter generates a false CTS packet in order to disguise attacker as legitimate relay node or destination node. The followings introduce these attacks according to the control packets of Synergy MAC in greater detail.

A. Source Attack using False RTS

The first security attack is that of virtual jamming by an attacker node which deliberately sends false RTS packet to relay node and destination node. Let us take the case of Fig. 2. As shown in Fig. 2, attacker node (A) sends the false RTS packet (RTS_A) to relay node (R) and destination node (D). A then waits for the CTS packet (CTS_R) from R and CTS packet (CTS_D) from D. RTS_A causes R and D to deny legal RTS packet (RTS_S) from source node (S). This means that because R and D have already received the RTS packet from A, they reject the additional RTS packet from S. Once A receives the CTS_R and the CTS_D, it starts transmitting its false data packets ($Data\_A$) to the R. Thus, this attack blocks the transmission of the RTS_S and the data packets ($Data\_S$) from S. Consequently, S can not start its data packets transmission to R.

Next, A sends the RTS_A to only D. This scenario is depicted in Fig. 3. Since the authentication(or integrity) mechanism is not applied to the control packets exchange

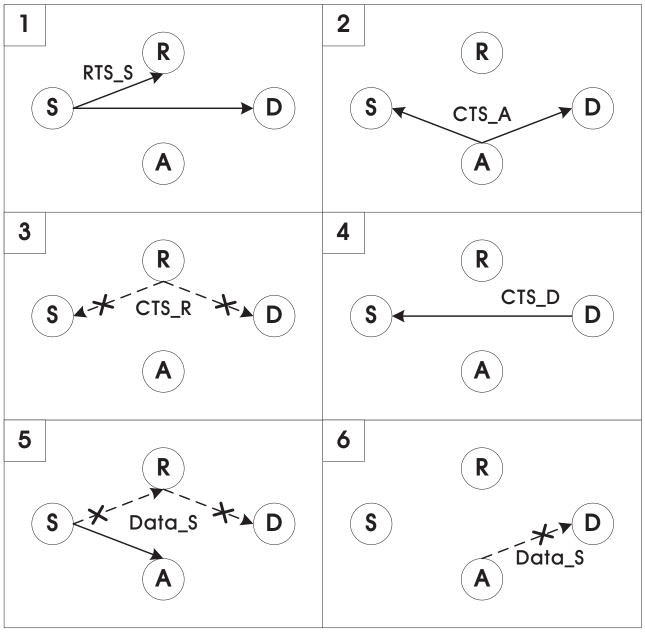


Figure 4. Relay attack: false CTS transmission to source and destination.

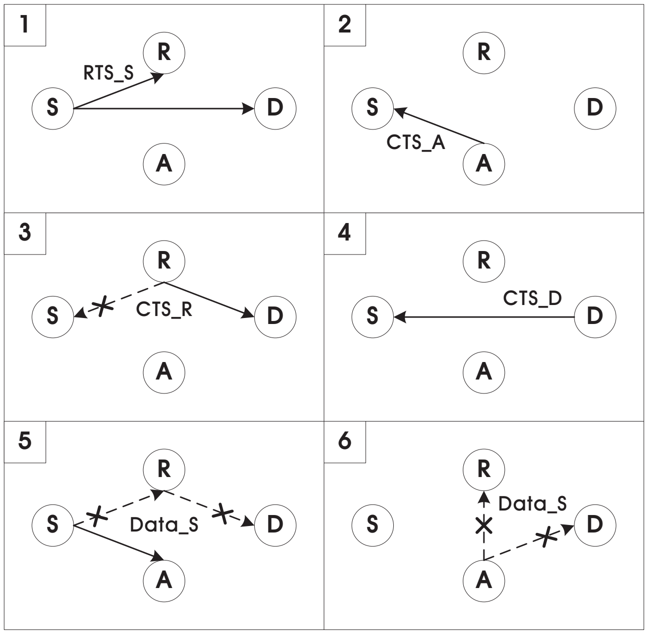


Figure 5. Relay attack: false CTS transmission to source.

between S and D, the legal RTS_S from S can be rejected by D due to an illegal previous RTS_A received from A. Accordingly, CTS_D is sent from the D to the A, not S. This means that the S continuously waits for the CTS_D from the D to finish the handshaking process. As a result, normal cooperative communication between S and D can not be guaranteed.

B. *Relay Attack using False CTS*

The second security attack is that of false CTS packet sending by the attacker node in order to disturb the relay node. An attacker node may try to deny relay node's legal CTS packet to source node and/or destination node by sending the false CTS packet, causing the source node and/or destination node to reject a legal CTS packet from relay node. As shown in Fig. 4, the false CTS packet (CTS_A) is sent from attacker node (A) to source node (S) and destination node (D). Accordingly, the legal CTS packet (CTS_R) from relay node (R) is denied by S and D. Then, D sends its CTS packet (CTS_D) to S. After receiving the CTS_A and CTS_D, S starts data packet ($Data\_S$) transmission to A, but R. A deliberately stops forwarding $Data\_S$ to R, which results in DoS attack caused by A. Due to this false transmission to A, cooperative communication between S and D via R is not established.

Fig. 5 illustrates another attack from attacker node's illegal CTS packet in the Synergy MAC. In the case of sending illegal CTS packet (CTS_A) to only source node (S), since the S is typically not come to know of this, although the legal CTS packet (CTS_R) is sent from relay node (R) to S, it is denied by S. Then, the destination node (D) sends a CTS packet (CTS_D) to S in order to notify that it successfully received the CTS_R. This also means that attacker node (A) is an intended legitimate relay node forwarding data packets ($Data\_S$). Therefore, S sends $Data\_S$ to A, not R. Finally, A denies cooperative communication to the S by simply dropping the $Data\_S$ it receives from S.

The potential relay node attack using illegal CTS packet is also shown in Fig. 6. Since the destination node (D) receives the illegal CTS packet (CTS_A) from attacker node (A), it rejects the legal CTS packet (CTS_R) from relay node (R). After receiving the CTS packet (CTS_D) from D, source node (S) sends its data packets ($Data\_S$) to R. If R receives the $Data\_S$ from S, it doesn't forward $Data\_S$ to D, but forwards it the A. A drops the $Data\_S$ received from R. It also spoofs an ACK, causing the S to wrongly conclude a successful cooperative transmission via R.

C. *Destination Attack using False CTS*

Fig. 7 shows a destination node attack which caused by the illegal CTS packet from attacker node. In this case, the attacker node (A) transmits a false CTS packet (CTS_A) to source node (S), informing the S that it is an intended recipient of future data packets ($Data\_S$). And, since the authentication(or integrity) mechanism is not applied to CTS_A, the legal CTS packet (CTS_D) from destination node (D) can be rejected by S due to a previous illegal CTS_A from A. Just after receiving the CTS_A from A, S transmits $Data\_S$ to relay node (R). Subsequently, the R receives the $Data\_S$ and then forwards it to A. The A may try to deny cooperative communication to S

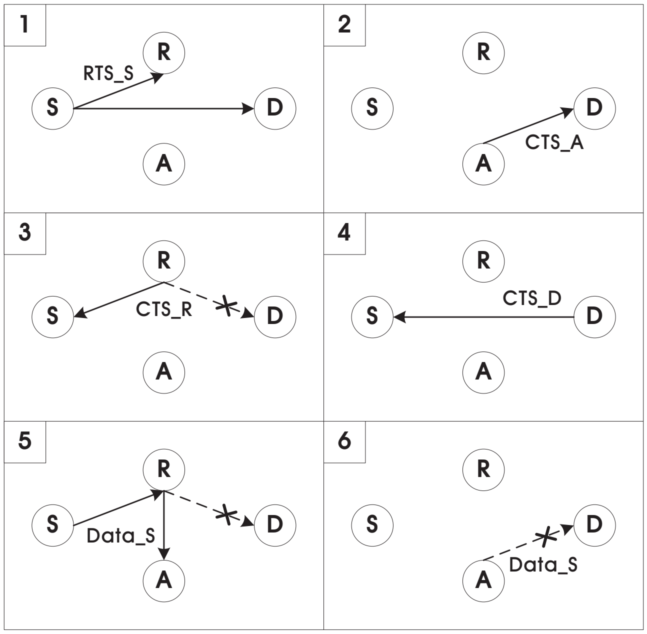


Figure 6. Relay attack: false CTS transmission to destination.

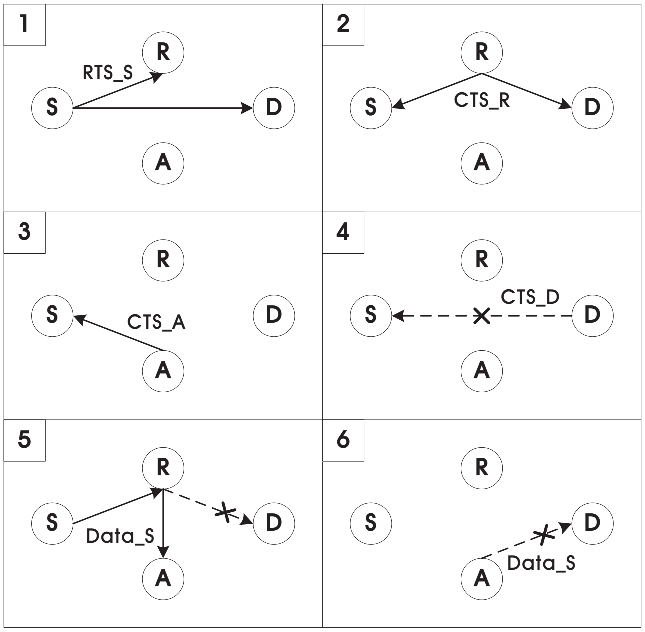


Figure 7. Destination attack: false CTS transmission to source.

by deliberately not forwarding $Data\_S$ received from R. Consequently, cooperative communication between S and D is not established.

V. Conclusion

This paper presented the case study of security attacks based on control packets (RTS and CTS) vulnerabilities in Synergy MAC. Furthermore, it analyzed security vulnerabilities at each handshaking stage while attacking control packets exchanged among nodes (source, destination, and relay). This study is the comprehensive analysis of security vulnerabilities caused by attacker node in Synergy MAC. It can be significant in the use of design of efficient authentication solutions for secure Synergy MAC. The analytical results can be extended to not only cooperative wireless network security, but also WSN security design in general.

As future work, the author plans to design and implement lightweight low-power authentication mechanism suitable for cooperative wireless networks. The plan is then to examine some effects with security cost, power consumption, and transmission performance using the proposed mechanism.

References

[1] A. Nosratinia, T. E. Hunter, and A. Hedayat, “Cooperative Communication in Wireless Networks,” IEEE Communication Magazine, Vol. 42, pp. 74–80, 2004.

[2] A. Bletsas, A. Khisti, D. P. Reed, and A. Lippman, “A Simple Cooperative Diversity Method Based on Network Path Selection,” IEEE Journal on Selected Areas in Communications, Vol. 24, pp. 659–672, 2006.

[3] H. Adam, C. Bettstetter, and S. M. Senouci, “Adaptive Relay Selection in Cooperative Wireless Networks,” IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, pp. 1–5, 2008.

[4] S. Shankar, C. T. Chou, and M. Ghosh, “Cooperative Communication MAC (CMAC) – A New MAC Protocol for Next Generation Wireless LANs,” IEEE International Conference on Wireless Networks, Communications and Mobile Computing, pp. 1–6, 2005

[5] S. Kulkarni, P. S. Prasad, and P. Agrawal, “Enabling Cooperation in Mobile Ad Hoc Networks,” IEEE Sarnoff Symposium, pp. 1–5, 2009.

[6] S. Kulkarni and P. Agrawal, “Safeguarding Cooperation in Synergy MAC,” IEEE Southeastern Symposium on System Theory, pp. 156–160, 2010.

[7] IEEE Std. 802.11b–1999, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: High Speed Physical Layer Extension in the 2.4GHz Band, 1999.

[8] P. Liu, Z. Tao, and S. Panwar, “A Cooperative MAC Protocol for Wireless Local Area Networks,” IEEE International Conference on Communications, pp. 2962–2968, 2005.

[9] IEEE Std. 802.11–1999, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, 1999.

[10] A. B. Nacef, S. M. Senouci, Y. Ghamri-Doudane, and A. L. Beylot, “COSMIC: A Cooperative MAC Protocol for WSN with Minimal Control Messages,” IFIP International Conference on New Technologies, Mobility and Security, pp. 1–5. 2011.

[11] IEEE 802.11e/D4.0, Draft Supplement to Part 11: Wireless Medim Access Control (MAC) and Physical Layer (PHY) Specifications: Medium Access Control (MAC) Enhancements for Quality of Service (QoS), 2002.

[12] C. T. Chou, J. Yang, and D. Wang, “Cooperative MAC Protocol with Automatic Relay Selection in Distributed Wireless Networks,” IEEE International Conference on Pervasive Computing and Communications Workshops, pp. 526–531, 2007.

[13] E. Perron, S. Diggavi, and E. Telatar, “On Cooperative Wireless Network Secrecy,” IEEE Conference on Computer Communications, pp. 1935–1943, 2009.

[14] A. Aksu, P. Krishnamurthy, D. Tipper, and O. Ercetin, “On Security and Reliability Using Cooperative Transmission in Sensor Networks,” IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing, pp. 1–10, 2010.

[15] Z. Han and Y. L. Sun, “Securing Cooperative Transmission in Wireless Communications,” IEEE International Conference on Mobile and Ubiquitous Systems: Networking & Services, pp. 1–6, 2007.

[16] H. Marques, J. Ribeiro, P. Marques, A. Zuquete, and J. Rodriguez, “A Security Framework for Cognitive Radio IP Based Cooperative Protocols,” IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications, pp. 2838–2842, 2009.

[17] T. O. Walker III, M. Tummala, and J. McEachen, “Security Vulnerabilities in Hybrid Flow-specific Traffic-adaptive Medium Access Control,” IEEE Hawaii International Conference on System Sciences, pp. 5649–5658, 2012.

[18] Y. Tan, S. Sengupta, and K. P. Subbalakshmi, “Analysis of Coordinated Denial-of-Service Attacks in IEEE 802.22 Networks,” IEEE Journal on Selected Areas in Communications, Vol. 29, pp. 890–902, 2011.

[19] IEEE P802.22/D0.1, Draft Standard for Wireless Regional Area Networks Part 22: Cognitive Wireless RAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Policies and Procedures for Operation in the TV Bands, 2006.

[20] S. Dehnie and S. Tomasin, “Detection of Selfish Nodes in Networks Using CoopMAC Protocol with ARQ,” IEEE Transactions on Wireless Communications, Vol. 9, pp. 2328–2337, 2010.

[21] K. H. Kim, “Analysis of Security Vulnerability in Cooperative Communication Networks,” IARIA International Conference on Networking and Services, pp. 80–84, 2011.

[22] S. Makda, A. Choudhary, N. Raman, T. Korakis, Z. Tao, and S. Panwar, “Security Implications of Cooperative Communications in Wireless Networks,” IEEE Sarnoff Symposium, pp. 1–6, 2008.

Codebook Subsampling and Rearrangement Method for Large Scale MIMO Systems

Tianxiao Zhang
School of Advanced Engineering
Beihang University
Beijing, China
anterzhang@126.com

Xin Su, Jie Zeng, Limin Xiao, Xibin Xu, Jingyu Li
Tsinghua National Laboratory for Information Science and Technology,
Tsinghua University
Beijing, China
{suxin, zengjie, xiaolimin, xuxibin, lijingyu}@mail.tsinghua.edu.cn

***Abstract*—In large scale multiple-input multiple-output (MIMO) systems, the size of codebook increases greatly when transmitters and receivers are equipped with more antennas. Thus, there are demands to select subsets of the codebook for usage to reduce the huge feedback overhead. In this paper, we propose a novel codebook subsampling method using chordal distance of different codewords and deleting them to affordable payload of Physical Uplink Control Channel (PUCCH). Besides, we design a related codebook rearrangement algorithm to mitigate the system performance loss when there are bit errors in the feedback channel.**

Keywords-Large scale MIMO; codebook subsampling; codebook rearrangement; PMI feedback

I. INTRODUCTION

The explosive growth of wireless data service calls for new spectrum resource. Meanwhile, the available spectrum for further wireless communication systems is very limited and expensive. Since the capacity of a multiple-input multiple-output (MIMO) system greatly increases with the minimum number of antennas at the transmitter and receiver sides under rich scattering environments [1], the large scale MIMO [2] shown in Fig. 1 is one of the most important techniques to address the issue of exponential increasing in wireless data service by using spatial multiplexing and interference mitigating.

For the consideration of practical application, the number of antennas on the terminal side is restricted, and thus the number of multiplexing layers is limited though the number of antennas on the base station (BS) could be very large. As a result, we should explore the large scale MIMO system potentials by utilizing beamforming technologies. The performance of beamforming relies on the accuracy of precoding. However the size of codebook can be very large when antennas are increased, considering that the payload capacity of Physical Uplink Control Channel (PUCCH) is limited to 11 bits [3]. To decrease the overhead in Channel State Information (CSI) feedback, the choice of codebook subsampling for transmission is necessary [4][5].

Several subsampling methods have been proposed in Rel-10 in 2,4,8 Tx scenario. The subset selection in this case naturally corresponds to the reduction of granularity in direction and/or phase offset [6], such as uniform subsampling or staggered subsampling which keeps better granularity. However, the codebook design for large scale MIMO system may not be based on direction for each polarization and phase offset between polarizations; hence the application above for large scale MIMO is restricted.

In this case, we propose a novel codebook subsampling method which applies to all kinds of codebook design, in which we select the subset of codebook using chordal distance of different codewords and delete them to affordable payload of PUCCH. In addition, to further optimize the performance of Precoding Matrix Indication (PMI) when errors occur, we propose the codebook rearrangement method to decrease the impact of mismatch between PMI and the channel.

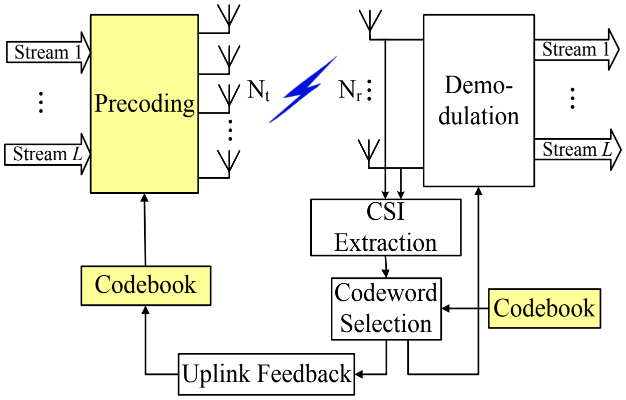


Figure.1 The close-loop MIMO system model

This paper is organized as follows: Section II introduces the model of the precoding system. Section III presents the codebook subsampling method and the codebook rearrangement method. Section IV shows the simulation results. And finally, Section V concludes the paper.

II. System Model

A. System Model

In this paper, we discuss about a close-loop MIMO system with N_t transmit antennas and N_r receive antennas depicted in Fig. 1. For massive MIMO system N_t could be very large.

Firstly, the data vector *S* in the system is demultiplexed into *L* streams. *L* is limited by $1 \le L \le \min(N_t, N_r)$. When *L*=1 we call the transmission beamforming, while *L*>1 we call it multiplexing. After the data vector *s* is preprocessed by a $N_t \times L$ precoding matrix W_i, we get a $N_t \times 1$ signal vector *x* for the N_t antennas to transmit:

$$x = Ws \tag{1}$$

Thus, after *x* passing the channel and being added the noise, we will get the received signal *y*, which can be expressed as:

$$y = HW_i s + v \tag{2}$$

where *H* ($H \in C^{N_r \times N_t}$) denotes the fading channel matrix with its entry H_{ij} denoting the channel response from the j^{th} transmit antenna to the i^{th} receive antenna, and *v* denotes the white complex Gaussian noise vector with covariance matrix $N_0 I_{N_r}$.

The precoding matrix is selected from the predesigned codebook which is known to the transmitter and the receiver. Taking the downlink as an example, when UEs have received the pilots from the BS, the receiver can choose the optimal codeword after the channel estimation. Then the receiver reports the PMI with limited bits to BS [7] through the uplink channel. If the feedback is limited to *B* bits, the size of codebook satisfies $N = 2^B$. Thus the transmitter can retrieve the precoding matrix and perform the precoding.

B. Kerdock Codebook

The basic idea of the kerdock codebook design is utilizing the feature of Mutually Unbiased Bases (MUB) to construct precoding matrices. The main characteristic of the kerdock codebook is that all the elements of the matrix are ±1 or ±j. Hence, the kerdock codebook has some advantages, such as low requirement for storage, low computational complexity for codeword search, and the simple systematic construction.

The MUB property is described as follows:

$S = \{s_1, ..., s_{N_t}\}$, $U = \{u_1, ..., u_{N_t}\}$ are two orthonormal bases with size $N_t \times N_t$. If the column vectors drawn from S and U satisfy $\left|\langle s_i, u_j \rangle\right| = 1/\sqrt{N_t}$, we can say that they have the mutually unbiased property [8].

An MUB is the set $S = \{s_1, ..., s_{N_t}\}$ satisfying the mutually unbiased property. The Kerdock codebook has several construction methods such as Sylvester–Hadamard construction and power construction. In this paper, we use the Sylvester–Hadamard construction:

First, we construct the generating matrices D_n ($N_t \times N_t$ diagonal matrices with ±1, ±j elements) for *n*=0,1,2…N_t-1 according to [9].

Then we construct the corresponding orthonormal matrix:

$$W_n = \frac{1}{\sqrt{N_t}} D_n \hat{H}_{N_t}, n = 0, 1, ..., N_t - 1 \tag{3}$$

where $\hat{H}_{N_t}$ is the $N_t \times N_t$ Sylvester–Hadamard matrix:

$$\hat{H}_{N_t} = \hat{H}_2 \otimes \hat{H}_2 .... \tag{4}$$

where $\hat{H}_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$

For the beamforming, we can construct the codebook by selecting each column of all the bases as the precoding vector:

$$C = \{f_1 = W_0^{\{1\}}, f_2 = W_0^{\{2\}}, ..., f_N = W_{N_t - 1}^{\{N_t\}}\} \tag{5}$$

And for a *L*-layer spatial multiplexing codebook, the largest codebook is derived by taking all *L*-column combinations from each W_n.

C. Codeword Search

We can choose the optimal main codeword from $\mathrm{K} = \{K_1, K_2, ..., K_N\}$ through the estimate of the channel. The codebook is shared by the transmitter and receiver.

Codeword selection criteria: for 1-layer beamforming, the beamformer that minimizes the probability of symbol error for maximum ratio combining receiver is expressed as [10]:

$$\hat{f}[i] = \arg\max_{f \in C} \left\| H[i] f \right\|_2^2 \tag{6}$$

where f denotes a $N_t \times 1$ matrix. For spatial multiplexing with a zero forcing receiver, the minimum singular value selection criterion is expressed as:

$$\hat{F}[i] = \arg\max_{K \in C} \lambda_{\min} \{H[i] K\} \tag{7}$$

ISBN: 978-1-61208-256-1

where $\lambda_{\min}$ denotes the minimum singular value of the argument. This selection criterion approximately maximizes the minimum sub-stream signal-to-noise ratio (SNR).

III. The Novel Codebook Subsampling and Rearrangement Method

A. Codebook Subsampling

To pursue the maximum SNR, we select the codeword with the smallest chordal distance from the transmission channel. The basic idea of the codebook subsampling method is to delete one codeword of the codeword pairs which have the smallest chordal distance. The chordal distance between two precoding vector is represented by

$$d_{chord}(f_i,f_j)=\sqrt{1-\left(\frac{\langle f_i,f_j\rangle}{\|f_i\|\|f_j\|}\right)^2} \qquad (8)$$

with ||.|| being the norm of the vector. If the chord distance between two codewords is the smallest among the codewords pool, we may reserve only one of them and delete another. Therefore we could decrease the overhead as well as remain the performance of precoding at the utmost.

The process of subsampling is shown as follows:

(1) Suppose the codebook includes *K* codewords. Divide the codewords into *g* groups.

a) Compute the chordal distance between any two codewords $d(f_i,f_j)$. Choose f_i and f_j as reference codewords if their distance is the largest.

b) Compute the chordal distance between the rest $f_L(L\in(0,K],L!=i\&\&L!=j)$ and reference codeword. If $d(f_i,f_L)<d(f_L,f_j)$, put f_L in f_i 's group. Otherwise put f_L in f_j 's group.

c) Repeat the procedure until the number of groups is *g*.

(2) Delete codewords and related PMI.

a) Compute the chordal distance between any two codewords in $l^{th}(l=1,2,\cdots,g)$ group. Find the $\min d_l(f_i,f_j)$

b) Choose the $\min\{\min d_l(f_i,f_j)\}(l=1,2,\cdots,g)$ as the codeword pair to deal with (suppose in the m^{th} group).

c) Compute the chordal distance between the rest $f_L(L!=i\&\&L!=j)$ in m^{th} group and reference codeword. If $\sum d(f_i,f_L)<\sum d(f_L,f_j)$, delete f_i, otherwise delete f_j.

d) Select the new $\min d_m(f_i,f_j)$. Back to a) until the number of codewords satisfy the requirement of PMI feedback.

The summary of the algorithm is given in Table I.

TABLE I. Simulated Subsampling Algorithm

```
//K: the total number of codewords
//B: cycling times
//N=2^B : groups of codewords
//n: current number of groups
//dis[i][j]: matrix of chordal distance between codewords
 n=1;
 loop for B cycles
    loop for n cycles
       calculate any of two codewords f_i, f_j with  dis[i][j]
       (f_i, f_j ∈ the n_k group)
       get the two codewords f_i, f_j with the largest codewords distance
       in the n_k group
       if dis(i,t)<dis(j,t) [t∈size(n_k), t≠i,j] then
          allocate the codeword to the group 2*(n_k-1)
       else
          allocate to the group 2*n_k
       end if
    end loop
    n=n*2
 end loop
loop for N cycle
   calculate the n_k group any of two codewords f_i, f_j  with dis{n_k}[i][j]
end loop
while(K>payload)
   dis(n_k)=min(dis{n_k}(:))
   m=argmin(dis(n_k))
   In the m group
   if ∑dis(i,t)<∑dis1(j,t) then
      delete the codeword f_i in the m group
   else
      delete the codeword f_j in the m group
   end if
   renew the m group with a minimum distance
end while
```

B. Codebook Rearrangement

The error in PMI feedback could lead to severe mismatch of precoding vector and user's channel, thus greatly decreasing transmission gain and increasing unreliability. By decreasing the mismatch caused by PMI transmission error, we could compensate for the performance loss, even the precoding vector is not optimal.

Therefore we rearrange the PMI, reduce the Hamming distance of binary indexes of codewords with high correlation. Consider one bit error in PMI feedback. When the two indexes with one bit of Hamming distance are arranged to codewords with high correlation, even if the error occurs and the base station uses the wrong codeword, the wrong codeword could still perform well due to the high correlation with the correct one, thus ensuring the compatibility with the channel and decreasing the gain loss.

The process of subsampling is presented in Table II, and the description is given as follows:

(1) Divide PMI into B PMI groups based on binary code weight. f_i denotes the original codeword associated with PMI w_i , and U_i denotes the new one. $U_0 = f_0$.

(2) Select one $w_{b,i}$ in $b^{th}(b = 1,2,\cdots,B)$PMI group. Find all the $w_{b-1,j}$ in $(b-1)^{th}$ PMI group that $d_{binary}(w_{b,i}, w_{b-1,j}) = 1$.

(3) In $l^{th}(l = 1,2,\cdots,g)$codeword group, compute $\sum_j d(f_k, f_{b-1,j}), f_k \in l^{th}$

If $\sum_j d(f_{k'}, f_{b-1,j}) = \min \sum_j d(f_k, f_{b-1,j})$

($f_k \in l^{th}$codeword group), then $U_{b,i} = f_{k'}$.

(4) If all codewords in l^{th}codeword group are rearranged with new PMI, turn to $(l+1)^{th}$codeword group.

(5) If all PMI in b^{th}PMI group are rearranged with new codeword, turn to $(b+1)^{th}$PMI group.

TABLE II. SIMULATED REARRANGEMENT ALGORITHM

```
//G=log2(payload)
//PMI_B : groups of PMI based on code weight
loop for payload cycles
    restore each PMI in PMI_B{i} with code weight i;
end loop
$D_0=f_0$
for i=1:G
    for j=1:size(PMI_B{i})
        $P_{i_j}$=PMI_B{i}( j)
        $U_{j_{i-1}}=\{ f_{j_{i-1}}, d(P_{i-1},P_{i_j})=1\}$
        $j_i'=\text{argmin}\sum d(f_m,U_{j_{i-1}})$ ($f_m \in$ the $n_k$ group)
        $D_{j_i}=f_{j_i'}$
        if the codeword in the $n_k$ group all have been
        allocated new PMI then
            go to next group
        else
            continue in this group
        end if
    end
end
```

IV. SIMULATION RESULTS

This section we present the simulation results under the configuration given in Table III. The simulation procedure follows the system model in Fig. 1.

TABLE III. SIMULATION PARAMETERS

Parameter	Value
Frequency	2.1 GHz
System bandwidth	10 MHz
Channel modelling	i.i.d, CN(0,1)
Number of BS antennas	32
Number of UE antennas	1
Channel estimation	Ideal
UE receiver	MMSE
SNR	10 dB

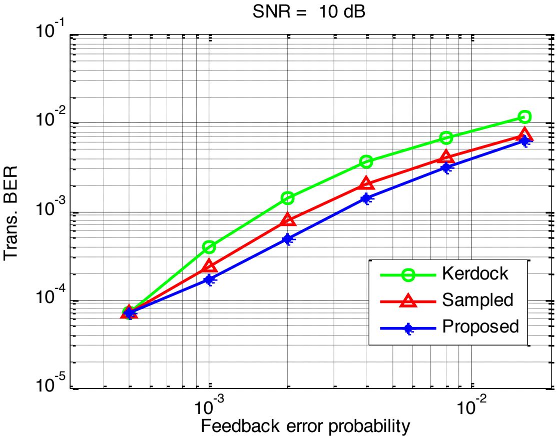


Figure.2 The BER performance of different codebooks

Fig. 2 shows the performance of downlink transmission BER vs. feedback error probability. The feedback error probability means the probability of each bit-error occurs in the PMI feedback, and the downlink transmission BER means the bit error rate of the downlink transmission. From the results, we can see the codebook after subsampling has significant BER performance gain compared with the original Kerdock codebook, because of the low probability of error occurrence due to the fewer bits for feedback. And the proposed codebook after rearrangement has the further BER performance gain since to configure high correlation codewords with reduced code distance, we decrease the performance loss of system when the mismatch of precoding vector and the channel occurs.

V. CONCLUSION

In this paper, we proposed a novel codebook subsampling method based on chordal distance as well as the related codebook rearrangement algorithm for codebook designs in large scale MIMO system. The codebook subsampling method can reduce the feedback overhead without impacting the system performance, and the rearrangement algorithm can significantly mitigate the system performance loss when errors in the feedback channel occur. Simulation results show that the Kerdock codebook after subsampling and rearrangement has significant performance gain under the non-ideal uplink feedback channel in large scale MIMO system.

ACKNOWLEDGMENT

This work was supported by Beijing Natural Science Foundation Funded Project (No.4110001), National S&T Major Project (No. 2011ZX03003-002), Tsinghua Independent Research (No. 2010TH203-02) and Samsung Company.

REFERENCES

[1] E. Telatar, “Capacity of multi-antenna Gaussian channels”, European Transactions on Telecommunications, vol. 10, no. 6, 1999, pp. 585-595.

[2] F. Rusek et al., “Scaling up MIMO: opportunities and challenges with very large arrays,” IEEE Signal Processing Magazine, 2012, vol. 30, no. 1, pp. 40 - 60.

[3] 3GPP TS 36.213: Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures, pp. 56-64.

[4] R1-104164 Way Forward on 8Tx Codebook for Rel.10 DL MIMO, CATT et al, RAN1 #61bis, June 2010, Dresden, Germany.

[5] R1-104259 Way Forward on CSI Feedback for Rel.10 DL MIMO, Alcatel-Lucent et al, RAN1 #61bis, June 2010, Dresden, Germany.

[6] R1-104901 8Tx Codebook Subsampling, Panasonic, August 2010, Madrid, Spain.

[7] 3GPP TS 36.211: Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation, pp. 17-20.

[8] A. Klappenecker and M. Rotteler, “Constructions of mutually unbiased bases”， Finite Fields Appl., 2004，pp. 137–144,.

[9] R. W. Heath, Jr., T. Strohmer, and A. J. Paulraj, “On quasi-orthogonal signatures for CDMA systems”， IEEE Trans. Inf. Theory, vol. 52, no.3, 2006，pp. 1217–1226.

[10] D. J. Love and R. W. Heath, Jr., “Limited feedback unitary precoding for spatial multiplexing systems”, IEEE Trans. Inf. Theory, vol. 51, no.8, 2005，pp. 2967-2976.

A Power-Aware Real-Time Routing Mechanism for Wireless Sensor Networks

Mohamed Aissani, Sofiane Bouznad, Badis Djamaa, and Ibrahim Tsabet

Research Unit in Computer Science, Ecole Militaire Polytechnique
P.O. Box 17 EMP, Bordj-El-Bahri, Algiers, Algeria
{maissani, bouznad.sofiane, badis.tos, utopia.ibrahim}@gmail.com

Abstract— **To optimally manage the limited energy of nodes without degrading efficiency of routing protocols in delivering real-time packets in wireless sensor networks, we propose in this paper an efficient power-aware real-time routing (PRR) mechanism. Firstly, it increases the network fluidity and saves more energy of nodes by removing early in network all useless data packets according to their residual deadline and expected end-to-end delay. Secondly, it reinforces the real-time behavior of the used routing protocol and preserves the network resources by selecting from the current-node queue the most urgent packet to be forwarded first. Finally, it saves energy of nodes without degrading the protocol efficiency in delivering real-time flows by combining adjusted transmission power of current node with relay speed of the forwarding candidate neighbors when selecting a next forwarder for the current packet. PRR is simple to implement and can be easily integrated in any geographic routing protocol. Associated with the well-know real-time routing protocol SPEED by using TinyOS, and evaluated in its embedded simulator TOSSIM, PRR achieved good performance in terms of network energy consumption, packet loss ratio, and node energy balancing.**

Keywords-wireless sensor networks; real-time routing; energy-aware routing; node energy balancing.

I. Introduction

Flexibility, fault tolerance, reduced production cost, high capture capacity, and rapid installation are characteristics that enabled a wireless sensor network (WSN) to have multiple application domains, such as disasters detection and monitoring, mapping biodiversity, intelligent building, precision agriculture, machinery monitoring and preventive maintenance, environmental control, logistics and intelligent transportation, and medicine. However, the WSN realization requires satisfaction of some constraints that arise from a number of factors guiding the design phase, such as fault tolerance, scalability, cost, durability, material and topology.

WSNs are often characterized by a dense and large scale deployment with limited processing, storage, transmission and energy resources. It is recognized that conserving energy is an unavoidable issue in the design of WSNs because it imposes strict constraints on network operations [1-3]. In fact, the energy consumed in sensor nodes has an important impact on network lifetime that has become the dominant performance criterion. Extending lifetime of a WSN is a shared objective by designers and researchers. It is necessary that routing algorithms use paths that save more energy of nodes.

Although existing works [4-12], summarized in Section II, play important roles in improving network performance, design of energy-aware real-time routing protocols is still a challenging area in WSNs. To contribute in this domain, we propose in this paper an efficient power-aware real-time routing (PRR) mechanism which:

- Increases the network fluidity and saves more energy of nodes by removing early in the network all useless packets according to their residual deadline and expected end-to-end delay.
- Reinforces the real-time behavior of the used routing protocol and preserves the network resources by selecting from the current-node queue the most urgent packet to be forwarded first.
- Saves energy of nodes without degrading the protocol efficiency in delivering real-time flows by combining adjusted transmission power with relay speed of the forwarding candidate neighbors when selecting a next forwarder.

The rest of the paper is organized as follows. Section II summarizes the related works. Section III describes the proposed PRR mechanism that aims to improve efficiency of real-time routing protocols based on geographic location of sensor nodes. Section IV evaluates and discusses performance of our proposal. Section V concludes the present paper.

II. Related Works

Some existing real-time routing protocols don’t consider explicitly the limited energy of sensor nodes [4-8]. RAP [4] is one of the earlier real-time routing protocols for WSNs. It provides service differentiation in the timeliness domain by using velocity-monotonic classification of data packets. It works only when most traffic is periodic and all periods are known previously. Also, it is not adaptable to dynamics of a network. SPEED [5] is designed to be a stateless, localized algorithm with minimal control overhead. It achieves an end-to-end soft real-time communication by maintaining a desired delivery speed across the sensor network through a novel combination of feedback control and stateless non-deterministic geographic forwarding. MMSPEED [6] extends SPEED to support different delivery velocities and levels of reliability. It provides QoS differentiation in two quality domains, namely, timeliness and reliability, so that packets can choose the most proper combination of service options depending on their timeliness and reliability requirements. THVR [7] adopts, like SPEED, the approach of mapping packet deadline to a velocity, which is known as a good metric to delay constrained packet delivery. However, its routing decisions are based on two-hop neighborhood information to achieve lower end-to-end deadline miss ratio and higher energy utilization efficiency. DMFR [8] routes data packets in five stages: initialization, packet transmission, jumping transmission, jumping probability adjustment and transmission finish. Transition from transmission stage to jumping stage occurs when a node is congested. To reduce the packet loss ratio, each sensor node dynamically adjusts its jumping probabilities.

However, some of other existing routing protocols use specific mechanisms to save energy of sensor nodes and/or to maximize the sensor network lifetime [9-12]. PATH [9] improves real-time routing performance by means of reducing the packet dropping in routing decisions. It is based on the concept of using two-hop neighbor information and power-control mechanism. The former is used for routing decisions and the latter is deployed to improve link quality as well as reducing the delay. The protocol dynamically adjusts transmitting power in order to reduce the probability of packet dropping and addresses practical issue like network holes, scalability and loss links in WSNs. EARTOR [10] is designed to route requests with specified end-to-end latency constraints, which strikes the elegant balance between the energy consumption and the end-to-end latency and aims to maximize the number of the requests realized in network. The core techniques adopted include the cross-layer design that incorporates the duty cycle, a bidding mechanism for each relay candidate that takes its residual energy, location information, and relay priority into consideration. EEOR [11] improves the sensor network throughput by allowing nodes that overhear the transmission and closer to the sink to participate in forwarding the packet, i.e., in forwarder list. The nodes in forwarder list are prioritized and the lower priority forwarder will discard the packet if the packet has been forwarded by a higher priority forwarder. One challenging problem is to select and prioritize forwarder list such that the energy consumptions by all nodes is optimized. Extensive simulations in simulator TOSSIM show that this protocol performs well in terms of energy consumption, packet loss ratio, and average delivery delay. TREE [12] is a routing strategy with guarantee of QoS for industrial wireless sensor networks by considering the real-time routing performance, transmission reliability, and energy efficiency. By using two-hop information, real-time data routes with lower energy cost and better transmission reliability are used in the proposed routing strategy.

III. PROPOSED POWER-AWARE MECHANISM

The proposed PRR mechanism routes data packets in three stages: useless packet remove (Section III-A), urgent packet selection (Section III-B) and next forwarder selection (Section III-C). Note that the calculus are done locally in the current node and the used information is either in the received packet to forward, such as previous hops' delay, geographic location of source and destination nodes, or inside the current node, such as its geographic location and those of its neighbors.

A. Useless packet remove

Many real-time routing protocols [4-12] forward, often over long distances, packets that have no chance to reach their destination because of its insufficient deadline. This is because the packet deadline information, which in important in this type of applications, is not exploited by these protocols. To save more energy of nodes, the proposed PRR mechanism ensures an early removal of any useless packet because it will not reach its destination. Indeed, only packets with sufficient residual deadline to reach the sink node are forwarded in network. Also, PRR increases the network fluidity and reinforces the real-time aspects of the routing protocol. To do this, PRR calculates the expected end-to-end delay allowing the current packet to reach its destination node, then decides whether to remove or not the current packet depending on both this expected end-to-end delay and constant threshold α related to application requirements in which the removal of delayed packets is performed.

1) Expected end-to-end delay: As shown in Figure 1, current node i calculates the expected end-to-end delay $T_{id}(p)$, allowing current packet p to reach its destination d, by using Formula (1). In this formula, $T_{si}(p)$ denotes the previous hops' delay since source node s, $D_{si}(p)$ is the geographic distance traveled by packet p until current node i, $D_{id}(p)$ is the remaining geographic distance to reach d.

$$T_{id}(p) = \frac{D_{id}(p)}{D_{si}(p)} * T_{si}(p) \qquad (1)$$

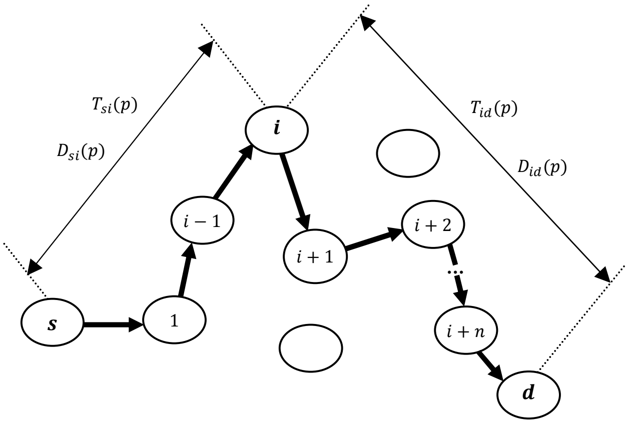


Figure 1. Expected end-to-end delay estimation.

2) Packet remove decision: Having the expected end-to-end delay $T_{id}(p)$ and to decide whether to remove or not packet p, current node i applies the decision rule shown in Figure 2, where $D_{sd}(p)$ is distance between source node s and destination node d, $AD(p)$ is the advance in distance of data packet p toward its destination which is given by

Formula (2) and showen in Figure 3, and α is a constant parameter set in the interval [0,1] according to the application requirement. The value of α must be close to 0 for energy-critical applications to maximize the network lifetime or close to 1 for time-critical applications to minimize the packet loss ratio.

```
IF $AD(p) > \alpha * D_{sd}(p)$ THEN
   IF $T_{id}(p) > Deadline(p)$ THEN Remove packet $p$; ENDIF
ELSE
   IF $\frac{AD(p)}{\alpha * D_{sd}(p)} T_{id}(p) > Deadline(p)$ THEN Remove $p$; ENDIF
ENDIF
```

Figure 2. The packet remove décision algorithm.

The packet decision remove algorithm (Figure 2) is explained as follow. If $AD(p)$ is greater than $\alpha * D_{sd}(p)$ then node i removes each delayed packet after the distance threshold $\alpha * D_{sd}(p)$. Otherwise, $T_{id}(p)$ is multiplied by $AD\,(p)/\big(\alpha * D_{sd}(p)\big)$ to give more chance to current packet p to advance in network. If the result exceeds $Deadline(p)$ despite the given chance then packet p is removed to save the energy of nodes and to increase the network fluidity. In our simulations given in Section IV parameter α is set at 0.5. Thus, each delayed packet p with an advance $AD(p)$ greater than 50% of the total distance $D_{sd}(p)$ is immediately removed by current node i in order to increase fluidity of links and to save resources of sensor nodes.

$$AD(p) = \begin{cases} \dfrac{D_{si}^2(p) - D_{id}^2(p) + D_{sd}^2(p)}{2D_{sd}(p)} & \text{IF } D_{si}(p) < D_{sd}(p) \\ D_{sd}(p) & \text{OTHERWISE} \end{cases} \qquad (2)$$

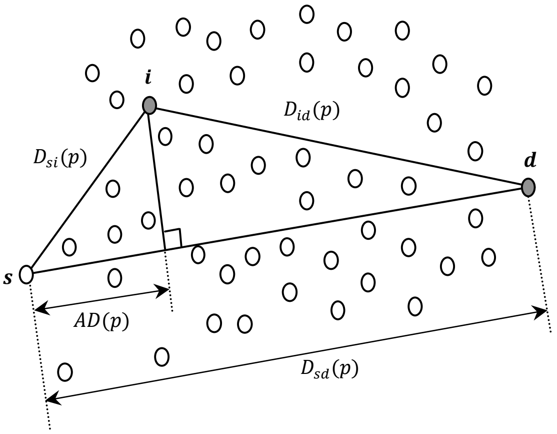


Figure 3. Advance in distance estimation.

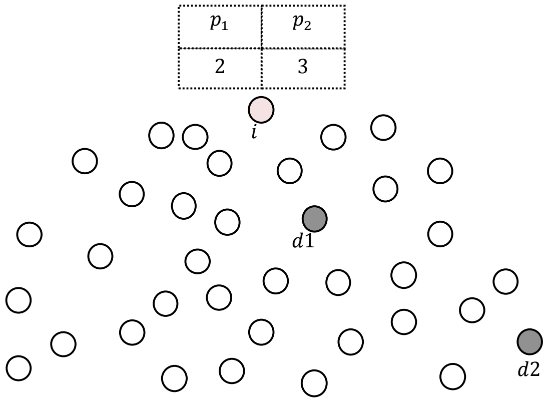
B. Urgent packet selection

Most existing real-time routing protocols use scheduling schemes based only on residual deadline of a data packet to forward [13]. However, these schemes may not be effective when packets are sent to different destinations; case of a network using several sinks. In fact, these schemes give forwarding priority to a packet whose deadline is the smallest although it is very close to its destination. In the example shown in Figure 4, node i has two packets to forward: p_1 for destination $d1$ with 2 ms (milliseconds) as deadline and p_2 for destination $d2$ with 3 ms as deadline. According to existing scheduling schemes based on residual deadline, node i will firstly forward packet p_1 and then probably causes the removal of packet p_2 because of its distance to destination $d2$. The proposed PRR mechanism provides an efficient solution to this problem because it performs as follows:

- For each data packet p_j in the queue of current node i, calculate the decision parameter $D(p_j)$ by Formula (3), where $T_{id}(p_j)$ is the expected end-to-end delay, obtained by Formula (1), allowing packet p_j to reach its destination d.
- Selects data packet p_k having the smallest decision parameter $D(p_k)$ by running Function (4).

$$D(p_j) = Deadline\,(p_j) - T_{id}(p_j) \qquad (3)$$

$$D(p_k) = Min\{\ D(p_j)\,;\ \forall p_j \in Queue(i)\ \} \qquad (4)$$



p_1	p_2
2	3

Figure 4. Most urgent packet selection by current node i.

Then, in the PRR mechanism, the two packets p_1 and p_2 (Figure 4) will probably reach their respective destination before their deadline expires because current node i will forward packet p_2 before packet p_1. Note that PRR is also valid for a network using one destination node (sink).

Note that the most urgent packet selection can be done in two ways: a) during the packet reception by a node where its queue is scheduled according to Formula (3) or b) during the forwarding process where the most urgent packet is timely chosen from the queue. In way (a), PRR minimizes calculations but loses reliability because the queuing delay of packet p_j is not considered when estimating its $T_{id}(p_j)$. But in way (b), the current-node queue is not scheduled and selection of the most urgent packet requires extraction of all packets belonging to this queue. Since PRR is designed to achieve lower loss ratio and to reduce energy consumption, way (b) has been implemented in our proposal, where a current node applies both the decision rule (Figure 2) and Function (4) on all packets in its queue in order to remove each delayed packet and to select the most urgent packet among the not delayed packets.

C. Next forwarder selection

Radio range adjustment has become possible in recent sensor nodes. The quantity of energy required to send a message is proportional to the transmission power used by the sender node [14]. Since the attenuation of the radio power of a wireless link is usually proportional to the square of the distance between the sender and the receiver, the proposed PRR mechanism uses this idea. Indeed, PRR adjusts the sender transmission power according to location of receiver node in order to reduce energy consumed during the routing process.

A power-based algorithm tries to minimize the quantity of power required to route a message between source and destination nodes. The most commonly used energy model [15] calculates by using Formula (5) the energetic cost of a message forwarded by node u to node v, that are separated by distance d_{uv}, by. In this formula, constant parameters α and c depend on the network environment: c represents the signal processing cost and α is the signal attenuation (α≥2). In our performance evaluation in Section IV, we have α=2 and c=0. The optimal cost of a link in terms of energy consumption is that minimizes Formula (5); a function that does not consider any real-time service.

$$Power\ (uv)\ = \begin{cases} d_{uv}^{\alpha}\ \ + c & \text{IF}\ \ d_{uv} \neq 0 \\ 0 & \text{OTHERWISE} \end{cases} \tag{5}$$

The same definition of FS (Forwarding candidate neighbors Set) introduced in SPEED (Figure 5) is used in our proposal, but an improved forwarding strategy has been proposed. In SPEED, the next forwarder of current packet p toward destination node d is selected by current node i from its FS, which is constructed from its NS (Neighbors Set), according to a relay speed metric. Node n_i is a forwarding candidate neighbor if $distance(n_i, d) < distance(i, d)$. In Figure 5, we have: $NS = \{ n_1, n_2, n_3, n_4, n_5, n_6, n_7 \}$ and $FS = \{n_1, n_2, n_3\}$. The relay speed provided by a neighbor n_i in FS is given in SPEED by Formula (6), where L is distance between node i and destination d , $Lnext$ is distance between neighbor n_i in FS and destination d, and $HopDelay(n_i)$ is the estimated delay of link (i, n_i).

$$S(n_i) = \frac{L - Lnext}{HopDelay(n_i)} \quad ; \quad n_i \in FS \tag{6}$$

To make SPEED energy-aware with high reliability in forwarding real-time flows, Formula (5) cannot be applied directly in PRR because our objective is to achieve lower packet loss ratio and higher energy utilization efficiency. To do this, PRR considers all neighbors in FS of current node i to select the next forwarder of current packet p. PRR combines the transmission power $P(n_i)$ required in node i to reach a neighbor n_i in FS, given by Formula (5), with the relay speed $S(n_i)$ of n_i, given by Formula (6). Neighbor n_k with the higher decision parameter $D(n_i)$ is selected by node i as next forwarder. Formally, node i applies Function (7), where $D(n_i)$ is given by Formula (8).

$$NextHop(p) = n_k \ ; \ \text{with} : D(n_k) = Max\, \{ D(n_i) \ ; \ \forall n_i \in FS \ \} \tag{7}$$

$$D(n_i) = \frac{S(n_i)}{P(n_i)} \quad ; \quad n_i \in FS \tag{8}$$

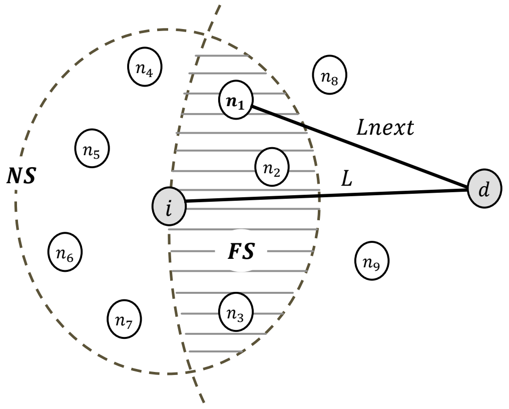


Figure 5. The sets NS and FS of a current node i in SPEED.

IV. PERFORMANCE EVALUATION

To evaluate performance of the PRR mechanism, we associate it with the well-known real-time routing protocol SPEED [5] and the resulting protocol is called PA-SPEED (Power-Aware SPEED). We change in SPEED only the SNGF (Stateless Nondeterministic Geographic Forwarding) component. In the PA-SPEED protocol, when a node has to forward a data packet it first removes all delayed packets from its queue, then selects the most urgent packet among the not delayed packets and finally forwards the selected urgent packet to the neighbor realizing the best tradeoff between transmission power and relay speed.

The protocols SPEED and PA-SPEED have been implemented in TinyOS [16] and evaluated in its embedded sensor network simulator TOSSIM [17]. Also, the recent existing routing protocol EEOR [11] has been evaluated in this simulator and in the same conditions. Since we are interested by real-time applications, we used a scenario of detecting events that occur randomly in a field. Once an event is detected, the information captured will be forwarded in a required deadline toward a sink which is usually connected to an actuator.

Our simulation scene uses a uniform random distribution of sensor nodes. We perform simulations on a terrain with size 500×500 meters and 625 deployed sensor nodes (with 12 neighbors per node as density). Two destination nodes are deployed and each one receives packets concerning particular event detection. At each time period, 20 randomly source nodes, equitably distributed on each side of the network, detect an event and forward corresponding information to one destination node (sink). Each simulation runs during 230 seconds. Parameters used in our simulations are given in TABLE I.

For each simulation, we set the packet deadline to 500 ms (milliseconds), we vary the source rate from 3 to 23 pps (packets per second) and we measure the performance of SPEED [5], EEOR [11] and PA-SPEED in terms of packet loss ratio, energy consumed per delivered packet, and energy balancing factor (ebf). The later represents variance in energy consumed by all sensors with the same initial energy. Formally, $ebf = (1/ns) * \sum_{k=1}^{ns}(ec_k - ec_{avr})^2$, where ec_k is the energy consumed by sensor k and ns is the number of deployed sensors, ec_{avr} is the average energy consumed by all deployed sensors.

TABLE I. SIMULATION PARAMETERS.

MAC layer	CSMA-TinyOS
Radio layer	CC2420 radio layer
Propagation model	log-normal path loss model
Queue size	50 packets
Transmission channel	WirelessChannel
Bandwidth	200 Kilobytes per second
Packet size	32 bytes
Energy model	PowerTOSSIMz model
Node radio range	40 meters

The obtained simulation results, given in the figures 6-8, show that the PRR mechanism, used in the PA-SPEED protocol, is efficient in terms of delivering real-time flows and managing energy of sensor nodes.

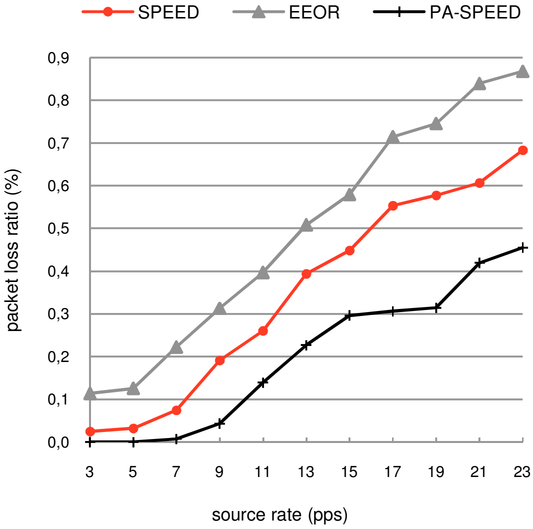


Figure 6. Success in delivering real-time packets.

Indeed, PA-SPEED loses less data packets (Figure 6) and consumes less energy of sensor nodes (Figure 7) than the protocols EEOR and SPEED. This is due to the PRR mechanism which first increases the network fluidity and saves energy of nodes by removing each packet having less chance to reach its destination according to its residual deadline and expected end-to-end delay, then reinforces the real-time behavior of the PA-SPEED protocol by selecting from the current-node queue the most urgent packet among the not delayed packets to be forwarded first, and finally forwards the selected urgent packet to the neighbor realizing the best tradeoff between transmission power of the current node and relay speed of the next forwarder neighbor. In application with high rate, the protocols EEOR and SPEED lose more packets because the deadline information is not used in their routing decisions.

Figure 8 shows that PA-SPEED outperforms SPEED and EEOR in balancing energy of nodes. This performance is due to the PRR mechanism which uses the SPEED load balancing metric, i.e. relay speed given in Formula (6), which is based on a hop delay estimation representing the links' fluidity.

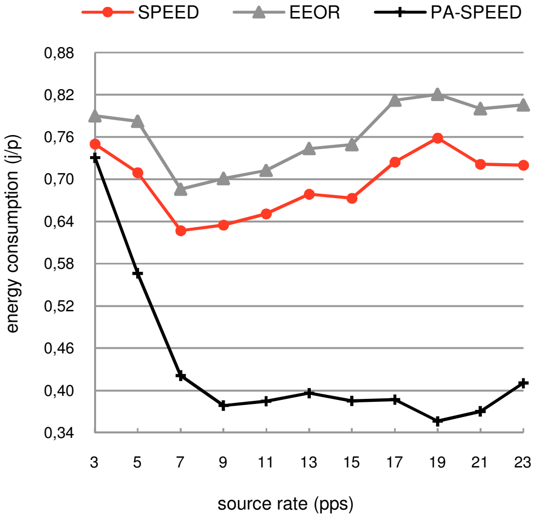


Figure 7. Average energy consumed per delivered packet.

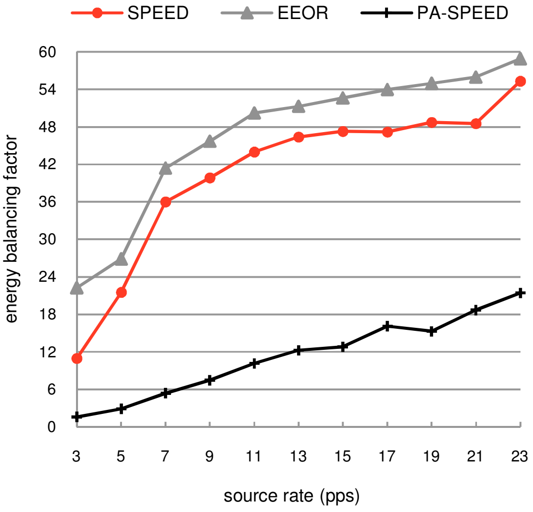


Figure 8. Performance in node energy balancing.

V. CONCLUSION

An efficient mechanism (PRR) that aims to improve energy managing and to deliver maximum real-time packets in wireless sensor networks has been proposed in this paper. In this power-aware mechanism, the current node removes from its queue all delayed packets to increase links' fluidity and to save nodes' energy, then selects from the list of not delayed packets the most urgent packet according to residual deadline and expected end-to-end delay to satisfy real-time application constraints, and finally, forwards the selected urgent packet to the neighbor realizing the best tradeoff between transmission power and relay speed.

Then, we have associated the PRR mechanism with the existing SPEED real-time routing protocol and the obtained protocol (PA-SPEED) has achieved good performance in terms of packet loss ratio, energy consumed per delivered packet, and node energy balancing.

Since we base dropping decisions concerning delayed packets simply on estimated travel times towards the sink, our future work will consider any kind of weights, urgencies, fairness, or importance values of packets in order to have a less aggressive approach. We also plan to put our source codes in Imote2 sensor nodes for experimental tests in order to consolidate the simulation results presented in the present paper.

REFERENCES

[1] K. Akkaya and M. Younis, "A Survey on Routing Protocols for Wireless Sensor Networks," Ad Hoc Networks, vol. 3(3), pp. 325-349, July 2005.

[2] S. Ehsan, and B. Hamdaoui, "A Survey on Energy-Efficient Routing Techniques with QoS Assurances for Wireless Multimedia Sensor Networks," IEEE Communications Surveys & Tutorials, vol. 14(2), pp. 265–278, May 2012.

[3] R. Marjan, D. Behnam, A. B. Kamalrulnizam, and L. Malrey, "Multipath Routing in Wireless Sensor Networks: Survey and Research Challenges," Sensors journal, vol. 12(1), pp. 650-685, May 2012.

[4] C. Lu, B.M. Blum, T.F. Abdelzaher, J.A. Stankovic, and T. He, "RAP: A Real-Time Communication Architecture for Large-Scale Wireless Sensor Networks," Proc. IEEE Real Time Technology and Applications Symposium (RTAS), IEEE Press, pp. 55-66, 2002, doi:10.1109/RTTAS.2002.1137381.

[5] T. He, J.A. Stankovic, C. Lu, and T. Abdelzaher, "A Spatiotemporal Communication Protocol for Wireless Sensor Networks," IEEE Transactions on Parallel and Distributed Systems, vol. 16(10), pp. 995-1006, 2005.

[6] E. Felemban, C.G. Lee, and E. Ekici, MMSPEED: Multipath Multi-SPEED Protocol for QoS Guarantee of Reliability and Timeliness in Wireless Sensor Networks, IEEE Transactions on Mobile Computing, vol. 5(6), pp. 738-754, 2006.

[7] Y. Li, C. Shue, Y.Q. Song, Z. Wang, and T. Sun, "Enhancing Real-Time Delivery in Wireless Sensor Networks With Two-Hop Information," IEEE Transactions on Industrial Informatics, vol. 5(2), pp. 113-122, 2009.

[8] C.L. Wu, F. Xia, L. Yao, H. Zhang, and B. Liu, "Dynamical Jumping Real-Time Fault-Tolerant Routing Protocol for Wireless Sensor Networks," Sensors Journal, vol. 10(3), pp. 2416-2437, 2010.

[9] P. Rezayat, M. Mahdavi, M. Ghasemzadeh, and M. Agha Sarram, "A Novel Real-Time Power Aware Routing Protocol in Wireless Sensor Networks," Journal of Computer Science, vol. 10(4), 300-305, 2010.

[10] W. Yang, W. Liang, and W. Dou, "Energy-Aware Real-Time Opportunistic Routing for Wireless Ad Hoc Networks," Proc. IEEE Global Telecommunications Conference, IEEE Press, pp. 1-6, December 2010, doi:10.1109/GLOCOM.2010.5683491.

[11] X. Mao, S. Tang, and X. Xu, "Energy Efficient Opportunistic Routing in Wireless Sensor Networks," IEEE Transactions on Parallel and Distributed Systems, vol. 22(11), pp.1934-1942 November 2011.

[12] L. Xue, X. Guan, Z. Liu, and B. Yang, "TREE: Routing strategy with guarantee of QoS for industrial wireless sensor networks," International Journal of Communication Systems (IJCS), May 2012, doi:10.1002/dac.2376.

[13] K. Liu, N. Abu-Ghazaleh, and K.D. Kang, "JiTS: Just-in-Time Scheduling for Real-Time Sensor Data Dissemination," Proc. 4th IEEE Annual Int'l Conference on Pervasive Computing and Communications, IEEE Press, pp. 46-50, March 2006, doi: 10.1109/PERCOM.2006.33.

[14] V. Shnayder, M. Hempstead, B. Chen, and G.W. Allen, M. Welsh, "Simulating the Power Consumption of Large-Scale Sensor Network Applications," Proc. 2nd ACM Conference on Embedded Networked Sensor Systems (SenSys), ACM Press, pp. 188-200, November 2004, doi:10.1145/1031495.1031518.

[15] V. Rodoplu and T. Meng, "Minimizing energy mobile wireless networks," IEEE Journal on Selected Areas, vol. 17(1), pp. 1333–1347, 1999.

[16] P. Levis and D. Gay, "TinyOS programming," Cambridge University Press, USA, April 2009.

[17] P. Levis, N. Lee, M. Welsh, and D. Culler, "TOSSIM: Accurate and Scalable Simulation of Entire TinyOS Applications," Proc. 1st ACM Conference on Embedded Networked Sensor Systems, ACM Press, pp. 126 - 137, November 2003, doi:10.1145/958491.958506.

Mobility-based Greedy Forwarding Mechanism for Wireless Sensor Networks

Riad Kouah, Mohamed Aissani
Research Unit in Computer Science
Ecole Militaire Polytechnique (EMP)
P.O. Box 17 Bordj-El-Bahri, Algiers, Algeria
riadkouah@hotmail.com, maissani@gmail.com

Samira Moussaoui
Computer Science Department
University of Science and Technology (USTHB)
P.O. Box 32, Bab-Ezzouar, Algiers, Algeria
smoussaoui@usthb.dz

Abstract—**Geographic routing in mobile sensor networks has attracted attention in recent years. When a sensor node has a packet to forward, it selects the closest available neighbor to the sink as next forwarder regarding only location parameter. However, this strategy does not consider the mobility of sensor nodes. To overcome this problem, we propose in this paper an efficient geographic routing mechanism based on a new next-hop selection metric. It combines the distance to the sink, the moving direction and the moving speed of the forwarding candidate neighbors. This moving direction is based on neighbor evolution in distance according to the sink calculated by a sender node between two successive received location beacons. Associated with the well-known GPSR routing protocol, our mechanism achieved good performance in both delivering data packets and conserving resources of sensor nodes.**

Keywords—mobile sensor networks; geographic routing; sensor mobility; greedy forwarding.

I. INTRODUCTION

Geographic routing in wireless sensor networks (WSNs), and especially greedy forwarding, is a new challenge when nodes move. In fact, the efficient and scalable greedy forwarding is a promising scheme for large-scale WSNs when location of each node is available [1, 2]. Indeed, the current packet is forwarded to a 1-hop neighbor who is closer to the sink than the sender node. This process is repeated until the data packet reaches the sink. Traditionally, the next-forwarder selection is based only on the location parameter. However, this can degrade the performance of geographic routing in mobile WSNs. The location failure, which is resulted from mobility of nodes, degrades the routing performance in delivering packets and wastes the limited energy of nodes.

Although existing works [3-9], summarized in Section II, play important roles in improving the performance of the geographic routing in mobile WSNs, design of new routing solutions is still a challenging research area. Thus, we first analyze in this paper the impact of node mobility on the routing performance and then we propose an efficient geographical routing mechanism, called MGF (Mobility-based Greedy Forwarding), which combines the distance to the sink, the moving direction and the moving speed of the forwarding candidate neighbors of a sender node in a new routing decision metric. This combination is also used the DGF (Direction-based Greedy Forwarding) mechanism that we have proposed in our previous work [10]. The difference between the two mechanisms is explained as follows. To calculate a neighbor moving direction according to the sink, the DGF mechanism uses the neighbor evolution in both distance and angle between two successive location beacons that receives a sender node, but the MGF mechanism uses the neighbor evolution in distance only.

We associate the proposed MGF mechanism with the well-known GPSR (Greedy Perimeter Stateless Routing) protocol [2] in order to improve its efficiency in mobile WSNs and the obtained protocol is called GPSR-MS (GPSR with Mobile Sensors). The major difference between GPSR-MS and existing protocols dedicated for mobile WSNs (Section II) includes the following aspects:

- GPSR-MS operates without organizing network in clusters, while the majority of existing routing protocols designed for mobile WSNs are cluster-based where the maintenance process consumes many resources of sensor nodes.
- In existing cluster-based routing protocols, the greedy forwarding mode is not applied, while the GPSR-MS protocol is based on this scalable and efficient mode.
- Few of existing routing protocols are designed for mobile WSNs. Therefore, the GPSR-MS protocol strengthens this class of protocols. Our objective is to maximize the packet delivery ratio with the minimum consumption of energy.

The rest of the paper is organized as follows. Section II presents the main routing schemes and protocols proposed in literature for mobile WSNs. Section III discusses the node mobility effect on the greedy forwarding performance. Section IV presents the proposed MGF mechanism. Section V evaluates performance of our proposal. Section VI concludes the paper.

II. RELATED WORKS

Node mobility has added several challenges in mobile WSNs. One of these challenges is routing. Existing routing protocols in mobile WSNs based on type of nodes are regrouped in three classes: protocols that support static

ISBN: 978-1-61208-256-1

nodes and mobile sink(s) [3-4], protocols that support mobile nodes and static sink(s) [5-8], and protocols that support mobile nodes and mobile sink(s) [9]. In literature, the majority of works have been focused on the first class of protocols. In other hand, there are less works concerning both the second and the third class.

Fodor et al. [3] propose the gradient-based routing protocol (GBRP) to use mobile sinks that move in order to decrease the energy consumption of the whole network. In GBRP, sensor nodes maintain a list of neighboring next hops that are in the right direction towards the closest sink. The protocol uses a restricted flooding to update locations of the mobile sinks. The principle behind is to register by each node the cost between the appropriate sink and the given node and to update only these routing entities where the relative change in cost is above a threshold. Wang et al. [4] propose a mobile sink routing protocol (MSRP) with registering in cluster-based mobile WSNs. The MSRP architecture consists of four phases: clustering, registering, dissemination and maintenance. The cluster-heads are elected and the network is divided into multiple clusters during the first phase. The mobile sink which comes into the communication range of a cluster-head is registered into this cluster using the second phase. Once the mobile sink is registered, it immediately receives from the cluster-head all sensed data in the cluster during the third phase. Possible new sensor nodes are added to the cluster and the cluster-head is then evaluated during the fourth phase. Yang et al. [5] propose a dynamic enclose cell (DEC) routing algorithm which decreases the control overhead by constructing cells to retain stable the network in high mobility. DEC groups the nodes into cells and develops the routing path using the cells boundaries. When the nodes are moving, only the adjacent cells of the moving nodes are reconstructed. In this way, the negative impact of the node mobility is minimized. Arboleda et al. [6] propose a cluster-based routing (CBR) protocol for mobile WSNs using zone-base information and a cluster-like communication between nodes. The CBR protocol is based on two stages: route creation and route preservation. Lambrou et al. [7] present a routing scheme for hybrid mobile WSNs that forwards packets to mobile nodes. The routing of data messages, containing position of each detected event, can be easily achieved using a geographic routing based on greedy techniques towards a fixed sink. Moreover, the sink can easily request information from a specific region or even a single static node using the position information. Santhosh-Kumar et al. [8] propose an adaptive cluster-based routing (ACBR) scheme for mobile WSNs by including mobility as a new criterion for creation and maintenance of clusters. Saad et al. [9] propose an energy efficient routing algorithm called Ellipse-Routing. Using a region-based routing, the proposed algorithm builds a virtual ellipse thanks to the source and the sink geographic positions. So, only nodes within this ellipse can forward a message towards the sink. Then, the algorithm was extended in order to take in account the errors that occur in node location.

III. Node Mobility effect on Greedy Forwarding

In greedy forwarding, the selected next forwarder is the closest neighbor to the sink in term of distance, projection, or direction based only on the nodes' location. But, mobility of nodes causes the problem of location information freshness inside the neighbors table of each sender node. This may result failures in routing decisions. This problem can be resolved by broadcasting location beacons. But when node mobility increases rapidly, the beaconing overhead (packet control) grows also rapidly.

When nodes move, the greedy forwarding mode does not often guaranty positive progression of data packets towards the sink. Thus, when a sender node selects its next forwarder, the later may be not available because it moved. In the other hand, another node can comes into the sender neighborhood, but it is not considered when selecting the next forwarder because it was not detected by the sender node. This situation has its importance when the non-detected node is the closest neighbor to the sink.

Figure 1 shows the impact of location beaconing period on greedy forwarding. When this period is long, table of neighbors of current node i will be obsolete due to movements of nodes y and z. At time t_0 (Figure 1-a), node y is leaving the communication range of node i and node z is coming into this range. At time t_1 (Figure 1-b), if node i selects as forwarder node y (i.e., a non-available neighbor), the packet will be lost due to link failure. In the other hand, the non-detected neighbor z is not considered by node i although it is the closest neighbor to sink s.

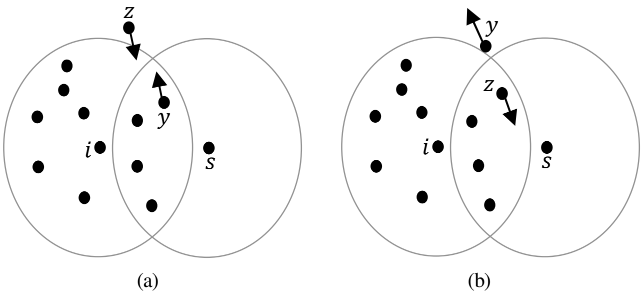


Figure 1. Beaconing period impact on greedy forwarding.

Figure 2 presents the nodes' moving direction impact on the greedy forwarding. At time t_0 (Figure 2-a), node z moves toward sink s and node y moves in the opposite sense of sink s. However, at time t_1 (Figure 2-b), node z will be the closest neighbor to sink s according to node y. But, the closest neighbor in the neighbors table of node i is always node y. Because the obsolete table of node i, the packets that should be sent to node z will be always sent to node y. This problem will be resolved in next broadcasting period. Figure 3 depicts the nodes' moving speed impact on the greedy forwarding. Current node i has two neighbors: node y and node z that move in the same direction regarding sink s, but node z is more speedy than node y. At time t_0 (Figure 3-a), node y is closest to sink s than node z. Then node y is the best forwarder node. At time t_1 (Figure 3-b), node z becomes closer to sink s than node y. But, node z will not be considered as the new best forwarder because the neighbors table of current node i is not yet updated.

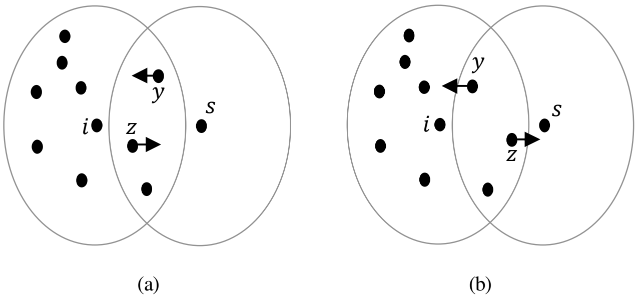


Figure 2. Node moving direction impact on greedy forwarding.

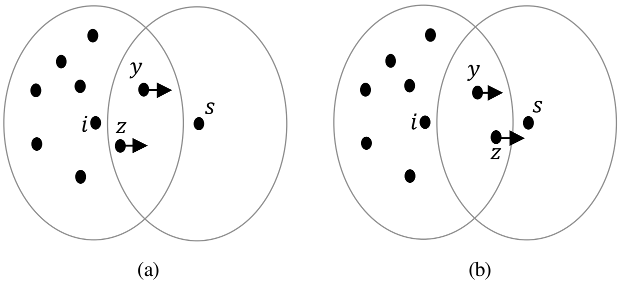


Figure 3. Nodes' moving speed impact on greedy forwarding.

Also, mobile nodes can repair voids that appear in a WSN due to their moving propriety. Consequently, greedy forwarding mode will be more preferment by using the shortest paths. With mobility of nodes, effect of the problem caused by voids (holes) on geographic routing performance is reduced [11]. In fact, the movement of sensor nodes can eliminate some voids created in WSNs. Thus, a geographic routing protocol, such as GPSR, reduces the use of bypassing mode where routing paths are long and then the energy consumption is excessive and the end-to-end delay is extended.

In greedy forwarding mode, the progress of packets toward the sink is rapid. Figure 4 shows the positive impact of nodes' mobility on routing-path length, between source node s and destination node z, by repairing a void without using a specific scheme. Consequently, GPSR will use reedy mode in the most cases. At time t_0 (Figure 4-a), a void appears in network. At time t_1 (Figure 4-b), this void is repaired thanks to movement of some nodes. Then, average path length, end-to-end delay and energy consumption will be reduced significantly.

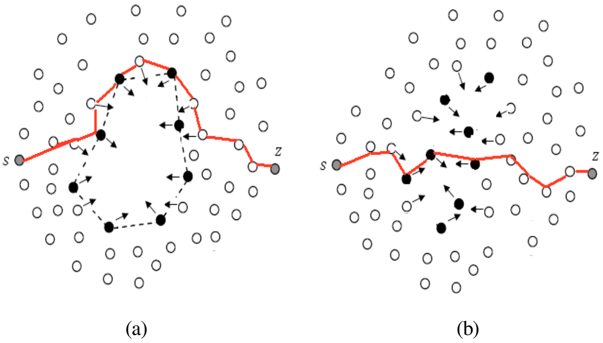


Figure 4. Nodes' movement impact on repairing voids.

IV. PROPOSED MGF MECHANISM

To be efficient in mobile WSNs, the proposed MGF mechanism uses a new decision metric when selecting the next forwarder of the current packet. This metric considers the moving direction, the moving speed, and the distance to the sink of forwarding candidate neighbors of the sender node. The MGF mechanism supposes that each node moves with a strict direction according to the sink. However, each neighbor of a node i can moves toward sink s, moves away from sink s, or stills static according to sink s. The moving direction is defined by the distance variation of the neighbors according to sink s, as shown in Figure 5. The moving direction of neighbor n, between two recent times t_0 and t_1, is calculated using its two last distances to sink s. Neighbor n may approaches (or far from) sink s in term of distance variation.

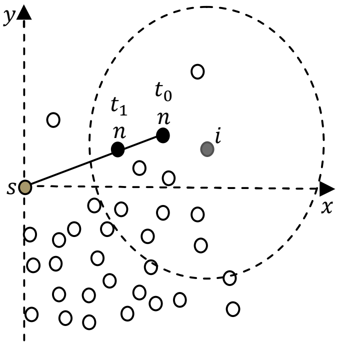


Figure 5. Node approaching to the sink in term of distance.

The greedy mode weaknesses, discussed in this section, induce packet losses, delivery delays and excessive energy consumption. Indeed, the use of only distance to select the intermediate forwarders has limits in dynamic environments caused by nodes' mobility. However, the use of periodic and frequent location beacons cannot resolve the problem because it creates packet collisions, overloads the network and consumes more energy. Consequently, some packets will be lost and other packets will be delayed. Therefore, the next forwarder selection in a node must consider multiple metrics of its neighbors, such as moving speed, moving direction and distance to the sink. The objective is to obtain a geographic routing protocol that maximizes the packet delivery ratio, minimizes the average path length and reduces the control packet overhead.

We suppose a WSN formed by one static sink and several mobile nodes. Thanks to network initialization phase, each node knows its position, positions of its neighbors, and the sink's position. Also, each node has a table (TABLE I) which contains information about its neighbors, such as location, moving speed and moving direction. The proposed MGF mechanism operates in two following phases: neighbors' information update and next-forwarder selection.

1) Neighbors' information update: Each sensor node broadcasts periodically a 1-hop location beacon informing its neighbors about its geographic position. The period of this beacon can be fixed according to the nodes' moving speed. Thanks to these beacons, each node updates a local table containing information about all neighbors. We added to these table two new fields to record moving speed and moving direction of each neighbor. TABLE I shows the structure of the neighbors' table of a node. We also added a specific field in a location beacon, where the structure is given in TABLE II, to convey the moving speed of a node to all its neighbors. When a node i receives a beacon B from its neighbor n, it checks the existence of n in its neighbors' table T. If node n does not exist, node i inserts information concerning n in T (TABLE I), else it calculates the new moving direction of n by using Formula (1), where $DT(n,s)$ represents the old distance separating $x_{n,T}$ from sink s calculated using T, $DB(n,s)$ is the new distance separating n from s calculated using B. The distances $DT(n,s)$ and $DB(n,s)$ are based on locations that are extracted from T, respectively from B, are given by the respective formulas (2) and (3). Note that $x_{n,T}$ and $y_{n,T}$ are locations of n in T, $x_{n,B}$ and $y_{n,B}$ are locations of n in B, x_s and y_s are locations of sink s in current node i. Once the above calculations are done by node i, it updates all information concerning each neighbor n in its table T.

$$Dir(n,s) = \frac{DT(n,s)}{DB(n,s)} \tag{1}$$

$$DT(n,s) = \sqrt{(x_{n,T} - x_s)^2 + (y_{n,T} - y_s)^2} \tag{2}$$

$$DB(n,s) = \sqrt{(x_{n,B} - x_s)^2 + (y_{n,B} - y_s)^2} \tag{3}$$

2) Next-forwarder selection: This phase aims to enhance the greedy mode of GPSR by handling parameters of the mobile nodes. Thus, we propose a new routing factor combining three parameters: 1) distance $DT(n,s)$ between neighbor n and sink s, 2) moving direction $Dir(n,s)$ of neighbor n and 3) moving speed $Speed(n)$ of neighbor n. When current node i has to send a packet to sink s, by using a greedy forwarding, it selects from its neighbors table a node n having the smallest $MGFactor(n,s)$ given by Formula (4), where direction $Dir(n,s)$ is given by Formula (1). Note that when $Dir(n,s)$ is equal to 1 then n is static, when it is greater than 1 then n approaches the sink, and when it is less than 1 then n moves away from the sink.

$$MGFactor(n,s) = \frac{DT(n,s) * Dir(n,s)}{Speed(n)} \tag{4}$$

TABLE I. STRUCTURE OF A NEIGHBORS TABLE.

Field	Mission/Content
ID	Identifier of a neighbor node
Position	Coordinates (x_j, y_j) of a neighbor j
Direction	Neighbor moving direction
Speed	Neighbor moving speed
ExpTime	Expire time of a neighbor in the table

TABLE II. STRUCTURE OF A LOCATION BEACON.

Field	Mission/Content
ID	Identifier of the node that sent a beacon
Position	Location of the node that sent a beacon
Speed	Moving speed of the node that sent a beacon

V. PERFORMANCE EVALUATION

We first implemented and evaluated the traditional GPSR protocol using the simulator NS2 [12] with mobility of nodes. Then we associated the proposed MGF mechanism with GPSR and evaluated in same conditions the resulting protocol (GPSR-MS). Since GPSR can handle mobility of nodes by reducing the location beacon period, we evaluate performance of this protocol under four values of this period (2, 3, 4 and 5 milliseconds) and obtained results are shown in the graphs as GPSR(2), GPSR(3), GPSR(4) and GPSR(5), respectively. This period is set to 5 milliseconds (ms) for the proposed GPSR-MS protocol.

For our simulations, we used a terrain 600×600 meters with 350 mobile sensors deployed randomly. Then they moves according to Random Waypoint Model (RWM) with a random speed in [5-20] mps (meter per second) to simulate the mobility in realistic environments. The sink is placed at the center of the terrain and 12 sources are selected randomly. Each source generates one CBR flow with a rate increased step by step from 1 to 12 pps (packet per second). For each rate and at the end of the simulation time, we measure the packet delivery ratio, the control packet overhead, the average path length and the network energy consumption per delivered packet.

Compared to the original GPSR protocol in Figure 6, the proposed GPSR-MS protocol achieves a better packet delivery ratio. Indeed, the number of packets dropped in

GPSR is important when a beaconing period is large (5 milliseconds). Also, Figure 7 shows a good performance of GPSR-MS in term of average path length compared to the original GPSR protocol. This is due to our MGF mechanism that dynamically selects as next forwarders the neighbors that move toward the sink.

Note that when the location beacon is not large (2 milliseconds) the average routing-path length is reduced in the original GPSR protocol because tables of neighbors of sensor nodes are frequently updated. Consequently, GPSR generates many location beacons that overload the network (Figure 8) and then consumes excessive energy of sensor nodes (Figure 9). On the other hand, the proposed GPSR-MS protocol delivers more data packets, generates less control overhead and optimally manages energy of nodes compared to all variants of the GPSR protocol.

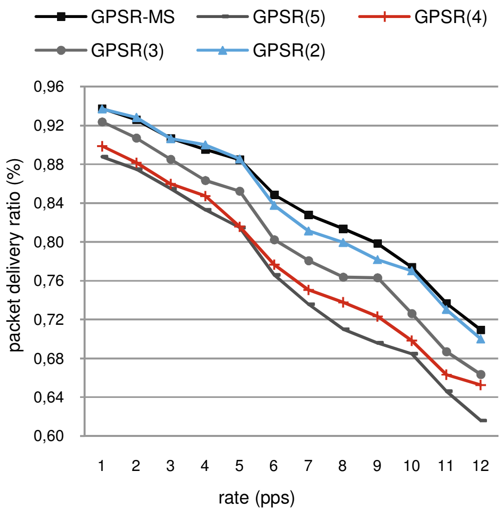


Figure 6. Packet delivery ratio *vs.* source rate.

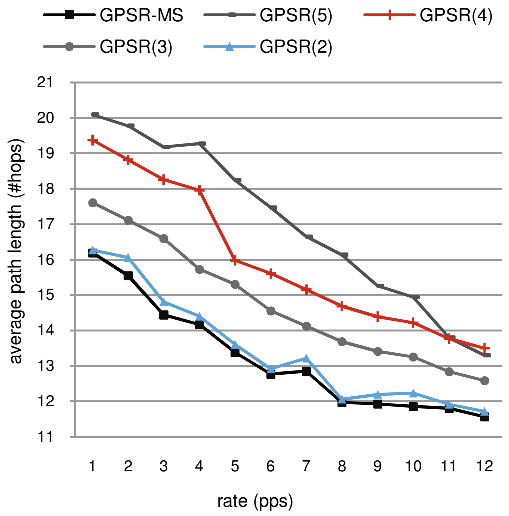


Figure 7. Average path length *vs.* source rate.

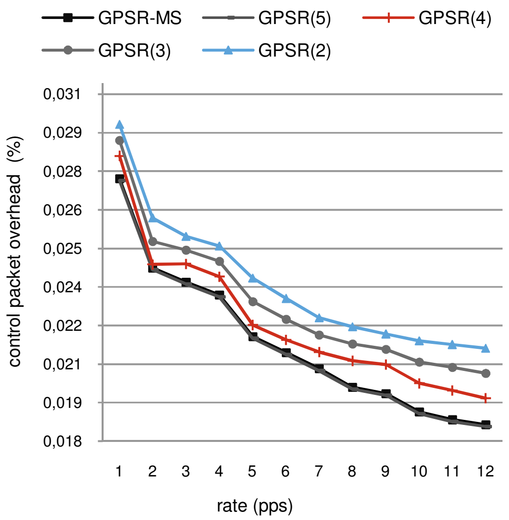


Figure 8. Control packet ovearhead *vs.* source rate.

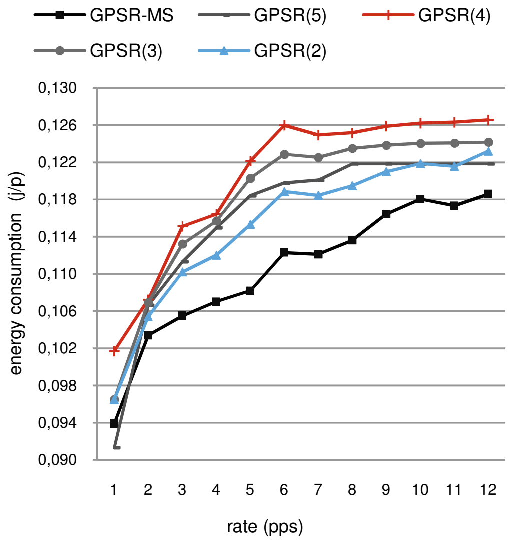


Figure 9. Energy consumption *vs.* source rate.

VI. CONCLUSION

Existing geographic schemes using greedy forwarding in mobile WSNs still have problems according mobility of sensor nodes. To contribute on solving these problems, we have proposed the MGF mechanism for mobile WSNs. It is simple to implement, saves the network resources and could be associated with various geographic routing protocols. The merit of our proposal is that the current packet is forwarded to the best neighbor node in terms of distance, moving direction and moving speed according to the static sink. We have associated the MGF mechanism with the well-known GPSR protocol and the resulting protocol, called GPSR-MS, has achieved good performance compared to different versions of the original GPSR. Indeed, GPSR-MS delivers more packets, broadcasts less control packets, uses the shortest routing paths and economizes much energy of sensor nodes. Our future work will evaluate performance of the GPSR-MS protocol with the group mobility concept.

REFERENCES

[1] Q. Fang, J. Gao, and L.J. Guibas, "Locating and bypassing holes in sensor networks," IEEE Mobile Networks and Applications, vol. 11(2), pp. 187–200, April 2006.

[2] B. Karp and H. Kung, "GPSR: Greedy perimeter stateless routing for wireless networks," Proc. of the ACM/IEEE Conference on Mobile Computing and Networking, pp. 243-254, Boston, Massachusetts, USA, August 6-11, 2000.

[3] K. Fodor and A. Vidacs, "Efficient Routing to Mobile Sinks in Wireless Sensor Networks," Proc. of the 2nd International Workshop on Performance Control in WSNs (PWSN), pp. 1–7, Austin, Texas, USA, October 23, 2007.

[4] Y.H. Wang, K.F. Huang, P.F. Fu, and J.X. Wang, "Mobile Sink Routing Protocol with Registering in Cluster-Based WSNs," Proc. of the 5th International Conference on Ubiquitous Intelligence and Computing, pp. 527-535, Oslo, Norway, June 23-25, 2008.

[5] Y. Yang, L. Dong-Hyun, P.K. Myong-Soon, and I.H. Peter, "Dynamic Enclose Cell Routing in Mobile Sensor Networks," Proc. of the Asia-Pacific Software Engineering Conference (APSEC), pp.736-737, Busan, Korea, Nov. 30 – Dec. 3, 2004.

[6] M. Liliana, C. Arboleda, and N. Nidal, "Cluster-based Routing Protocol for Mobile Sensor Networks," Proc. of the 3rd International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks, pp. 24-28, Canada, August 7–9, 2006.

[7] C.G. Panayiotou, T. Theofanis, and P. Lambrou, "A Survey on Routing Techniques supporting Mobility in Sensor Networks," Proc. of the 5th International Conference on Mobile Ad-hoc and Sensor Networks, pp. 78–85, China, Dec. 14-16, 2009.

[8] G.S. Kumar, A. Sitara, and K.P. Jacob, "An adaptive cluster based routing scheme for mobile wireless sensor networks," Proc. of the 2nd International Conference on Computing, Communications and Networking Technologies, pp. 1-5, Karur, India, July 29-31, 2010.

[9] C. Saad, A. Benslimane, J. Champ, and J.C. Konig, "Ellipse routing: A Geographic Routing Protocol for Mobile Sensor Networks with uncertain positions," Proc. of the 2nd Int'l Conference on Future Generation Communication and Networking, pp. 1–5, Sanya, China, December 13-15, 2008.

[10] R. Kouah, S. Moussaoui and M. Aissani, "Direction-based Greedy Forwarding in Mobile Wireless Sensor Networks", Proc. of the Eighth Advanced International Conference on Telecommunications (AICT), Stuttgart, Germany, May 27-June 1, 2012.

[11] M. Aissani, A. Mellouk, N. Badache, and M. Boumaza, "A Novel Approach for Void Avoidance in Wireless Sensor Networks, International Journal of Communication Systems (IJCS), vol. 23(8), pp. 945–962, 2010.

[12] Collaboration between researchers at UC Berkeley, LBL, USC/ISI, and Xerox PARC. The ns Manual, on line at: *http://www.isi.edu/nsnam/ns/*, last access in Jan. 2013.

Evaluation of Data Center Network Structures Considering Routing Methods

Yuta Shimotsuma, Yuya Trutani, Yuichi Ohsita, and Masayuki Murata
Graduate School of Information Science and Technology, Osaka University
Osaka, Japan
{y-shimotsuma, y-tarutn, y-ohsita, murata}@ist.osaka-u.ac.jp

***Abstract*—In a data center, servers communicate with each other to handle a large amount of data, and the network within the data center should provide sufficient bandwidth. In addition, the traffic pattern in a data center changes in a short interval, and the data center network should accommodate such frequently changing traffic. Since it is hard to obtain traffic information of the whole network in a short interval, the routing methods using local traffic information are suitable for a large data center network. Though there are many researches to construct data center networks, none of them discuss the characteristics of the data center network structures that can provide sufficient bandwidth considering the routing methods. In this paper, we evaluate the network structures constructed by setting various parameters of the Generalized Flattened Butterfly (GFB), FatTree and Torus considering the routing methods. The results show that the network constructed in a hierarchical manner and having multiple links from a node in each layer can provide sufficient bandwidth between all servers.**

***Keywords*—data center; network structure; routing method**

I. Introduction

In recent years, online services such as cloud computing have become popular, and the amount of data, required to be processed by such online services, is increasing. To handle such a large amount of data, large data centers with hundreds of thousands of servers have been built.

In a data center, servers handle a large amount of data by communicating with each other. The performance of the data center depends on the network connecting between servers. There are many researches to construct data center networks [1-5]. A network structure called FatTree, which can provide large bandwidth between all servers using commodity switches with a small number of ports, has been proposed by M. Al-Fares et al. [1]. A cost-efficient structure for a high-radix network, which can provide large bandwidth and a small average number of hops between all servers, has also been proposed [2]. The methods to connect a large number of servers with a small number of switches have also been proposed by C. Guo et al. [3, 4]. In addition, we have proposed the method, named GFB, which can construct appropriate network structures by setting parameters to meet the demands of applications in a data center [5]. However, these papers aimed to propose the specific network structures, and did not discuss the characteristics of network structures suitable to a data center sufficiently.

In a data center, handling significant traffic changes and failures is another problem. Traffic in a large data center changes frequently [6]. Failures are also common in a large data center [7]. Even when such traffic changes or failures occur, we should keep the performance of the data center. One of the important methods to keep the performance of the data center is the routing methods. By changing the routes based on the current traffic and network status, we can keep the large bandwidth between servers even in case of traffic changes or failures.

To calculate the optimal routes of the traffic between all server pairs, we require the traffic information of the whole network. In a larger data center, however, it is difficult to collect the traffic information of the whole network in a short interval, while the traffic changes frequently. Thus, in a data center, a routing method should control routes between servers based on the local information which can be obtained by each switch. There are several researches to control routes between servers for a data center [8]. A. Greenberg et al. proposed a method to distribute loads by randomly choosing switches in high layers at the Tree structure [8, 1]. M. Al-Fares et al. use the Equal-Cost Multi-Path routing (ECMP) to distribute loads to multiple paths in the FatTree [1]. However, these methods consider only a particular network structure and none of them clarify the network structure suitable to the routing methods.

In this paper, we evaluate the various network structures considering the routing methods for a data center. Then, we clarify the characteristics of the network structures where the routing methods can provide a large bandwidth between all servers.

The rest of this paper is organized as follow. Sections II and III explain the network structures and the routing methods used in this paper, respectively. In Section IV, we describe the evaluation process. Then we discuss the results of the evaluation in Section V. Finally, we conclude this paper in Section VI.

II. Network structure

In this paper, we compare the performance of various network structures constructed with switches which have the same number of 10 Gbps ports, and clarify the characteristics of the network structures that can provide sufficient bandwidth between all servers. In this paper, we use the following network structures.

A. GFB

The GFB [5] is a network structure which is constructed hierarchically; the upper-layer GFB is constructed by connecting the lower-layer GFBs. The GFB has the following parameters.

- Number of layers: k
- Number of links per node used to construct layer-k GFB: L_k

Copyright (c) IARIA, 2013.   ISBN: 978-1-61208-256-1

- Number of layer-$(k-1)$ GFBs used to construct layer-k GFB: N_k

By setting these parameters, we can construct various network structures.

The layer-k GFB is constructed based on the ID assigned for each layer-$(k-1)$ GFB. First, the GFBs having the nearest ID are connected to construct a ring topology. Then the residual links are used to connect the layer-$(k-1)$ GFBs so that the interval of the IDs of the layer-$(k-1)$ GFBs connected to a certain layer-$(k-1)$ GFB is equal.

In this paper, we construct and compare the various network structures by setting the parameters of the GFB.

B. FatTree

The method to construct the topology called *FatTree* by using switches with small number of ports was proposed by Al-Fares et al. [1]. The FatTree is a tree structure including multiple roots and multiple pods constructed of multiple switches.

Each pod is regarded as the switch with a large number of ports constructed by multiple switches with a small number of ports. Pods are constructed as the butterfly topology, where each switch uses a half of its ports to connect it to the switches of the upper layer, and the other half of its ports to connect it to the switches of the lower layer. The switches at the lowest layers are connected to the servers.

Though the method proposed by Al-Fares et al. [1] constructs the 3-layer FatTree, which is constructed of the root switches and the pods with two layers, we can construct the FatTree topologies with more layers. The k-layer FatTree constructed of switches with n ports includes $(2k-1)\frac{n}{2}^{k-1}$ switches.

C. Torus

Torus is a network structure constructed by locating switches in multidimensional grid. The n-dimensional Torus is constructed based on the IDs of the switch which are the n-dimensional vectors. In the n-dimensional Torus, the switch A is connected to the switch B whose ID is next to the ID of the switch A in a dimension and equals the ID of the switch A in the other dimensions. The n-dimensional Torus requires switches with $2n$ ports.

III. Routing method

A routing method selects the route passed by each packet from the candidates of the routes by using the traffic information. In this paper, we classify the routing methods based on the traffic information used by them.

The first one is a method that does not use the traffic information. In this method, each switch selects the next node passed by a packet randomly from the candidates of the next hop. We call this method the *random routing*.

The second one uses only the local traffic information that can be obtained by each switch. In this method, each switch selects the next hop whose corresponding link has the least utilization among the candidates. We call this method the *local routing*.

The third method uses traffic information of the whole network. In this routing method, a server, which is used to calculate a route of traffic, collects the overall traffic information and determines a route of traffic based on this information. We call this method the *global routing*.

In this paper, we use the above three types of the routing methods. In each method, we use the two types of the candidates of the routes; the first type of the candidates includes only the shortest paths, and the second one includes the shortest paths and the paths which are one hop longer than the shortest path.

The random routing and the local routing immediately adapt the route of a packet so as to suite the current traffic without collecting overall traffic information when traffic changes. On the other hand, the global routing cannot change the routes before the traffic information of the whole network is collected. For the global routing, we evaluate both cases that the traffic information after the traffic changes is collected or only the traffic information before the traffic changes is obtained.

In this paper, we also evaluate the case that link failure occurs. In all of the routing methods used in this paper, we calculate the routes after eliminating the routes including the failed links from the candidates. If no candidates remain, the traffic cannot be accommodated.

IV. Evaluation process

A. Overview

In this paper, we evaluate the combination of the network structures and the routing methods. In this evaluation, we generate traffic between all servers. Then, the network structure accommodates the traffic along the routes calculated by the routing methods. If we cannot find the routes without congestion, the traffic cannot be accommodated. In this paper, we focus on the case of the maximum amount of traffic that can be accommodated by the combination of the network structures and the routing methods, to evaluate the maximum bandwidth provided between all servers.

B. Traffic

A data center has two kinds of traffic; the mice traffic and the elephant traffic. The mice traffic is generated when a server exchanges a small message with the other servers. The elephant traffic is generated when a server exchanges a big data, such as files.

In this paper, the traffic is generated as the combination of the mice and elephant traffic. The mice traffic is generated between all servers and the amounts of the mice traffic are set to the random values so as to make the mice traffic occupying around 10 % of the bandwidth of all links in a network. The elephant traffic is generated between randomly selected server pairs. The amount of the elephant traffic is set to the largest value that can be accommodated by the combination of the network structure and the routing method. To set the amount

TABLE I
NETWORK STRUCTURES WITH VARIOUS NUMBERS OF LINKS AT EACH LAYER

name	N_0	N_1	N_2	L_0	L_1	L_2
GFB224	4	6	14	2	2	4
GFB242	4	6	14	2	4	2
GFB314	4	6	14	3	1	4
GFB323	4	6	14	3	2	3

of elephant traffic, we increase the amount of the elephant traffic unless the congestion occurs.

In this evaluation, we generate the traffic change by regenerating the mice traffic and newly selecting the server pairs where the elephant traffic is generated.

C. Metrics

In this paper, we focus on the case that the largest amount of the elephant traffic is generated. Then we investigate the smallest amount of the elephant traffic between server pair among the server pairs where the elephant traffic is generated. By investigating the smallest amount of the elephant traffic, we compare the bandwidth between servers that can be provided at least. In addition, when we generate link failures, we also investigate the communication failure ratio which is defined by the ratio of flows between switch pairs that have no routes from the source server to the destination server.

V. EVALUATION

In this section, we discuss the characteristics of a network structures which can provide a large bandwidth between all servers. In this evaluation, to clarify the characteristics of the network structures suitable to a data center, we compare the network structures constructed by setting various parameters of the GFB. The GFB constructs various network structures by setting its parameters. By comparing the network structures constructed by setting the parameters of the GFB, we clarify the characteristics of the network structures suitable to a data center. First, we compare the performances of network structures with various numbers of links at each layer. Then, we compare the performances of network structures with various numbers of layers.

We also compare the performances of network structures when link failures occur. Finally we compare the performances of the GFB with that of the Torus and the FatTree.

A. Comparison of network structures with various link at each layer

In this subsection, we use the network structures shown in Table I, constructed by connecting 336 switches with 8 ports. The network structures are constructed by setting the parameters in the GFB. All of the network structures used in this subsection, N_i are set to the same value. We change L_i to clarify the impacts of the number of the links in each layer on the bandwidth provided between servers. In this subsection, we refer each network structure by the number of links at each layer. The elephant traffic to one destination server per one switch is generated. The server pairs where the elephant traffic is generated are selected randomly.

Figure 1 shows the minimum amount of the elephant traffic between a certain server pair which can be accommodated to each network structure by using each routing method. As shown in this figure, the global routing can provide the largest bandwidth between all servers at any network structures if the accurate traffic information is collected. The global routing, however, can provide only as large bandwidth between all servers as the random routing if we have only the traffic information before the traffic change. It is difficult to collect accurate traffic information of all links in a short time interval, though traffic in a large data center changes frequently[6]. Therefore, in the data center, the routes should be calculated based on the local traffic information that can be obtained by each switch.

As shown in Fig. 1, the local routing can provide as large bandwidth between all servers as the random routing. This is because each switch does not have the information of the links connected to the next switches. Thus, each switch cannot know whether the next switch has the congested links, and may select the next switch having the congested links. As a result, the local routing provides only the similar bandwidth to the random routing.

This figure also indicates that the candidate routes including the one hop longer paths provide larger bandwidth. This is because each switch has more candidates by including one hop longer path. Thus, it is easy to avoid the congested links.

In the following evaluation, we focus on the characteristics of each network structure. We compare the bandwidth provided by the random routing with the candidates including the shortest path and the one hop longer path.

As shown in Fig. 1, GFB224 can provide the largest bandwidth between all servers. The largest bandwidth between all servers depends on the maximum link utilization. To discuss the link utilization, we model the probability that traffic between each switch pair passes the link by Eqs. (1) and (2). In this model, we assume that the link passed by the traffic is selected randomly among the link in the network. We also assume that the probability that the traffic between each switch pair passes the link depends only on the layer of the link, because each layer of the GFB is symmetric.

P_k is the probability to select a link in the layer k as the link passed by the traffic. P_1 is calculated by

$$P_1 = p \times \left(1 - \prod_{k=0}^{H_1-1}\left(1 - \frac{1}{l-k}\right)\right), \qquad (1)$$

where p is the probability that the flow passes a particular layer-1 GFB, l is the number of links at the layer-1 GFB and H_1 is the number of hops at the layer-1 GFB.

The layer-2 GFB uses $(N_1 \times L_2)$ links per layer-1 GFB for the connection between the layer-1 GFBs. We assume that the sufficient number of links are used to connect the layer-1 GFBs, and the layer-1 GFBs are fully connected. Thus, no traffic passes multiple links between layer-1 GFBs. Therefore,

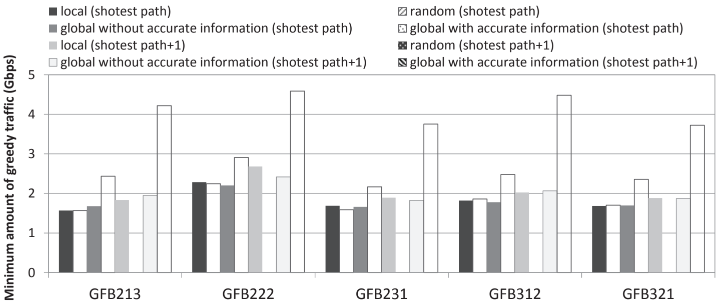


Fig. 1. Minimum amount of the elephant traffic between a certain server pair which can be accommodated by each network structure with various numbers of links at each layer

P_2 is calculated by

$$P_2 = q \times \frac{1}{l_{GFB}}, \tag{2}$$

where q is the probability that the flow passes a particular link between layer-1 GFB-pair and l_{GFB} is the number of the links between a certain layer-1 GFB-pair.

A network structure whose P_1 and P_2 in these equations is small can provide large bandwidth between all servers. As shown in Eq. (1), to make P_1 smaller, a network structure should have smaller H_1. We can reduce H_1 by reducing the number of switches in the layer-1 GFB or increasing the number of the links in the layer-1 GFB since the maximum number of hops at the layer-1 GFB is calculated by $\left\lceil \frac{N_1}{2\times(L_1-1)} \right\rceil$ according to Tarutani et al. [5]. Increasing the number of links at the higher layer also reduces H_1, because if multiple switches are connected to the laye-1 GFB including the destination, the flow goes out the GFB including the source from the switch nearest to the source among the switches connected to the destination GFB.

As shown in Eq. (2), to make P_2 smaller, a network structure should have large l_{GFB}. To increase l_{GFB}, we need to increase the number of links at layer-2 GFB, or reduce the number of the layer-1 GFBs connected at the layer-2 GFB because l_{GFB} is calculated by $\frac{N_1 \times L_2}{N_2}$.

The above discussion is also applicable when the number of layers is more than 2. The probability that the flow passes the links at the higher layer is small when the number of the links between the GFB pair is large.

According to the above discussion, to provide a large bandwidth, the number of links at each layer L_k should be sufficiently large compared with N_k. In this subsection, we set N_1 to 4, N_2 to 6, and N_3 to 14. In the GFB242, the number of links at the layer 3, L_3 is small compared with $N_3 = 14$. In the GFB314, the number of links at the layer 2, L_2 is small compared with $N_2 = 6$. Thus, these network structures cannot provide a large bandwidth between servers.

The GFB224 provides a larger bandwidth than the GFB323. This is because the number of links at the lowest layer is sufficient even when $L_1 = 2$, since the GFB of the lowest layer includes only 4 switches. In this case, by adding more links to the upper layers, we reduce the number of hops to the other layer-1 GFBs, and increase the bandwidth provided between servers compared with adding more links to the lowest layer.

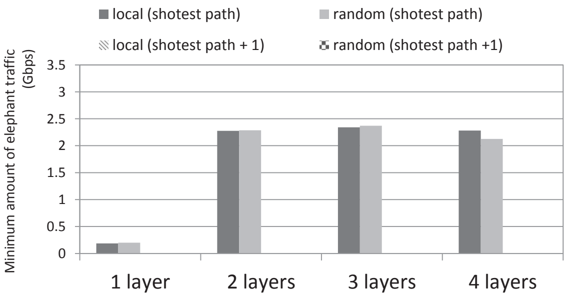


Fig. 2. Minimum amount of the elephant traffic between a certain server pair which can be accommodated by each network structure with various numbers of layers

To summarize the above discussions, the GFBs where the number of links at a particular layer is small cannot provide large bandwidth between all servers. Therefore we need a network structure which has sufficient number of links at all layers.

B. *Comparison of network structures with various numbers of layers*

In this subsection, we evaluate the performances of the network structures when we change the number of layers. In this evaluation, we use the network structures shown in Table II. In all network structures used in this subsection are constructed of 336 switches with 8 ports. The network structures constructed with various numbers of layers by setting the parameters in GFB. In this subsection, we refer each network structure by the number of layers. In this subsection, the elephant traffic to one destination server per one switch is generated. The server pairs where the elephant traffic is generated are selected randomly.

Figure 2 shows the minimum amount of elephant traffic which can be accommodated by each network structure. As shown in Fig. 2, GFB(1 layer) can provide the smallest bandwidth between all servers. GFB(2 layers) and GFB(3 layers) can provide lager bandwidth between all servers. This is because the large number of hops between servers in the

TABLE II
NETWORK STRUCTURES WITH VARIOUS NUMBERS OF LAYERS

name	the number of layer	N_0	N_1	N_2	N_3	L_0	L_1	L_2	L_3
GFB (1 layer)	1	336	-	-	-	8	-	-	-
GFB (2 layers)	2	14	24	-	-	4	4	-	-
GFB (3 layers)	3	4	6	14	-	2	2	4	-
GFB (4 layers)	4	3	4	4	7	2	2	2	2

1-layer GFB increases the link utilization.

The number of hops between a switch pair is calculated as follows. $H_k^{(i)}$ is the number of hops between a switch pair at the layer-k GFB with ID(i), and P is the set of ID of the layer-$(k-1)$ GFBs where a flow passes.

$$H_k = \sum_{i \in P} (H_{k-1}^{(i)} + 1) - 1 \tag{3}$$

As shown in Eq. (3), the number of hops between a switch pair depends on the number of GFBs at the lower layer passed by the traffic and the number of hops at the lower layer GFBs. The 1-layer GFB is constructed by adding links to a ring topology so that the interval of the IDs of the switches connected to a certain switch is equal. When the number of switches is 336, the interval of the IDs is 56, and the number of hops between switches is large. As shown in Eq. (1), the large number of hops causes the high probability that a flow passes a particular link. As a result, the 1-layer GFB cannot provide the smallest bandwidth.

In the 2-layer GFB used in this evaluation, the layer-1 GFB includes 14 switches, and the layer-2 GFB includes 24 fully connected layer-1 GFBs. Thus, the number of hops in each layer is significantly smaller than the 1-layer GFB, and the number of hops between a switch pair is small. As a result, the probability that a flow passes a particular link is smaller, and a network structures constructed in a hierarchical manner can provide larger bandwidth between all servers than the 1-layer GFB.

In the 4-layer GFB, we cannot provide as large bandwidth as the 2-layer or 3-layer GFB. This is because the number of hops in the 4-layer GFB becomes larger than that in the 2-layer or 3-layer GFB. As shown in Eq. (3), the number of hops between a switch pair is calculated by adding the number of hops at lower layer recursively. The number of the recursive calculations increases if the number of the layers increases. If the increase of the number of the layers does not reduce the number of hops in each layer sufficiently, the increase of the number of the layers makes the number of hops between a switch pair large. As a result, though the 4-layer GFB can provide larger bandwidth than the 1-layer GFB, the 4-layer GFB can provide only smaller bandwidth than the 2-layer or 3-layer GFB.

As discussed above, to provide larger bandwidth between all servers, the network structure should connect any switch pairs with a small number of hops. It is effective to reduce the number of GFBs at each layer by constructed in a hierarchical manner. However, if the number of layers is too large, the traffic between a switch pair passes many layers, and its number of hops becomes large. Therefore, we need to set the number of layer to as small value as possible without a large number of hops in each layer.

C. Evaluation in the case of link failures

In a large data center, failures such as link failure are common. Thus, the network should be robust to failures to keep the service provided in a data center even when failure occurs. In this subsection, we evaluate the communication failure ratio of various network structures when links have failed, and discuss the characteristics of a network structure robust to failures.

Figure 3 and 4 show the communication failure ratio of the network structures shown at Tables I and II when several links failed. As shown in Figs. 3 and 4, each switch having the candidates of the routes including the one hop longer path can achieve the smaller communication failure ratio than the case of the candidates including only the shortest path. This is because each switch has more candidates by including one hop longer path and it is easy to bypath the failed links.

As shown in Fig. 3, the communication failure ratio in the GFB224 is the smallest. As discussed in the previous subsection, the GFB242, the GFB314, and the GFB323 have the links passed by many flows. The failures of such links passed by many flows cause the large communication failure ratio in these network structures.

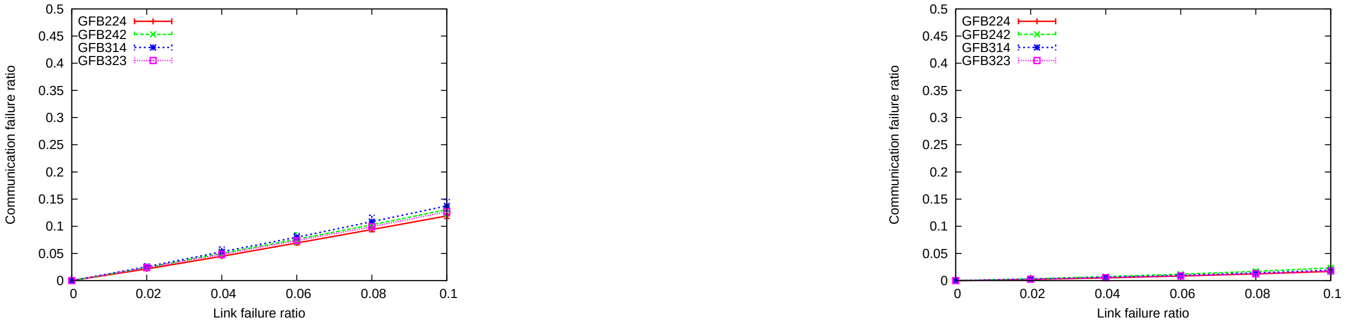
When the candidate routes include only the shortest paths, the communication failure ratio is the lowest in the 4-layer GFB. This is because the number of routes between switches is large in the 4-layer GFB. In the 2-layer GFB or the 3-layer GFB, though the number of hops is smaller than the 4-layer GFB, the number of the shortest paths between servers is small.

By including the one hop longer paths in the candidates of the routes, the 2-layer GFB and the 3-layer GFB also achieves the similar communication failure ratio to the 4-layer GFB. However, the communication failure ratio of the 1-layer GFB is large even when the candidate routes include the one hop longer paths. This is because the probability that each flow passes the link is large in the 1-layer GFB. Thus, the failure of the links in the 1-layer GFB has a large impact on many flows, and causes the large communication failure ratio.

As discussed above, a network structure, which has multiple routes between servers and the small probability that each flow passes each link, are robust to failures.

D. Comparison of GFB, Torus and FatTree

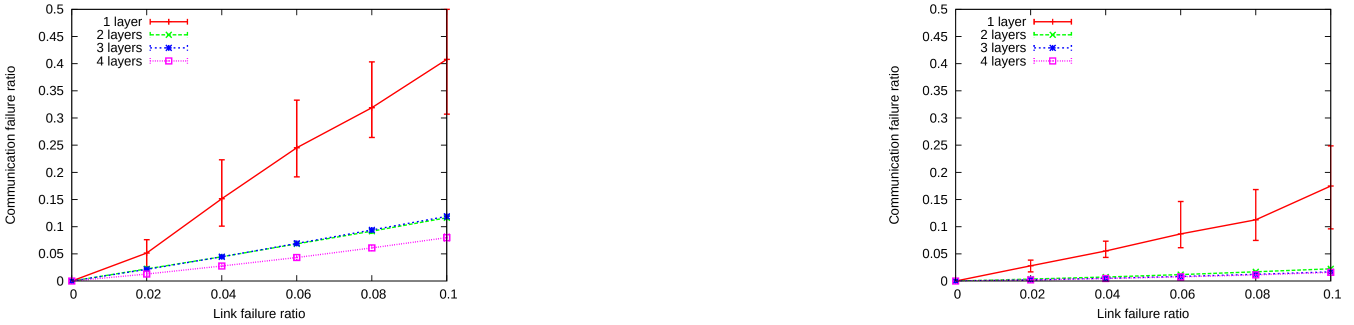
In this subsection, we compare the performances of the network structures constructed by setting parameters in the



(a) In the case of the candidate routes including only the shortest paths

(b) In the case of the candidate routes including one hop longer paths

Fig. 3. Communication failure ratio of network structures with various numbers of links at each layer in the case of link failures



(a) The case of candidate routes including only the shortest paths

(b) The case of candidate routes including one hop longer paths

Fig. 4. Communication failure ratio of network structures with various numbers of layers in the case of link failures

TABLE III
COMPARISON OF THE NETWORK STRUCTURE USED IN OUR EVALUATION

name	the number of switches	the number of links	average hops	maximum hops
GFB(5,5,6)	150	450	4.09	7
Torus	150	450	4.93	8
FatTree	189	486	6.62	7

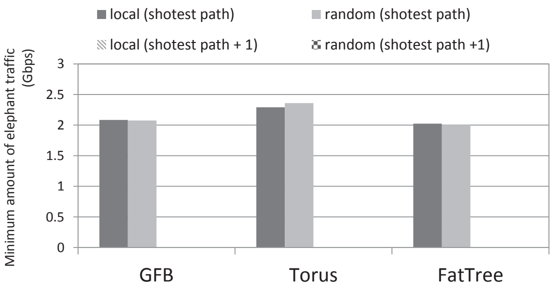


Fig. 5. Minimum amount of elephant traffic which can be accommodated by GFB, Torus and FatTree when the elephant traffic to one destination server per one switch is generated

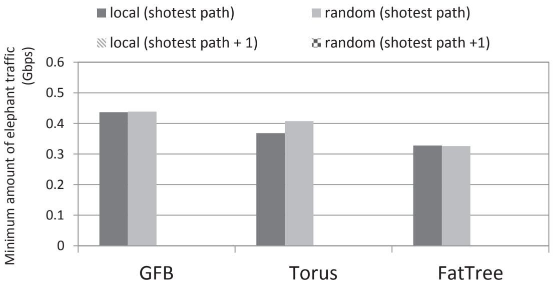


Fig. 6. Minimum amount of elephant traffic which can be accommodated by GFB, Torus and FatTree when the elephant traffic to ten destination server per one switch is generated

GFB with that of the Torus and the FatTree. The Torus is the well-studied network structure. Compared with the GFB, the Torus has a large number of hops, but it has more routes between a switch pair. The FatTree is the network structure used in the existing data centers. The FatTree has the same number of maximum number of hops as the GFB, but the

ISBN: 978-1-61208-256-1

average number of hops between servers is larger than the GFB or the Torus. Similar to the Torus, the FatTree has multiple routes between server pairs. By comparing the GFB with these network structures, we evaluate the impacts of the average number of hops and the number of routes between switch pairs.

In this evaluation, we use the network structures shown in Table III. Table III also includes the average number of hops and the max number of hops between all switch in each network structure. The GFB and the Torus are constructed of 150 switches with 6 ports, and the FatTree is constructed of 189 switches with 6 ports. In this evaluation we use a network structure by setting parameters $[N_0 = 5, N_1 = 5, N_2 = 6, L_0 = 2, L_1 = 2, L_2 = 2]$ in GFB. Also we use the 3-dimentional 5 × 5 × 6 Torus, and the 4-layer FatTree which is constructed by allocating 54 switches at the lowest layer. In the FatTree, only the switches at the lowest layer are connected to servers. In the FatTree, we generate the same number of flows as in the GFB and the Torus. In the FatTree, we obtain only the case that each switch has the candidates of the routes including only the shortest paths because the one hop longer path does not exist in the FatTree.

In this subsection, we obtain the evaluation results of two cases of the ratios of the generated elephant traffic. In the first case, we generate that the elephant traffic to one destination server per one switch. In the other case, we generate the elephant traffic to ten destination servers per one switch is generated. Figure 5 shows the minimum amount of the elephant traffic which can be accommodated by each network structure when the elephant traffic to one destination server per one switch is generated. Figure 6 shows the minimum amount of the elephant traffic which can be accommodated by each network structure when the elephant traffic to ten destination servers per one switch is generated. As shown in Figs. 5 and 6, the GFB and the Torus can provide larger bandwidth between all servers than the FatTree. This is caused by the large average number of hops of the FatTree. In the FatTree, each flow passes more links, and requires the bandwidth of a large number of links. As a result, the bandwidth of each link is occupied with a small amount of traffic between servers.

Fig. 5 also indicates that the Torus can accommodate more traffic than the GFB, if the candidate routes include only the shortest path. This is because the Torus has a larger number of the shortest paths between switches than the GFB. However, in case of the candidate routes including one hop longer path, the GFB also has the sufficient number of routes between switches, and can accommodate the similar amount of traffic to the Torus.

As shown in Fig. 6, the GFB can provide the largest bandwidth between all servers by using the local routing with the candidates including one hop longer path. Though the Torus provides the largest bandwidth in Fig. 5, the Torus cannot provide as large bandwidth as the GFB in Fig. 6. This is because that the number of hops between switch pairs is larger in the Torus than the GFB. Thus, in the Torus, the bandwidth of each link is occupied with a small amount of traffic between servers.

As discussed above, the network structure with a large number of candidate routes between servers is required to provide a large bandwidth between servers when the number of flows in the network is small. However, when the number of flows in the network is large, the average number of hops of the flow becomes more important. Thus, the network structures should have small average number of hops to provide a large bandwidth when the number of flows is large.

VI. Conclusion

In this paper, we evaluated the data center networks considering the routing methods. According to the results, to provide a large bandwidth between servers, we should make the number of hops small by constructing the network in a hierarchical manner, and make the number of routes between servers large by adding multiple links from a node in each layer.

Acknowledgments

This work is a part of ”Research & Development of Basic Technologies for High Performance Opto-electronic Hybrid Packet Router” supported by National Institute of Information and Communications Technology (NICT).

References

[1] M. Al-Fares, A. Loukissas, and A. Vahdat, “A Scalable, Commodity Data Center Network Architecture,” ACM SIGCOMM Computer Communication Review, vol. 38, Oct. 2008, pp. 63–74.

[2] J. Kim, W. J. Dally, and D. Abts, “Flattened butterfly: a cost-efficient topology for high-radix networks,” in Proceedings of the 34th annual international symposium on Computer architecture, vol. 35, Jun. 2007, pp. 126–137.

[3] C. Guo, H. Wu, K. Tan, L. Shi, Y. Zhang, and S. Lu, “DCell: A scalable and fault-tolerant network structure for data centers,” ACM SIGCOMM Computer Communication Review, vol. 38, Aug. 2008, pp. 75–86.

[4] C. G. et al., “BCube:A high performance, server-centric network architecture for modular data centers,” ACM SIGCOMM Computer Communication Review, vol. 39, Aug. 2009, pp. 63–74.

[5] Y. Tarutani, Y. Ohsita, and M. Murata, “A Virtual Network to Achieve Low Energy Consumption in Optical Large-scale Datacenter,” in Proceedings of the 13th International Conference on Communication Systems IEEE ICCS 2012, Nov. 2012.

[6] T. Benson, A. Anand, A. Akella, and M. Zhang, “MicroTE: Fine Grained Traffic Engineering for Data,” in Proceedings of ACM CoNEXT, Dec. 2011, pp. 1–12.

[7] P. Gill, N. Jain, and N. Nagappan, “Understanding network failures in data centers: measurement, analysis, and implications,” ACM SIGCOMM Computer Communication Review, vol. 41, Aug. 2011, pp. 350–361.

[8] A. G. et al., “VL2: A scalable and flexible data center network,” ACM SIGCOMM Computer Communication Review, vol. 39, Aug. 2009, pp. 51–62.

Economically Efficient Interdomain Overlay Network Based on ISP Alliance

Xun Shao
Graduate School of Information Science and Technology
Osaka University
Osaka, Japan
Email: x-shao@ist.osaka-u.ac-jp

Go Hasegawa, Yoshiaki Taniguchi, Hirotaka Nakano
Cybermedia Center, Osaka University
Osaka, Japan
Email: (hasegawa, y-tanigu, nakano)@cmc.osaka-u.ac.jp

***Abstract*—As interdomain routing protocol, BGP is a fairly simple, and allows plenty of policies based on ISPs' preferences. However, recent studies show that BGP routes are often non-optimal in end-to-end performance, due to technological and economic reasons. To obtain improved end-to-end performance, overlay routing, which can change traffic routing in application layer, has gained attention. However, overlay routing often violates BGP routing policies and harms ISPs' interest. In order to take the advantage of overlay to improve the end-to-end performance, while overcome the disadvantages, we propose a novel interdomain overlay structure, in which overlay nodes are operated by ISPs within an ISP alliance. The traffic between ISPs within the alliance could be routed by overlay routing, and the other traffic is still routed by BGP. As economic structure plays very important role in interdomain routing, we then propose an effective and fair charging and pricing scheme within the ISP alliance in correspondence with the overlay routing structure. At last, we give a simple pricing algorithm, with which ISPs can find the optimal prices in the practice. By mathematical analysis and numerical experiments, we show the correctness and convergence of the pricing algorithm.**

***Keywords*-BGP; interdomain; overlay routing; charging; pricing**

I. Introduction

The Internet is composed of thousands of networks owned by Internet Service Providers (ISPs), which are selfish, often competing economic entities. The task of establishing routes between ISPs is called interdomain routing. The standard interdomain routing protocol is the Border Gateway Protocol (BGP), which is a path-vector protocol. BGP allows routing policies to override distance-based metrics with policy-based metrics. ISPs often wish to control next hop selection so as to reflect agreements or relationships they have with their neighbors. Two common relationships ISPs have are: customer-provider, where one ISP pays another to forward its traffic, peer-peer, where two ISPs agree that connecting directly to each other would mutually benefit both. ISPs often prefer customer-learned routes over routes learned from peers and providers when both are available. This is because sending traffic through customers generates revenue for the ISP while sending traffic through providers costs the ISP money.

Although BGP is the sole interdomain routing protocol currently, the authors in [1] found that the default BGP paths are not often optimal with respect to end-to-end performance. At any time, for 30 to 80 percent of the paths we can find there are alternative paths with significantly improved measures of quality. There are both technical and economic reasons to expect that BGP routing is non-optimal. Theoretically, the BGP uses "shortest" path routing, where paths are chosen to minimize hop count. However, hop count correlates less well with performance than explicit measurements. Moreover, economic considerations can also limit routing options. Routing policies are driven by many concerns especially the contracts with neighbor ISPs and monetary prices.

In order to realize better end-to-end performance, overlay networks [2]–[6] have recently gained attention as a viable alternative to overcome functionality limitations of BGP. The basic idea of overlay networks is to form a virtual network on top of the physical networks so that overlay nodes can be customized to incorporate complex functionality without modifying the native IP network. Typically, these overlays route packets over paths made up of one or more overlay links to achieve a specific end-to-end objective.

Routing in overlay networks often violates BGP routing policies [7]–[11]. Consider, for example, a hypothetical ISP-level connectivity graph as shown in Fig. I. In that figure, overlay nodes exist in a, b, d and e. Overlay nodes are trying to obtain the best possible route to each other. Overlay node in b can route data to overlay node e using the overlay path bdce, which results in dfs ISP being used for transiting traffic. This is a violation of the ISP's transit policy at d. From an economic perspective, we see that the performance improvement comes at the expense of d, as d has to pay b and c for the overlay traffic through the illegitimate path. Because overlays operate at the application layer, the violations typically go undetected by the native layer. In order to take the advantage of overlay to improve the end-to-end performance, while maintaining the ISPs' benefit, we propose an economically efficient interdomain overlay structure operated by ISPs based on ISP alliance. The ISP alliance in this paper is formed by adjacent ISPs. Each ISP in the alliance operates one or more overlay nodes, and all the overlay nodes form an overlay network. The traffic between ISPs in the alliance can be routed by overlay routing

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-256-1

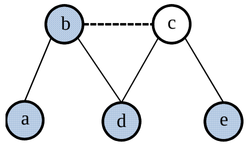


Figure 1. An example of routing policy violation. Solid circles represent ISPs with overlay nodes in their domains, empty circle represents ISP with no overlay node in its domain. Solid lines represent transit relation, and dashed line represents peering relation

for better end-to-end performance. The routes are chosen by traffic source ISPs, and multiple path routing is also employed. The alliance is limited to adjacent ISPs for three reasons: first, according to the results in [12] and [6], the alliance formed by regional ISPs can improve the end-to-end performance significantly, second, it is easy for regional ISP alliance to set up, manage and maintain compared with global one, third, the ISPs' loss of interest caused by policy violation can be avoided, which we go into detail in Section III. In the overlay network, in order to take full advantage of bandwidth resource, the business relationships such as provider-customer and peer-peer do not exist, neither BGP routing policies. In the overlay network, each ISP is responsible for transiting traffic across its network for its neighbors. As reward, it can receive money from the ISPs who send traffic. Within the ISP alliance, we avoid harming ISPs' interests caused by policy violation by introducing novel economic structure.

As ISPs are individual economic entities, we cannot separate routing from economic. ISPs always have dual roles: when sending traffic, they are customers, who pay the transit providers; when transiting traffic across their network, they are providers who charge the traffic sender. As customer, ISPs prefer the paths with better performance and lower price; while as provider, ISPs make pricing decision to maximize revenue. In this paper, we deal with the two roles in a unified effective economic structure. By the word "effective", we mean that the ISP who is willing to pay more money can enjoy better routes. On the other hand, if specific route has better performance, the ISPs along it should gain more revenue by making optimal pricing decision. Besides effectiveness, fairness among ISPs along identical route is also important.

In the above routing and charging structure, we model the relationship between ISPs' routing decision and properties of routes – performance and price. As customer, ISPs' routing decision is decided by route performances, prices and ISP's own property. The decision includes which path to choose, and how much traffic to send. Based on this model, we study ISPs' pricing scheme as provider, and obtain the optimal price to maximize the revenue. In order to realize the optimal price in the practice, we study the non-cooperative pricing game [13] played by individual ISPs, and find that it is neither effective nor fair. We believe that if ISPs realize the undesired properties of the non-cooperative pricing game, they would seek cooperation. We then propose a pricing scheme based on route bundle – a bundle of routes having the same entrance ISP with each other – and prove that it is a better pricing scheme than non-cooperative pricing game. At last, we give a simple algorithm for route bundles to find the optimal prices, which can maximize the revenue. According to mathematical analysis and numerical experiments, we show that our pricing algorithm is correct, and can always converge to the optimal price.

The remainder of this paper proceeds as follows. Section II gives ISP alliance based overlay structure, routing and charging scheme. Section III goes into detail with ISPs' routing, charging and pricing. We conclude in Section IV.

II. ISP ALLIANCE BASED INTERDOMAIN OVERLAY NETWORK STRUCTURE

In order to take the advantages of overlay networks while overcome the disadvantages, we propose an interdomain overlay network, in which overlay nodes are operated by ISPs belonging to the same alliance. In this section, we elaborate the structure of the ISP alliance, and make a brief discussion with the routing and charging scheme within the alliance.

A. Overlay network structure

An ISP alliance is formed by adjacent ISPs by bilateral contract. An example of interdomain overlay network based on ISP alliance is shown in Fig. 2. In this figure, we only show the border routers of each ISP. ISP1, ISP2 and ISP3 form an alliance, while ISP4 and ISP5 do not belong to the alliance. The three ISPs in the alliance construct an overlay network by setting virtual links between border routers. If the traffic demand is between two ISPs in the alliance, then it could be routed by overlay network with overlay routing. Otherwise the traffic demand is routed by the origin BGP routing. The two routing schemes co-exist, and can be

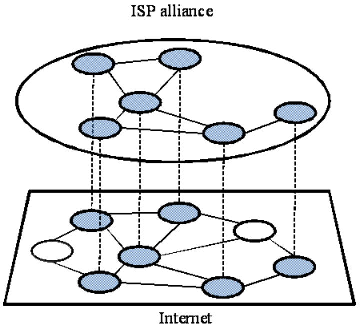


Figure 2. Overlay network based on an ISP alliance. The solid circles are ISPs within the alliance, and the empty circles are ISPs not in the alliance.

applied for different kinds of traffic. That is, our approach does not preclude the Internet as it is today neither does it exclude BGP policies. Instead of competing with BGP, our architecture can be seen as a complementary tool for ISPs.

Note that the ISP alliance can only be formed by adjacent ISPs. An ISP with no direct connection to any ISP in a specific alliance cannot be accepted. By this limitation, ISPs' loss of interest can be avoided. For example, in Fig. I, suppose a, b and d form an ISP alliance, if e is accepted, then d may suffer a loss of interest as illustrated in Section I. Within the ISP alliance, we avoid harming ISPs' economic loss caused by policy violation with effective and fair charging scheme in correspondence with the overlay routing structure. We make a brief introduction of the charging scheme, and go into detail in Section III.

Two charging schemes co-exist in the ISP alliance. One is the origin Internet charging scheme, in which two ISPs make a contract of either provider-customer (transit) or peer-peer (peering). With transit contract, the customer pays the provider for both up-streaming and down-streaming traffic, while with peering contract, the traffic transport is for free in both directions. The BGP charging scheme is applied for the traffic with source ISP or destination ISP outside the alliance. The other pricing scheme is applied for the overlay network. In the overlay network, as every ISP provides transit service, ISPs act as providers when they transit the traffic for their neighbors, and charge the traffic sender. When they send traffic to the other ISPs in the alliance, they are customers, and pay the ISPs along the routes they use.

B. Comparison of routing and charging in and outside the ISP alliance

In order to make the intra-alliance routing and charging scheme more clearly, we make a brief comparison with the Internet. Fig. 3 shows the summery of the comparison. First, as the Internet is very huge and ISPs are located all across the world, hierarchical routing structure is adopted. Geographic distributed stub ISPs can connect to each other only with the transit service of local ISPs and the backbone. However, our ISP alliance is supposed to construct with tens of ISPs near each other geographically, so that a simple but effective flat routing structure is adopted. Second, the business relationships in the Internet include transit and peering. As known to us, customer ISPs do not transit traffic for their providers, and peering ISPs do not provide transit service for each other. It turns out that some routes are illegal because they may violate the routing polices even if they have better performance. As comparison, in our ISP alliance, every ISP provides transit service for all its neighbors in order to take advantage of all potential routes. As compensation, the ISP who provides transit service will be paid by the traffic sender. Third, we design a charging scheme intra-alliance, which is different from the charging scheme of the Internet. In the intra-alliance charging scheme, the traffic sender s pays the other ISPs along the route to t.

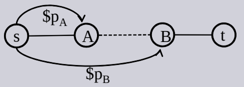
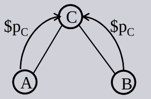
	ISP alliance	Internet
Routing structure	Flat routing structure is adopted because an ISP alliance is assumed to consist of tens of ISPs. Source routing is employed, and multiple paths are allowed.	Hierarchical routing structure is adopted because the Internet is very large. Single path routing is generally used.
Business relations	Every ISP in the alliance provides transit service to all its neighbors.	Transit and peering. Routes violating the policies are not permitted.
Charging	Traffic users pay every ISP along the routes that their traffic traverse.  	Customer ISP pays its providers. 

Figure 3. Routing structure and policies in and outside of ISP alliance

Note that with BGP routing and charging structure, a source ISP can only decide the next hop ISP, and has no control to the rest of the route. It is not necessary that the money the source ISP pays to the next hop ISP is positively correlated with the whole route performance. But with the routing and charging structure we propose, the correlation between source ISPs' routing decision, route performance and price is created. In the next section, we go into detail with the charging and pricing scheme.

III. Routing, charging and pricing within the ISP alliance

A. ISPs' routing decision and pricing strategies

The point to propose effective charging and pricing scheme is well capturing the properties of ISPs' routing decision. In the prominent work of [14], the authors introduce a model to capture the relationship between traffic demand and prices of routes. Suppose the price of a route r is p_r, which is the sum of prices determined by every ISP along r. Then the relationship is abstractly modeled by a demand function $d_r(p_r)$, which is strictly decreasing and differentiable. Moreover, if a function $g_r(p_r)$ is defined as $g_r(p_r) = -d_r(p_r)/d'_r(p_r)$, then $g_r(p_r)$ must be decreasing with respect to p_r. With this restriction on $g_r(p_r)$, the demand is inelastic when price is low, which means the demand is dominated by ISPs' need to communicate; but when prices increases, the demand becomes elastic, which means price becomes a more important factor in ISPs' decisions once price passes a certain threshold. This model succeeds in grabbing the properties of Internet service, however, it can only be used in single path routing system. Moreover, in this model, price is the only factor to affect ISPs' routing decision. In the overlay network in our work, multi-path routing is supposed in order to make full use of network resources. When making routing decisions, ISPs do not only consider the prices, but also the performance. In the rest of this section, we introduce our method to model the relationships among ISPs' routing decision, price and performance of routes.

Suppose there is only one route R_1 from source ISP s to destination t, then s has no choice but to send the traffic through R_1. Denote the price of R_1 as p_1, then the traffic volume is $d(p_1)$, where d is the aggregate traffic demand function. We assume d is decreasing, differentiable, and $-d(p)/d'(p)$ is decreasing with respect to p as in [14]. Now, if a better route R_2 is added with price $p_2 > p_1$, then $d(p_2)$ traffic would change to R_2, $d(p_1) - d(p_2)$ traffic will remain on R_1, and the total traffic volume remains $d(p_1)$. Now suppose there are m routes $R_1, ..., R_m$ between source ISP s and one destination t. The performance indicator of R_i is Per_i and the price is p_i. The performance indicator is logical, and larger Per_i indicates better performance. Without loss of generality, we assume $Per_1 < Per_2 < ... < Per_m$, and $p_1 < p_2 < ... < p_m$ correspondingly. The traffic demand from s to t will be $d(p_1)$, because p_1 is the lowest price of all routes. The traffic volume through R_i is $d(p_i) - d(p_{i+1})$. We can see that the traffic volume on R_i is dependent on the traffic volume on R_{i+1}. The only route on which the traffic volume does not depend on any other routes is R_m, and the traffic volume $f_m = d(p_m)$.

Denote the revenue obtained from R_m as Re_m, then $Re_m = p_m d(p_m)$. The ISPs on R_m can set price p_m to maximize Re_m independent to the other routes. The first order condition of Re_m with respect to p_m is $Re'_m(p_m) = d(p_m) + p_m d'(p_m)$. Let $Re'_m(p_m) = 0$, then we have $p_m = -d(p_m)/d'(p_m)$. As $-d(p_m)/d'(p_m)$ is decreasing, the unique solution exists for the optimization problem. Denote the optimal price of R_m is p^*_m, then revenue of R_{m-1} is $Re_{m-1} = p_{m-1}(d(p_{m-1}) - d(p^*_m))$. The first order condition of Re_{m-1} with respect to p_{m-1} is $Re'_{m-1}(p_{m-1}) = d(p_{m-1}) + p_{m-1}d'(p_{m-1}) - d(p^*_m)$. Let $Re'_{m-1}(p_{m-1}) = 0$, then we have $p_{m-1} = -d(p_{m-1})/d'(p_{m-1}) + d(p^*_m)$. As $-d(p_{m-1})/d'(p_{m-1})$ is decreasing with respect to p_{m-1}, the unique solution exists to the optimization problem. The optimal prices of the other routes can be obtained in the same way as above.

We can see that in this model, better route can decide optimal price with higher priority, and the optimal price of worse route always depends on the price of better route. The best route can decide optimal price independently to any other route. We believe that this model is more efficient than the models in which routing decision is not correlated with performance.

B. Analysis of route based pricing strategies

Although the charging scheme in Section III-A seems ideal, it is difficult to realize it in practice, because ISPs are selfish, and global cooperation cannot be expected. A very natural and easy way to realize the route based pricing scheme is non-cooperative pricing game, in which prices are determined for every individual route by the ISPs on those routes independently. We illustrate this scheme with a simple network example shown in Fig. 4. In the figure, s is an ISP who sends traffic to t. A, B, and C are intermediate ISPs. There are two routes for s to reach t. One is $ABCt$, which is denoted as R_1, and the other is ACt which is denoted as R_2. With the route based pricing, prices are determined based on routes. As the hierarchical structure does not exist, the commodity is specific route, the customer is the ISP who sends traffic through that route, and the provider being paid is every ISP on that routes. With non-cooperative pricing game, each AS could decide price for each route in a non-cooperative way to maximize the revenue obtained from that route. It seems natural and easy to realize because no cooperation among ASes is needed. But in fact, we find that this method is nether effective nor fair.

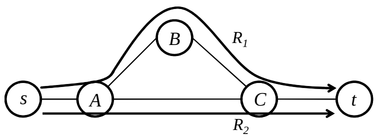


Figure 4. A simple network example

In Fig. 4, suppose route R_1 is better than R_2. Denote p_{A1} as A's price on R_1, p_{A2} as A's price on R_2, p_{B1} as B's price on R_1, p_{C1} as C's price on R_1, and p_{C2} as C's price on R_2. p_1 is the price of R_1, and $p_1 = p_{A1} + p_{B1} + p_{C1}$. p_2 is the price of R_2, and $p_2 = p_{A2} + p_{C2}$. f_1 is the traffic volume through $R1$, and f_2 is the traffic volume through R_2. The demand function is $d(p) = exp(-p^2)$, which is continuous, deceasing, and $-d(p)/d'(p)$ is also decreasing. According to the model in Section III-A, $f_1 = d(p_1)$, and $f_2 = d(p_2) - d(p_1)$. If the ISPs on R_1 and R_2 play a non-cooperative pricing game fairly, the prices can be obtained as follows:

For ISP A:

$$\begin{aligned} &\max Re_{A1} = p_{A1} d(p_{A1} + p_{B1} + p_{C1}) \\ &\max Re_{A2} = p_{A2}(d(p_{A2} + p_{C2}) - d(p_{A1} + p_{B1} + p_{C1})), \end{aligned} \tag{1}$$

where Re_{A1} is A's revenue obtained from R_1, and Re_{A2} is A's revenue obtained from R_2.

For ISP B:

$$\max Re_{B1} = p_{B1} d(p_{A1} + p_{B1} + p_{C1}), \tag{2}$$

where Re_{B1} is B's revenue obtained from R_1.

For ISP C:

$$\begin{aligned} &\max Re_{C1} = p_{C1} d(p_{A1} + p_{B1} + p_{C1}) \\ &\max Re_{C2} = p_{C2}(d(p_{A2} + p_{C2}) - d(p_{A1} + p_{B1} + p_{C1})), \end{aligned} \tag{3}$$

where Re_{C1} is C's revenue obtained from R_1, and Re_{C2} is C's revenue obtained from R_2. Then the only Nash equilibrium is achieved when $p_{A1} = p_{B1} = p_{C1} = 0.24$, and $p_{A2} = p_{C2} = 0.15$. The traffic through R_1 is $f_1 = 0.61$, the traffic through R_2 is $f_2 = 0.31$. A's revenue is 0.19, B's revenue is 0.15, and C's revenue is 0.19.

In the above example, each ISP plays the game by considering R_1 and R_2 separately, and the result is efficient and fair for ISPs on the same route. But if, for example, A, realizes that it is disjoint point of R_1 and R_2, it would change to an alternative behavior as follows:

$$\begin{aligned} \max R_A =& R_{A2} + R_{A1} \\ =& p_{A1} d(p_{A1} + p_{B1} + p_{C1}) + p_{A2}(d(p_{A2} + p_{C2}) \\ & - d(p_{A1} + p_{B1} + p_{C1})). \end{aligned} \tag{4}$$

When Nash equilibrium is achieved, $p_{A1} = 0.82$, $p_{A2} = 0.34$, $p_{B1} = 0.12$, $p_{C1} = 0.12$, and $p_{C2} = 0.34$. The traffic though R_1 is $f_1 = 0.11$, and the traffic through R_2 is $f_2 = 0.40$. The revenue of A is 0.23, the revenue of B is 0.01, and C's revenue is 0.06. From the above results, we can find that the traffic through the better route R_1 decreases dramatically, which reduces the efficiency of the traffic routing. Moreover, on both R_1 and R_2, A obtains more revenue than the other ISPs on the identical route, which is unfair to the other ISPs. As above, the non-cooperative pricing game based on route would not be acceptable. If the ISPs realize the undesirable properties of non-cooperative pricing game, they will look for some kind of cooperation. In the next section, we give our pricing scheme based on route bundle.

C. Pricing based on route bundle

In this paper, route bundle is defined as a set of routes having the same entrance ISP with each other. For example, in Fig. 5, R_1 and R_2 have the same entrance A, so that they are in the same route bundle RB_1. R_3 has different entrance from routes in RB_1, so that R_3 itself is route bundle RB_2. In fact, the inefficiency and unfairness in the non-cooperative route based pricing only happens at the disjoint point of multiple routes within identical route bundle. With pricing based on route bundle, the price is determined for route bundle, rather than individual route, so that the undesirable properties with route based pricing do not exist. In order to realize bundle based pricing scheme, cooperation with ISPs in the same bundle is required. Source ISP s would be noticed by the entrance A and D the price for RB_1 and RB_2 respectively, and decides how to route traffic. The traffic sent to RB_1 also has two options R_1 and R_2, and ISPs can choose a better one freely. The accounting can be done as follows. As source routing is employed, the route information can be found in the head of the packet. When a packet with entrance A and destination t enters A, A could write the price in the head of the packet, and forward it. Thus, every ISP on the route can keep record of the price and the packet amount. In the end of the contract cycle, the ISPs can share the revenue obtained from routes in identical route bundle. The share of each ISP can be calculated with bilateral negotiation. Although in the overlay network, the hierarchical structure does not exist, in fact, neighboring ISPs do not really have equal position. In practice, the two ISPs have either customer-provider contract or peering contract, so that ISPs may not be satisfied to share the revenue equally. One possible negotiation is, neighboring ISPs bargain with each other to decide the relative sharing. After every pair of ISPs finish the bargaining, the share of every ISP can be calculated.

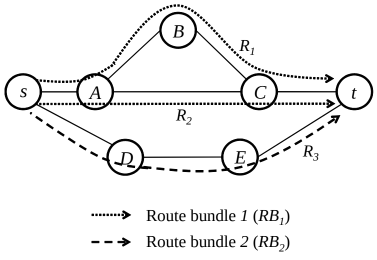


Figure 5. A network example with route bundles

D. Pricing algorithm

Section III-C showed that the price of a specific route bundle is decided by the entrance ISP of the bundle. In fact, what the entrance ISP faces is simple optimization problem with just a single variable. Although the objective function may be neither convex nor concave, we have shown that it has a unique optimal point in Section III-A. Therefore, it can be solved by a one-dimensional search method. The entrance ISP could set a starting price from the empirical value p_0, and then update it periodically. Supposing prices are updated in steps of u, the ISP can update the price as follows:

1) Set the price p to the empirical value p_0
2) Loop step 3 to step 5 periodically until the optimal price being found
3) Increase p by one unit. If the revenue decreases, go to step 5. Else, go to step 4
4) Keep increasing p, until revenue begins decreasing
5) Keep decreasing p, until revenue begins decreasing

This method is valid for the following reason. Suppose a set of route bundles $RB_1, ..., RB_n$ are competing for traffic with each other. Without loss of generality, we assume the route bundles are in ascending order with respect to performance. The revenue of a specific route bundle RB_i can be represented by $Re_i = p_i(d(p_i) - d(p^*_{i+1}))$. The first order condition is

$$Re'_i(p_i) = (p_i + \frac{d(p_i)}{d'(p_i)} - \frac{d(p^*_{i+1})}{d'(p_i)})d'(p_i), \tag{5}$$

where p^*_{i+1} is the optimal price of RB_{i+1}. As $-\frac{d(p_i)-d(p_{i+1})^*}{d'(p_i)}$ is decreasing, a unique solution to maximize Re_i exists, which is denoted by p_i^*. If $p_i \leq p_i^*$, then $Re_i'(p_i) \geq 0$, which means that Re_i increases with respect to p_i in $(0, p_i^*]$. If $p_i > p_i^*$, then $Re_i'(p_i) < 0$, which implies that Re_i decreases with respect to p_i. The validity of the pricing method can then be proved straightforwardly. We also find that, with this method, entrance ISPs can determine the optimal prices without knowing the exact formula for the demand function d.

Note that, if multiple route bundles have the same performance, we need to make a tie–breaking rule. In this work, the traffic source ISP should choose any one of the route bundles to transmit traffic.

E. Numerical experiments

In this section, we describe numerical experiments for showing the validity and convergence of our pricing method. We conduct experiments based on a network with as shown in Fig. 6.

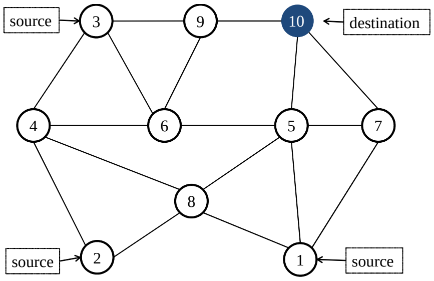
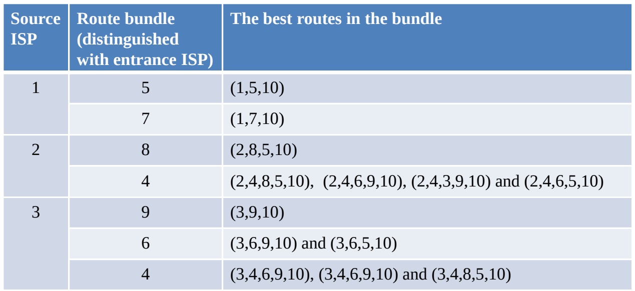


Figure 6. Network for experiment. Circles represent ISPs



Source ISP	Route bundle (distinguished with entrance ISP)	The best routes in the bundle
1	5	(1,5,10)
	7	(1,7,10)
2	8	(2,8,5,10)
	4	(2,4,8,5,10), (2,4,6,9,10), (2,4,3,9,10) and (2,4,6,5,10)
3	9	(3,9,10)
	6	(3,6,9,10) and (3,6,5,10)
	4	(3,4,6,9,10), (3,4,6,9,10) and (3,4,8,5,10)

Figure 7. Route bundles and routes they contain

In the figure, ISP 1, 2 and 3 are source ISPs transmitting traffic to ISP 10. We assume links have the same propagation delay, and queuing delay is not considered. Therefore, the hop count can represent the latency, and latency is taken as the performance indicator in the experiments. The route bundles and routes they contain in Fig. 6 can be summarized as Fig. 7. At the beginning of the experiments, entrance ISPs set prices based on values from previous experience, and then adjust the prices periodically and independently. To make the experiments more clear, we assume that competing route bundles adjust prices in turn. Prices are assumed to be adjusted in steps of 1.0. Changes in price and revenue with respect to time are shown in Figs. 8(a), 9(a), 8(b), 9(b), 8(c) and 9(c).

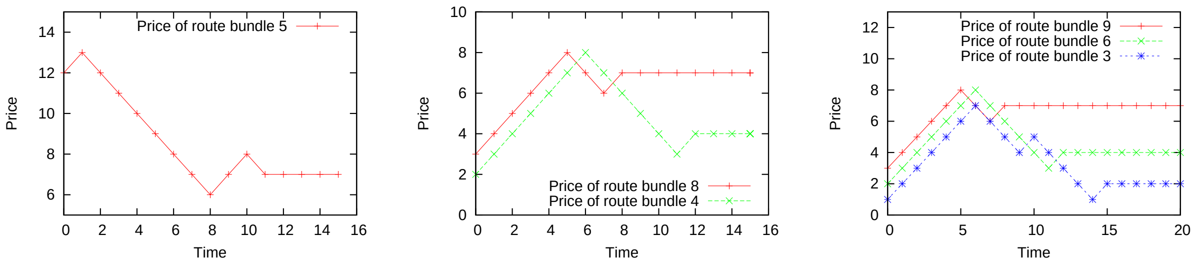
Note that between ISP 1 and 10, there are two route bundles with entrance ISP 5 and 7, which have the same latency. According to our tie–breaking rule, 1 can choose any route bundle to transmit traffic. We assume route bundle with entrance ISP 5 (route bundle 5) is chosen. The initial price is set as 12.0 which is higher than the optimal price. After some steps of adjusting, the optimal price 7.0 is found (Fig. 8(a)), and the revenue achieves the highest (Fig. 9(a)). Between ISP 2 and 10, there are also two route bundles 8 and 4. The route in route bundle 8 has less hop count than the routes in route bundle 4, which indicates route bundle 8 is better than 4. At the beginning, route bundle 8 initializes p_0 as 2.0 and route bundle 4 initializes p_0 as 1.0. Both of the prices are lower than the optimal prices. The price adjusting process is shown in Fig. 8(b). In Figs. 9(b), we can find that route bundle 4 receives 0 revenue in a period of time. This is because during that period, route bundle 4 sets higher price than route bundle 8, so that ISP 2 transmits all the traffic through route bundle 8. From Figs. 8(b) and 8(c), we can also find that the convergence of route bundles depends on the converge of better route bundles. The price adjusting of a route bundle can not converge before all the better route bundles finish adjusting prices.

IV. Conclusion

In this paper, we propose an interdomain overlay network in which nodes are operated by ISPs within an ISP alliance. The traffic between ISPs within the alliance could be routed by overlay routing to overcome the functionality limitations of BGP. According to the definition of the ISP alliance and the economic structure within the alliance, the BGP policy violation problem can also be addressed.

As ISPs are individual economic entities, interdomain routing issues cannot be separated from economic factor. We study ISPs' routing decision facing multiple routes, and model the relationship between ISPs' routing decision and route properties – performance and price. Based on this model, we obtain the optimal price for each route to maximize the revenue.

Although the optimal price exists, it is difficult to realize it in practice. We show that a non-cooperative pricing game by selfish ISPs would lead to ineffective and unfair result. We believe that if ISPs realize the above fact, they would seek cooperation. We then propose a pricing scheme based on route bundle – a bundle of routes having the same entrance ISP with each other – and show that it is better than the non-cooperative pricing game. At last, we give a simple pricing algorithm with which ISPs can find the optimal prices without precise knowledge of traffic source ISPs. With

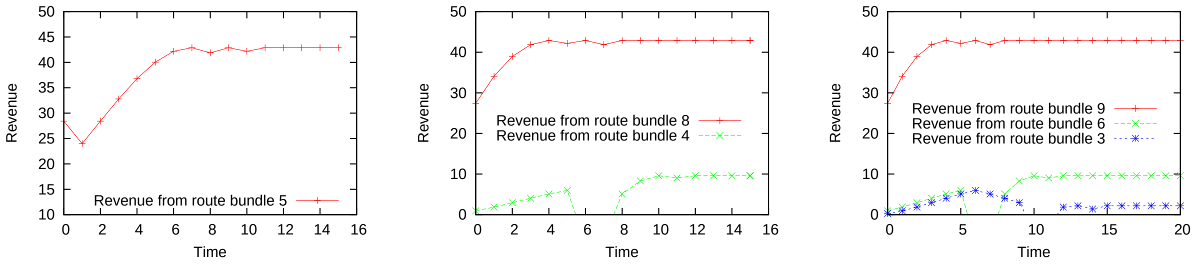


(a) Price of route bundle 5

(b) Price of route bundles 8 and 4

(c) Price of route bundles 9, 6 and 3

Figure 8. Price of route bundles



(a) Revenue of route bundle 5

(b) Revenue of route bundles 8 and 4

(c) Route bundle 3

Figure 9. Revenue of route bundles 9, 6 and 3

mathematical analysis and numerical experiments, we show the correctness and convergence of the pricing algorithm.

References

[1] S. Savage, T. Anderson, A. Aggarwal, D. Becker, N. Cardwell, A. Collins, E. Hoffman, J. Snell, A. Vahdat, G. Voelker, and J. Zahorjan, "Detour: A Case for Informed Internet Routing and Transport," *IEEE Micro*, vol. 19, pp. 50–59, January 1999.

[2] D. Anderson, H. Balakrishnan, M. F. Kaashoek, and R. Morris, "Resillient Overlay Networks," in *Proceeding of the ACM SIGCOMM*, November 2001.

[3] Z. Duan, Z. L. Zhang, and Y. T. Hou, "Service Overlay Networks: SLAs, QoS, and Bandwidth Provisioning," *IEEE/ACM Transactions on Networking*, vol. 11, pp. 870–883, December 2003.

[4] H. T. Tran and T. Ziegler, "A Design Framework towards the Profitable Operation of Service Overlay Networks," *Computer Networks*, vol. 51, pp. 94–113, January 2007.

[5] A. Capone, J. Elias, and F. Martignon, "Routing and Resource Optimization in Service Overlay Networks," *Computer Networks*, vol. 53, pp. 180–190, October 2009.

[6] Y. Hei, A. Nakao, T. Ogishi, T. Hasegawa, and S. Yamamoto, "AS Alliance for Resilient Communication over the Internet," *IEICE Transactions on Communication*, vol. E93-B, no. 10, pp. 2706–2714, October 2010.

[7] J. H. Wang, D. M. C. Chiu, and J. C. S. Lui, "Modeling the Peering and Routing Tussle between ISPs and P2P Applications," in *Proceedings of IEEE IWQoS 2006*, June 2006, pp. 51–59.

[8] S. Seetharaman and M. Ammar, "Characterizing and Mitigating Inter-Domain Policy Violations in Overlay Routes," in *Proceedings of IEEE ICEP 2006*, November 2006, pp. 259–268.

[9] G. Hasegawa, Y. Hiraoka, and M. Murata, "Evaluation of Free-Riding Traffic Problem in Overlay Routing and Its Mitigation Method," *IEICE Transactions on Communication*, vol. E92-B, no. 12, pp. 3774–3783, December 2009.

[10] J. H. Wang, D. M. Chiu, and J. C. S. Liu, "A Game-Theoretic Analysis of the Implications of Overlay Network Traffic on ISP Peering," *Computer Networks*, vol. 52, pp. 2961–2974, October 2008.

[11] X. Shao, G. Hasegawa, Y. Taniguchi, and H. Nakano, "The Implication of Overlay Routing on ISPs' Connecting Strategies," in *Proceedings of ITC 2011*, no. September, 2011, pp. 286–293.

[12] R. Winter, "Modeling the Internet Routing Topology with a Known Degree of Accuracy," in *Proceedings of ACM/IEEE/SCS PADS 2009*, June 2009.

[13] G. Fudenberg and J. Tirole, *Game theory*. MIT Press, 2000.

[14] L. He and J. Walrand, "Pricing and Revenue Sharing Strategies for Internet Service Providers," *IEEE Journal on Selected Areas in Communications*, vol. 24, pp. 942–951, May 2006.

Evaluation of Reliable Multicast Implementations with Proposed Adaptation for Delivery of Big Data in Wide Area Networks

Aleksandr Bakharev
Siberian State University of Telecommunication and Information Sciences
Novosibirsk, Russia
a.bakharev@emw.hs-anhalt.de

Eduard Siemens
Anhalt University of Applied Sciences
Koethen, Germany
e.siemens@emw.hs-anhalt.de

Abstract*—This paper describes the state of contemporary open source reliable multicast solutions and reveals deficiencies regarding their use for massive data transport in Content Delivery Networks (CDN). A performance evaluation of the three most popular open-source implementations - *UDP-based File Transport Protocol, NACK-oriented Reliable Multicast* and *Pragmatic General Multicast* in multi-gigabit IP-based networks was performed in the 10Gigabit-WAN laboratory of the Communications Group of Anhalt University of Applied Sciences. This evaluation was completed under the real-world scenario of heavy-weight content distribution in Wide Area Networks. The performance evaluation presented in this paper reveals bottlenecks and deficiencies in current approaches and the paper proposes ideas for improvements and further development of the reliable multicast data delivery family. The defined test scenario was limited to three recipients for the following two reasons: Big data distribution does not imply a large number of recipients, and the goal of this work was to determine upper performance bounds even in a quite simple scenario as a starting point for further investigations. This investigation identified three main challenges: congestion control, losses recovery management and send/receive buffer management. The investigations presented have been performed in the course of a research project in which reliable point-to-multipoint IP-based data transport solution will be proposed. The goal is to achieve data rates of up to 1 Gbit/s per stream with up to ten simultaneous streams from one content server, even in presence of high RTT delays and packet losses in the network.

Keywords-CDN; reliable multicast; network performance; cloud computing; big data

I. INTRODUCTION

According to a report of the IEEE Ethernet Working Group in [1], in the time period from 2013 to 2018, world traffic will grow by a factor of ten in comparison to the 2010 value. Such a rapid growth in network traffic means improving existing networking technologies and seeking new approaches to data distribution in the core IP network. Challenges such as effective utilization of available bandwidth become crucial. One of the technologies that addresses this issue is multicast networking [2].

In general, the idea of reliable multicast networking aims at achieving maximum utilization of bandwidth whilst avoiding unnecessary duplication of data. In classic unicast networking, each IP packet is sent by a host to exactly one recipient. In the case of multicast networking, data sources deal with groups of recipients and always send only one packet to the entire group. The packets are then duplicated by intermediate network devices such as IP routers and switches. This packet duplication is only performed when the network device knows that it is no longer possible to use one packet for the entire recipient group. Consequently, on all common parts of a network path between a sender and a receiver, the number of packet duplications is minimized.

First standardized in 1986, IP multicast protocols were originally an unreliable data transport solution [2]. One of the first worldwide multicast implementations was *Mbone* [3] with its multicast protocol family such as IGMP or PIM, released in the early 1990s. This protocol family was fairly well adapted to the needs of multimedia applications such as conferencing and live messaging, and, for a long time, multimedia communication was the only application of multicast data transmission. However current use of multicast communication has significantly widened. With the rapid growth of the amount of Internet traffic around the globe, simultaneous point-to-multipoint data delivery is becoming crucial in large Content Delivery Networks (CDNs) and cloud infrastructures. Therefore, distribution of large amounts of content is an ongoing task for most large CDNs. Replications of databases, HD-video delivery, online gaming etc. require high network performance and transmission efficiency. For example, Felix Baumgartner’s recent ultrasonic jump was watched in nearly real-time by more than 8 million people; a world record for the number of simultaneous video streams.

Such simultaneous data delivery is one of the big challenges in reliable multicast networking. Raising efficiency of content distribution within CDNs is one of the purposes of reliable multicast communication. For example, the Akamai CDN uses IP multicast technology to provide subscription-based media streaming for consumers. The Amazon cloud constantly receives customers’ requests to enable multicast on its EC2 clouds and is currently planning to implement it. The emergence of enormous online gaming services such as the PlayStation Network with over 90 million [4] connected unique consoles (members) must also maintain reliable multicast sessions. These cases clearly demonstrate the need for modern networking in terms of transmission session management for multiple recipients.

This paper is organized as follows. Section I gave an introduction to the research field. Section II gives a brief description of the setup for testing the performance measurements of the selected multicast solutions. Section III describes the approaches of the evaluated protocols and presents data related to the results. Section IV describes the revealed deficiencies and Section V proposes an improvements' plan for the solutions considered. In Section VI, we conclude and propose an agenda for further investigations and implementation of higher-speed reliable multicast data transport.

II. Testbed Definition

The chosen testbed is based on facilities of the 10 Gbit/s test lab installed at Anhalt University of Applied Sciences in Koethen, Germany. All test cases were performed on 64-bit OpenSuSE Linux PC systems. The test network comprises one sending server and three recipients that belong to one multicast group. The work was performed using only three recipients because the ultimate goal of tests was to evaluate upper maximal data rate limit for current reliable multicast approaches. Because this was the main goal, there was no sense in deploying a larger and more complex topology with multiple recipients at this stage. The test network is very well-scalable due to the use of the hardware-based 10G WAN impairment emulator Netropy 10G [5], with which a total data throughput of up to 21 Gbit/s can be achieved. With this device, WAN-sized networks can be emulated and network parameters of the emulated channels - delay, jitter, packet loss – adjusted with an accuracy of about 20 ns. Our test scenarios assume a transmission of one 10 Gbyte file over the emulated WAN to the tree recipients under different network conditions, whereby Round Trip Time (RTT) and packet loss rates are increased up to 50 ms and 0.3% respectively. The topology of the test setup is shown in Figure 1. In general, this topology assumes inhomogeneous delays among emulated links. However, for current tests, we emulated a simplified case with similar RTT values and packet loss rates on each emulated link.

III. Protocols Overview

The following three solutions were considered: *NACK-oriented Reliable Multicast (NORM)*, *UDP-based File Transport Protocol (UFTP)*, and *Pragmatic General Multicast* (PGM). These solutions were chosen due to their high popularity and the availability of a ready to use transport application built upon the respective reliable multicast transport protocol. PGM does not contribute a ready-to-use data transport application, though the protocol stack is used in different production environments such as the one at TIBCO [6], which uses PGM for discovering new members of computing cluster.

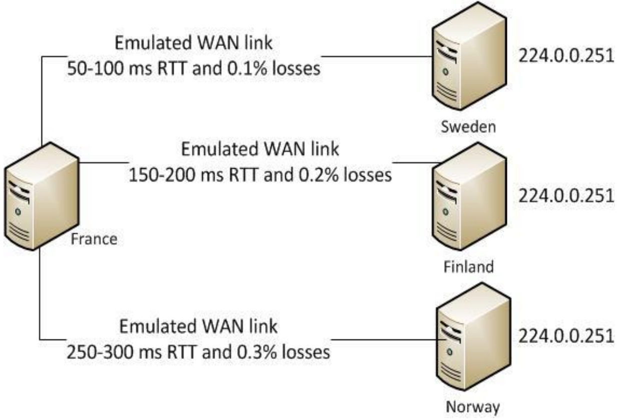


Figure 1. HSA test installation

A. NACK-Oriented Reliable Multicast (NORM)

The *NORM* protocol was defined within RFC 5740 [7] in 2009. The source code of a reference implementation of NORM is maintained by the Naval Research Laboratory [8]. As well as being a transport protocol, the protocol provides a ready-to-use application that can be compiled from available C source code on Linux. Based on Berkeley UDP sockets, the NORM application offers features such as TCP friendly congestion control, which provides fair sharing of available bandwidth between multiple data streams. NORM uses selective negative acknowledgements (NACKs) to provide reliability. *NORM* can also be used in conjunction with Forward Error Correction (FEC), which is currently only an on-demand feature.

As shown in Figure 2, the data rate decreases fast even with very few impairments to the link. This rapid decrease means that the NORM maximal data rate is very sensitive to retransmissions caused by network losses. According to protocol specification, users have to enable FEC to minimize the amount of active NACKs in the network. It also means that the NORM algorithm does not focus on the improvement of NACK management efficiency. Instead, it focuses on improving reliability through the FEC mechanism. However, the problem is that FEC is a difficult approach for big data transmission because there are huge increases in the FEC overhead, even on links in good condition. FEC redundancy issues will be discussed in more detail in Section IV.A. NACK-based reliability, as implemented by default in NORM, enables the receiver to send NACKs at any time – in fact, as soon as a loss has been detected. For bulk data transmission, this causes an enormous batch of NACKs

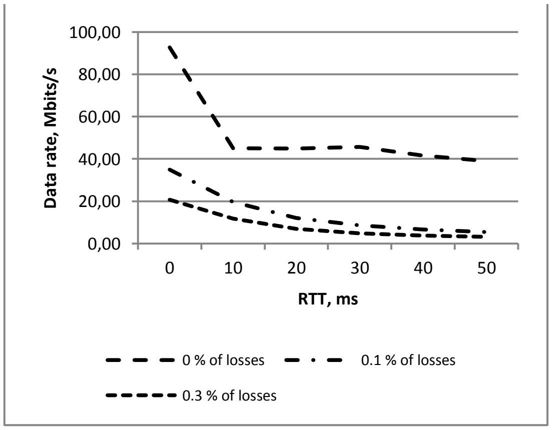


Figure 2. NORM performance dependency on RTT and losses

And, therefore, leads to a decrease in data rate as well. Consequently, protocol parameterization, as currently used in NORM, requires significant tuning to raise transport data rate. NORM's RFC would allow a suitable configuration of the transmission settings e.g. by using minimal inter-NACKS intervals or by consolidating NACKs from multiple packets into one packet.

B. UDP-Based File Transfer Protocol (UFTP)

UFTP is also a reliable multicast protocol with a corresponding end-user application and can be considered as a successor to the *Starburst Multicast FTP* (*MFTP)* [9] proposed in 2004. It provides reliable multicast file transfer through UDP transport. The protocol is currently in use in the production of the Wall Street Journal for transporting WSJ pages to their remote printing plants via satellite [9].

UFTP uses a specific scheme of data transmission organization. First of all, the protocol decides how to divide input data into data sets. Input data are split into blocks, whereby one block is always sent within one UDP packet. Since these blocks are, in turn, logically grouped into sections, the sender just sends a section to a multicast group. As soon as the transmission of a section is finished, the sender requests the current status of received data from each multicast receiver and receives a batch of packets containing a list of the packets missing at the site of each recipient. On reception of all NACKs, missed blocks are retransmitted in the unicast way to the requesting recipient. The sender begins to transmit a new section only after all the recipients in the multicast group have confirmed the reception of all blocks of the previous section. This type of data transmission organization results in protocol performance being significantly increased compared to *NORM*. Figure 3 shows the data rate evaluation results in the same testbed as for *NORM*. The results reveal that UFTP has a high loss tolerance and that recovery of lost packets does not reduce the overall data rate as significantly as does *NORM*. However, in both cases, a significant data rate reduction with an increased RTT can be observed.

The obtained results reveal a significantly more efficient sections-based data transmission method than the classic one in the NORM (NACK packet if reception failure revealed). However, when packet loss occurs, the long retransmission periods mean that section-based acknowledgment shows significant dependency on the RTT. Also, transmitting data in this way represents NACKs consolidation, since the UFTP receiver sends NACKs that contain information about multiple missed packets.

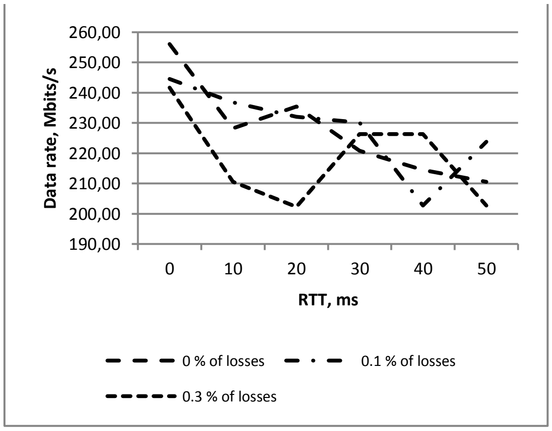


Figure 3. UFTP performance dependency on RTT and losses

C. Pragmatic Reliable Multicast (PGM)

The Pragmatic Reliable Multicast (*PGM)* protocol is described in RFC 3208 [10] and is officially supported by IP routers of Cisco Systems (beginning from Cisco IOS Software Releases 12.0 T). This protocol has been developed with the ultimate goal of providing reliable data transmission service for as many recipients as possible. This design automatically means dispensing with ACKs in favor of NACKs, since using ACKs implosion [11] significantly reduces the scalability of the end application and the entire protocol. The retransmission window has to be defined by the user within the configuration of the reliable multicast session. *PGM* assumes allocated disk space (in the form of a buffer) as a window size with a default value of 10 MB. As an option, the retransmission window size can be configured for dynamic adjustment based on NACK-silence times. *PGM* operates over classic IP multicast stack and does not deal with group management, delegating this tasks directly to IGMP, By comparison, the previously described protocols deal with group instances themselves. So, PGM works as a superstructure (in form of raw socket), over UDP and IP multicast stack.

An open source implementation of *PGM* is *openPGM*, which is a framework for the development of new reliable multicast applications. Since there is no ready-to-use *openPGM* application, we had to develop our own test application for sending and receiving files via *PGM*. Through contact with a *PGM* development and maintenance team, we were told that *openPGM* is not designed to be a file transfer protocol. Suggestions for adapting *openPGM* for big data transmission depended on using FEC and Lower-Density-Parity-Codes (LDPC). However it was important for us to get some exact values on possible data rate with the *openPGM* solution. Simple tests revealed an end-to-end data throughput of 27.1 Mbits/s without the packet loss and emulated packet delays in the 1-to-3 multicast scenario that had been found in the two previous tests. Even on RTTs of greater than 10 ms, the data transmission almost stalls. Initially, the idea of the protocol was to multicast very short data blocks such as market quotes and trades. Because we had very specific demands on big data transmission, we decided not to perform further exhaustive tests with openPGM. However, for our research agenda, the protocol provides interesting algorithms and possibilities for session management and dealing with NACKs. This information could be valuable for future work, at least in terms of quick NACKs processing.

IV. Problems and Solutions

Summarizing, it can be stated that on networks with no packet loss and with low round trip delays of up to 20 ms, UFTP provides reliable data delivery in a multicast fashion with up to 250 Mbit/s. However, the data rate is significantly impacted by increasing network impairments such as delay and, especially, packet loss. For a further increase of data transfer rates using multicast, at least three significant problems must be addressed:

A. Congestion control schemes used for the rate control
B. Improvement of packet loss recovery algorithms
C. Send and receive buffer management (Section V)

A. *Congestion control*

Regarding congestion control, the main consideration is whether the receiver or the sender should be responsible for congestion control. For instance, the *Source Adaptive Multi-layered Multicast* (SOAMM) [12] algorithm proposes adjusting video encoding settings at source as a reaction to continuous congestion control feedback. *Receiver-driven Layered Control* (RLC) [12] represents receiver based congestion control. It functions completely source-independent and a participant joins the multicast group accordingly to its own available resources. Such an approach assumes multi-layered multicast with different subscription levels. Available subscriptions - which in IP multicast refers to a multicast group - are to be advertised by the sender, which uses special Synchronization Points (SPs) for this purpose.

Another important challenge here is how to decide between window-based and rate based congestion control schemas.

Due to scalability issues, the classic idea of window-based schemes, such as ones used in TCP, do not fit the requirements of modern reliable multicast communications. With increasing multicast group size, the probability of an acknowledgements' implosion problem also rises. Such an implosion can itself significantly slow down a multicast session. In this scenario, bottlenecks will be on the sender site, and this effect is known as "crying baby problem" [13]. Due to the mentioned ACKs implosion in window-based schemes, most of the contemporary reliable multicast implementations deal with rate-based schemes. However, there is a big difference in comparison to the unicast case. The system of metrics used in a reliable unicast transmission with rate based congestion control is fairly easy - upper data rate limit and appropriate adoption of the data rate on ARQ. However, in multicast transmission, we deal with fairly difficult network paths with couples of branches in which we have to evaluate the entire pattern of multicast tree efficiency. For this purpose, at least two special prediction metrics are proposed [14]:

1. Analysis of multicast tree shape with computation of graph edges weights.
2. Group size as a determining factor [15].

B. *Error recovery*

Three basic schemes are widely used for error correction today:

1. ARQ-based ones with acknowledgements of received data packets, retransmission schemes and timers for retransmissions.
2. The well-known FEC schemes with redundancy in each data packet
3. Error Resilient Source Coding (ERSC), which, in fact, just conceal losses at the receiver site.

Each scheme is used in special cases. Thus, ARQ-based schemes are mainly targeted at delaying insensitive applications, while FEC is mainly used in delay-sensitive applications. It is worth noticing that FEC could be implemented in two different ways: redundant symbols are either transmitted in a separate data packet or within regular data packets. However, for redundancy reasons, FEC is often disabled or even not implemented in contemporary multicast protocols, since redundant packets often make transmission very bulky. Since packet losses in packet-switched networks come in bursts and affect hundreds or thousands of packets, the FEC algorithm will generate so much redundant data that it will aggravate network conditions. As shown in [16], even at a link with a 0.1 % loss rate, the number of required redundant symbols grows exponentially. This result was found for HDTV streaming with a data rate of 1.5 Gbits/s. This work was done as a laboratory case, while real-environment conditions assume loss rates of up to 5% for intercontinental links [17]. The most popular codes for FEC are the Red-Solomon Code and the Tornado Code [16]. ERSC, in turn, is well suited to live video streaming but does not provide full reliability for each sent bit and is, therefore, not suitable for static data transmission.

V. IMPROVEMENTS PLAN

The problems and findings described in Section IV point to ideas for optimizing existing solutions and developing entirely new solutions for big data transmission.

In the future, we initially propose dealing with effective data transmission; for instance, by separating entire data array by packets with a further grouping of packets to sections, similar to UFTP implementation with a fairly high upper data rate limit. This mechanism would work fairly well with a buffering of NACKs. The problem of NACKs buffering was initially raised in RFC 3269 [17]. NACKs buffering was aimed at minimizing the amount of NACKs in the network without increasing transmission latency. This challenge is like "walking on the razor's edge", but we are convinced that exhaustive tests and precise adjustments will help us find the most effective NACKs' buffer size.

In the field of congestion control, we are working on multicast-adapted rate-based congestion control with the prediction of network behavior by defined metrics (tree shape and group size).

The error recovery scheme shall be kept NACK-based in order to avoid the ACK implosion problem. We have also decided to dispense with FEC due to the high FEC overhead when losses come in bursts.

As shown in [18], losses in L2 and L3 caused by buffers' overflow prevail over BER-caused losses. For efficient buffer management implementation, we propose designing a novel send and receive buffer implementation adapted to reliable multicast constraints. We are planning to reach data rates of 1Gbit/s per stream in presence of up to 10 destinations within a session. At such high data rates, the ability to read and write data in the most effective fashion becomes crucial. Generally, the idea is to assume dynamic memory allocation for each stream with further re-allocation of available memory among other streams.

VI. Conclusion

A general overview of contemporary reliable multicast implementations is given in the paper. Our research reveals that, even with quite a small number of recipients, the upper limit of throughput on reliable data transport is currently not more than 250 Mbit/s. Performance results also revealed that packet loss causes the most significant decrease of transmission data rate. Thus, future work will focus more on improving error recovery schemes. Analysis of considered protocols revealed possible algorithm improvement for raising data rate performance. Our work reveals a few trends that could potentially be implemented in a reliable multicast scheme with the primary goal of achieving a data rate of 1 Gbit/s per reliable stream with at least up to 10 destinations.

References

[1] IEEE 802.3 Ethernet Working Group, "IEEE Industry Connections Ethernet Bandwidth Assessment". San Diego : IEEE 802.3 Plenary meeting, 2012.

[2] S. Deering, "Host Extensions for IP Multicasting", RFC-1112, 1989.

[3] K. Savetz, N. Randall, and Y. Lepage, "MBONE: Multicasting Tomorrow's Internet", John Wiley & Sons Inc (Computers), ISBN 978-1568847238, 1996.

[4] A. Osborn, "Number of Registered PlayStation Network Accounts Reaches 90 Million". March, 2012, http://www.playstationlifestyle.net/2012/03/07/number-of-registered-playstation-network-accounts-reaches-90-million/. [retrieved: January 2013]

[5] Apposite Technologies oficial web site, http://www.apposite-tech.com/index.html. [retrieved: January 2013]

[6] TIBCO official web site, http://www.tibco.com/. [retrieved: January 2013]

[7] B. Adamson and C. Bormann, "NACK-Oriented Reliable Multicast (NORM) Transport Protocol", RFC-5740, 2009.

[8] Naval Research Laboratory, http://www.nrl.navy.mil/. [retrieved: January 2013]

[9] K. Miller and K. Robertson, "StarBurst Multicast File Transfer Protocol (MFTP) specification", Internet-draft, 1997.

[10] T. Speakman and J. Crowcroft, "PGM Reliable Transport Protocol Specification" RFC-3208, 2001.

[11] D. S. Vijayakumar and S. V. Ram, "A network processor implementation for solving the ACK implosion problem", ACM-SE 44. Proceedings of the 44th annual Southeast regional conference, New York, March, 2006, pp. 732-733.

[12] D. Constantinescu, D. Erman, and D. Ilie, "Congestion and Error Control in Overlay Networks", Blekinge Institute of Technology, Sweden, 2007.

[13] H. Holbrook, S. Singhal, and D. R. Cheriton, "Log-Based Receiver-Reliable Multicast for Distributed Interactive Simulation". ACM SIG-COMM-95, Cambridge, August, 1995, pp. 328-341.

[14] R. C. Chalmers and K. C. Almeroth, "Developing a Multicast Metric. Global Telecommunications Conference", IEEE GLOBECOM '00, San Francisco, December, 2000, pp. 382-386.

[15] J. Chuang and M. Sirbu, "Pricing multicast communication: A cost based approach", INET’98, Geneva, July, 1998, pp. 281-297.

[16] S. Senda, H. Masuyama, S. Kasahara, and Y. Takahashi, "FEC Performance in Large File Transfer over Bursty Channels", Proceedings of the 4th International Working Conference on Performance Modelling and. Evaluation of Heterogeneous Networks (HET-NETs'06), D. Kouvatsos (ed), West Yorkshire, U.K., September 10-13, 2006, P07/1-10.

[17] Y. Angela Wang, C. Huang, J. Li, and K. W. Ross, "Queen: Estimating Packet Loss Rate between Arbitrary Internet Hosts". Proceedings of the 10th International Conference on Passive and Active Network Measurement, April, 2009, pp. 57-66.

[18] S. Dixit and T. Wu, "Content Networking in the Mobile Internet". s.l. : John Wiley and Sons, Inc., ISBN 0-471-46618-2, 2004.

[19] R. Kermode and L. Vicisano, "Author Guidelines for Reliable Multicast Transport (RMT) Building Blocks and Protocol Instantiation documents", RFC-3269, 2002.

[20] E. Siemens, R. Einhorn, and A. Aust, "Multi-Gigabit Challenges: Similarities between Scientific Environments and Media Production". ACIT - Information and Communication Technology, June, 2010.

Optimizing Multicast Content Delivery over Novel Mobile Networks

Tien-Thinh Nguyen
Department of Mobile Communications
EURECOM
Sophia Antipolis, France
E-mail: Tien-Thinh.Nguyen@eurecom.fr

Christian Bonnet
Department of Mobile Communications
EURECOM
Sophia Antipolis, France
E-mail: Christian.Bonnet@eurecom.fr

Abstract—**The rapid growth in mobile traffic leads to the current evolution trend of mobile networks towards a flat architecture. However, the centralized mobility management protocols (e.g. MIPv6, PMIPv6) are not optimized for the flat architecture due to their limitations e.g. complex tunnel management, scalability issue, etc. Hence, a novel mobility management has been proposed for the flat architecture, called distributed mobility management (DMM). IP multicast, an effective mechanism for traffic delivery, can be enabled in DMM by deploying MLD Proxy function at mobile access routers (MARs) with the upstream interface being configured to the multicast infrastructure (before mobility) or to the tunnel towards the mobile node's mobility anchor (after mobility) (namely tunnel-based approach). In case of mobility, the utilization of the tunnel may result in the tunnel convergence problem when the multiple instances of the same multicast traffic converges to a MAR due to the multiple tunnels established with several mobility anchors (leading to the redundant traffic at the MARs). Compared to PMIPv6, the tunnel convergence problem may become much more severe, especially in highly mobile regime. In this paper, we propose some mechanisms to greatly reduce the amount of redundant traffic at the MARs with a minor increase of service disruption time compared to the tunnel-based approach.**

Keywords-Future Internet; IP multicast; multicast mobility; tunnel convergence problem; handover delay; Distributed Mobility Management.

I. INTRODUCTION

The explosion of wireless devices like smartphones, tablets makes a dramatic increase in mobile traffic [1]. How to manage a large number of mobile terminals as well as a huge mobile traffic increase becomes a major challenge to network operators. Also, the evolution of wireless application and services lead to new requirements such as seamless mobility across the heterogeneous access technologies (session continuity, application transparency), consistent quality of experience and stringent delay constrains.

With the evolution of wireless technology, heterogeneous networks provide the possibility for great capacity increase at a low cost. However, only increasing capacity is unable to address the network challenge as well as to meet the new service requirements. In this context, several strategies have been proposed for efficiently delivering the traffic such as traffic offloading e.g. Local IP Access (LIPA), Selected IP Traffic Offload (SIPTO) [2] and Content Delivery Networks (CDNs) mechanisms [3]. They reflect the current evolution trend of mobile networks - shift to a flat IP architecture to lower costs, reduce system latency, and decouple radio access and core network evolution [4].

Still, the current IP mobility management protocols like Mobile IPv6 (MIPv6) [5], Proxy Mobile IPv6 (PMIPv6) [6] do not work perfectly with such a flat architecture due to their limitations e.g. complex tunnel management, poor performance (like non-optimal route, tunneling overhead) and scalability issue [4][7]). Thus, a novel approach, called distributed mobility management (DMM) [8][9], has been proposed to cope with the flat architecture and overcome the limitations of centralized mobility management. The idea is that the mobility anchors are placed closer to the user; the control and data plane are distributed among the network entities. In addition, mobility service is provided dynamically to the terminal/service that really needs to simplify the network and lower the cost. As a result, the DMM concept enables networks to be scaled up cost-effectively as data increases. DMM is currently a quite hot topic in the IETF and 3GPP.

In the future, multimedia will be indeed a main service as well as a major challenge of the networks [1]. Thus, how to efficiently distribute this type of traffic becomes one of the key questions. In this context, IP multicast which provides an effective mechanism for video delivery plays a very important role.

Regarding the multicast over DMM environments, multicast mobility support can be enabled by deploying Multicast Listener Discovery (MLD) Proxy function [10] at mobile access routers (MARs). When an MN starts a multicast session at the current MAR, it receives the multicast traffic from the multicast infrastructure via the current MAR. In case of mobility, the traffic will be forwarded via the tunnel from the previous to the current MAR. This resembles the tunnel-based approach in PMIPv6 [11]. This scheme can be applied for both multicast source and listener in DMM. However, in this paper, we mainly focus on the multicast listener support.

Although this simple scheme can bring multicast listener support into DMM environments, there are some issues e.g. tunnel convergence problem and sub-optimal routing, among others [12]. Since the objective of DMM is moving the mobility anchors from the core to the edge of the networks, the number of mobility anchors in a DMM domain (anchoring MAR) will be much more than that in a PMIPv6 domain (LMAs - in the core network). Thus, the tunnel convergence problem may get more serious than that in PMIPv6 especially in highly mobile regime. This problem can be eliminated by using the native multicast infrastructure for delivering multicast traffic (direct routing approach). However due to the delay related to multicast join process; it may cause significant service disruption (large handover delay and number of packet loss) during handover.

In this paper, we propose two mechanisms which are able to reduce the impact of tunnel convergence problem (redundant traffic at MARs) with an acceptable service disruption time. The first proposal is a trade-off between direct routing and tunnel-based approach. The DMM domain is divided into "virtual multicast domains" (m-domains) in which the MARs are configured to the same upstream multicast router (MR). When an MN moves between MARs in the same m-domain, the direct routing takes place; while the tunnel-based approach is applied for handovers between MARs in different m-domains. As a result, it can significantly reduce the utilization of mobility tunnel for delivering multicast traffic, and reduce redundant traffic at MARs accordingly, with a minor increase of service disruption time compared to the tunnel-based approach. The second proposal uses a single multicast mobility anchor (MMA) for all attached listeners in a DMM domain, similar to [13]. This solution eliminates the redundant traffic but may cause a noticeable service disruption during handover when considering a large domain.

The rest of this paper is organized as follows. Section II describes related work on the mobility management and multicast mobility. In section III, the solutions including different approaches are introduced. Section IV provides performance analysis in terms of redundant traffic and service disruption time. Section V shows numerical results taking into account the impact of different factors. Eventually, section VI concludes the paper and provides perspectives for the future work.

II. RELATED WORK

A. Distributed Mobility Management (DMM)

Due to the lack of DMM standard, in this paper, a generic approach considers that a DMM domain consists of the MARs which implement the functionality of a plain access router, a mobile access gateway (MAG), and a local mobility anchor (LMA) [9][14]. In a DMM domain, an MN gets a different set of IP addresses when changing its point of attachment. In case of mobility, the MN's flows are anchored (if necessary) at the MAR in which the using MN's prefix is allocated. Hence, the packets can be redirected via the tunnel from the previous to the current MAR. Distributed mobility management can be applied fully where both data and control plans are distributed; or partially where the central mobility anchor is still present, but for control plane only.

B. Multicast Mobility

Multicast support for mobile listener can be enabled within a PMIPv6 domain by deploying MLD Proxy function at MAGs while LMA provides multicast router or MLD Proxy function. In this scenario, the upstream interface of an MLD Proxy instance at MAG is configured to the tunnel towards the corresponding mobile node's LMA (called tunnel-based solution) [11]. The presence of the tunnel raises the issues of tunneling overhead, non-optimal route and tunnel convergence problem. Another possibility for multicast support is the direct routing approach [13] that takes advantage of the native multicast infrastructure for delivering multicast traffic, thus avoiding tunnel convergence problem. Yet, this approach may require the multicast tree reconstruction during handover, which may result in a significant service disruption.

Regarding multicast in DMM environments, there is no detailed solution for multicast support, since the DMM is still in its infancy. In [15], the authors provide different use cases for IP multicast support as well as mention about the issues when IP multicast is applied in DMM paradigm. Two scenarios are considered regarding the multicast functionality deployed in the MAR: MLD Proxy or multicast router.

In the first scenario, the direct routing approach is used for new multicast sessions while the tunnel-based is used for the sessions after mobility (handoff sessions). When an MN initiates a multicast session at the current MAR, the multicast traffic will be delivered from the multicast infrastructure to the MAR. Thus, the upstream interface of an MLD Proxy instance at MAR is configured towards the multicast infrastructure. Once the MN moves to a new MAR (nMAR), an MLD Proxy instance at the nMAR adds the downstream interface to the MN and configures its upstream interface to the bi-directional tunnel towards the previous MAR (pMAR). Then, the multicast traffic is routed from the pMAR to the nMAR. It is noted that the tunnel can be dynamically created or pre-established for sharing between MNs as similar as in PMIPv6 [6].

Nevertheless, this scheme does not address any specific optimizations and performances issues such as tunnel convergence, sub-optimal routing, and service disruption. In particular, the tunnel convergence problem becomes a severe issue since the number of mobility anchors in a DMM domain is supposed to be increased. Also, tunneling encapsulations impact the overall network performance and incur delays in multicast packet delivery [14].

In the second scenario, the multicast router function is deployed at all MARs that allows them to select the upstream multicast router based on multicast routing information and/or network management criteria. Thus, the tunnel convergence problem and sub-optimal routing are avoided. However, due to its implementation or operational costs, operators may not want to support multicast routing on MAR. For that reason, in this paper we focus on the case where MAR acts as an MLD Proxy.

III. DESCRIPTION OF THE SOLUTIONS

As described in the previous section, in DMM environments, the tunnel convergence problem becomes more severe compared to that in PMIPv6 especially in highly mobile environment. In this paper, we propose two solutions to address this problem taking into account the service disruption time. Both solutions are considered in two schemes: fully and partially distributed.

- Optimizing multicast content delivery solution (in short OMCD): Similar to PMIPv6, there are two possible approaches for multicast mobility support in DMM environments: direct routing and tunnel-based. The direct routing can helps avoid the limitations of the tunnel-based approach (e.g. tunnel convergence problem, tunnel overhead and sub-optimal routing) but can cause significant service disruption time. Thus, we propose a hybrid solution: direct routing for handoffs inside an m-domain, tunnel-based for handoffs between m-domains. This solution can bring some benefits like reducing tunnel convergence problem and tunnel overhead (compared to tunnel-based approach); and decreasing service disruption time (compared to direct routing approach).
- Multicast Mobility Anchor in DMM (MMA-DMM): A network entity called multicast mobility anchor (MMA) is introduced to provide multicast service access to all attached listeners in a DMM domain, similar to [13]. This simple method helps to avoid the tunnel convergence problem but may result in a significant service disruption during handover.

A. Optimizing multicast content delivery (OMCD)

The DMM domain is divided into m-domains in which the MARs have the same upper MR. When an MN moves between MARs in the same m-domain, the direct routing approach is applied. Otherwise, the tunnel-based takes place (for handoff between m-domains). It should be noted that the using of the mobility tunnel for delivering multicast traffic is temporary and it is kept till the new MAR starts receiving packets from the multicast infrastructure.

The decision to apply which approaches will be based on the comparison between the addresses of the upstream multicast router (UMRA) of the MARs (pMAR, nMAR). In partially distributed scheme, the decision will be made by a Multicast Mobility Control (MMC) which acts as a mobility signaling relay [9]. The address of multicast upstream router of all MARs needs to be stored at MMC. It can be done by a static configuration or during Proxy Binding Update (PBU) / Proxy Binding Acknowledgement (PBA) messages exchanging between MARs and MMC (PBU/PBA need to be extended to convey the address of the MAR's upstream MR). In fully distributed scheme, the nMAR which deploys an enhanced function called Mobility Decision Function (MDF) can make the decision.

The solution is described in Fig 1. In this figure, MAR1 and MAR2 belong to the m-domain 1 (with the common upstream multicast router MR1); while MAR3 and MAR4 belong to the m-domain 2 (MR2's m-domain). A listener (MN1) subscribes to a multicast channel (S, G) at MAR1 and latter moves from MAR1 to other MARs. The operations of the solution are briefly described as follows:

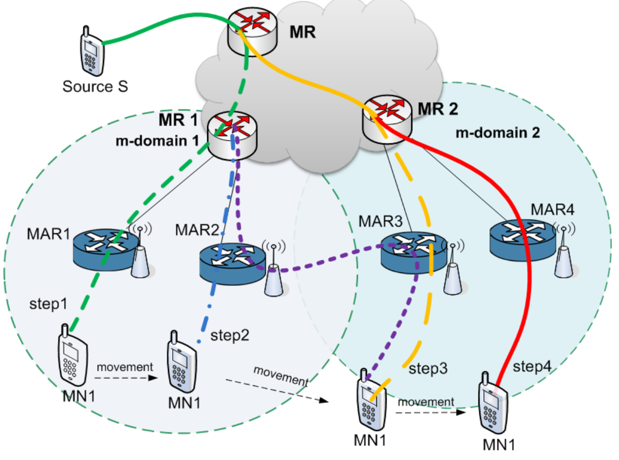


Figure 1. Demonstration of OMCD solution.

- Step1: When the MN1 starts a multicast session at MAR1, the multicast traffic is routed directly from the native multicast infrastructure to MAR1 (following the route S-MR1-MAR1-MN1).
- Step2: The MN1 moves to MAR2 (handoff inside an m-domain). Thus, direct routing scheme is applied. The MAR2 configures its upstream interface to the common MR in its m-domain (MR1) and receives the multicast traffic from this MR (S-MR1-MAR2-MN1).
- Step 3: Then, the MN1 moves to MAR3 which belongs to the m-domain 2. First, MAR3 configures its upstream interface to MAR2 and receives multicast traffic for (S, G) from MAR2 (S-MR1-MAR2-MAR3-MN1). Then, MAR3 sends an aggregated MLD Report to its default MR (MR2) to get the traffic from the multicast infrastructure (S-MR2-MAR3-MN1). Once MAR3 receives multicast packet from MR2, it sends a MLD report to MAR2 to discontinue receiving multicast traffic from the tunnel between them (MAR2-MAR3). These operations of MAR3 can be done by using a MLD Proxy with multiple upstream interfaces [16].
- Step 4: Again, when the MN moves inside an m-domain from MAR3 to MAR4, the direct routing scheme takes place to deliver the multicast traffic from the native multicast infrastructure to MAR4 (S-MR2-MAR4-MN1).

1) Partially distributed scheme

Once an MN attaches to a MAR, it acquires an IP address issued from the prefix (Pref1) which is allocated by the current MAR. It then can use this address to initiate new multicast sessions. The current MAR will receive multicast traffic from the multicast infrastructure then forwards them to the MN as described in the previous section.

When the MN moves to a new MAR (nMAR), the nMAR allocates a new prefix (Pref2) for the MN and sends a PBU to the MMC (see Fig. 2, Fig. 3). After checking its database, the MMC forwards it to the previous MAR (pMAR) which then replies by a PBA. After checking the UMRAs of pMAR and nMAR, the MMC send a PBA to the nMAR which consists of pMAR's address and an addition (M) flag. The flag M is set to 1 if two UMRAs are the same,

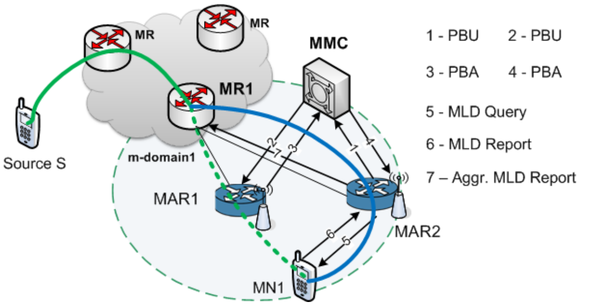


Figure 2. Handover inside an m-domain (partially distributed scheme).

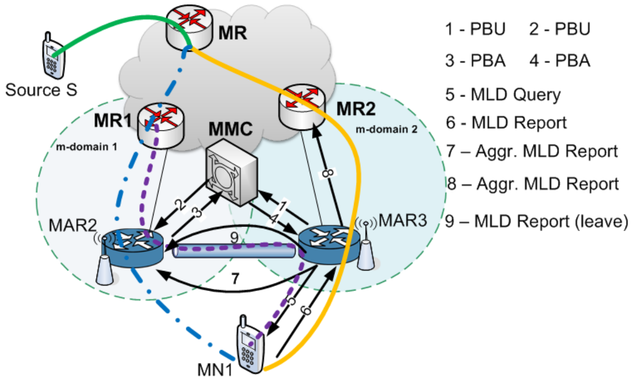


Figure 3. Handover between m-domains (partially distributed scheme).

otherwise 0. Then a tunnel is established between two MARs to route the unicast traffic from/to MN1 using Pref1.

For multicast service, after obtaining the MN multicast subscription information by using a regular MLD Query/Report procedure, and checking the M flag, the nMAR will decide to configure its upstream interface towards the pMAR or the multicast infrastructure. If the flag M is equal to 1, the nMAR then sends an aggregated MLD Report to the upper MR, otherwise to the pMAR (M=0), in order to subscribe to the necessary multicast groups on behalf of the MN. In case two MARs belong to different m-domains (M=0), the nMAR also sends an aggregated MLD Report to the MR in the multicast infrastructure to get multicast traffic from this MR (direct routing approach). Thus, the using of the tunnel between the nMAR and the previous one is temporary and it will be kept till the nMAR starts receiving packet from the native multicast infrastructure. Upon receiving multicast packets, the nMAR will check the sequence number of the packet from the tunnel. If there is any missing packet, it will wait till the packet is forwarded from the pMAR. It then requests to leave the multicast groups from the pMAR.

2) *Fully distributed scheme*

In fully distributed scheme, it is supposed that the nMAR knows the address of the previous one. There are several methods to get this address such as using a layer 2 handover infrastructure (e.g. IEEE 802.21), or using a distributed LMA-discovery mechanism. The exact process to get this address is out of scope of this paper.

Similar to partially distributed scheme, when an MN initiates a new multicast session, the multicast traffic is transmitted from the native multicast infrastructure to the MN (direct routing approach). When the MN moves to a new MAR (nMAR), the nMAR sends a PBU message to the previous one (pMAR). The pMAR then replies by a PBA that contains its UMRA. Upon receiving the PBA and checking its UMRA, the nMAR will make a decision to configure its upstream interface to the tunnel towards the pMAR; or towards the multicast infrastructure (MR2) as described in the previous section (see Fig. 4, Fig .5).

B. *Multicast Mobility Anchor in DMM (MMA-DMM)*

Serving as a mobility anchor for multicast traffic for all MARs in a DMM domain, the MMA can act as an additional MLD Proxy or a multicast router [13]. In this scenario, an MLD Proxy instance is deployed at each MAR with the upstream interface being configured to the MMA. The operations for both partially and fully distributed scheme are the same, and as follows.

When an MN starts a new multicast session, the current MAR sends an aggregated MLD Report to the MMA which then subscribes to the multicast group (if necessary) and forwards multicast traffic to the MAR. When the MN moves to nMAR, the similar processes are executed allowing the nMAR to receive multicast traffic from the MMA.

Since the MARs only receives the multicast traffic from the MMA, the tunnel convergence problem is avoided. However, the requirement of DMM for the distributed deployment (traffic does not need to traverse central deployed mobility anchors) cannot be respected [17]. Again, it raises the problem of single point of failure and sub-optimal routing. These problems can be slightly reduced by deploying several MMAs in which each MMA serves one or several multicast channels.

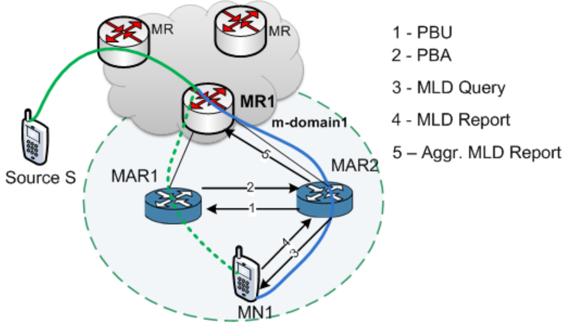


Figure 4. Handover inside an m-domain (fully distributed scheme).

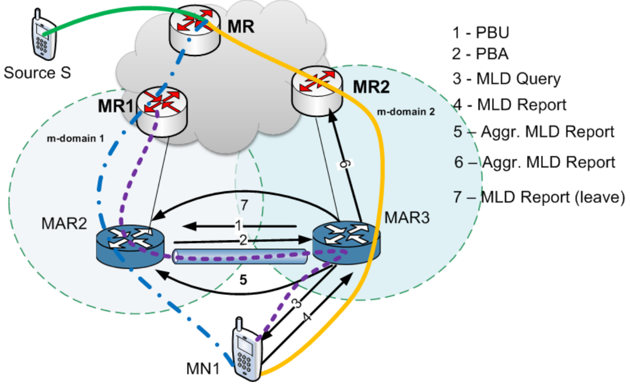


Figure 5. Handover between m-domains (fully distributed scheme).

IV. Performance Analysis

A. Comparison of Tunnel convergence problem

To measure the solutions proposed, it is assumed that an MN starts at least one new multicast session when it moves to a new MAR.

If the direct routing approach is used for both new and handoff sessions, the multicast traffic is always routed from the multicast infrastructure to the MAR. As a result, the tunnel convergence problem is eliminated. Similarly, the MMA-DMM solution also helps to avoid the tunnel convergence problem. In this section, we compare the number of tunnels established for multicast traffic between one MAR to the others or to the multicast infrastructure for the same multicast group (namely N_t) in case of tunnel-based approach (TB) and OMCD solution via the ratio between them (ϴ). ϴ is calculated as $\Theta = N_{t\,(OMCD)} / N_{t\,(TB)}$. Thus ϴ can be used to illustrate how efficient OMCD is, compared to tunnel-based solution in terms of reducing number of redundant traffic (tunnel convergence problem).

If the tunnel-based approach is used for handoff sessions, each time the MN moves to a new MAR, a new mobility tunnel will be established between this MAR and the previous one to redirect the multicast traffic to the current location of the MN. Consequently, the number of tunnels established (multicast tunnel) is proportional to the number of handoffs between MARs (proportion is α). Also, the number of multicast tunnels established in OMCD solution is proportional to the number of "virtual handoffs" between m-domains. Let E_{TB}, E_{OMCD} denote the expected number of handoffs between MARs and between m-domains, respectively. Each m-domain coverage area is supposed to be circular with n subnets (n MARs). Let m denote the number of MARs in the DMM domain. Then we have:

$$N_{t\,(TB)} = \alpha\, E_{TB} / m, \tag{1}$$

$$N_{t\,(OMCD)} = \alpha\, E_{OMCD} / m. \tag{2}$$

According to the [18], $E_{OMCD} = E_{TB} / \sqrt{n}$. Then we obtain:

$$\Theta = 1 / \sqrt{n}. \tag{3}$$

B. Comparison of Service disruption time

A service disruption time analysis has been done in [19] taking into account the different schemes (fully and partially distributed; reactive and proactive handover). However, only tunnel-based approach is considered. In this section, three approaches - tunnel-based (TB), MMA-DMM and OMCD are considered for both partially distributed (PD) and fully distributed scheme (FD).

Fig. 6 shows a reference topology for performance analysis. The delay between the entities is defined as follows:

- t_{wl}: the delay between the MN and access router (AR) (wireless connection).
- t_{am}: the delay between the AR and MAR.
- t_{mm}: the delay between two MARs.
- t_{mc}: the delay between the MAR and MMC.
- t_{ma}: the delay between the MAR and MMA.

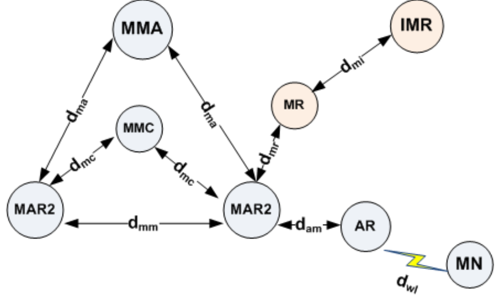


Figure 6. Reference topology for performance analysis.

- t_{mr}: the delay between the MAR and its upstream multicast router.

Similar to [19] the service disruption time is studied based on a well-known factor, called session-to-mobility ratio (SMR) that represents the relative ratio of session arrival rate to the user mobility rate. It is assumed that the subnet residence time (MAR subnet) and multicast session duration follow an exponential distribution with parameter η and μ, respectively. Hence, SMR is defined as $\rho = \eta / \mu$ [18]. Since each m-domain coverage area is supposed to be circular with n subnets (n MARs), the handoff probability between MARs in the same m-domain and between MARs in different m-domains are defined as $\rho_{MAR} = 1 / (1+\rho)$ and $\rho_{MR} = 1 / (1+\rho\sqrt{n})$, respectively as in the literature [18].

The average service disruption time for handoff between MARs is calculated as $T = D * \rho_{MAR}$ where D is the service disruption time. Let t_{L2} denote the Layer 2 handover delay. Assuming that the delay associated with the processing of the messages in the network entities (e.g. time for PBU processing and updating binding cache in pMAR) is included in the total value of each variable. Then the service disruption time is given detailed as:

$$D_{TB\text{-}PD} = t_{L2} + 3t_{am} + 3t_{wl} + 4t_{mc} + 2t_{mm}, \tag{4}$$

$$D_{TB\text{-}FD} = t_{L2} + 3t_{am} + 3t_{wl} + 4t_{mm}, \tag{5}$$

$$D_{MMA\text{-}PD} = t_{L2} + 3t_{am} + 3t_{wl} + 4t_{mc} + 2t_{ma}, \tag{6}$$

$$D_{MMA\text{-}FD} = t_{L2} + 3t_{am} +3t_{wl} + 2t_{mm} + 2t_{ma}. \tag{7}$$

In the direct routing approach (DR), the nMAR's upstream MR needs to join and get multicast traffic from a multicast router in the multicast infrastructure that already had multicast forwarding states for this group (called common multicast router or CMR). Thus, an additional delay is taken into account: $2t_{mi}$. The service disruption time in the direct routing approach is calculated as follows:

$$D_{DR\text{-}PD} = t_{L2} +3t_{am} + 3t_{wl} + 4t_{mc} + 2t_{mr} + 2t_{mi}, \tag{8}$$

$$D_{DR\text{-}FD} = t_{L2} + 3t_{am} + 3t_{wl} + 2t_{mm} + 2t_{mr} + 2t_{mi}. \tag{9}$$

ISBN: 978-1-61208-256-1

When the MN performs handoffs inside an m-domain, the MR of this m-domain has already subscribed to the multicast group (the MR and CMR located at the same entity), thus $t_{mi} = 0$. We obtain the value of delay for direct routing approach when the MN moves inside an m-domain in case of partially and fully distributed scheme, called $D^{*}_{DR\text{-}PD}$, $D^{*}_{DR\text{-}FD}$ respectively.

Since in OMCD solution, the direct routing approach is applied when an MN performs handoffs inside an m-domain while tunnel-based takes place for inter m-domain handover. Thus, the average service disruption time is calculated as:

$$T_{OMCD\text{-}PD} = (\rho_{MAR} - \rho_{MR})\, D^{*}_{DR\text{-}PD} + \rho_{MR}\, D_{TB\text{-}PD}, \quad (10)$$

$$T_{OMCD\text{-}FD} = (\rho_{MAR} - \rho_{MR})\, D^{*}_{DR\text{-}FD} + \rho_{MR}\, D_{TB\text{-}FD}. \quad (11)$$

C. Comparison of End-to-End delay

In the direct routing approach, the end-to-end delay is calculated as $D_{e\ (DR)} = t_{S,\ MAR} + t_{MAR,\ MN}$. The delay between MN and MARs are supposed to be the same ($t_{MN,\ pMAR} = t_{MN,\ nMAR}$). If the tunnel-based approach is used, after handover, there is an additional delay compared to the direct routing approach: $t_{S,pMAR} - t_{S,nMAR} + t_{pMAR,\ nMAR}$. With a large delay between two MARs (tunnel delay), the end-to-end delay is significantly increased. In average, the end-to-end delay of the tunnel-based approach is increased $t_{MAR\text{-}MAR}$ that is the average delay between two MARs compared to that of the direct routing.

In the MMA-DMM solution, the end-to-end delay depends on the position of MMA. In a significant large domain, it may be much higher than that of the direct routing approach. For the OMCD solution, the using the tunnel pMAR-nMAR is temporary, thus, in average, the end-to-end delay is almost the same as in the direct routing approach.

V. Numerical Results

This section presents the numerical results based on the analysis given in the previous section. The default parameter values for the analysis are introduced in TABLE I, in which some parameters are taken from [19].

TABLE I. PARAMETERS FOR PERFORMANCE ANALYSIS

Parameters	*Values*	*Parameters*	*Values*	*Parameters*	*Values*
t_{L2}	100ms	t_{wl}	5ms	t_{am}	2ms
t_{mm}	2ms	t_{mc}	3ms	t_{mr}	5ms
t_{ma}	20ms	t_{mi}	0ms	n	32

Fig. 7 shows how efficient OMCD is in comparison with the tunnel-based solution in terms of reducing number of redundant traffic at MARs (tunnel convergence problem). As n increases, the amount of redundant traffic decreases. When all MARs in a DMM domain belong to only one m-domain (n = m), there is no redundant traffic at MARs (OMCD becomes MMA-DMM solution).

The average service disruption time as a function of SMR (ρ) is illustrated in Fig. 8, when the number of MARs in an m-domain is fixed to 32. The service disruption time of MMA-DMM solution is definitely higher than that of the others. Although the service disruption time of OMCD solution is a bit higher than that of tunnel-based approach, the difference between them is negligible.

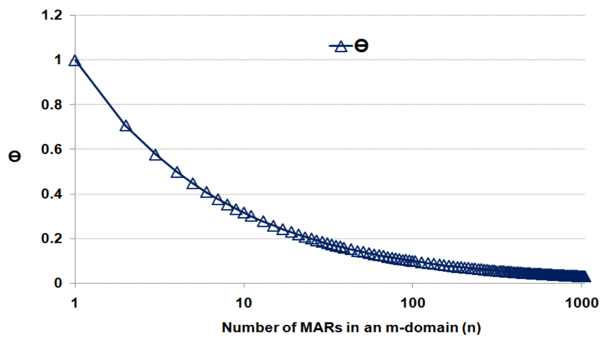


Figure 7. Ratio between number of redundant traffic in the OMCD solution and in the tunnel-based approach (Θ).

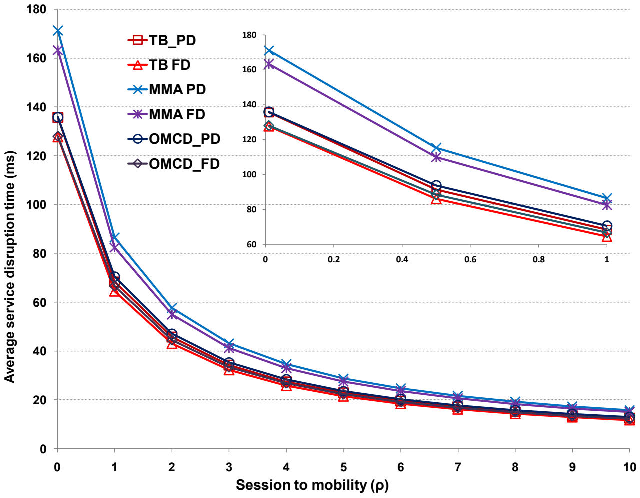


Figure 8. Average service disruption time as a function of SMR (ρ).

Now, the service disruption is considered when the number of MARs in an m-domain (n) is varied. Since the delay between two nodes depends on the bandwidth, the propagation delay and the distance between them, for simplicity, we suppose that the delay is proportional to the distance (proportion is τ). It is assumed that the architecture of an m-domain is hierarchically formed as a binary tree with a d_{mr}-layer [20]. Therefore, t_{mr} is calculated as $t_{mr} = \tau \log_2(n)$. It is noted that when n is equal to number of MARs in the network (n = m), OMCD becomes MMA-DMM solution. Fig .9 describes the average service disruption time as a function of number of MARs in an m-domain when ρ = 0.1 and τ = 2. The average service disruption time in the OMCD solution is slightly increased when the number of MARs is increased as a result of the trade-off with the decreased of the redundant multicast traffic.

Regarding the tunnel delay impact, the value for t_{mm} is varied over a range from 0.1 to 30ms. In Fig .10, we can see how the different solutions are dependent on the mobility

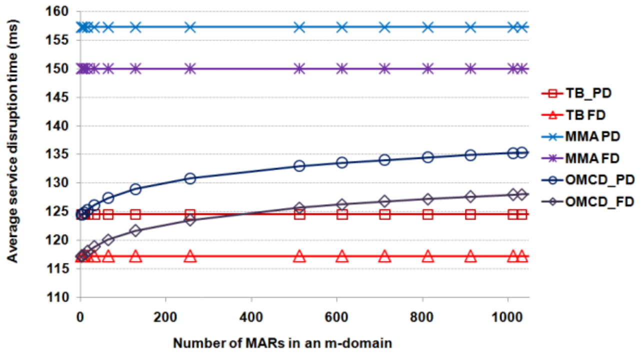


Figure 9. Average service disruption time as a function of n.

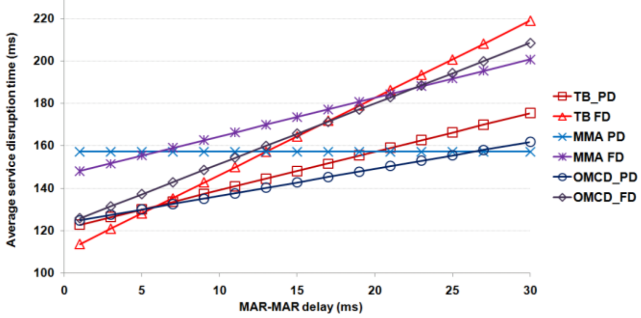


Figure 10. Tunnel delay effect.

tunnel. As t_{mm} increases, the average service disruption time for all approaches (except MMA_PD) increases. It is worth noting that if the tunnel delay is larger than a specific value, the OMCD becomes better than the tunnel-based solution. The MMA-DMM becomes the best solution in terms of service disruption time if the tunnel delay continues increasing.

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed two solutions to address the tunnel convergence problem as a result of multicast listener mobility over DMM environments. The first solution helps to greatly reduce the number of redundant traffic caused by the tunnel convergence problem with a minor increase of service disruption time compared to the tunnel-based approach. The second one uses a network entity (or several) serving multicast service for all attached listeners in a DMM domain. It is an easy way to solve the tunnel convergence problem but may cause a significant service disruption.

In the future, the multicast source mobility will be considered in DMM environments. Also, the simulations will be made based on the Network Simulator NS-3 and a DMM implementation (extended version of OAI PMIP [21]) to better evaluate the performance of different approaches.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Community Seventh Framework Programme (FP7-ICT-2009-5) under grant agreement n. 258053 (MEDIEVAL project).

REFERENCES

[1] Cisco White Paper, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2011-2016", February 2012.

[2] 3GPP TR 23.829, "Local IP Access and Selected IP Traffic Offload (LIPA-SIPTO)", Release 10, Mars 2011.

[3] R. Costa, T. Melia, D. Munaretto, and M. Zorzi, "When Mobile Networks Meet Content Delivery Networks: Challenges and Opportunities", MobiArch, August 2012.

[4] H. Chan. Yokota, J. Xie, P. Seite, and D. Liu, "Distributed and Dynamic Mobility Management in Mobile Internet: Current Approaches and Issues", Journal of Communications, Vol.6, No.1, February 2011.

[5] C. Perkins, D. Johnson, and J. Arkko, "Mobility Support in IPv6", RFC 3775, July 2011.

[6] S. Gundavelli, K. Leung, V. Devarapalli, K. Chowdhury, and B. Patil, "Proxy Mobile IPv6", RFC 5213, August 2008.

[7] H. Chan (Ed.), "Problem Statement for Distributed and Dynamic Mobility Management", IETF-Draft (work-in-progress), October 2011.

[8] H. Yokota, "Use case scenarios for Distributed Mobility Management", IETF-Draft (work-in-progress), September 2010.

[9] CJ. Bernardos, A. de la Oliva, F. Giust, T. Melia, and R. Costa, "A PMIPv6-based solution for Distributed Mobility Management", IETF Draft (work-in-progress), March 2012.

[10] B. Fenner, H. He, B. Haberman, and H. Sandick, "Internet Group Management Protocol (IGMP) /Multicast Listener Discovery (MLD)-Based Multicast Forwarding (IGMP/MLD Proxying)", RFC 4605, August 2006.

[11] T. Schmidt, M. Waehlisch, and S. Krishnan, "Base Deployment for Multicast Listener Support in PMIPv6 Domains", RFC 6224, April 2011.

[12] I. Romdhani, M. Kellil, H. Lach, A. Bouabdallah, and H. Bettahar, "IP Mobile Multicast: Challenges and Solutions", IEEE Communications Surveys & Tutorials, vol. 6, no. 1, pp. 18-41, 2004.

[13] JC. Zuniga, LM. Contreras, CJ. Bernardos, S. Jeon, and Y. Kim, "Multicast Mobility Routing Optimizations for Proxy Mobile IPv6", IETF Draft (work-in-progress), March 2012.

[14] P. Seite and P. Bertin, "Distributed Mobility Anchoring", IETF Draft (work-in-progress), July 2012.

[15] S. Figueiredo, S. Jeon, and R. L. Aguiar, "IP Multicast Use Case Analysis for PMIPv6.based Distributed Mobility Management", July 16, 2012.

[16] H. Asaeda and S. Jeon, "Multiple Upstream Interfaces Support for IGMP/MLD Proxy", IETF Draft (work-in-progress), October 2012.

[17] H. Chan (Ed.), "Requirements for Distributed Mobility Management", IETF Draft (work-in-progress), September 2012.

[18] C. Makaya and S. Pierre, "An Analytical Framework for Performance Evaluation of IPv6-Based Mobility Management Protocols", IEEE Transaction on Wireless Communications, vol. 7, no. 3, pp. 972-983, March 2008.

[19] S. Figueiredo, S. Jeon, and R. Aguiar, "Use-cases Analysis for Multicast Listener Support in Network-based Distributed Mobility Management", PIMRC, September 2012.

[20] Y. Min-hua, Y. Lv-yun, L.Yu, and Z. Hui-min, "The implementation of Multicast in Mobile IP", WCNC, March 2003.

[21] OAI PMIP, http://www.openairinterface.org/openairinterface-proxy-mobile-ipv6-oai-pmipv6.

Wireless Sensor Actor Networks For Industry Control

Yoshihiro Nozaki, Nirmala Shenoy
Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, NY, USA
yxn4279@rit.edu, nxsvks@rit.edu

Qian Li
CAST-Telecommunications Engineering Technology
Rochester Institute of Technology
Rochester, NY, USA
qxl2571@rit.edu

***Abstract*— A robust and reliable architecture for wireless sensor actor networks for industry control is discussed and described in this paper. The stringent physical constraints in an industry environment are taken into consideration. A combination of MAC and routing protocol to support reliable and robust transportation of data is described.**

Keywords-Sensor Actor Networks; Industry Control; Robust and Reliable Architectures.

I. Introduction

Wireless Sensor-Actuator Networks (WSAN) comprise of wireless sensors and actuators (or actors). Sensors are low-processing, low-energy devices that sense data such as temperature, pressure and so on. The sensed data is gathered at a *sink* to be analyzed and acted upon. Typically sensors are low-cost disposable devices. Based on the sensed data, actuators make decisions and take action. Actuators have higher processing capacity and are not energy constrained. They may also perform the functions of a *sink*.

Significant technology advances have resulted in major cost reductions in sensors and actuators. This coupled with elegant techniques to overcome challenges in wireless transmissions make WSANs attractive and viable for many applications. Examples are environment / habitat monitoring and control, battlefield surveillance, industry control and automation. In WSAN for environment and habitat monitoring and control, and battlefield surveillance, a large number of sensors are randomly deployed in potentially inaccessible areas, hence they be disposable and highly energy conserving. Multi-hop data collection paths, self-configuration and self-healing are predominant features of WSAN in such applications. Importance of security in such WSANs depends on the applications.

Considering a *Wireless Sensor-Actuator Network for Industry Control* (WSANIC), high survivability and ability to support data, event and task prioritization are predominant requirements. Security is important because of the critical nature of the application. For example explosives, high power and chemical industries could have serious detrimental effects in terms of cost and/or human loss if tampered with. The fact that sensors and actuators could be placed in least human-frequented areas makes them highly vulnerable to security attacks.

In contrast to the distinctive features mentioned earlier for WSANs, in a WSANIC, sensors and actuators are manually placed, resulting in a more stationary and deterministic topology. Self-configuration and self-healing are required upon device failures or environmental changes. Devices may not be disposable and batteries can be charged or changed regularly. Thus, some issues that pose serious challenges in WSAN are less problematic in WSANIC [3]. Robustness, interference in communications and data reliability are of major concern in a WSANIC. To improve robustness one has to look for options other than using powerful antennas as high power transmissions pose danger in inflammable spaces and increase interference effects [2]. In an industry environment, *high electromagnetic fields* due to heavy electrical devices and power cables are normal to expect, which negates the use of low power transmissions by sensor and actors. Communications interference is also caused due to events such as environment conditions, moving people and objects all of which can impact timely data transmission. Data reliability is critical as corrupted data could result in improper control of machinery and processes, which could be catastrophic.

Section II describes current industry control networks. Related works that are addressing WSANIC issues is provided in Section III. Section IV describes about WSANIC. Section V introduces our proposed architecture and Section VI analyses the result of simulations. Section VII provides the conclusions.

II. Control Networks in Industry

Wired *Control Networks* (CN) are adequately supporting industry control requirements today. However, in industries dealing with explosives, moving, or rotating machinery, some locations are inaccessible or highly inconvenient to monitor using wired systems. The cabling and conduits for wired sensors and actuators besides being vulnerable to damage can be cost prohibitive - ranging typically to as much as one third to one half of the total system cost [1]. Industrial sensors have seen a steady decrease in costs and the eventual driving cost factor becomes cabling rather than the sensor or actuator cost. A low cost wireless sensor-actuator system with reasonable battery life to provide reliable data collection spanning an entire industry plant, while meeting cost objectives could create a paradigm shift in industry maintenance and control [1]. Such systems would also allow computing power in locations that previously would have been cost-prohibitive [4].

A. Wired Control Network

A *Process Control System* in an industry uses sensors to measure the process parameters and actuators to adjust the

ISBN: 978-1-61208-256-1

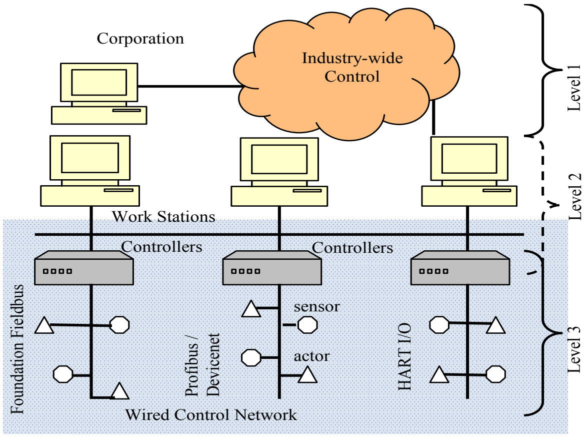


Figure 1. Wired Industry Control Network Architecture

operation of the process. Control action can be inbuilt into actuators or can be in separate entities called controllers. In industry control, it is convenient to have controllers separate from actuators as the controllers collect data from several sensors, make decision on an appropriate action to take (like *proportional, integral, derivative* or combinations of these) and actuate several actuators [3].

In Fig. 1, a typical wired industry-wide control network is shown. It has three levels of hierarchical control. The network at level 3 *that connects the sensors and actuators to the controllers* is of interest to us and we use the term ***wired CN*** for this segment. In this article, we analyze a ***wireless CN*** (WSANIC) that can replace the wired CN.

At level 3, *Foundation Fieldbus* (FF), *Profibus* and *Devicenet* are some of the wired CN industry standards being used [2]. The standards assume inherently *high predictability* and *reliability* as they operate over wired networks and target *real-time* data delivery. Real-time and reliable data delivery is very important in industry control, since loss of scheduled data could result in costly consequences [3]. Other performance affecting factors to consider are data rates, distance and transmission ranges. For example at the physical layer of FF, the official data rate is 31.25 Kbps. A process unit in a plant could span tens to hundreds of meters. Depending on the cable types and whether the controller is mounted close to the sensor/actuator or in a remote room, the distance range of FF is expected to be from 200 to 1900 meters [3]. *As a promising alternative to industry control, a WSANIC should have capabilities similar to the wired CN.*

B. Wireless Control Network

The frequency spectrum used in current wireless networks, can support high data rates. However, long transmission ranges are difficult to achieve given that high power transmissions are undesirable. In [4], Enwall T. provides statistics from studies conducted on suitability of major wireless network standards like 802.11g, 802.11s, Zigbee 802.15.4 and WiMax for industry control as per ISA-SP100. From the statistics it is clear that none of the above standards come close to doing what they need to do to fully support industrial applications. However, combining Zigbee with a *service broker* [4] improved its rating considerably, though it still fell short in several aspects such as network and messaging security, adequate reporting rates, quality of service in terms of timeliness, delivery ordering and recovery actions among others.

III. Related Work

A survey of related literature reveals that there are few contributions that address WSANIC issues [1 - 4]. The prime focus in these articles are on how best to replace the FF or other similar wired CN [3] with a wireless counterpart.

From an industry and standards perspective, several wireless organizations are investigating solutions and pursuing adoption of wireless standards promoted by them. Of these WINA, Zigbee, ISA wireless system for automation, wireless HART are some major ones [2]. However none of these efforts takes into consideration industry environmental, placement and access restrictions.

In [8], the authors observe that “a WSAN should be robust to node failures and in general exhibit fast dynamic response to changes”. In [9], researchers at *Massachusetts Institute of Technology* harnessed the robustness inherent in mesh topologies in a WSANIC test bed. These observations indicate that topology and architectural issues are important to consider in a WSANIC architecture. High survivability and security are of also very important. These are best addressed via suitable architectures and/or topology.

IV. Wireless Sensor Actuator Networks for Industry Control

We start with three main devices essential in a WSANIC, namely sensors, actuators and controllers and distinguish their functions in an industry control environment. Without loss of generality, it is assumed that sensors and actuators are distinct devices. Sensors are end devices that collect and transmit data while actuators are end devices that receive data and actuate a lever or valve. The controller, which we henceforth call an *Access Control Point* (ACP) is the data collection device that collects data from several sensors and is the source point of control data to several actuators. Inter-ACP communication required for industry wide control may be over wireless or wired links is out-of-scope in this work. ACPs will be limited in number and positioned at specific locations. Hence it may not be possible for all sensors and actuators to have line of sight communications path to an ACP. For robustness in connectivity it is further essential that sensors and actuators have routes to multiple ACPs.

A. The Architecture

To overcome the physical issues due to communications range, line of sight and to provision multiple paths between ACPs and sensor/actuators special devices called ‘relays’ are introduced. Relays forward data for other devices and will provide multiple paths of communications. It has been observed [5] that multiple types of devices result in complex management due to diversity in techniques, data collection methods and protocols. In the proposed architecture, multiple types of devices are necessary to provide robustness and adaptability. However complex communications and management are avoided by using a set of medium access and routing protocols common to all devices.

The architecture comprising of ACPs, sensors, actuators and the relay mesh that emerges from the discussions thus, far is pictured in Fig. 2. The emphasis is on WSANIC at level 3 that will embed into the 3-level hierarchy from Fig. 1. As per the architecture, relays and an Access Control Point (ACP) are used besides sensors and actors. The ACP is responsible for implementing the proportional, integral and/or derivative control depending on the process. The control action is then conveyed to the actuators. The relays facilitate robust connectivity between the ACPs and actuators; ACPs and sensors by providing redundant paths. They are also useful to keep the transmission power low, and facilitate multi-hop communications when two nodes are distant to one another.

B. *The Protocols*

In a typical wired CN standard like the FF, the protocol stack is derived from the OSI 7 layer model, where only the lower two layers namely the physical and the data-link are specified; the network, transport and session layers are removed[3]. The proposed protocol stack for WSANIC also has two layers. The lower layer is the physical layer, which is not the focus of this article, and the layer above i.e. layer 2, has integrated medium access control (MAC) and routing functions that operate off a single header. This is very attractive in wireless networks as it reduces header overhead, processing requirements and its associated delays, while allowing MAC and routing functions to interwork closely.

C. *The MAC Functions*

A MAC protocol for WSANIC should provide timely and near-lossless data delivery that is comparable to wired CN. In wired CN, it is naturally assumed that priority data carrying vital information under alarm conditions will be delivered reliably and in time. However, this assumption is not valid in wireless networks and sensitive, urgent data has to be handled specially to facilitate timely and reliable delivery.

Timely delivery can be achieved through preemptive priority. Preemption requires abortion / delay of other transmissions or receptions on the arrival of high priority data. This capability can be provisioned through the use of a dual channel MAC (one channel to carry high priority data and another for normal data) where the MAC switches the local processing to handle high priority data on its arrival.

Reliability can be achieved through retransmissions on loss of acknowledgements, if accomplished within acceptable latency limits or in the routing functions through the use of concurrent multipath transmissions of critical data to increase the probability of its delivery.

Normally a scheduled MAC is considered suitable for reliable and timely delivery of data. However, we advocated a multi-hop mesh topology which makes it difficult if not impossible to implement scheduled MAC due to synchronizations issue. Moreover in industry environment, an unscheduled MAC will have more flexibility to provide combinations of periodic, event-based and query-based data collection / delivery. If an unscheduled MAC is used, reliability of data delivery has to be achieved via acknowledgements and retransmissions. Given the frequency spectrum used in current wireless networks, the data rates achieved are very high compared to a wired CN data rates (like FF) and retransmissions on loss of acknowledgements can be processed within acceptable latency limits. The routing scheme to be presented next also support timely and reliable data delivery, as it has the capability to send priority data concurrently on proactively maintained multiple paths.

D. *Routing Functions*

ACPs, sensors and actuators in WSANIC can be stationary or mobile. The set of relays that forward data from sensors to actuators can vary due to mobility of ACPs, sensors, and actuators; battery drain at relays or environmental changes. A single route is not advisable as data loss due to route failure could occur. Multiple routes from sensors to ACPs and ACPs to actuators can alleviate this problem. Delays due to new route discovery also cannot be tolerated in critical applications. Hence a robust proactive multipath routing scheme with low overheads would be ideally suited. The *Multi Meshed Tree* (MMT) routing [6] [7] has these desirable features.

E. *MAC and Routing Protocols*

The MAC protocol uses carrier sensing similar to 802.11, but adopts a more deterministic medium access approach. In this new approach, nodes take turns to access the media, based on neighbor knowledge and is called the *Neighbor Turn Taking* (NTT) MAC protocol [10]. This protocol has been previously shown via simulation to perform better than IEEE 802.11 CSMA/CA in terms of end-to-end packet latency and rate of successfully transmitted packets under saturated traffic conditions [11]. The proposed routing scheme sets up overlapping (meshed) trees originating at the ACPs and ending at the sensors and actuator. The meshed trees provide multiple robust routes. They also use neighbor knowledge and are based on the MMT algorithm.

V. IMPLEMENTATION

In this section, we describe the integrated NTT and MMT (NTT-MAC) operation.

A. *The Semi-Automated Architecture*

Fig. 2 shows the semi-automated architecture [12] with relays, sensors, actuators, and ACPs. In this architecture, sensors send data to ACPs, and collected data is processed at

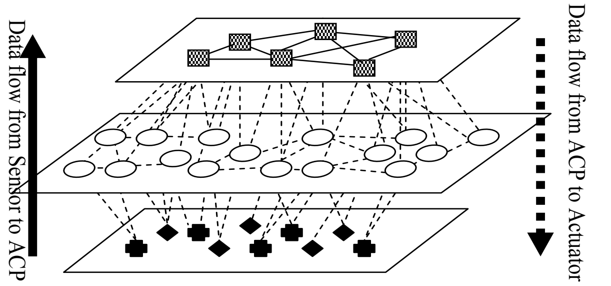


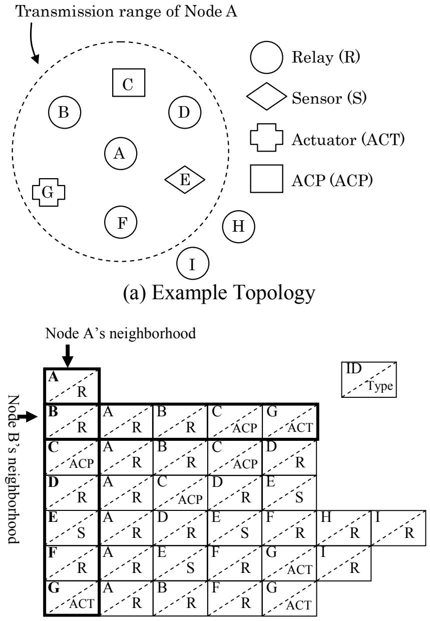
Figure 2. The Semi-automated Architecture

ACPs. The architecture shows 3 layers; the top layer is mesh of ACPs. The middle layer is a mesh of relay nodes, and the bottom layer comprises of sensors and actuators. All nodes in this architecture communicate over a wireless media except for the ACP mesh which could be wire connected. After the data is processed in ACPs, ACPs decide on the proper actuators that are to be activated and communicate to them. In the semi-automated architecture, route maintenance for both sensors-ACPs and ACPs-actuators routes is required. This will result in two way communications along the routes established. Hence, a MAC with low collisions low latency and a robust routing protocol are essentail.

B. *Neighbor Turn Taking Medium Access Control*

NTT-MAC uses a distributed loosely scheduled approach based on neighbor knowledge and their activities. NTT operation requires two processes, 'neighbor sensing' and 'turn scheduling'. Because there are four different types of nodes sensors, relays, actuators, and ACPs, the NTT-MAC proposed in [10] was customized to the new architecture.

1) Neighbor Sensing: Each node overhears the neighbor nodes to calculate its turn to access the medium next. To accomplish this, all nodes in the network advertise themselves and their 1-hop neighbors periodically. Nodes thus, know their neighbor's neighbor information i.e. 2-hops neighbor information. In addition, node type such as sensor, relay, actuator, and/or ACP is also advertised. Fig. 3(b) shows an example of neighbor knowledge of the topology in Fig. 3(a). Nodes B, C, D, E, F, and G are neighbors of Node A. In Fig. 3(b), the left most column in the table represents Node A's neighbor list and each row represents each neighbor's neighbor list. For example, Node B's neighbors are nodes A, C, G and their node types are relay (R), ACP, and actuator (ACT).



(a) Example Topology

A R						
B R	A R	B R	C ACP	G ACT		
C ACP	A R	B R	C ACP	D R		
D R	A R	C ACP	D R	E S		
E S	A R	D R	E S	F R	H R	I R
F R	A R	E S	F R	G ACT	I R	
G ACT	A R	B R	F R	G ACT		

(b) Neighbor table of node A

Figure 3. Neighbor knowledge example

2) Turn Scheduling: Turn scheduling is achieved based on neighbor table and their activities as described next.

a) Neighbor Activities: Each node calculates its next turn based on the sender node's neighbor list which it overhears from its neighbors transmissions. For example, if Node B in Fig. 3 (a) sends a packet, all neighbors nodes A, C, and G hear the transmission of Node B. They will then calculate their next turn by looking up Node B's neighbor list. The neighbor list indicates the order of each node's turn. Therefore, the next sender from Node B will be Node C, and second sender will be Node G, third will be Node A. In order to synchronize their turns, the order in each neighbor list has to be the same with all neighbors. In this work, ACK is used for DATA, and hence each node computes their turn to transmit based on the type of message they overhear.

b) Node's activities: The turn calculation is based on a node's neighbor size. For example, Node B calculates its next turn to be 4th because its neighbor size is 3.

c) Updating: Each node has one next turn scheduled at any time. Thus, each node compares previous turn scheduling time and new turn scheduling time after every turn calculation, and applies the latest scheduled one.

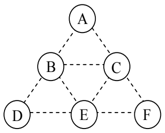
C. *Multi Meshed Tree Routing*

For routing, the Multi-Meshed Tree (MMT) protocol is used to create logical meshed trees in the network. These trees are rooted at the ACP, and the ACTs and sensors are the leaf nodes. Since the semi-automated architecture has two-way data flow, sensor nodes need routes to ACPs and ACPs need routes to actuators. In addition, a sensor can communicate with any ACP and any ACP can communicate with any actuator. Hence, both sensors and ACPs are required to maintain routing information. As a result, route maintenance can become complicated and difficult. Most well-known routing protocols (proactive and reactive) in wireless ad hoc networks such as *Dynamic Source Routing* (DSR) and *Optimized Link State Routing* (OLSR) are required to maintain routing information at sender nodes. MMT requires only ACPs to maintain route information to ACTs. Sensors have the route information to ACPs, which is inherent in their allocated virtual IDs (VIDs). By nature of MMT, leaf nodes in the trees such as sensors and actuators can know routes to the root nodes of the trees once they joined the trees as this information is inherent in the assigned VIDs to the leaf nodes. Likewise, the root nodes such as ACPs know routes for both sensors and actuators. Therefore, sensors do not require to maintain routing information. Because the logical trees are meshed, MMT protocol provides not only overlapping coverage, but also route robustness while avoiding loops in the meshed topology. An optimized version of the MMT algorithm as presented in [7] is used to reduce control packets of MMT in this work.

1) Multi-Meshed Trees (MMT)

As mentioned above, trees are grown from root nodes (ACPs) to leaf nodes (i.e. sensors and actuators) through relay nodes. Each meshed-tree can be viewed as a cluster and the ACP is the cluster head (CH) and all other nodes are the

ISBN: 978-1-61208-256-1



	A	B	C	D	E	F
VID 1	1	11	12	111	112	123
VID 2		121	113	1211	122	1132
VID 3		1221	1123	1121	1212	1122
VID 4				1222	1131	1223
VID 5					1111	
VID 6					1231	

Figure 4. Example of MMT (Hop limit = 3)

cluster clients. 3-ways handshake is adopted by nodes when during the joining process. The ACP or CH initiates tree creation by broadcasting an advertisement (AD) containing its VID. On hearing the AD packet, neighbor nodes which want to join the tree will send a join request (JR) to the sender of AD packet i.e. the parent node. The parent then records the new VID into a JR message and forwards to the CH, which register the new VID to its cluster member. Because the child node can hear the forwarded JR message, the child can know the new VID assigned to it at the time. The CH replies with a join acceptance (JA) packet to the parent after registering the new VID. Finally, the parent sends the JA to the child. And then, the child node starts to advertise its new VID to its neighbors. The new VID for a child node is one additional digit appended to the parent's VID. Fig. 4 shows an example topology and VIDs in MMT. For example, if a CH node A has VID *1*, the child VID can be between *11* and *19*. So, Node B and C will get VID *11* and *12*. Since Node C has *12*, its child can be between *121 – 129*. In this manner, the VID carries the route information. The total number of digits in a VID indicates the hop distance from CH, and also route to CH. The process continues until the tree encounters defined limits such as maximum hop count, cluster size or reaching edge nodes.

To avoid loops in trees, VIDs are not assigned if there is already a child-parent relationship with a particular VID. This VID acceptance rule applies for not only direct parent-child, but also for any grandparents or grand children.

We include the knowledge from NTT into the joining process by combining JR and JA during the 3-ways handshake as shown in Fig. 5. Nodes B and C are neighbors of Node A which has VID *111*. After Node A broadcasts its VID, Node A calculates its next turn based on its neighbor table. Node B and C overhear Node A's AD packet and calculate their next turn based on Node A's neighbor table. As their turn scheduling is based on Node A's neighbor table, next turn scheduling time of Node B and C are the time before Node A's next turn. Hence, join request from all Node A's neighbor can be received before Node A's next turn to transmit. Therefore, Node A can combine all JR messages from its neighbors ideally and assign new VIDs for all the children nodes when Node A gets its next turn and forward the combined JR to the CH. The CH node returns a combined join acceptance (combined JA) to Node A after new VIDs are registered.

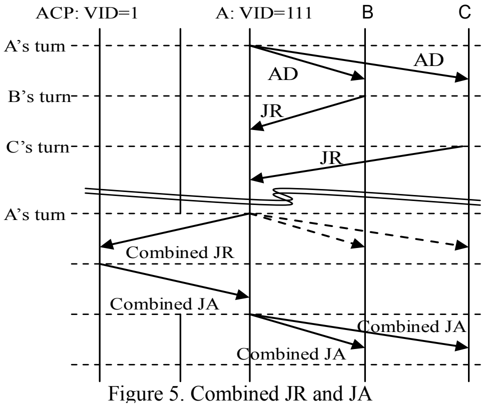
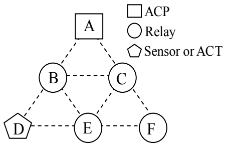


Figure 5. Combined JR and JA



	A	B	C	D	E	F
VID 1	1	11	12	111	112	
VID 2		121	113	1211	122	
VID 3				1121		
VID 4				1221		

Figure 6. MMT for the semi-automated architecture (Hop limit = 3)

D. Interaction Between NTT and MMT

Since MMT uses neighbor knowledge for optimized cluster joining process, MMT interacts with NTT to look up neighbor table. Each node maintains neighbor knowledge which includes not only node ID but also node types. MMT helps set up routes between sensor to ACP and ACP to actuator. If HOP_LIMIT is 5 and a parent VID is *1111*, the parent is located 3 hops from the CH 1. If a child node joined this VID, the child node will be at the 4th hop from CH 1 and the child's child node will be at the 5th hop (last hop). Therefore, if the child node does not have SENSOR or ACT node in its neighbor, the child VID will be meaningless and would use up one node cluster client position wastefully. Thus, a child node which does not have any SENSOR or ACT (actuator) in its neighbors' neighbor will not send JR to the parent if the parent VID is already HOP_LIMIT - 2. Fig. 6 shows optimized MMT for the sample topology. SENSOR and ACT do not allow having child node, so Node D does not have child VID. Because Node F does not have any SENSOR and ACT in its neighbor and Node C and E have already reached 2 hops from the CH, Node F does not join any tree. As a result, total number of control packets is reduced significantly because total number of VIDs is reduced. On the other hand, NTT interacts with MMT to identify sender and destination nodes from VIDs and to calculate turn scheduling from neighbor table and own VIDs.

VI. ANALYSIS RESULTS

A. The Topology

Fig. 7 is the topology used in the OPNET simulations [13]. The topology shows relative placement of the sensors and actuators with respect to the ACPs which is similar to semi-automated industry architecture.

The topology places the relays, sensors and actuators around the ACPs but with the relays between the sensors / actors and ACPs. Several simulations were conducted by varying the number of sensors / actors and with different simulation seed values and the results averaged. The tests were repeated using DSR and 802.11 CSMA/CA for comparison between them. At the ACPs and the sensors, data was generated at the rate of one packet in 0.05 seconds. The data packet size was maintained at 500 bits.

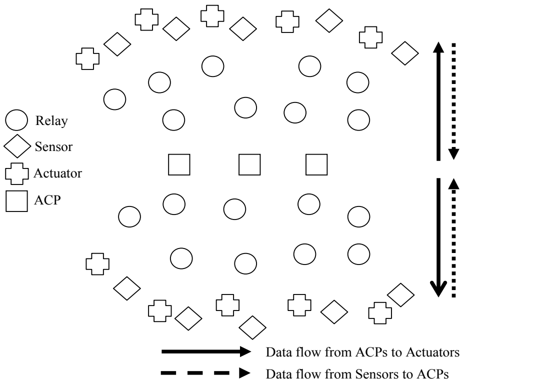

Relay
Sensor
Actuator
ACP
Data flow from ACPs to Actuators
Data flow from Sensors to ACPs

Figure 7. Relative placement of sensors, actuators, relays, and ACPs

B. Performance Metrics

1) Average End to End Latency is the time taken from transmission of a data packet at the sender to its reception at the receiver.

2) Success rate is calculated as the ratio of total number of packets received correctly at the destination node to the total number of packets sent by the sender node.

Two different situations for MMT based systems in each scenario were simulated. One is with 'route salvage' option which has salvage function. The other one doesn't have route salvage. DSR has 'route salvage' implemented. As can be seen in table 1, in all scenarios MMT/NTT based solutions has a consistently higher success rate of over 98%.

The latency was also recorded against the number of hops between the sending and receiving nodes. In most cases, the average end to end delivery latency for MMT/NTT is lower than DSR/CSMA/CA despite the fact that the MMT/NTT delivered more packets.

TABLE I. PERFORMANCE COMPARISON MMT-NTT VS DSR

Metrics \ Scenarios			5 ACT/SENSOR	10 ACT/SENSOR
Success Rate	MMT-NTT (route salvage)		99.104930	99.802880
	MMT-NTT (no route salvage)		98.310980	99.568000
	DSR		95.330350	92.823010
Packet Latency	MMT-NTT (route salvage)	1-hop	N/A	N/A
		2-hop	0.000917	0.002599
		3-hop	0.002643	0.005812
		4-hop	0.002836	0.009116
		5-hop	N/A	0.006059
	MMT-NTT (no route salvage)	1-hop	N/A	N/A
		2-hop	0.001000	0.002100
		3-hop	0.002754	0.005468
		4-hop	0.002820	0.007900
		5-hop	N/A	N/A
	DSR	1-hop	N/A	0.000759
		2-hop	0.001719	0.002600
		3-hop	0.004089	0.004930
		4-hop	0.007069	0.010240
		5-hop	0.008872	0.014348

VII. CONCLUSIONS

The aim of this article was to provide insights into architectural and design issues that could affect the design of a wireless sensor-actuator networks for industry control. Towards this we described the physical constraints encountered in a wireless industry environment and proposed a suitable topology and an architecture that would address survivability and security. We then highlighted MAC functions essential to handle data, task and event prioritization, which is vital for wireless industry control. Lastly we identified a secure routing scheme that complements and integrates into the MAC, to provide the requisite connectivity robustness.

The NTT-MAC is contention based but uses a loosely scheduled medium access scheme that does not require strict time synchronization or a central server because it schedules based on neighbor activity. The main performance aspect we targeted when we developed NTT-MAC scheme was to achieve reduced latency, higher success rate and fairness in medium access among contending users. We also introduced a routing protocol based on the MMT algorithm, which is a proactive routing protocol at layer 2 along with the NTT MAC. MMT is developed to support high route robustness with a quick and easy forwarding approach based on virtual IDs. In industry control, Wireless Sensor-Actuator Ad-hoc Network using NTT-MAC algorithm and MMT-routing algorithm will provide high quality of performance. The performance metrics focused were success rate and packet delivery latency. The simulation results show improved performance of MMT-NTT in terms of success rate and end to end latency than DSR operating with 802.11 MAC.

REFERENCES

[1] T. Brooks, "Wireless technology for industrial sensor and control networks," Sensors for Industry, 2001. Proceedings of the First ISA/IEEE Conference, 2001, pp.73-77, doi: 10.1109/SFICON.2001.968502.

[2] J. Song, A. K. Mok, D. Chen, and M. Nixon, "Challenges of wireless control in process industry," in Workshop on Research Directions for Security and Networking in Critical Real-Time and Embedded Systems, San Jose, CA, 2006, http://moss.csc.ncsu.edu/~mueller/ftp/pub/mueller/papers/cps06.pdf. (accessed March 2013)

[3] D. Chen, M. Nixon, T. Aneweer, R. Shepard, and A. K. Mok, "Middleware for wireless process control systems," Workshop on Architectures for Cooperative Embedded Real-Time Systems, 2004, http://wacerts.di.fc.ul.pt/papers/Session1-ChenMok.pdf. (accessed March 2013)

[4] T. Enwall, "Deploying Wireless Sensor Networks for Industrial Automation and Control," http://www.eetimes.com/design/industrial-control/4013661/Deploying-Wireless-Sensor-Networks-for-Industrial-Automation-Control.(accessed March 2013)

[5] I. F. Akyildiz and I. H. Kasimoglu, "Wireless sensor and actor networks: research challenges," Ad Hoc Networks, Volume 2, Issue 4, October 2004, pp. 351-367.

[6] N. Shenoy, Y. Pan, and V. G. Reddy, "Quality of Service in Internet MANETs", Personal, Indoor and Mobile Radio Communications, 2005. PIMRC 2005. IEEE 16th International Symposium on , vol.3, 2005, pp. 1823-1829.

[7] N. Shenoy, Y. Pan, D. Narayan, D. Ross and C. Lutzer, "Route robustness of a multi-meshed tree routing scheme for Internet MANETs," Global Telecommunications Conference, 2005. GLOBECOM '05. IEEE, vol.6, 2005, pp. 3351-3356.

[8] B. P. Gerkey and M. J. Mataric, "A market-based formulation of sensor-actuator network coordination," in Proceedings of

the AAAI Spring Symposium on Intelligent Embedded and Distributed Systems, Palo Alto, California, March 25-27 2002, pp. 21-26.

[9] P. Robert, "Wireless Mesh Networks", http://www.sensorsmag.com/networking-communications/standards-protocols/wireless-mesh-networks-968. (accessed March 2013)

[10] N. Shenoy, C. Xiaojun, Y. Nozaki, S. Hild and P. Chou, "Neighbor Turn Taking MAC - A Loosely Scheduled Access Protocol for Wireless Networks," Personal, Indoor and Mobile Radio Communications, 2007. PIMRC 2007. IEEE 18th International Symposium on, 2007, pp. 1-5.

[11] E. F. Golen, Y. Nozaki and N. Shenoy, "An analytical model for the Neighbor Turn Taking MAC protocol," Military Communications Conference, 2008. MILCOM 2008. IEEE, 2008, pp. 1-7.

[12] L. Barolli, T. Yang, G. Mino, F. Xhafa and A. Durresi., "Routing efficiency in wireless sensor-actor networks considering semi-automated architecture," J. Mob. Multimed, vol.6, 2010, pp. 97-113.

[13] OPNET modeler, http://www.opnet.com/. (accessed March 2013)

An Analysis of Impact of IPv6 on QoS in LTE Networks

Tao Zheng, Daqing Gu
Orange Labs Beijing
France Telecom Group
Beijing, China
e-mail: {tao.zheng; daqing.gu}@orange.com

***Abstract* - As a network evolution goal, IPv6 will be deployed in Long Term Evolution (LTE) networks, including access network, core network, mobile carrier IP network and service-related networks. IPv6 introduction in LTE will affect he Quality of Service (QoS) mechanism. In this paper, the impact of IPv6 on LTE's end-to-end QoS is analyzed in some related aspects and some actions are proposed to improve the QoS according to the analysis results. In addition, some IPv6 transition solutions might affect QoS mechanisms in LTE are analyzed to reduce the negative impacts while IPv6 introduction.**

Keywords-LTE; QoS; TFT; PCC; flow label

I. INTRODUCTION

With the LTE evolution and the rapid development of mobile Internet and multimedia services, QoS, especially Internet Protocol (IP) QoS is becoming more and more important in mobile networks. IPv6 introduction in LTE network can impact on QoS of mobile services.

There are negative and positive impacts of IPv6. Negative impacts include more overhead, tunneling and translation caused by transition solutions, etc.. Positive impacts include the potential application of the IPv6 Flow Label field [1], no NAT required for forwarding IPv6 traffic, etc..

LTE defines a class-based QoS concept, which reduces the implementing complexity while still allowing enough differentiation of traffic handling by mobile operators. Carrier bearers can be classified into two categories based on the QoS they provide: Minimum Guaranteed Bit Rate (GBR) bearers and Non-GBR bearers. Fig. 1 shows the QoS architecture in LTE [2].

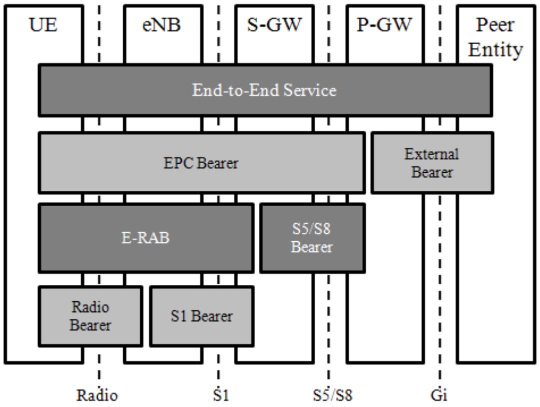


Figure 1. QoS architecture in LTE

The LTE's end-to-end QoS service consists of the external bearer service and the Evolved Packet System (EPS) bearer service. The former is used to carry out services between the LTE core network and external network nodes; and the latter is further subdivided into the EPS Radio Bearer service and the S1 Bearer service (EPS Access Bearer service) and the S5/S8 Bearer Service (Core Network Bearer Service).

The Radio Bearer service is used to transport the EPS bearer service data units between the eNodeB and the User Equipment (UE) according to the requested QoS. Moreover, it supports IP header compression and user plane encryption functions, and provides mapping and multiplexing information for UEs.

The S1 Bearer service implements the transport of EPS bearer service data units between the Serving GateWay (S-GW) and the eNodeB according to the requested QoS, and provides QoS guarantees for end-to-end IP traffic flows, while multiplexing multiple Service Data Flows (SDFs) onto the same EPS bearer.

The S5/S8 Bearer service controls and utilizes the backbone network in order to provide the transport of EPS bearer service date units among EPS Core network nodes.

Similar to the QoS mechanism adopted in Universal Mobile Telecommunications System (UMTS), Traffic Flow Template (TFT) [3] mechanism is used to provide QoS guarantee in LTE. The TFTs contain packet filter information that allows the UE and Public Data Network Gateway (P-GW) to identify the packets belonging to a certain IP packet flow aggregate. Fig. 2[1] describes the TFT architecture in LTE. The UE and the P-GW or Serving Gateway (S-GW) use packet filters to map IP traffic onto the different bearers. The TFTs are typically created when a new EPS bearer is established, and they can be modified during the lifetime of the EPS bearer.

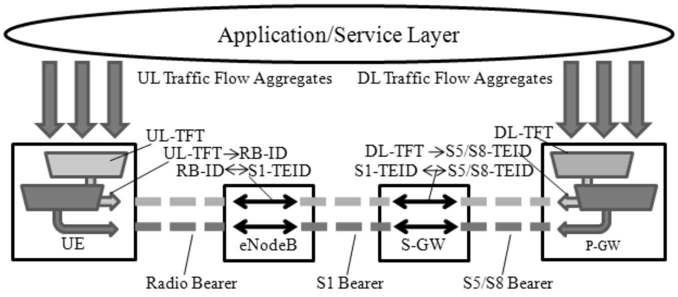


Figure 2. TFT reference network architecture

ISBN: 978-1-61208-256-1

Policy and charging control (PCC) [4] is another QoS related mechanism that provides operators with advanced tools for service-aware QoS and charging control. Different services have very different requirements of QoS, which are needed for the packet transport. PCC enables a centralized control to ensure that the service sessions are provided with the appropriate transport, for example, in terms of bandwidth and QoS treatment. The PCC architecture enables control of the media plane for both the IP Multimedia Subsystem (IMS) and non-IMS services. Furthermore, PCC also provides the means to control charging on a per-service basis. The reference network architecture for PCC in EPS is shown in Fig. 3.

In the LTE IP carrier network, Differentiated Services (Diffserv) mechanisms or the Resource Reservation Protocol (RSVP) protocol still can be used for QoS policy enforcement and resource reservation purposes now. Multiple Protocol Label Switching Virtual Private Network/Traffic Engineering (MPLS VPN/TE) also can be used to provide QoS in IP carrier network.

From the above introduction, we can conclude that the current QoS mechanism used in LTE network is based on bearers (in access and core mobile networks) and Diffserv or RSVP techniques (in IP carrier network). With the deployment of LTE and the explosion of IP-based services, such as Voice over IP (VoIP), Video on demand (VoD), etc., QoS becomes very important. After IPv6 introduction in LTE, some new scenarios are proposed and the QoS mechanisms will face new challenges. It's necessary to analyze the impact of IPv6 on QoS in LTE networks and take some actions.

This paper is organized as follows. In Section 2, we introduce end-to-end QoS Model in LTE networks. In Section 3, impacts and actions of IPv6 on end-to-end QoS Model are provided. In Section 4, impact of IPv6 transition solutions on QoS mechanisms are analyzed. Finally, Section 5 summarizes the conclusions.

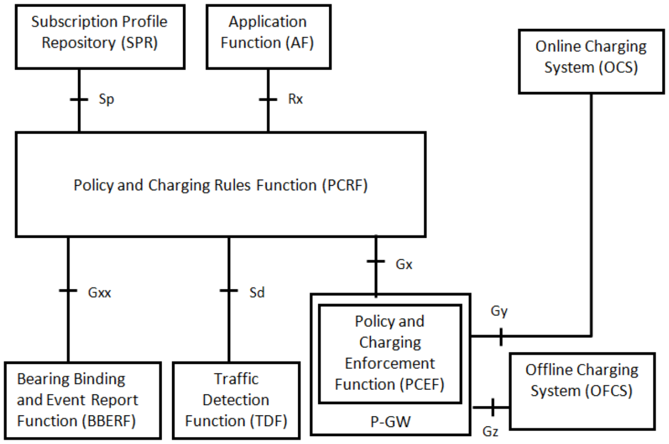

Subscription Profile Repository (SPR)
Application Function (AF)
Online Charging System (OCS)
Sp
Rx
Policy and Charging Rules Function (PCRF)
Gx
Gxx
Sd
Gy
Policy and Charging Enforcement Function (PCEF)
P-GW
Offline Charging System (OFCS)
Gz
Bearing Binding and Event Report Function (BBERF)
Traffic Detection Function (TDF)

Figure 3. PCC reference network architecture

II. End-to-End QoS Model

The critical factors determining QoS in mobile networks are:

- Radio network performance
- Network capacity
- Network design--delay in the system and sufficient capacity available end to end
- Application and service characteristics

The end-user performance is affected by every protocol layer and network element in the connection path, from one UE to other UE or server in the remote end of the network. As shown in Fig. 4 (only the user plane is considered), it is useful to analyze and estimate the end-user experience following a bottom-up approach, starting from the lower levels of the layer architecture and considering a cumulative degradation of the performance based on the effects of the different layers and their interactions. The ideal throughput provided by layer one (physical layer) is considered initially as the starting point, and then the performance degradation introduced by each of the upper layers in the protocol stack is estimated.

Compared with IPv4, IPv6 has some differences that may affect the QoS, such as:

- Packet header--the nominal IPv6 header is twice the size of the IPv4 header.
- Additional system messages, such as Neighbor Advertisement, Neighbor Solicitation, Router Solicitation, Router Advertisement, and Redirect.
- Network Address Translator (NAT)--Due to lack of public IPv4 addresses, NATs are deployed in some IPv4 core network; IPv6 eliminates the primary need for NATs, but translation or tunnel will be needed in a hybrid IPv4/IPv6 network without full dual-stack deployment.

Next section will discuss the impact of IPv6 on in the end-to-end QoS model showed in Fig. 4.

III. Impact on End-to-End QoS Model

In this section, we analyzed the impact of IPv6 on end-to-end QoS following the path from an UE to another UE or a server according to the models showed in Fig. 1 and Fig. 4.

A. UE Local Bearer Service

UE local bearer service is determined by the signaling and data processes embedded in the UE. So the processing capacity of the UE will affect the QoS.

In IPv6, there are some additional system messages and data processing compared to IPv4. For example, the interface identifier (IID) part of the IPv6 address assigned to the cellular interface of the UE may be changed periodically and randomly[5] hence making it more difficult to identify the terminal, please see next section “PDCP Layer” for details. Changing the IID of an IPv6 address randomly requires more processing capacity and resources in the UE.

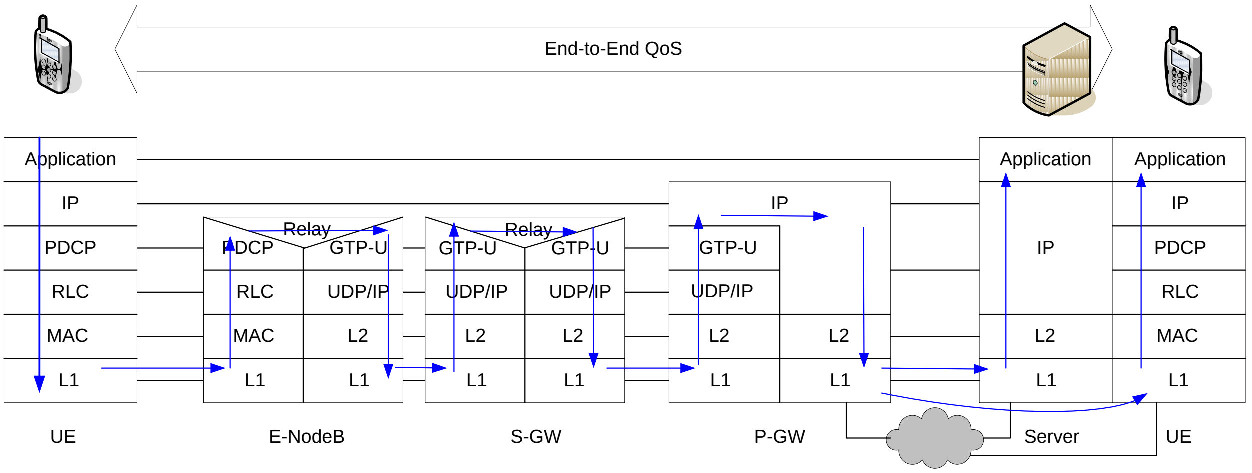


Figure 4. Layers affecting end-to-end QoS

During the transition period, the Packet Data Protocol (PDP) bearer will be dual-stack (one IPv4v6 bearer or two individual IPv4 and IPv6 bearers) or IPv6-only. If the UE sets up dual-stack bearers, more resources are required than for a single-stack bearer.

B. Radio Bearer Service

In the radio bearer service, the LTE Layer 2 user-plane protocol stack is composed of three layers as shown in Fig. 4, Packet Data Convergence Protocol (PDCP) layer, Radio Link Control (RLC) layer and Medium Access Control (MAC) layer.

In these layers, an important design feature is that all the Protocol Data Units (PDUs) and Service Data Unit (SDUs) are byte-aligned which means that the lengths of the PDUs and SDUs are multiples of 8 bits. This is to facilitate handling by microprocessors, which are normally defined to handle packets in units of bytes. This implies that sometimes unused padding bits are needed, and thus the cost of design for efficient processing is that a small amount of potentially-available capacity is wasted.

The size of an IPv6 packet nominal header is twice the size of an IPv4 packet header. So the padding is likely to be different and then the potentially-available capacity is different too. There may be negative and positive impacts on the QoS of the radio bearer service when IPv6 is deployed.

1) PDCP Layer

One of the main functions of PDCP is header compression using the RObust Header Compression (ROHC) [6] protocol defined by the IETF. In LTE, header compression is very important because there is no support for the transport of voice services via the Circuit-Switched (CS) domain. The different packet header between IPv4 and IPv6 will impact the overhead of packets, especially for small packet services, such as VoIP.

The main principle of header compression is to avoid sending fields which do not change between consecutive packets. ROHC is able to reduce a Real Time Protocol (RTP)/ User Datagram Protocol (UDP)/IPv6 header from 60 bytes to 3 or 4 bytes. This means that the 12.2 kbps AMR speech bit rate would be increased to 14.6 kbps with IPv6.

During the IPv6 address allocation process, The P-GW allocates a globally unique /64 IPv6 prefix via Router Advertisement to a given UE [5]. After the UE has received the Router Advertisement message, it constructs a full IPv6 address via IPv6 Stateless Address autoconfiguration. To ensure that the link-local address generated by the UE does not collide with the link-local address of the PDN GW, the PDN GW shall provide an IID to the UE and the UE shall use this to configure its link-local address. For stateless address autoconfiguration however, the UE can choose any interface identifier to generate IPv6 addresses, other than link-local, without involving the network.

If the same IID is used in multiple contexts, it becomes possible for that the identifier to be used to correlate seemingly unrelated activity. For example, a network sniffer placed strategically on a link across which all traffic to/from a particular host crosses could keep track of which destinations a node communicated with and at what times.

Due to privacy consideration, the UE may change its IID which is part of IPv6 address periodically and randomly. That means the IPv6 address of UE will be changed periodically and should be sent to network when ROHC applied. Then the relative overhead in ROHC will be higher when the IID changed. So for VoIP or Machine to Machine (M2M) services with small Maximum Transmission Units (MTUs) and belonging to closed services whose security can be guaranteed, a non-changing IID is suggested to be used for the sake of lower overhead.

2) RLC Layer

The main functions of the RLC layer are segmentation and reassembly of upper layer packets in order to adapt them to the size which can actually be transmitted over the radio interface. For radio bearers which need error-free transmission, the RLC layer also performs retransmission to recover from packet losses.

The functions of the RLC layer are performed by "RLC entities". An RLC entity is configured in one of three data transmission modes: Transparent Mode (TM), Unacknowledged Mode (UM) and Acknowledged Mode (AM).

In AM, special functions are defined to support retransmission. When UM or AM is used, the choice between the two modes is made by the eNodeB during the RRC radio bearer setup procedure, based on the QoS requirements of the EPS bearer.

Compared with IPv4, IPv6 have some Internet Control Message Protocol version 6 (ICMPv6) messages, such as Neighbor Discovery, Auto-configuration, Multicast Listener Discovery, and Path MTU Discovery. These messages should be transmitted on a separate Radio Access Bearer using AM RLC mode for error-free transmission.

3) MAC Layer

This layer performs multiplexing of data from different radio bearers. Therefore there is only one MAC entity per UE. By deciding the amount of data that can be transmitted from each radio bearer and instructing the RLC layer as to the size of packets to provide, the MAC layer aims to achieve the negotiated QoS for each radio bearer.

In MAC layer, there are no difference between transmitting IPv4 packets and IPv6 packets.

C. S1 Bearer Service

According to LTE layer architecture defined by 3rd Generation Partnership Project (3GPP), the Radio Access Network (RAN) access bearer service between the RAN and the core network in user plane is transmitted by IP protocols. The LTE radio network is connected to the Evolved Packet Core (EPC) through the S1 interface. The S1 interface is divided into two parts:

- S1-MME carries signaling messages between the base station and the Mobility Management Entity (MME).
- S1-U carries user data between the base station and the S-GW.

There are two IP headers separated by GTP Tunnel Header and which correspond to the transport layer and the end-user IP packet respectively. If the transport layer is still IPv4 and the EPS bearer is IPv6, the overhead of IP header in S1-U is 1.5 times the overhead of the IPv4 header. If the transport layer and EPS bearer are both IPv6, the overhead of the IPv6 header in S1-U is twice the overhead of the IPv4 header. The available bandwidth at the S1-U interface will consequently be affected when IPv6 is introduced in the RAN access bearer.

Due to the greater size of the IPv6 packet header, the bandwidth between the RAN and the Core network (S1-U interface) should be dimensioned accordingly so as to limit the risk of congestion occurrences.

D. Core Network

For the General Packet Radio Service (GPRS) architecture, until Release 9, two PDP types are defined, namely IPv4 and IPv6. This means that an UE can only request one type of IP address per PDP context. With the introduction of the EPC system in Release 8, a new PDP type has been introduced called IPv4v6, and which enables the UEs to use a single Dual-stack bearer. This PDP type is available starting from Rel-9 for GPRS.

During the transition period, if the UE is dual-stack enabled, two individual IPv4 and IPv6 bearers (case where the UE doesn't support the IPv4v6 PDP type) or one IPv4v6 bearer may be set up in the S-GW and P-GW. For the S-GW and P-GW, the maximum number of simultaneous bearers that can be supported at any given time is a performance criterion. If some UEs set up two individual IPv4 and IPv6 bearers, the capacity of the S-GW and P-GW will be affected.

E. Backbone Bearer Service

The architecture of EPS bearer transport for GTP Tunnels in the backbone bearer is similar to that of S1 bearer. Obviously the problems are the same as in the S1 bearer when IPv6 is introduced.

Due to the greater size of the IPv6 packet header, the bandwidth of the core network should be dimensioned accordingly so as to limit the risk of congestion occurrences.

F. External Bearer Service

The external bearer service is related with external networks. There maybe several ways to transmit IPv6 traffic in external networks:

- If the external network is IPv6-capable or dual stack, no impacts.
- If the external network remains IPv4 and is MPLS-capable, the ideal solution is 6PE (IPv6 over MPLS)/6VPE (IPv6 over MPLS VPN). The 6PE approach allows existing IP/MPLS networks to carry the IPv6 packets using MPLS labels; hence only the PE devices need to support the IPv6 protocol and addressing. In 6PE/6VPE, the QoS of IPv6 traffic is guaranteed by MPLS.
- The IPv6 traffic can be transmitted in tunnels, if the external network is IPv4-only. Tunneling allows for the transport of the encapsulated data unit across the encapsulating protocol’s transport network. Typically, when employed as part of an IPv6 transition mechanism, the existing IPv4 transport infrastructure is used to encapsulate IPv6 packets, thereby using the existing IPv4 infrastructure to provide basic IPv6 connectivity.

The overhead of the encapsulation scheme should be considered to dimension the bandwidth used by external bearers. And the time to establish and activate the tunnels will influence the delay related to the forwarding of IPv6 traffic.

If the opposite end-points is of a different IP version or addresses need to be translated due to the deployment of a transition solution, a translation node is needed somewhere in the forwarding path. Address translation (usually with Application-Level Gateway (ALG)) is harmful to QoS, but is necessary for some transition solutions.

G. Impacts and Actions

Through the above analysis about impact of IPv6 integration on QoS, some parts in mobile network can be affected after IPv6 deployed. Fig. 5 shows the impacts on QoS for different parts of mobile networks.

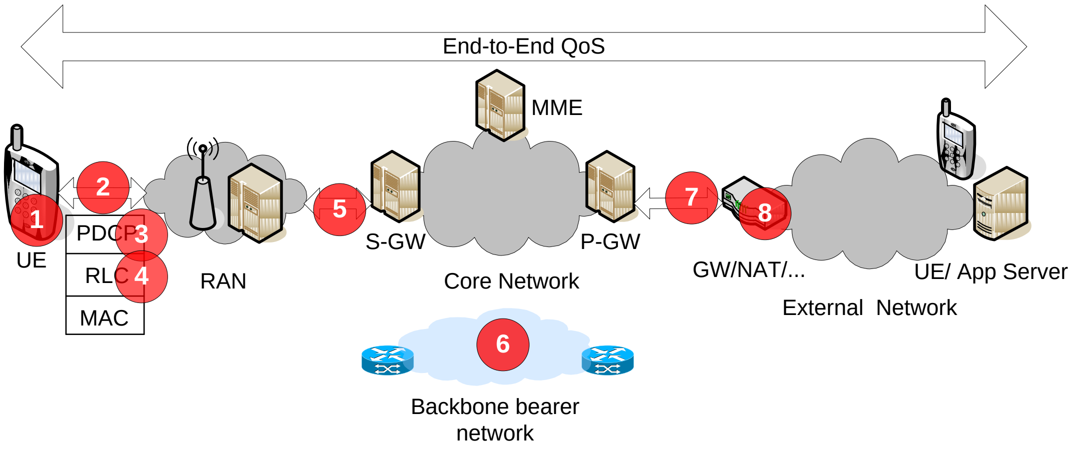


Figure 5. Localization of the impacts of IPv6 integration on QoS in LTE

Impact #1: Compared with IPv4, IPv6 and dual-stack need more processing capacity and resources in the UE.

Impact #2: When IPv6 applied to PDUs and SDUs of L2 layers, the different unused padding bits will affect the potentially-available capacity and the QoS of radio bearer services when IPv6 is deployed.

Impact #3: The overhead in ROHC will be higher when the UE's IID changes. So for VoIP or M2M services with small MTUs and belonging to closed services whose security can be guaranteed, a non-changing IID is suggested to be used for the sake of lower overhead.

Impact #4: Some ICMPv6 messages should be transmitted on a separate RAB using AM RLC mode for error-free transmission.

Impact #5: Due to longer length of packet header and additional system messages, the bandwidth between the RAN and the Core should be augmented to avoid congestion when IPv6 is introduced.

Impact #6: Due to longer length of packet header and additional system messages, the bandwidth of the core network should be augmented to avoid congestion when IPv6 is introduced.

Impact #7: The overhead of tunnels should be considered in the bandwidth of external bearers. The processing time of start and end points of tunnels will also influence the delay of IPv6 traffic.

Impact #8: Address translation (usually with ALG) is harmful to QoS, but is a necessity in some transition solutions.

The following table shows the QoS characteristics effected by above 8 impacts.

TABLE I. QoS Characteristics Affected by Impacts from IPv6 Introduction

Impacts	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Throughput/Bandwidth	√	√	√		√	√	√	
Packet Loss Rate				√				
Delay(Latency)	√	√	√				√	√

By analyzing the impacts listed above, four actions are proposed for better QoS:

Action #1: Improve the capacity and processing power of devices. Apply to Impacts #1;

Action #2: Optimize related parameters and increase resource utilization rate. Apply to Impacts #2;

Action #3: Increase bandwidth of networks. Apply to Impacts #2, #5, #6, and #7;

Action #4: Minimize the number of translating or tunneling on end-to-end path. Apply to Impacts #7, #8.

IV. Impact of Transition Solutions and Flow Label

Several transition solutions have been proposed to solve IPv4 public address exhaustion and to introduce IPv6 in mobile networks. These solutions employ translation and tunneling techniques and introduce some functional elements to network, which impacts on the QoS mechanisms, such as TFT, PCC.

A. Impact on TFT

When IPv6 is introduced in mobile network, the direct impact on TFT mechanism is that the filter information may contain the IPv6 attributes: IPv6 Next Header, Traffic class and Flow Label. That means that TFT should support IPv6 filter attribute combinations. For example, the new filter combination: Remote address and subnet mask, IPv6 traffic class and flow label. Compared with the other two combinations, this one gives us a fine-grained filter to tell different QoS flows by using IPv6 Flow Label instead of port range or IPSec Security Parameter Index (SPI) attributes.

Because the TFT mechanism works between UE and P-GW, the TFT mechanism will be impacted when a translation or a tunnel function is required in the UE for a transition solution. For tunnel-based solutions, the TFT filter attributes will be taken from the tunnel IP header instead of the internal IP header. So there should have some mechanisms to map the internal filter attributes to external filter attributes of tunnel, e.g., by copying the IPv6 Flow Label. For translation-based solutions, the filter attributes for the original IP packet should be translated to the new IP packet to ensure integrity of filters.

Transition solutions can be cataloged two classes:

- Impact on TFT: Bump-in-the-Host (BIH) [7] (translation in UE), Address Plus Port (A+P) [8] and Dual-Stack Lite (DS-Lite) [9] (tunnel between UE and Carrier Grade NAT (CGN), except for GTP encapsulation)
- No impact: Gateway-Initiated Dual-Stack Lite (GI-DS-Lite) [10], NAT64/DNS64 [11][12], Per-Interface NAT [13], Overlapping clusters, v4 on demand

B. Impact on PCC

The PCC rules should support IPv6 IP Connectivity Access Network (CAN) sessions and service data flow filters. When the PCC architecture is used and a Policy and Charging Rules Function (PCRF) function is added, this function and the corresponding Gx interface must also support IPv6-enabled PCC rules. The Rx interface and the Application function should take into account IPv6 flows and addresses.

There are two aspects PCC mechanisms impacted by transition solutions:

Some transition solutions, such as GI-DS-Lite, overlapping clusters, Per-Interface NAT and NAT64, etc., introduce a translation function (NAT/CGN) between the P-GW and the service platforms if UEs access IPv4 services. In this case, an impact on Rx interface between the PCRF and the Application Function (AF) in the PCC architecture will need to be addressed, because the UE's address information exchanged on Rx interface stays the payload of packets. So the NAT/CGN should include an ALG which understand the protocol on the Rx Interface.

Another problem is that multiple customers share the same public IPv4 address (for most of transition solutions) or even the same private IPv4 address (for DS-Lite, GI-DS-Lite and Per-Interface NAT). In the PCC mechanism, the PCC rules including QoS parameters will be bound with IP CAN sessions. Shared IPv4 address blocks the binding process between an AF session and an IP Can session. Other identifications, such as IPv6 tunnel address or tunnel ID, are required to help to find the right bearer.

C. Impact of Flow Label

IPv6 has introduced a field named Flow Label to identify and mark a flow. General rules for the Flow Label field have been documented in [14], but how to apply this field in real-world network is still an open issue.

Under the current LTE architecture, the traffic is encapsulated in GTP tunnel and transported based upon the EPC bearer. The IPv6 Flow Label can be used to replace bearer identification to improve the current QoS mechanisms in mobile networks. In current mobile networks, the QoS granularity is based upon the bearer characteristics as defined by the PCC and TFT QoS mechanisms. The fine granularity of QoS can be implemented with the help of Flow Label [1].

In addition, the IPv6 Flow Label can be used for QoS provision in the IP backbone network carrying mobile networks. At present, the IP network can use Diffserv mechanisms or the RSVP protocol for QoS policy enforcement and resource reservation purposes. Such QoS policy and RSVP design is usually engineered in advance (e.g. configuring the actions to be performed by a router that supports the Expedited Forwarding Per-Hop Behavior). If the IP backbone network is IPv6-capable, it can perceive the bearers in mobile networks by a mapping mechanism between the bearer identification and the IPv6 Flow Label, and then provide QoS guarantee by combining Flow Label with other technology such as Diffserv. This QoS policy can be adjusted by changing the mapping between the bearer ID and the IPv6 Flow Label according to the mobile services type and traffic. When congestion is occurring, it is possible to drop the packets in same bearers for minimizing the affected bearers with the help of Flow Label.

V. Conclusion and Future Work

In this paper, we provided an in-depth analysis of the end-to-end QoS impacts by IPv6 introduction in LTE networks and provide recommendations about the QoS impacts and improvement by the IPv6 introduction in mobile networks. Negative impacts should be reduced for the sake of QoS improvement by adopting appropriate actions as mentioned in section III. Positive impacts should be taken into account to improve QoS. In the future work, we consider researching on impacts of Mobile IPv6 in mobile networks.

References

[1] T. Zheng, L. Wang, and D. Gu, A flow label based QoS scheme for end-to-end mobile services, Proc. The Eighth International Conference on Networking and Services (ICNS 2012), March 2012, pp. 169-174.

[2] D. Flore, LTE RAN architecture aspects, 3GPP IMT-Advanced Evaluation, December 2009.

[3] 3GPP TS 23.401, General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access, September 2012.

[4] 3GPP TS 23.203, Policy and charging control architecture, September 2012.

[5] T. Narten, R. Draves, and S. Krishnan, Privacy Extensions for Stateless Address Autoconfiguration in IPv6, IETF RFC 4941, September 2007.

[6] K. Sandlund and G. Pelletier The RObust Header Compression (ROHC) framework, IETF RFC 5795, March 2010.

[7] B. Huang, H. Deng, and T. Savolainen, Dual-Stack hosts using "Bump-in-the-Host" (BIH), IETF RFC 6535, February 2012.

[8] R. Bush, The Address plus Port (A+P) approach to the IPv4 address shortage, IETF RFC 6346, August 2011.

[9] A. Durand, R. Droms, J. Woodyatt, and Y. Lee, Dual-Stack Lite broadband deployments following IPv4 exhaustion, IETF RFC 6333, August 2011.

[10] F. Brockners, S. Gundavelli, S. Speicher, and D. Ward, Gateway-Initiated Dual-Stack Lite deployment, IETF RFC 6674, July 2012.

[11] X. Li, C. Bao, and F. Baker, IP/ICMP translation algorithm, IETF RFC 6145, April 2011.

[12] M. Bagnulo, A. Sullivan, P. Matthews, and I. van Beijnum, DNS64: DNS extensions for network address translation from IPv6 clients to IPv4 servers, IETF RFC 6147, April 2011.

[13] J. Arkko, L. Eggert, and M. Townsley, Scalable operation of address translators with Per-Interface bindings, IETF RFC 6619, June 2012.

[14] J. Rajahalme, A. Conta, B. Carpenter, and S. Deering, IPv6 flow label specification, IETF RFC 3697, March 2004.

Smart Network Selection and Packet Loss Improvement during Handover in Heterogeneous Environment

Ahmad Rahil, Nader Mbarek, Olivier Togni
Laboratoire d'Electronique, Informatique et Image, UMR 6306
University of Burgundy
Dijon, France
ahmad.rahil | nader.mbarek | olivier.togni@u-bourgogne.fr

***Abstract*—Seamless Handover between networks in heterogeneous environment is essential to guarantee end-to-end QoS for mobile users. A key requirement is the ability to select the next best network. Currently, the implementation of the IEEE 802.21 standard by National Institute of Standards and Technology (NIST) considers only the signal strength as a parameter to determine the best network. In this paper, we propose including additional parameters such as available bandwidth, mobile node speed and type of network during selecting a new network to improve the QoS for mobile user's application. The results of the experiments that we performed using Network Simulator show that there is a need for a new framework taking into account these parameters to guide network selection process during handover and to provide mobile users with QoS guarantee.**

Keywords-Seamless vertical handover; QoS parameters; IEEE 802.21 MIH

I. INTRODUCTION

Communicating from anywhere at any time is becoming a requirement of great importance for mobile users. However, the rapid expansion of wireless network technologies creates a heterogeneous environment. Nowadays, mobile users would like to acquire, directly from their device, different kinds of services like Internet, audio and video conferencing which sometimes require switching between different operators. Moreover, user preferences differ, some are interested in service costs only; others will be satisfied with broadband networks that cover large geographic areas, etc. Consequently, to satisfy the above requirements, user mobility should be covered by a set of different overlapping networks forming a heterogeneous environment. Mobile device should be able to choose, from all available networks in its environment, the one that meets its needs and ensures accordingly the transition from one cell to another in the same technology (horizontal handover) or between different types of technologies (vertical handover). During this period of handover the challenge is to conserve the QoS parameters guarantee.

The remainder of this paper is organized as follows: Section II describes the background. Section III describes the main components of IEEE 802.21 standard and its implementation using NS2 simulator. Section IV provides an overview of wireless protocols used in our simulation environment. Section V describes the simulation scenarios and results and we conclude in Section VI.

II. BACKGROUND

The IEEE 802.21 [1] is an emerging standard, also known as Media Independent Handover (MIH) that supports management of seamless handover between different networks in a heterogeneous environment. The current implementation of the IEEE 802.21 standard for the network simulator NS by National Institute of Standards and Technology (NIST) based on draft 3 [2][3] considers only the signal strength as a unique parameter to determine the best network [4]. We argue in this paper that this parameter alone is not sufficient to satisfy user requirement. Indeed, signal strength, available bandwidth, traffic on the serving network and packet loss ratio are among the other parameters that affect the requirement of mobile user in terms of QoS guarantee. For example, a bad QoS, when using a real time application in a handover process, may be due to a lack of available bandwidth because of high load in the visited network while the signal strength is good.

Several attempts have been made to improve the handover within the MIH framework. Chandavarkar et al. [5] proposed an algorithm for network selection based on the energy of the battery, the speed of the mobile, and the coverage radius of the network in order to avoid power loss during handover and to improve the efficiency of seamless handover. Siddiqui et al. [6] proposed a new algorithm named TAILOR that uses different parameters of QoS with the user preferences to select the destination network. Also this algorithm optimizes the power consumption.

Jiadi et al. [7][8], modified the Media Independent Handover component (MIH) where handover is performed in three steps: initiation, selection and execution. The proposed process aims to improve the handover delay by adding new events to the initiation step that can be generated from the application layer instead of lower layer upon the user's satisfaction. Moreover they added a new algorithm at the selection step based on price, delay, Jitter, Signal Noise Ratio (SNR) and available data rate within the MADM (Multi Attribute Decision Making) function to improve the QoS during selection process.

The research work initiated in [9][10] proposed a selection algorithm based on the willingness of users to pay for a given service, while Cicconetti et al. [11] provided an algorithm based on three parameters: connectivity graph, connectivity table between nodes and the current geographical position of the serving network. The proposed algorithm reduces the handover time and the energy consumption of mobile node due to scanning.

The MIIS component (see Section III) of MIH is not fully implemented by NIST. Arraezet al. [12] implement this service and install it on each access point allowing user to save the energy of the battery by just activating a single interface. According to the IEEE 802.21 standard, an MIH user communicates, through the link layer, with its MIHF which sends a query to MIIS to retrieve the list of all networks in the vicinity. Alternatively the authors of [13][14] developed a new method to communicate with the MIIS through the upper layers using Web Services.

Moreover, 802.11 protocols defined 11 channels for communication and force the MN during the handover to scan all channels looking for the active one. Khan et al. [15], proposed a new algorithm based on the Media Independent Information Server (MIIS), to provide user with a list of only active channel to be scanned in order to save time during handover.

An et al. [16] added two new parameters to MIH that allow FMIPV6 to save the steps of proxy router solicitation and advertisement (RtSolPr/PrRtAdv). This resulted in a decrease of handover latency and improvement of packet loss ratio.

In this paper, we investigate the effect of the inclusion of three parameters with the signal strength into the destination network selection mechanism during handover. These parameters are: Available Bandwidth, type of network and mobile speed. As far as we know, these parameters have not been investigated at the same time before. As it will be detailed in Section V, our first experiment will show that by including the available bandwidth parameter (ABW) the packet loss will be improved. The second experiment will show that upon the type (WI-FI, WIMAX) of the current and destination network we can save on packet loss. The third experiment will show that it is worthily significant to consider the velocity of MN while selecting new network during handover.

III. IEEE 802.21 STANDARD

User mobility can be achieved at different levels of the protocol stack. The IEEE802.21 standard, also known as Media Independent Handover (MIH), provide mobility management at layer 2.5, by being inserted between layer 2 and layer 3. As depicted in Fig. 1, the Media Independent Handover Function (MIHF) is the main entity of the standard that allows communication in both directions between lower and upper layers through three services: event (MIES), command (MICS) and information (MIES) [3][17]**.**

A. Media Independent Event Services, MIES

This service detects changes in the lower layers (physical and link) to determine if it needs to perform handover. Two types of events can occur: "MIH Event" sent by the MIHF to the upper layers (3 +), and "Link Event" that spreads from the lower layers to the MIHF.

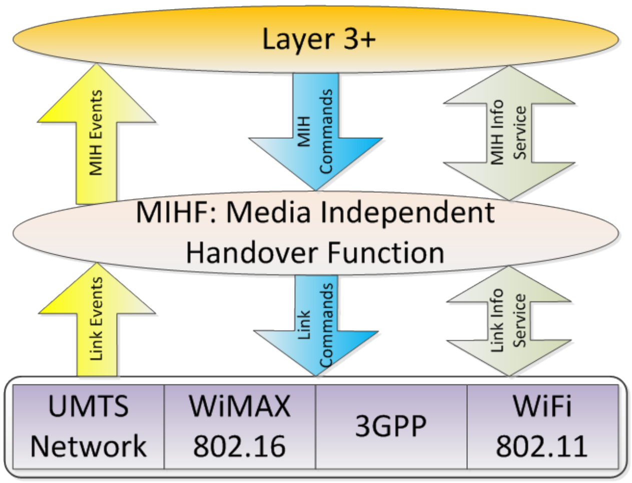


Figure 1. MIH architecture.

B. Media Independent Command Services, MICS

This service uses two types of events. The "MIH Commands" transmitted by the user towards the MIHF and "Link Commands" sent by MIHF to lower layers.

C. Media Independent Information Services, MIIS

The MIIS let the mobile user discover and collect information about features and services offered by neighboring networks such as network type, operator ID, network ID, cost, and network QoS, etc. This information helps doing a more efficient handover decision across heterogeneous networks.

IV. WI-FI AND WIMAX STANDARDS

A. IEEE 802.11, WIFI

IEEE 802.11, Wireless Fidelity (WI-FI) [18], is a wireless local network technology designed for a private LAN with a small coverage area (hundreds of meters). Different versions of 802.11 exist and communicate on different frequency bands with a different bit rate. In all simulations that we performed in this paper we use 802.11b. Mobility support in conventional IEEE 802.11 standard is not a prior consideration and horizontal handover procedure does not meet the needs of real time traffic [19]. WI-FI's QoS is limited in supporting multimedia or Voice over Internet Protocol (VoIP) traffic and several research activities have been carried out in an attempt to overcome this short fall[20].

B. IEEE 802.16, WIMAX

IEEE 802.16, WIMAX (Worldwide Interoperability for Microwave Access), technology is for metropolitan area network (MAN) covering a wide area at very high speed. QoS in WI-FI is relative to packet flow and similar to fixed Ethernet while WIMAX define a packet classification and scheduling mechanism with four classes to guarantee QoS for each flow: Unsolicited Grant Service (UGS), Real-Time Polling Service (RTPS), non-real-Time Polling Services (nrtPS) and Best Effort (BE). WIMAX mobile (802.16e) adds a fifth one called extended real-time Polling System (ertPS) [21]. WIMAX supports three handover methods: Hard Handover (HHO), Fast Base Station Switching (FBSS) and Macro-Diversity Handover (MDHO). The HO process [22] is composed of several phases: network topology advertisement, MS scanning, cell reselection, HO decision and initiation and network re-entry [23][24].

V. MIH PERFORMANCE EVALUATION

In this section, we will present three scenarios to evaluate the impact of the available bandwidth, type of network, and user velocity on selecting a destination network during handover.

A. *Simulation Environment*

To show the limits of using one parameter to select an access network and to motivate the need of advanced selection methods that combine several constraints, we present in this section several simulation scenarios using NS2, v2.29, which support the Media Independent Handover (MIH) module implemented by National Institute of Standards and Technology (NIST).

The studied scenarios focus on the importance of some criteria other than radio signal strength while evaluating network in the vicinity for handover. The First scenario studies the impact of the selected network available bandwidth. The second one tryout the type of destination network, and the third scenario experiments the speed effect of the MN on QoS during handover.

Simulation parameters are shown in Table I. Traffic used is a CBR (Constant Bit Rate), packet size is always constant to 1500 bytes and the throughput is determined by varying the interval of sending packet during simulation.

B. *Scenario I : NIST Selection Weakness*

1) Topology Description: Topology of this scenario, shown in Fig. 2, consists of two WLAN Access Points AP1 and AP2 (802.11b) located inside an 802.16 base station (BS) coverage area and one Mobile Node (MN) equipped with multiples interfaces. It is important to note that other stream of traffic source is connected to AP2 consuming its bandwidth. At the beginning, MN connected to AP1, starts moving to the center of the BS and on its way detect AP2. According to the NIST handover algorithm, that selects a new network based on the Radio Signal Strength only, AP2 is considered a better network than WIMAX and the MN will make a handover from AP1 to AP2. Once the MN reaches the limit coverage area of AP2, the handover to WIMAX base station occurs.

2) Scenario I Results: By increasing the throughput generated by the CBR application on the MN, we observe a greater number of packet loss overall scenarios. Fig. 3 shows the packet loss during HO. When a MN loses the signal on AP1 it needs to make a HO to another network, it has 2 choices: handover to AP2 or to WIMAX. According to NIST algorithm, which selects a new network based on the signal strength only, AP2 is selected and Fig. 3 shows the number of packet loss during handover AP1-AP2. When the MN reaches the limit coverage area of AP2, it makes the handover to WIMAX and we observe another amount of PL during HO AP2-WIMAX. 3) *Critics of the NIST algorithm*: select a destination network based on the signal strength received by the mobile node still unsatisfactory. Indeed, a mobile node, near to an overloaded base station, receives a strong signal. According to NIST algorithm, the MN handover to this base station and meet a high packet loss rate due to a lack of available bandwidth.

TABLE I. SIMULATION PARAMETERS

WI-FI Access Point AP1 and AP2 Parameters	
Transmission Power (Pt_)	0.027 W
Receiving Threshold (RXThresh)	1.17557e-10 W
Carrier Sending Threshold (CXTresh)	1.058.13 e-10 W
Coverage Radius	150 meters
Radio Propagation Model	Two-RayGround
Frequency (Freq)	2.4 GHz
Sensitivity to link degradation (lgd_factor_)	1.2
WIMAX Parameters	
Transmission Power (Pt_)	30 W
Receiving Threshold (RXThresh)	3e-11 W
Carrier Sending Threshold (CXTresh)	2.4 e-11 W
Coverage Radius	1500 meters
Radio Propagation Model	Two-RayGround
Frequency (Freq)	3.5 GHz
Sensitivity to link degradation (lgd_factor_)	1.2
Antenna Type	Omni Antenna
Modulation	OFDM

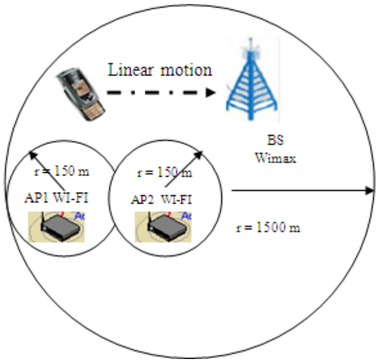


Figure 2. Scenario I topology.

C. *Multi Criteria Selection Algorithm*

In this section, we propose a new selection algorithm named Multi Criteria Selection Algorithm (MCSA) which is a modified version of the algorithm proposed by NIST to select a destination network based on two criteria: Radio Signal Strength (RSS) and available bandwidth (ABW) of destination network. We assume that the user preference consists of selecting a network with the largest available bandwith whatever the cost is. Then, we compare the number of packet loss during HO between MCSA and NIST algorithm.

1) Strategy of MCSA: A Mobile Node (MN) that is connected to a serving network receives beacons and router advertisement (RA) from Wi-Fi and WIMAX network in the vicinity. According to our proposed algorithm, MN will select the network that has the biggest available bandwidth (ABW). In order to get the value of ABW to the mobile node, we needed to change the structure of the beacons and router advertisement in NS2 by adding a new field that holds the value of ABW.

2) *MCSA results*: in order to compare MCSA and NIST results, we use the same topology of simulation cited in Fig. 2. By using our proposed MCSA algorithm, which aims to find among the visible list of networks, the one that have the largest available bandwidth (ABW), WIMAX is selected instead of AP2 and the total number of handovers decreases improving the total number of PL and the Quality of Service is preserved during the mobility of the MN

For a user who gives importance for the number of Packet Loss rather than type of network (WIFI or WIMAX), it is better to follow the strategy of our proposed MCSA algorithm that improves the packet loss ratio by 33%. Table II shows the improvement in number of HO and PL with MSCA for a given throughput.

We can conclude that selecting a destination network using only RSS as indicator does not meet the needs of all users. More accurate choice of destination network during handover would consider the ABW of the destination network. A new framework is needed to consider the values of different criteria to take a decision and make a better choice concerning the destination network during handover.

In order to better understand the sequence of events that a MN and Network perform during successful HO, we provide a short description of messages sequence chart in Fig. 4. The dashed and non-dashed bloc represents the flow of handover messages according to NIST and MCSA algorithm. By using our MCSA algorithm, we can save all messages in the dashed bloc which enables less signaling over the network and improvement in number of packet loss for a better QoS guarantee provided to a mobile user.

A detailed description of the events sequence according to the implementation of the IEEE802.21 standard by NIST is as follow:

1) MIH user on the MN sends MIH Capability Discovery Request to discover link capability supported (events and commands) for each mac of each node.

2) MIH user on the MN sends MIH Register Request to register to the local and remote MIHF.

3) MIH User on the MN sends MIH Get Status requesting the available network interface; it discovers the presence of 2 interfaces (WIFI and WIMAX) both interfaces support events and commands services of MIHF.

4) MIH user on the MN sends MIH Event Subscribe request to subscribe to the events on the given links for local and remote MIHF. This latter send MIH Event Subscribe response to the MIH User of the mobile node

5) Since the BS decides of the reservation of bandwidth, it informs the MN of the frame structure in the uplink and downlink. It sends the DL-MAP/UL-MAP to the WIMAX interface of the mobile node MN. The WIMAX base station is detected and generates a link up event toward the MIHF of MN.

6) MIHF of the MN order the WIMAX interface of MN to connect to the BS.

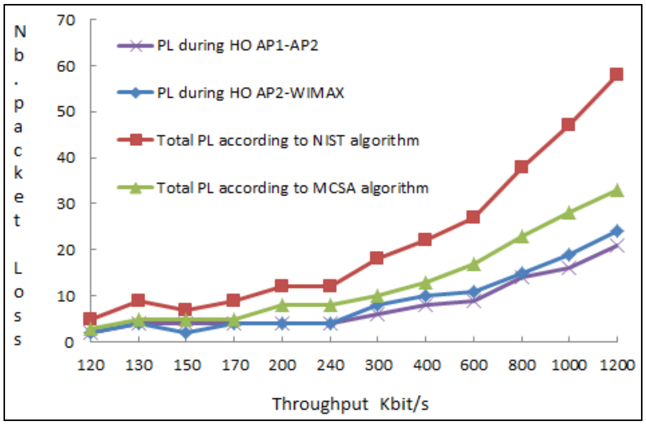


Figure 3. Packet loss according to NIST and MCSA algorithm.

TABLE II. COMPARISON OF HO NUMBER AND PL WITH EACH ALGORITHM

According to NIST algorithm		According to MCSA algorithm	
Number of HO	***Total PL***	***Number of HO***	***Total PL loss***
2 (AP1 to AP2 andAP2 to WIMAX)	20 AP1 to AP2:9 and AP2 to WIMAX:11	1 (AP1 to WIMAX)	10 AP1 to WIMAX:10

7) In this case, a router solicitation is sent form the MIPV6 module of MN to the neighbor discovery module of the BS.

8) Neighbor discovery module of BS reply by sending a router advertisement (RA) to the MIPV6 module of MN with the network prefix of WIMAX base station = 3.0.0; router-life time = 1800s and advertisement interval = 10s.

9) MN’s WIFI interface receive a beacon message with a power above the threshold value and trigger a link Detect event; the available bandwidth of AP1 is largely available (not consumed by any other traffic), according to the both algorithm MCSA and NIST, AP1 is considered as a better network.

10) MIHF of MN sends a link connect message to the WiFi interface of MN; exchange of association Request/Response between MN and AP1.

11) The WIFI interface of the MN send a link up message to the MIHF and MIH user of MN.

12) Exchange of router advertisement and router solicitation between the MIPV6 of MN and the neighbor discovery module of AP1 (first WIFI access point).

13) Starting of traffic flow between the WIFI interface of the MN and the correspondent node through the AP1 access point.

14) Once MN reaches the limit coverage of AP1, it starts receiving the beacon message coming from AP2. Detect the presence of a beacon power above the defined threshold.

15) WIFI interface of MN sends a link going down and link down to the MIH user of MN through the MIHF

16) MIH user of MN sends a link scan request to the MIHF of MN.

17) The WIFI interface of MN send a probe request and start scanning the 11 channels of WIFI interface looking for an active one.

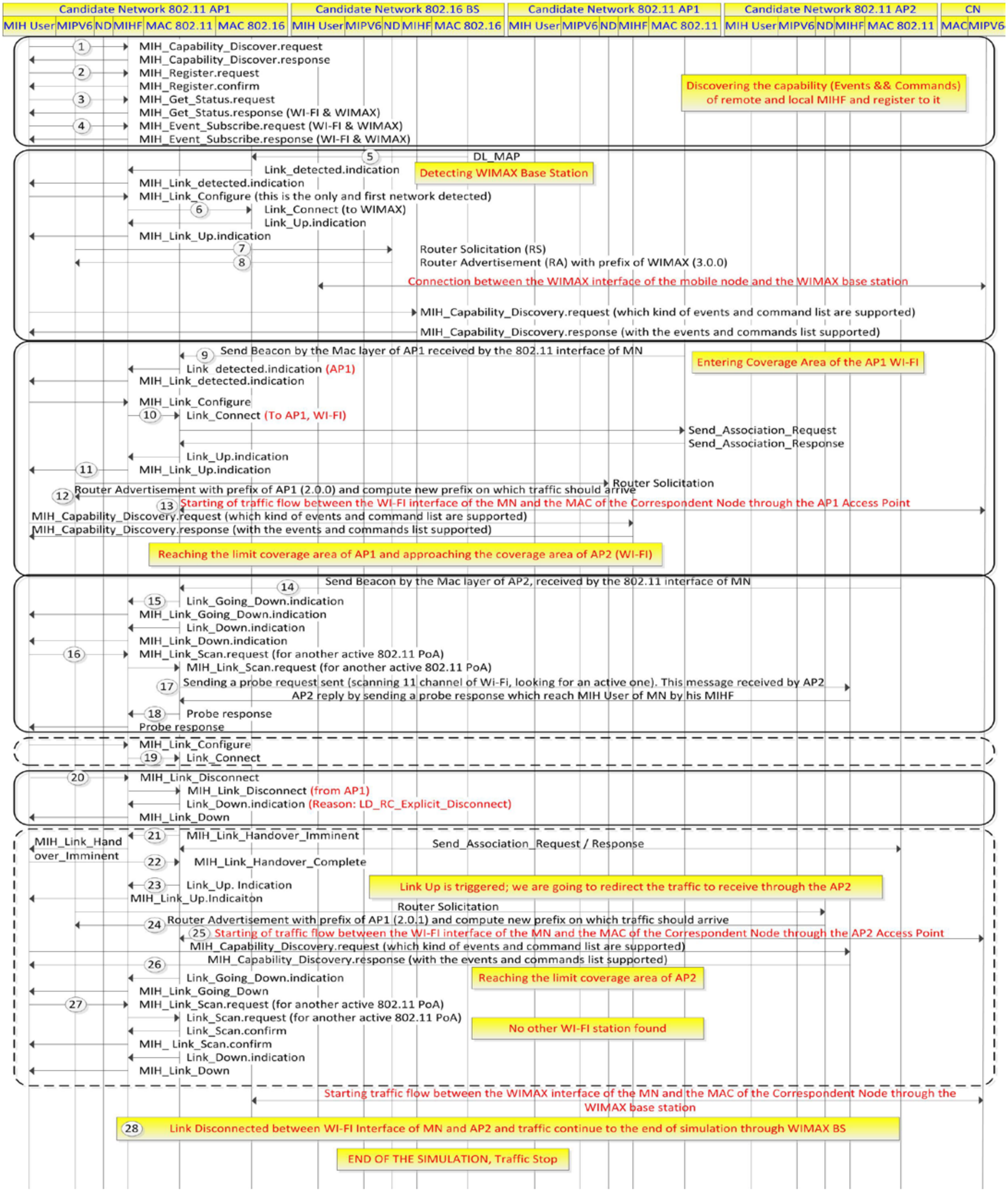

Candidate Network 802.11 AP1
Candidate Network 802.16 BS
Candidate Network 802.11 AP1
Candidate Network 802.11 AP2
CN
MIH User MIPV6 ND MIHF MAC 802.11 MAC 802.16
MIH User MIPV6 ND MIHF MAC 802.16
MIH User MIPV6 ND MIHF MAC 802.11
MIH User MIPV6 ND MIHF MAC 802.11
MAC MIPV6
1 MIH_Capability_Discover.request
MIH_Capability_Discover.response
2 MIH_Register.request
MIH_Register.confirm
3 MIH_Get_Status.request
MIH_Get_Status.response (WI-FI & WIMAX)
4 MIH_Event_Subscribe.request (WI-FI & WIMAX)
MIH_Event_Subscribe.response (WI-FI & WIMAX)
Discovering the capability (Events && Commands) of remote and local MIHF and register to it
5 DL_MAP
Link_detected.indication
Detecting WIMAX Base Station
MIH_Link_detected.indication
MIH_Link_Configure (this is the only and first network detected)
6 Link_Connect (to WIMAX)
Link_Up.indication
MIH_Link_Up.indication
7 Router Solicitation (RS)
8 Router Advertisement (RA) with prefix of WIMAX (3.0.0)
Connection between the WIMAX interface of the mobile node and the WIMAX base station
MIH_Capability_Discovery.request (which kind of events and command list are supported)
MIH_Capability_Discovery.response (with the events and commands list supported)
9 Send Beacon by the Mac layer of AP1 received by the 802.11 interface of MN
Entering Coverage Area of the AP1 WI-FI
Link_detected.indication (AP1)
MIH_Link_detected.indication
MIH_Link_Configure
10 Link_Connect (To AP1, WI-FI)
Send_Association_Request
Send_Association_Response
Link_Up.indication
11 MIH_Link_Up.indication
Router Solicitation
12 Router Advertisement with prefix of AP1 (2.0.0) and compute new prefix on which traffic should arrive
13 Starting of traffic flow between the WI-FI interface of the MN and the MAC of the Correspondent Node through the AP1 Access Point
MIH_Capability_Discovery.request (which kind of events and command list are supported)
MIH_Capability_Discovery.response (with the events and commands list supported)
Reaching the limit coverage area of AP1 and approaching the coverage area of AP2 (WI-FI)
14 Send Beacon by the Mac layer of AP2, received by the 802.11 interface of MN
15 Link_Going_Down.indication
MIH_Link_Going_Down.indication
Link_Down.indication
MIH_Link_Down.indication
16 MIH_Link_Scan.request (for another active 802.11 PoA)
MIH_Link_Scan.request (for another active 802.11 PoA)
17 Sending a probe request sent (scanning 11 channel of Wi-Fi, looking for an active one). This message received by AP2
AP2 reply by sending a probe response which reach MIH User of MN by his MIHF
18 Probe response
Probe response
MIH_Link_Configure
19 Link_Connect
20 MIH_Link_Disconnect
MIH_Link_Disconnect (from AP1)
Link_Down.indication (Reason: LD_RC_Explicit_Disconnect)
MIH_Link_Down
MIH_Link_Hand over_Imminent
21 MIH_Link_Handover_Imminent
Send_Association_Request / Response
22 MIH_Link_Handover_Complete
23 Link_Up. Indication
MIH_Link_Up.Indicaiton
Link Up is triggered; we are going to redirect the traffic to receive through the AP2
Router Solicitation
24 Router Advertisement with prefix of AP1 (2.0.1) and compute new prefix on which traffic should arrive
25 Starting of traffic flow between the WI-FI interface of the MN and the MAC of the Correspondent Node through the AP2 Access Point
MIH_Capability_Discovery.request (which kind of events and command list are supported)
26 MIH_Capability_Discovery.response (with the events and commands list supported)
Link_Going_Down.indication
Reaching the limit coverage area of AP2
MIH_Link_Going_Down
27 MIH_Link_Scan.request (for another active 802.11 PoA)
Link_Scan.request (for another active 802.11 PoA)
No other WI-FI station found
Link_Scan.confirm
MIH_ Link_Scan.confirm
Link_Down.indication
MIH_Link_Down
Starting traffic flow between the WIMAX interface of the MN and the MAC of the Correspondent Node through the WIMAX base station
28 Link Disconnected between WI-FI Interface of MN and AP2 and traffic continue to the end of simulation through WIMAX BS
END OF THE SIMULATION, Traffic Stop

Figure 4. Handover Flow Chart Messages.

18) This message received by AP2 which reply by sending a probe response to the MIH user of MN through its MIHF. MIH user of MN detects the presence of AP2.
According to NIST algorithm, that considers this access point as better network decide to handover to it (and continue with step 19). But according to MCSA algorithm which evaluate the available bandwidth of AP2 before handover to it, find its available bandwidth, consumed by other traffic, very small comparatively to WIMAX, ignore this network and handover to WIMAX directly (jump to step number 20).

19) MIH user sends to MIHF an MIH Link ConFig. this generates a Link Connect to the WIFI interface of MN (connection to AP2).

20) MIH user sends to the MIHF a MIH Link Disconnect which disconnects the connection between the WIFI interface of MN and AP1. According to NIST algorithm, we continue with step 21 and according to MCSA we jump to step 28 saving by that all steps between 21 and 27.

21) The WIFI interface of MN sends a link handover imminent message to the MIHF of MN.

22) MIH user of MN sends link handover complete to the MHIF of MN.

23) WIFI interface of MN send link up indication event to the MIH user of MN through his MIHF announcing the detection of AP2 (second WIFI access point).

24) MIPV6 module of MN sends router solicitation to the WIFI interface of AP2 which answer by a router advertisement with the new prefix (2.0.1).

25) Starting of traffic between the WIFI interface of MN and correspondent node (CN) through AP2.

26) MIH user sends the MIH Capability Discovery Request and response to the Mac layer of AP2 testing if the Events and Commands events list is supported.

27) The MN reaches the limit coverage of AP2, start a link going down event, the WIFI interface of MN send a link scan event looking for others network (delaying the connection to WIMAX) don't find anyone else WIMAX.

28) MN connect to WIMAX and a link disconnect event with WIFI is triggered and traffic continue to the end of the simulation through WIMAX.

D. *Scenario II : Type of Network Impact*

1) Topology Description: Fig. 4 illustrates the topology of scenario II. During this simulation we compare the delay taken by MN when it makes a HO from WI-FI to WIMAX (Fig. 5a) versus handover from WIMAX to WI-FI (Fig. 5b). Measurements are done according to handover algorithm of NIST only.

During the simulation, the MN moves from WI-FI (AP1) toward the center of BS. Once it reaches the limit coverage of AP1, a "link going down" trigger is fired announcing the need for handover. Since the only available network is 802.16 (WIMAX), the handover is made to this network. We also study the same simulation when the mobile moves from WIMAX to WI-FI.

2) Scenario II Results: Fig. 6 shows a decreasing curve of the handover delay as a function of the throughput generated by the MN application. Handover delay is the time difference between the first packet received on the destination network and the last packet received on the current served network. When we increase the throughput, the time between two consecutive packets is smaller and packets reach the destination network earlier, which explains the appearance of the downward curves of handover delay in Fig. 6.

Handover delay from WIMAX to WI-FI is less than the handover delay from WI-FI to WIMAX. When the MN connected to AP1 moves to the center of BS (Fig. 5a), it reaches the limit coverage area of AP1 and generates a "link going down" trigger. In this case, a scan process starts looking for a new network delaying the connection to BS (Fig. 6). While for handover from WIMAX to WI-FI network (Fig. 5b), the MN don't trigger this event because it is still in the coverage area of WIMAX (no loss of WIMAX signal) that's why we have less handover time (Fig. 6).

As a conclusion of this experiment, we can say that upon the type of destination network, we can have different values of handover delay and as a consequence different value of PL.

As shown in Fig. 6, we can note that by varying the throughput values between 120Kbit/s and 170Kbit/s, the handover time varies between 275ms and 200ms hence exceeding the maximum acceptable value of the QoS end-to-end delay parameter (150ms) for real time application. This criterion is worthy to be considered when selecting a new network during HO.

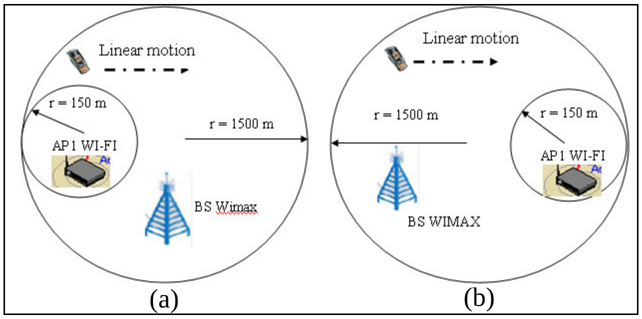


Figure 5. (a) Handover WI-FI-WIMAX, and (b) Handover WIMAX-WI-FI

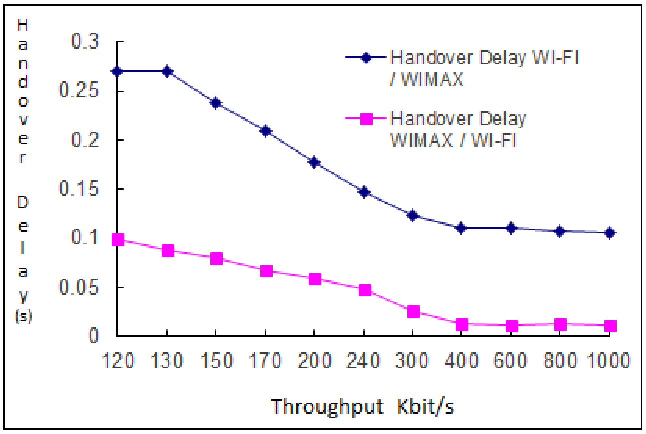


Figure 6. Handover Delay Curves

E. Scenario III : Speed Impact

1) Topology Description: In this scenario, shown in Fig. 7, we study the effect of MN speed on the packet loss during HO. At the beginning, the MN connected to WIMAX, moves to the center of the BS, resulting on a handover to AP1 and AP2 according to NIST algorithm. Once the MN reaches the limit coverage of AP2, it returns to WIMAX network.

2) Scenario III Results: For the three different experimented speeds the packet loss on WIMAX is null because 802.16e WIMAX is designed to support high speed mobile users [25]. Once a MN starts moving toward the center of the BS, it detects the presence of AP1, and according to NIST algorithm it makes a HO to AP1. Some PL happen during this HO and the value of this PL increases with mobile speed (Fig. 8) because WI-FI, unlike WIMAX, is limited in high-speed transport communications environment [26]; and doesn’t support high speed mobility, e.g., for a speed of 20m/s we can see a great impact of Doppler Effect on the system performance [27].

The same process happens during handover from AP1 to AP2 as we experienced other number of packet loss that increases with mobile speed. Also when the MN handover from AP2 to WIMAX some packet loss occur whose number increase with mobile speed. Accordingly, we conclude that users who place importance on the number of packet loss and MN speed would prefer to stay on WIMAX and never stream through AP1 or AP2. Moreover, we concluded that NIST fails to meet the requirement of mobile user moving at a speed higher than the pedestrian speed (1m/s). Thus, we argue that there is a need for a new framework that takes into account the user speed.

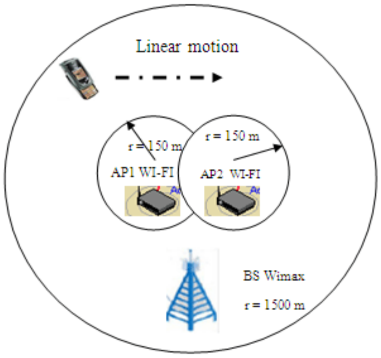


Figure 7. Scenario III topology

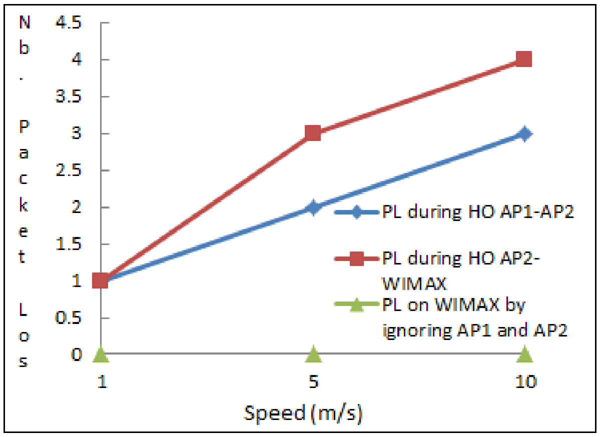


Figure 8. Packet loss as a function of mobile speed

VI. CONCLUSION AND FUTURE WORK

In this paper, we have evaluated the effect of some parameters like Radio Signal Strength, available bandwidth, type of network (802.11 or 802.16) and mobile speed for choosing the best network in the vicinity. We conclude that choosing a network based on the Radio Signal Strength only is not always a good strategy. The experiments that we conducted using the NS-2 simulator showed that the inclusion of additional parameters significantly improves the packet loss ratio and so the QoS guarantee for mobile users. In future work, we will propose a framework with a generic model that takes into consideration different levels of constraints such as network parameters with users and operators preferences to improve the selection of the best candidate network and optimize QoS parameters in terms of packet loss ratio, delay and jitter for real time applications.

REFERENCES

[1] IEEE 802.21, Media Independent Handover Standard http://standards.ieee.org/ getieee802/ download /802.21-2008.pdf. 09.11.2012.

[2] IEEE P802.21/D03.00, “The Network Simulator NS-2 NIST add-on—IEEE 802.21 model (based on IEEE P802.21/D03.00)”, National Institute of Standards and Technology (NIST), January 2007.

[3] M. M. Rehan, “Investigation of IEEE 802.21 Media Independent Handover”, PhD thesis, Mohammad Ali Jinnah University, 2009.

[4] K. Taniuchi, Y. Ohba, V. Fajardo, S. Das, M. Tauil, Y. Cheng, A. Dutta, D. Baker, M. Yajnik, D. Famolari, “IEEE 802.21: Media Independent Handover: Features, Applicability, and Realization”, IEEE Communications Magazine, vol. 47, Jan. 2009, pp. 112–120, doi: 10.1109/MCOM.2009.4752687.

[5] B. R. Chandavarkar and D. G. Reddy, “Improvement in Packet Drop during Handover between WiFi and WiMax”, International Conference on Network and Electronics Engineering vol. 11, Sep. 2011, IPCSIT, pp. **71-75,** doi: 10.7763/IPCSIT.

[6] F. Siddiqui, S. Zeadally, H. El-Sayed and N. Chilamkurti: “A dynamic network discovery and selection method for heterogeneous wireless networks” Int. J Internet Protocol Technology, Vol.4, Jul. 2009. pp. 99-114, doi: 10.1504/IJIPT.2009.027335

[7] F. Jiadi, J. Hong, and L. Xi, “User-Adaptive Vertical Handover Scheme Based on MIH for Heterogeneous Wireless Networks”, Wireless Communications, Networking and Mobile Computing, WiCom 09. 5th International Conference. Sep. 2009, pp. 1-4, doi: 10.1109/WICOM.2009.5302424.

[8] W. Ying, Z. Yun, Y. Jun, and Z. Ping, “An Enhanced Media Independent Handover Framework for Heterogeneous Networks”, IEEE Vehicular Technology Conference, VTC Spring 2008, pp. 2306–2310, doi: 10.1109/VETECS.2008.512.

[9] O. Ormond, G. Muntean, and J. Murphy, “Network Selection Strategy in Heterogeneous Wireless Networks”, Proc. of IT&T 2005: Information Technology and Telecommunications, Oct. 2005.

[10] E. Bircher and T. Braun, “An Agent-Based Architecture for Service Discovery and Negotiations in Wireless Networks”, 2nd Intl. Conf. on Wired/Wireless Internet Communications, vol. 2957, Feb. 2004, , pp. 295-306, doi :10.1007/978-3-540-24643-5_26.

[11] C. Cicconetti, F. Galeassi, and R. Mambrini, “Network-Assisted Handover for Heterogeneous Wireless Networks”, GLOBECOM Workshops (GC Wkshps), IEEE Dec. 2010, pp. 1-5, doi: 10.1109/GLOCOMW.2010.5700294.

[12] J. M. Arraez, M. Esseghir, and L. M. Boulahia, “An Implementation of Media Independent Information Services for the Network Simulator NS-2”, the 8th Annual IEEE Consumer Communications and Networking Conference - Wireless Consumer Communication and Networking, Jan. 2011, pp. 492–496, doi: 10.1109/CCNC.2011.5766519.

[13] V. Andrei, E. C. Popovici, and O. Fratu, “Solution for Implementing IEEE 802.21 Media Independent Information Service”, 8th International Communications Conference on, IEEE Jun. 2010, pp. 519–522, doi: 10.1109/ICCOMM.2010.5509008.

[14] V. Andrei, E. C. Popovici, O. Fratu, and S. V. Halunga, “Development of an IEEE 802.21 Media Independent Information Service”, Automation Quality and Testing Robotics (AQTR), IEEE International Conference on, 2010, vol. 2, pp. 1-6, doi: 10.1109/AQTR.2010.5520819.

[15] M. Q. Khan and S. H. Andresen, “An Intelligent Scan Mechanism for 802.11 Networks by Using Media Independent Information Server (MIIS)”, Advanced Information Networking and Applications (WAINA), IEEE Workshops of International Conference on Mar. 2011, pp. 221–225, doi: 10.1109/WAINA.2011.26.

[16] Y. Y. An, B. H. Yae, K. W. Lee, Y. Z. Cho, and W. Y. Jung, “Reduction of Handover Latency Using MIH Services in MIPv6”, Proc. of the 20th IEEE International Conference on Advanced Information Networking and Applications (AINA 06), vol. 2, Apr. 2006, pp. 229-234, doi: 10.1109/AINA.2006.283.

[17] H. Silva, L. Figueiredo, C. Rabadão, and A. Pereira, “Wireless Networks Interoperability - Wi-Fi Wimax Handover”, Proc. Systems and Networks Communications, Fourth International Conference on, (ICSNC 09), Sep. 2009, pp. 100–104, doi: 10.1109/ICSNC.2009.99.

[18] Wi-Fi Alliance, http://www.wi-fi.org 09.11.2012.

[19] H. Velayos and G. Karlsson, “Techniques to reduce the IEEE 802.11b handoff time”, Proc. Communications, IEEE International Conference on, Jul. 2004, pp. 3844–3848, doi: 10.1109/ICC.2004.1313272.

[20] N. T. Dao, R. A. Malaney, E. Exposito, and X. Wei, “Differential VoIP Service in Wi-Fi Networks and Priority QoS Maps”, IEEE Globecom, vol. 5, Dec. 2005, pp. 2653-2657, doi: 10.1109/GLOCOM.2005.1578241.

[21] B. XieI, W.Zhou, and J.Zeng, “A Novel Cross-Layer Design with QoS Guarantee for WiMAX System”, Pervasive Computing and Applications, ICPCA, Third International Conference on, vol. 2, Oct. 2008, pp. 835-840, doi: 10.1109/ICPCA.2008.4783726.

[22] IEEE P802.16, “IEEE Standard for Local and metropolitan area networks, Part 16: Air Interface for Fixed Broadband Wireless Access Systems”, February 2009.

[23] IEEE P802.16m/D4, “Air Interface for Fixed and Mobile Broadband Wireless Access Systems: Standard IEEE P802.16e”, 2010.

[24] M. A. Awal and L. Boukhatem, “WiMAX and End-to-End QoS Support”, WhitePaper, Univ. of Paris-Sud 11, CNRS. May 2009.

[25] S. Murawwat and T. Javaid, “Speed & Service based handover Mechanism for cellular WIMAX”, Computer Engineering and Technology (ICCET), 2nd International Conference on, vol. 1, April 2010, pp. V1-418-V1-422, doi: 10.1109/ICCET.2010.5486067.

[26] Z. ZHAO, “Wi-Fi in High-Speed Transport Communications”, Intelligent Transport Systems Telecommunications, (ITST), 9thInternational Conference on, Oct. 2009, pp. 430–434, doi: 10.1109/ITST.2009.5399314.

[27] M. Thaalbi and N. Tabbane, “Vertical Handover between WiFi Network and WiMAX Network According to IEEE 802.21 Standard”, Technological Developments in Networking, Education and Automation, 2010, pp. 533-537, doi: 10.1007/978-90-481-9151-2_93.

A High-precision Time Handling Library

Irina Fedotova
Faculty of Information science and Computer Engineering
Siberian State University of Telecommunication and Information Sciences, Novosibirsk, Russia
i.fedotova@emw.hs-anhalt.de

Eduard Siemens, Hao Hu
Faculty of Electrical, Mechanical and Industrial Engineering
Anhalt University of Applied Sciences
Koethen, Germany
{e.siemens, h.hu}@emw.hs-anhalt.de

***Abstract*—An appropriate assessment of end-to-end network performance presumes highly efficient time tracking and measurement with precise time control of the stopping and resuming of program operation. In this paper, a novel approach to solving the problems of highly efficient and precise time measurements on PC-platforms and on ARM-architectures is proposed. A new unified *High Performance Timer* and a corresponding software library offer a unified interface to the known time counters and automatically identify the fastest and most reliable time source, available in the user space of a computing system. The research is focused on developing an approach of unified time acquisition from the PC hardware and accordingly substituting the common way of getting the time value through Linux system calls. The presented approach provides a much faster means of obtaining the time values with a nanosecond precision than by using conventional means. Moreover, it is capable of handling the sequential time value, precise sleep functions and process resuming. This ability means the reduction of wasting computer resources during the execution of a sleeping process from 100% (busy-wait) to 1-1.5%, whereas the benefits of very accurate process resuming times on long waits are maintained.**

Keywords-high-performance computing; network measurement; timestamp precision; time-keeping; wall clock.

I. INTRODUCTION

Estimation of the achieved quality of the network performance requires high-resolution, low CPU-cost time interval measurements along with an efficient handling of process delays and sleeps [1][2]. The importance on controlling these parameters can be shown on the example of a transport layer protocol. Its implementation may need up to 10 time fetches and time operations per transmitted and received data packet. However, performing accurate time interval measurements, even on high-end computing systems, faces significant challenges.

Even though Linux (and in general UNIX timing subsystems) uses auto-identification of the available hardware time source and provides nanosecond resolution, these interfaces are always accessed from user space applications through system calls. Thus it costs extra time in the range of up to a few microseconds – even on contemporary high-end PCs [3]. Therefore, direct interaction with the timing hardware from the user space can help to reduce time fetching overhead from the user space and to increase timing precision. The Linux kernel can use different hardware sources, whereby time acquisition capabilities depend on the actual hardware environment and kernel boot parameterization. While the time acquisition of some time sources costs up to 2 microseconds, others need about 20 nanoseconds. In the course of this work, a new *High Performance Timer* and a corresponding library *HighPerTimer* have been developed. They provide a unified user-space interface to time counters available in the system and automatically identify the fastest and the most reliable time source (e.g. Time Stamp Counter (TSC) [4][5] or High-Performance Event Counter (HPET) [6][7]). In the context of this paper, the expression *time source* means one of the available time hardware or alternatively the native timer of the operating system, usually provided by the standard C library.

Linux (as well as other UNIX operating systems) faces a significant problem of inaccurate sleep time, which is known for many years, especially in older kernel versions, when Linux has provided a process sleep time resolution of 10 msec. This leads to a minimum sleep time of about 20 msec [8]. Even nowadays, when Linux kernels usually reduce this resolution down to 1 msec, waking up from sleeps can take up to 1-2 msec. With kernel 2.6 the timer handling under Linux has been changed significantly. This change has reduced the wakeup misses of sleep calls to 51 µsec on average and to 200-300 µs in peaks. However, for many soft-real-time and high-performance applications, this reduction is not sufficient. Presented *High Performance Timer* not only significantly improves the time fetching accuracy, but also addresses the problem of those imprecise wakeups from sleep calls under Linux.

These precision issues lead to the fact that, for high-precision timing within state machines and communication protocols, busy-waiting loops are currently commonly used for waits, preventing other threads from using the given CPU. The approach of the *High Performance Timer* library aims at reducing the CPU load down to an average of 1-1.5% within the sleep calls of the library and at raising the wakeup precision to 70-160 nsec. Reaching these values enables users of this library to implement many protocols and state machines with soft real-time requirements in user space.

The remainder of the paper is organized as follows. In Section II, related work is described. Section III shows the specific details of each time source within the suggested single unified High-Performance Timer class interface. In Section IV, we briefly describe the implemented library interface. Some experimental results of identifying appropriate timer source along with their performance characteristics are shown in Section V. In Section VI, precise process sleeping aspects are shown. Finally, Section VII describes next steps and future work in our effort to develop a tool for highly efficient high-performance network measurements.

II. Related Work

Since the problem of inefficient time keeping in Linux operating system implementation has become apparent, several research projects have suggested to access the timing hardware directly from user space [1][9][10]. However, most of this research considers handling of a single time hardware source only, predominantly the Time Stamp Counter [1][9][11]. Other solutions provide just wrappers around timer-related system calls and so inherit their disadvantages such as the high time overhead [12][13]. In other proposals, the entire time capturing process is integrated into dedicated hardware devices [14][15]. Most of this research focuses only on a subset of the problems, addressed in this work. Our work with the *HighPerTimer* library improves timing support by eliminating the system call overhead and also by application of more precise process sleep techniques.

III. Unified Time Source of the *HighPerTimer* Library

While most of the current software solutions on Linux and Unix use the timing interface by issuing *clock_gettime()* or *gettimeofday()* system calls, *HighPerTimer* tries to invoke the most reliable time source directly from the user space. Towards the user, the library provides a unified timing interface for time period computation methods along with sleep and wakeup interfaces, independently from the used underlying time hardware. So, the user sees a “unified time source” that accesses the best possible on the underlying hardware, and that generally avoids system call overheads. The *HighPerTimer* interface supports access the mostly used time counters: TSC, HPET and, as the last alternative, the native timer of the operating system, through one of the said Unix system calls. The latter time source we call the *OS Timer.*

Using the *Time Stamp Counter* is the fastest way of getting CPU time. It has the lowest CPU overhead and provides the highest possible time resolution, available for the particular processor. Therefore, in the context of our library, the TSC is the most preferable time source. In newer PC systems, the TSC may support an enhancement, referred to as an *Invariant TSC* feature. *Invariant TSC* is not tightly bound to a particular processor core and has, in contrary to many older processor families, a guaranteed constant frequency [16]. The presence of the *Invariant TSC* feature in the system can be tested by the *Invariant TSC* flag, indicated by the *cpuid* processor instruction. For most cases, the presence of this *Invariant TSC* flag is essential in order to accept it as a *HighPerTimer* source.

Formerly referred by Intel as a Multimedia Timer [7], the *High Precision Event Timer* is another hardware timer used in personal computers. The HPET circuit is integrated into the south bridge chip and consists of a 64-bit or 32-bit main counter register counting at a given constant frequency between 10 MHz and 25 MHz. Difficulties are faced when the HPET main counter register is running in 32-bit mode because overflows of the main counter arise at least every 7.16 minutes. With a frequency of 25 MHz, register overflows would occur even within less than 3 minutes. So, time periods longer than 3 minutes can’t reliably measured in 32 bit mode. So, in the *HighPerTimer* library, we decided to generally avoid using the HPET time source in case of a 32-bit main counter.

For systems, on which neither TSC nor HPET are accessible or TSC is unreliable, an alternative option of using the OS Timer is envisaged. This alternative is a wrapper issuing the system call *clock_gettime()*. This source is safely accessible on any platform. However, it has the lowest priority because it issues system calls, with their time costs of up to 2 microseconds in worst case [17][18].Depending on the particular computer architecture and used OS, these costs can be less due to the support of the so-called virtual system calls. These calls provide faster access to time hardware and avoid expensive context switches between user and kernel modes [19]. Nevertheless, invocation of *clock_gettime()* through a virtual system call is still slower than the acquisition time value from current time hardware directly. The difference between getting the time value using virtual system calls and getting the time values directly from the hardware is about 3 to 17 nsec, as measurement results, discussed in Section V, show.

IV. the HighPerTimer Interface

The common guidelines on designing any interfaces cover efficiency, encapsulation, maintainability and extensibility. Accordingly, the implementation of the *HighPerTimer* library pays particular attention to these aspects. Using the new C++11 programming language standard [20], the library achieves high efficiency and easy code maintainability. Furthermore, regarding the platform-specific aspects, *HighPerTimer* runs on different 64-bit and 32-bit processors of Intel, AMD, VIA and ARM, and considers their general features along with specialties of time keeping.

However, some attention must be paid to obtaining a clean encapsulation of hardware access when using C++. For this encapsulation, the *HighPerTimer* library comprises two header files and two implementation files called *HighPerTimer* and *TimeHardware*. Each of them contains three classes. *HighPerTimer* files contain *HighPerTimer*, *HPTimerInitAndClear* and *AccessTimeHardware* classes, as described below. In *TimeHardware* files, the classes *TSCTimer*, *HPETTimer* and *OSTimer* corresponding to the respective time sources TSC, HPET and the OS source have been implemented. Through an assembly code within the C++ methods, they provide direct access to the timer hardware, initialize the respective timer source, retrieve their time value and are at only *HighPerTimer* class’s disposal. Dependencies between the classes are presented in Fig. 1.

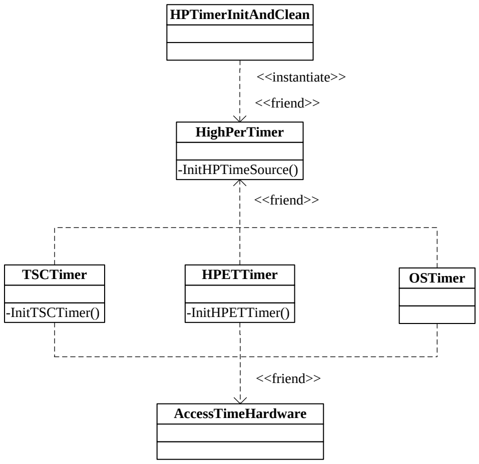


Figure 1. Simplified class diagram of *HighPerTimer* library

TSCTimer, *HPETTimer* and *OSTimer* classes have a “friend” relationship with the *HighPerTimer* class, which means that *HighPerTimer* places their private and protected methods and members at friend classes’ disposal. For safety and security reasons, we protect the hardware access from use by application users directly and permit access only from special classes. An *AccessTimeHardware* class provides a limited read-only access to some information on CPU and specific time hardware features, obtained in a protected interface. For example, some advanced users can find out failure reasons of the initialization routine of the HPET device and get a corresponding error message:

```
std::cout << AccessTimeHardware::HpetFailReason();
```

However, all the routines of time handling along with access to the actual timer attributes such as clock frequency are accessed by the library users via the *HighPerTimer* class. For interfacing with other time formats, *HighPerTimer* class provides a set of constructors that sets its object to the given time provided in seconds, nanoseconds or in the native clock ticks of the used time source. Via specific constructor, a time value in a Unix-specific time format [21] can also be assigned to a *HighPerTimer* object. The current time value is retrieved using the following piece of code:

```
// declare HighPerTimer objects
HighPerTimer timer1, timer2;
HighPerTimer::Now (timer1);
// measured operation
HighPerTimer::Now (timer2);
```

Comparison operators allow effective comparison to be performed using the main counter values. Some of these methods are declared as follows:

```
bool operator>= (const HighPerTimer& timer) const;
bool operator<= (const HighPerTimer& timer) const;
bool operator!= (const HighPerTimer& timer) const;
```

The user can also set the value of a timer object explicitly to zero and add or subtract the time values in terms of timer objects, tics, nanoseconds or seconds. Since the main “time” capability of a timer object is kept in the main counter only, the comparison operations between timer objects, as well as arithmetical operations on them, are nearly as fast as comparisons and elementary arithmetical operations on two int64 variables. Recalculations between tics, seconds, microseconds and nanoseconds are only done in the “late initialization” fashion when string representations of the timer object or seconds, microseconds or nanoseconds of the object are explicitly requested via the class interface:

```
// subtract from timer object
HighPerTimer & SecSub (const uint64_t Seconds);
HighPerTimer & USecSub (const uint64_t USeconds);
HighPerTimer & NSecSub (const uint64_t NSeconds);
HighPerTimer & TicSub (const uint64_t Tics);

// add to timer object
HighPerTimer & SecAdd (const uint64_t Seconds);
HighPerTimer & USecAdd (const uint64_t USeconds);
HighPerTimer & NSecAdd (const uint64_t NSeconds);
HighPerTimer & TicAdd (const uint64_t Tics);
```

Assignment operators allow a *HighPerTimer* object to be set from the Unix-format of time values - `timeval` or `timespec` structs [21]. Both of these structures represent time, elapsed since 00:00:00 UTC on 01.01.1970. They consist of two elements: the number of seconds and the rest of the elapsed time represented either in microseconds (in case of `timeval)` or in nanoseconds (in case of `timespec)`:

```
struct timeval {
   long tv_sec;  /* seconds */
   long tv_usec; /* microseconds */
}

struct timespec {
   long tv_sec;  /* seconds */
   long tv_nsec; /* nanoseconds */
};
```

Assignment to these structures is also possible with *HighPerTimer* objects through copying or moving:

```
const HighPerTimer & operator= (const struct
timeval & TV);
const HighPerTimer & operator= (const struct
timespec & TS);
const HighPerTimer & operator= (const HighPerTimer
& Timer);
HighPerTimer & operator= (HighPerTimer && Timer);
```

This way, the *HighPerTimer* library provides a fast and efficient way to handle time values by operating main counter value and seconds and nanoseconds values only on demand. It also relieves users from the manual handling of specific two-value structures such as `timeval` or `timespec`.

However, for the whole routine of handling time values, some central parameterization of the library must be performed at the initialization time of the library. Primarily, this is the initialization of the *HighPerTimer* source, which is accomplished on the basis of the appropriate method calls from the *TimeHardware* file. Especially, *InitHPTimeSource()* calls *InitTSCTimer()* and *InitHPETTimer()* methods, which attempt to initialize respective time hardware and return true on success or false on failure (see Fig. 1).

Before using any timer object, the following global parameters must be measured and set: the frequency of the main counter as a double precision floating point value and as a number of ticks of the main counter within one microsecond, the value of the shift of the main timer counter against Unix Epoch, the maximum and minimum values of *HighPerTimer* for the given hardware-specific main counter frequency, and the specified HZ frequency of the kernel. The value of HZ is defined as the system timer interrupt rate and varies across kernel versions and hardware platforms. In the context of the library, the value of HZ is used for the implementation of an accurate sleep mechanism, see Section VI. The strict sequence of the initialization process is determined within an additional *HPTimerInitAndClean* service class (see Fig. 1) by invoking corresponding *HighPerTimer* initialization methods through their "friend" relationship. A strict order of initialization of the given global variables must be assured, which is somewhat tricky since all the variables must be declared static and must be initialized before entering the main routine of the application.

Despite the advantage of automatic detection of the appropriate time source, situations sometimes arise when an application programmer prefers to use a different time source than the one automatically chosen at library initialization time. To account for this, a special ability to change the default timer is provided. This change causes a recalculation process for most of the timer attributes:

```
// create variable for a new value of time source
TimeSource MySource;
MySource = TimeSource::HPET;
HighPerTimer::SetTimerSource ( MySource );
```

However, since this change leads to invalidation of all the already existing timer objects within the executed program, this feature should be used with caution and only at the system initialization time, and definitely before instantiation of the first *HighPerTimer* object.

V. TIME FETCHING PERFORMANCE RESULTS

Table I shows the performance results when getting the time values using the *HighPerTimer* library as measured on different processor families. The mean and standard deviation values of the costs of setting a *HighPerTimer* object are shown. For this investigation, time was fetched in a loop of 100 million consecutive runs and set to a *HighPerTimer* object. Since we are interested here in measuring the time interval between two consecutive time fetches only, without any interruption in between, we filter out all outlying peaks. These peaks are most probably caused by process interruption by the scheduler or by an interrupt service routine. Thus, filtering out such outliers allows us to get rid of the bias caused by physical phenomena, which are outside the scope of this investigation.

TABLE I. COSTS OF SETTING TIMER ON DIFFERENT PROCESSORS

Processor (CPU)	**Time source**	**Mean, nsec**	**St. deviation, nsec**
Intel ® Core ™ i7-2600, 1600 MHz	TSC	16.941	0.1231
VIA Nano X2 U4025, 1067 MHz	TSC	38.203	0.3134
Athlon ™ X2 Dual Core BE-2350, 1000 MHz	HPET	1063.3	207.92

The following two examples demonstrate the behavior of *HighPerTimer* sources in more detail and allow a comparison of their reliability and costs depending on the particular processor conditions. Although Table I shows the results for all three processors, later investigations are shown only for less powerful systems. It makes sense to examine in more depth those systems, where for example, TSC is unstable or does not possess *Invariant TSC* flag (see Section III).

In the first case, processor VIA Nano X2 has TSC as a current time source. Costs of time fetching here are about 38 nsec. Since TSC source has the highest priority and has been initialized successfully, the HPET device check is not necessary and so omitted here. Moreover, on this processor, the Linux kernel is also using TSC as its time source and so, within the *clock_gettime()* call, the kernel is also fetching the TSC register of the CPU. Fig. 2 shows the relation between the TSC, OS and HPET timers on this processor. Similarity between TSC and OS costs are seen very clearly. As seen in Table 2, the difference between the mean value of time fetching between OS Timer and TSC Timer is 64 nsec. Each system call with a context switch would last at least ten times longer, thus we can conclude that, on this system, a virtual system call is issued by *clock_gettime()* instead of a real system call with a context switch. HPET source for the library can be set by the static method *HighPerTimer::SetTimerSrouce*. However, we would expect here much slower time operations, as seen in Table 2.

TABLE II. MEAN AND STANDARD DEVIATION VALUES OF HPET, TSC AND OS TIMER COSTS ON THE VIA NANO X2 PROCESSOR

Timer source	**Mean, nsec**	**Standard deviation, nsec**
TSC Timer	38.23	0.3134
HPET Timer	598.72	76.015
OS Timer	102.20	0.5253

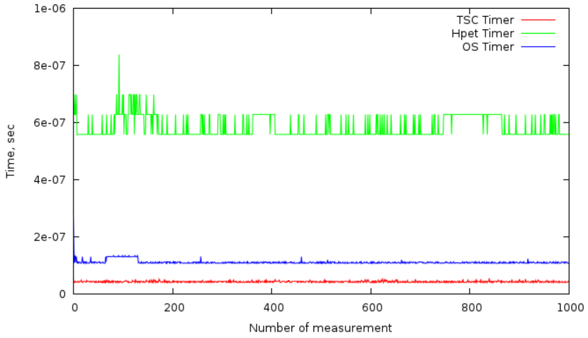


Figure 2. Measurements of TSC, HPET and OS Timer costs on the VIA Nano X2 processor

The next example illustrates another case of a dependence on the OS Timer from the current time source. For the processor AMD Athlon X2 Dual Core, the TSC initialization routine fails because TSC is unstable here. However, since the HPET device is accessible, there are two more options for the time source for *HighPerTimer* – HPET or OS Timers - and it is necessary to check the mean costs of getting the ticks of both timers.

Although the mean value of time fetching for TSC can be significantly lower than for HPET, the *HighPerTimer* library considers the TSC to be a non-stable, unreliable time source since the *Invariant TSC* flag (see Section III above) is not available and the TSC constancy is not identified by additional library checks. So, it must be assumed that TSC frequency changes from time to time due to power saving or other techniques of the CPU manufacturers. In the next step, HPET and OS Timer characteristics must be considered. The difference between the mean values of HPET and OS Timer is about 54.1 nsec, which is not enough for a system call with a context switch. Thus we conclude that *clock_gettime()* also uses the HPET timer and passes it to the user via a virtual system call. However, to provide an appropriate level of reliability, we also evaluate numbers through their deviation values. For this evaluation, a threshold for the difference of mean values was chosen. When the difference of the mean values of HPET and OS Timer is no more than 25%, we also take into account standard deviation values of time fetching and so check the temporal stability of the considered time source. Consequently, when the mean time fetching value of the two time sources is similar, the *HighPerTimer* library would give precedence to the time source with a less standard deviation of the time fetching costs.

TABLE III. MEAN AND STANDARD DEVIATION VALUES OF HPET, TSC AND OS TIMER COSTS ON THE AMD ATHLON PROCESSOR

Timer source	**Mean, µsec**	**Standard deviation, µsec**
TSC Timer	0.0251	0.0015
HPET Timer	1.0633	0.2079
OS Timer	1.1174	0.3743

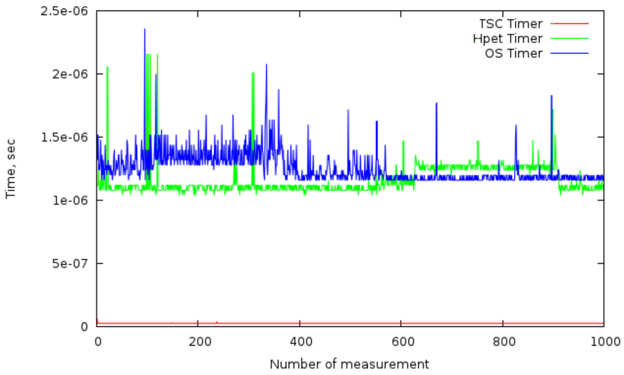


Figure 3. Measurements of TSC, HPET and OS Timer costs on the AMD Athlon processor

VI. PRECISE PROCESS SLEEPING ASPECTS

For process sleeping or suspension, Linux provides the sleep function (implemented in the standard C library). Dependent on the sleep duration, the function either suspends from the CPU or waits in the busy-waiting mode (sometimes also called spinning wait). However, measurements performed in this work revealed that the sleep function of the standard C library misses the target wake-up time by more than 50 microseconds on average. Such an imprecision however is unacceptable for high-accuracy program sleeps. By comparison, pure busy wait implementations within an application miss the target return time by about 100 nanoseconds, but keep the CPU busy throughout the wait time.

Unlike the C library's sleep call, the sleep of the *HighPerTimer* library combines these two ways of sleeping. It has very short miss times on waking up with a minimum CPU utilization at the same time. This improvement provides a big competitive advantage over the predecessor solutions.

HighPerTimer provides a wide range of functions for making a process sleep. For example, the user can define the specific sleep time, given purely in seconds, in microseconds or nanoseconds. A process suspension with a nanosecond resolution can be done as follows:

```
HighPerTimer timer1;
uint32_t SleepTimeNs(14500);
// sleep in nanoseconds
timer1.NSecSleep(SleepTimeNs);
```

Alternatively, the time value of a *HighPerTimer* object can be set to a specific time value at which the process shall wake up. On the call of *SleepToThis()*, the process will then be suspended till the system time has reached the value of that object :

```
//declare timer object equaled to 10 sec, 500 nsec
HighPerTimer timer2 (10, 500);
timer2.SleepToThis();
```

Table IV shows the precision of sleeps and busy-waits using different methods. Miss values are here the respective differences between the targeted wakeup time and real times of wakeups measured in our tests. However, the miss values of sleep times heavily depend on the fact, whether target sleep interval was shorter or longer, than time between two timer interrupts. So, Table IV consists of two parts – one where sleep time is longer then 1/HZ, and one where it is less than 1/HZ. Thus, the left column shows results for waits lasting longer than a period of two kernel timer interrupts. The right column shows the results for the scenario, in which the sleep call lasts less than the interval between two kernel timer interrupts. These measurements have been performed on the Intel Core–i7 processor. Other than in measurements from Section V, in this case it makes sense to show results on a more stable and powerful system. Moreover, it was expected that the accuracy of sleeps would be higher on the newer Linux kernel versions where time handling has been changed significantly. However, as the measurements below show, these kernel changes are still not sufficient.

In this test scenario, we have issued the respective sleep method within a loop of 100000 sleeps with different sleep times between 0.25 sec and 1 µsec, and then the mean value of the sleep duration miss has been calculated.

TABLE IV. THE COMPARISON OF MISS VALUES OF DIFFERENT METHODS OF SLEEPING, PERFORMED WITH TSC ON THE INTEL CORE –I7 PROCESSOR

	Sleep time >= 1/HZ	**Sleep time < 1/HZ**
	Mean miss, µsec	**Mean miss, µsec**
System sleep	61.985	50.879
Busy-waiting loop	0.160	0.070
HighPerTimer sleep	0.258	0.095

The above experiment took about 830 minutes, so the upper limit of the range for sleep time value was reduced to 0.25 sec. The chart in Fig. 4 demonstrates more detailed results of this experiment and shows the dependency of miss against the target sleep time in dependence from sleep duration. To track this dependency more deeply, here the range of sleep time value was increased and is taken between 10 sec and 1 µsec.

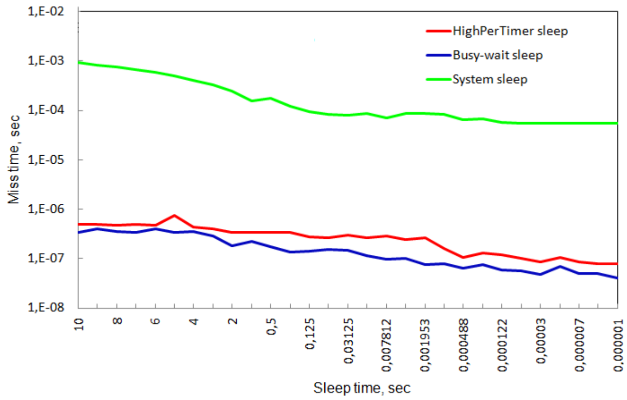


Figure 4. Dependency of miss on the target time from sleep time, performed with TSC on the Intel Core –i7 processor, HZ = 1000

In the next step, we measured the CPU consumption of the respective sleep routine. In the busy-waiting loop, the total CPU consumption during the test achieves, as expected, almost 100%. For the sleep function of the standard C library, it tends to zero. In the case of sleeping using the *HighPerTimer* library, the overall CPU consumption during the test was 1.89%, which can be considered as a good tradeoff between precision of waking up time and CPU consumption during the sleep.

VII. CONCLUSION AND FUTURE WORK

In accordance with the requirements of advanced high-speed data networks, we showed an approach for the unified high performance timer library that successfully solves two significant problems. Firstly, *HighPerTimer* allows identification of the most efficient and reliable way for time acquisition on a system and for avoiding system calls invocation on time acquisition. Secondly, it solves the problem of precise sleeping aspects and provides new advanced sleeping and resuming methods.

The *HighPerTimer* library has the potential to become widely used in estimation network packet dynamics, particularly when conducting high-accuracy and high-precision measurements of network performance. At this stage, the integration of the suggested solution into the appropriate tool for distributed network performance measurement [22] is in progress. Moreover, to the next steps, the better support of the ARM processor will be addressed. Since the ARM processor possesses neither HPET nor TSC, the only way to support ARM at this stage is to select OS Timer. Presumably, an invocation of the initial ARM system timer can afford to save several additional microseconds and improve the timer accuracy.

REFERENCES

[1] R. Takano, T. Kudoh, Y. Kodama, and F. Okazaki, "High-resolution Timer-based Packet Pacing Mechanism on the Linux Operating System," IEICE Transactions on Communication, Tokyo, vol. E94-B, no. 8, pp. 2199-2207, Nov., 2011.

[2] J. Micheel, S. Donnelly, and I. Graham, "Precision time stamping of network packets," Proc. of the 1st ACM SIGCOMM Workshop on Internet Measurement, San Francisco, CA, USA, pp. 273-277, Nov., 2001.

[3] H. Hu, "Untersuchung und prototypische Implementierung von Methoden zur hochperformanten Zeitmessung unter Linux," (German), Bachelor Thesis, Anhalt University of Applied Sciences, Koethen, Germany, Nov., 2011.

[4] Performance monitoring with the RDTSC instruction. URL: http://www.ccsl.carleton.ca/~jamuir/rdtscpm1.pdf, retrieved: Jan., 2013.

[5] E. Corell, P. Saxholm, and D. Veitch, "A user friendly TSC clock," Proc. PAM, Adelaide, Australia, pp. 141-150, Mar., 2006.

[6] S. Siddha, V. Pallipadi, and D. Ven, "Getting maximum mileage out of tickles," in Proc. of the 2007 Linux Symposium, pp. 201-208, 2007.

[7] Intel IA-PC HPET (High Precision Event Timers) Specification. URL: http://www.intel.com/content/dam/www/public/us/en/documents/technical-specifications/software-developers-hpet-spec-1-0a.pdf, retrieved: Jan., 2013.

[8] D. Kang, W. Lee, and C. Park, "Kernel Thread Scheduling in Real-Time Linux for Wearable Computers," ETRI Journal, Daejeon, Korea, vol. 29, no. 3, June 2007, pp. 270-280, doi: 10.4218/etrij.07.0506.0019.

[9] P. Orosz and T.Skopko, "Performance Evaluation of a High Precision Software-based Timestamping Solution," International Journal on Advances in Software, ISSN 1942-2628, 2011, vol. 4, no. 1, pp.181-188.

[10] T. Gleixner and D. Niehaus, "Hrtimers and Beyond: Transforming the Linux Time Subsystems," The Linux Symposium, Ottawa, Canada, 2006, vol. 1, pp. 333-346.

[11] D. Kachan, E.Siemens, and H.Hu, “Tools for the high-accuracy time measurement in computer systems,” (Russian), 6th Industrial Scientific Conference “Information Society Technologies”, Moscow, Russia, 2012, pp.22-25.

[12] Intel Trace Collector Reference Guide, p. 5.2.5. URL: http://software.intel.com/sites/products/documentation/hpc/ics/itac/81/ITC_Reference_Guide/ITC_Reference_Guide.pdf, retrieved: Jan., 2013.

[13] D. Grove and P. Coddington, “Precise MPI Performance Measurement Using MPIBench,” Proc. of HPC Asia, pp. 1-14, Gold Coast, Australia, 2001.

[14] The Dag project. URL: http://www.endace.com, retrieved: Jan., 2013.

[15] A. Pásztor and D. Veitch, “PC based precision timing without GPS,” The 2002 ACM SIGMETRICS international conference on Measurement and modeling of systems, Marina Del Rey California, USA, vol. 30, no. 1, pp. 1-10, June, 2002, doi: 10.1145/511334.511336.

[16] Intel 64 and IA-32 Architectures, Software Developer‘s Manual, vol. 3B 17-36. URL: http://download.intel.com/products/processor/manual/253669.pdf, retrieved: Jan., 2013.

[17] J. Dike, “A user-mode port of the Linux kernel,” USENIX Association Berkeley, pp. 63-72, California, USA, 2000.

[18] K. Jain and R. Sekar, “User-Level Infrastructure for System Call Interposition: A Platform for Intrusion Detection and Confinement,” Proc. of the ISOC Symposium on Network and Distributed System Security, pp.19-34, Feb., 2000.

[19] J. Corbet, “On vsyscalls and the vDSO. Kernel development news,” Linux news site LWN. URL: http://lwn.net/Articles/446125/, retrieved: Jan., 2013.

[20] Online C++11 standard library reference, URL: cppreference.com, retrieved: Jan., 2013.

[21] GNU Operating System Manual, “Elapsed Time”. URL: http://www.gnu.org/software/libc/manual/html_node/Elapsed-Time.html, retrieved: Jan., 2013.

[22] E.Siemens, S.Piger, C. Grimm, and M. Fromme, “LTest – A Tool for Distributed Network Performance Measurement,” Consumer Communications and Networking Conference, Las Vegas, NV, USA, 2004, pp. 234-244, doi: 10.1109/CCNC.2004.1286865.

Optimization of Server Locations in Server Migration Service

Yukinobu Fukushima
The Graduate School of Natural Science and Technology
Okayama University
Okayama, Japan
fukusima@okayama-u.ac.jp

Tutomu Murase
Cloud System Research Laboratories
NEC Corporation
Kanagawa, Japan
t-murase@ap.jp.nec.com

Tokumi Yokohira
The Graduate School of Natural Science and Technology
Okayama University
Okayama, Japan
yokohira@okayama-u.ac.jp

Tatsuya Suda
University Netgroup Inc.
Irvine, USA
tatsuyasuda@gmail.com

***Abstract*—In server migration service (SMS), a work place (WP) refers to a computer that runs a virtual machine, and a server refers to a virtual machine that runs a server-side application of a network application (NW-App). In SMS, WPs are deployed at various locations in a network, and servers may migrate between WPs towards the users of the NW-App to achieve better QoS for the users. In SMS, an SMS provider tries to provide an NW-App provider with a certain level of QoS that they agree upon, and if the SMS provider fails, it pays penalty (e.g., reimbursement of a part of service charges to users) depending on the degree and length of the QoS violation. Thus, the SMS provider is incentivized to migrate servers between WPs to satisfy the agreed-upon QoS level to reduce the penalty. On the other hand, an SMS provider also needs to be moderate in performing server migrations to avoid degradation of network QoS (i.e., QoS of background traffic). This is because a server is typically large in size and the server generates a large amount of traffic when it migrates, resulting in increasing delay and loss for its background traffic in a network. This paper formulates an integer-programming model for the off-line server locations decision (i.e., when and to which WP server should migrate) where the penalty associated with NW-App's QoS violations is minimized, keeping the number and distance of server migrations below a given level. This paper also compares the minimum penalty obtained through solving the integer-programming model against the penalty obtained with a greedy on-line server locations decision algorithm, which migrates a server to a WP that minimizes the current penalty with no consideration of the penalty that will arise in the future. Numerical examples show that the integer-programming model achieves 36% to 49% lower penalty than the greedy algorithm when the degradation of network QoS is little acceptable.**

Keywords-cloud computing*; *server migration service*; *integer-programming model*; *penalty

I. INTRODUCTION

Cloud computing [1] is emerging as a new computing paradigm. As one of its service models, IaaS (Infrastructure as a Service) cloud service (e.g., Amazon EC2 [2]) is attracting attention from the cloud research community. In IaaS cloud service, customers may operate their virtual machines (VMs) at IaaS cloud service provider's data center on demand with little initial capital investment and operation complexity.

In IaaS cloud service, the location of a VM is fixed at an IaaS provider's data center. In supporting highly interactive network applications (NW-Apps) such as network games application that consists of one or more server-side applications and client-side applications, if the QoS of the NW-App's communication degrades because of some reasons (e.g., there is a significant physical distance between a server-side application and its client-side applications), it is difficult for an IaaS provider to provide a NW-App with a desired level of QoS such as low delay and high throughput.

QoS of NW-Apps in IaaS cloud service may improve by adopting the server migration service (SMS) [3] (also referred to as micro data centers [4]). In SMS, work places (WPs) that run virtual machines are deployed at various locations, and a virtual machine (server) that runs a server-side application of an NW-App can migrate between WPs towards the users of the NW-Apps to achieve better QoS for the users.

In SMS, an SMS provider tries to provide an NW-App provider with a certain level of QoS that they agree upon, and if the SMS provider fails, it pays penalty (e.g., reimbursement of a part of service charges to users) depending on the degree and length of the QoS violation. Thus, the SMS provider is incentivized to migrate servers between WPs to satisfy the agreed-upon QoS level to reduce the penalty. On the other hand, an SMS provider also needs to be moderate in performing server migrations to avoid degradation of network QoS (i.e., QoS of background traffic). This is because a server is typically large in size (e.g., the storage size of a VM in IaaS can be a few hundreds of gigabytes) and the server generates a large amount of traffic when it migrates, resulting in increasing delay and loss for its

ISBN: 978-1-61208-256-1

background traffic in a network.

This paper formulates an integer-programming model for the off-line server locations decision (i.e., when and to which WP server should migrate) where the penalty associated with NW-App's QoS violations is minimized, keeping the number and distance of server migrations below a given level. This paper also compares the minimum penalty obtained through solving the integer-programming model against the penalty obtained with a greedy on-line server locations decision algorithm, which migrates a server to a WP that minimizes the current penalty with no consideration of the penalty that will arise in the future.

Previous work related to server migration service includes server migration within a single data center [5] and migration of databases (DBs) [6]. The paper [5] proposes an algorithm which enables a server dynamically migrate among different physical hosts within a single data center according to their workloads. The algorithm proposed in [5] reduces the number of servers required to achieve a given server response time. However, the paper focuses on server migration within a single data center and does not need to consider degradation of network QoS due to the server migration. Our paper considers server migration across a network and considers degradation of network QoS due to the server migration unlike the paper [5]. The paper [6] proposes a DB-migration scheduling algorithm and achieves the shortest communication time between DB servers and their clients by optimally determining locations of DB servers for a given query sequence. However, degradation of network QoS due to a DB server migration is not considered. Our paper considers degradation of network QoS due to the server migration.

The rest of the paper is organized as follows. Section II explains an NW-App model and a network model as well as the server migration service that are considered in this paper. Section III describes how server locations are determined and formulates it as an integer programming model. In Section IV, the numerical examples are presented. Section V concludes the paper.

II. A MODEL FOR THE SERVER MIGRATION SERVICE

A. A model for a network application

As shown in Fig. 1, a NW-App consists of one or more server-side applications and client-side applications, and the former run on a virtual machine (hereafter referred to as "server") that is operated at WPs, while the latter run on a user-terminal (hereafter referred to as "client") such as a note PC and a smart phone. It is assumed that a server runs a single server-side application of NW-App and when a server-side application needs to migrate to the new WP, the server that runs the application also migrates to the new WP. It is assumed that server S_i (i = $1, 2, \cdots, n$) communicates with a pre-determined set of clients C_i^j (i = $1, 2, \cdots, n$, and j = 1, $2, \cdots, m_i$). If a NW-App operates multiple servers for some purpose such as load balancing, server S_i also communicates with other servers S_ks ($k \neq i$) in the same NW-App. It is assumed that clients do not communicate with other clients.

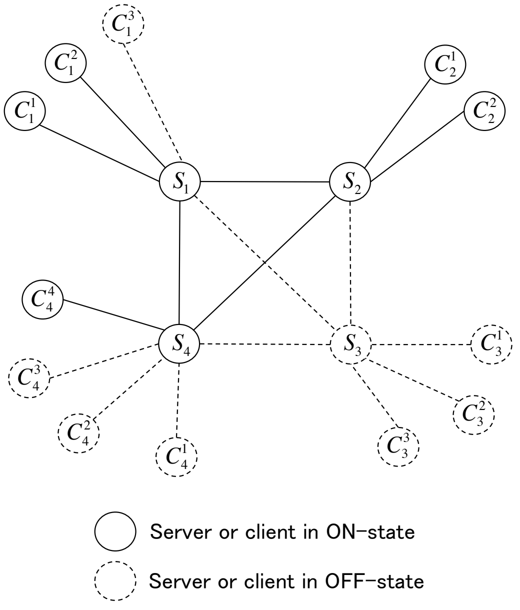


Figure 1. Communications in a network application

A server and a client each has two states, ON-state (i.e., running a server/client-side application) and OFF-state (i.e., not running a server/client-side application). The communication described above among servers and clients only occur when they are in ON-state. A client changes from OFF-state to ON-state when it starts to execute the client-side application, and it returns to OFF-state when its execution completes. A server is in ON-state only when at least one of its clients is in ON-state. For example, S_4 is in ON-state because its client C_4^4 is in ON-state, and S_3 is in OFF-state because all of its clients are in OFF-state.

B. A network model

Fig. 2 shows an example of a network considered in this paper. In Fig. 2, $R_1 \sim R_6$ are routers. Client C_i^j is connected to a router, and client-router association does not change throughout the time period of the interest of this paper. A work place (WP) is a physical computer and can operate a VM (server) that run a server-side application of a NW-App. A server in ON-state uses the resources of a WP (e.g., CPU and memory) for communication. Thus, the number of ON-state servers in a WP should be bounded due to the limitation of the resources (e.g., CPU time and memory space) of a WP. This paper assumes that the number of ON-state servers that a WP can support is bounded at a pre-determined threshold (refereed to as the capacity of a WP).

C. Server migration service

SMS can be classified into two service models: integrated model and overlay model depending on whether an SMS provider owns a network or not. In an integrated model, an

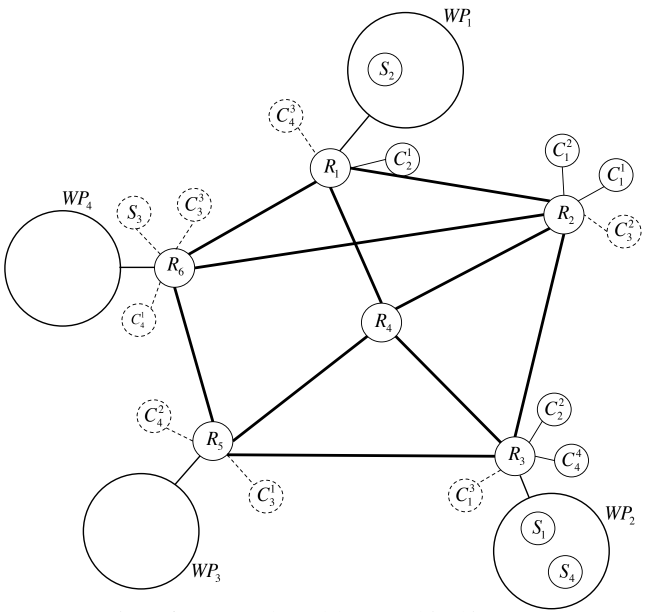


Figure 2. Network model assumed in this paper

SMS provider owns WPs and a network that connects among the WPs (e.g., carrier cloud with SMS), and consequently the objectives of the SMS provider will be 1) providing NW-Apps with good QoS and 2) keeping the network QoS good. In an overlay model, on the other hand, an SMS provider owns WPs only and rents network capacity from network operators to achieve the reachability among WPs, and consequently the objectives of the SMS provider will be 1) providing NW-Apps with good QoS and 2) reducing the network rental fees paid for network operators. Although our server locations decision model can cope with both the models, this paper focuses on the integrated model.

Fig. 3 shows SMS based on the integrated model. The intended customers of the SMS are NW-App developers who hope to run their NW-Apps with the desired level of QoS. In the SMS, prior to the start of the service, an NW-App developer and the SMS provider make an agreement regarding NW-App's QoS (such as communication delays between a server and a client, communication delays between servers, throughputs, and packet loss) of the service that the SMS provider provides the NW-App provider with. Based on the agreement, the NW-App developer pays fees to the SMS provider, and the SMS provider provides the NW-App developer with its service. If the NW-App's QoS is violated, the SMS provider pays the penalty for the NW-App developer. The penalty will be calculated based on the degree of the NW-App's QoS violation and its duration (i.e., how long the NW-App's QoS had been violated).

In order to provide QoS specified in the agreement and minimize the penalty, an SMS provider may migrate servers among WPs. Server migrations may be performed when the NW-App's QoS is violated (reactive migration) or in order to prevent the QoS from being violated (proactive migration). When a server migration is performed, the server needs to be migrated from a WP to another WP, creating additional traffic and resulting in possible QoS degradation of the network. Thus, in performing server migration, the SMS provider also needs to minimize the network QoS degradation caused by the additional traffic associated with server migration.

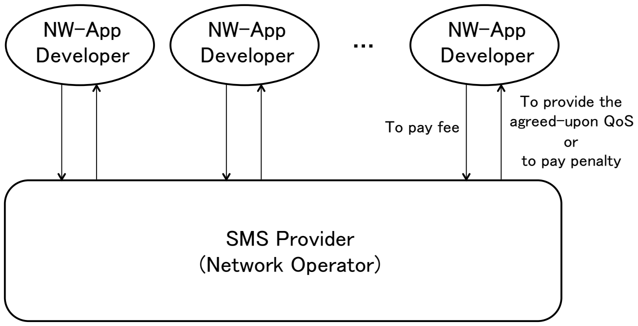

NW-App Developer
NW-App Developer
...
NW-App Developer
To pay fee
To provide the agreed-upon QoS or to pay penalty
SMS Provider (Network Operator)

Figure 3. Server migration service based on the integrated model

To decide when and to which WP servers migrate so that both the penalty and the network QoS degradation are minimized in SMS, two server locations decision approaches can be considered: off-line approach and on-line approach. An off-line approach makes decisions of server locations based on client ON/OFF state transitions in the past, present and future. It can be applied to NW-Apps whose clients show regular and easily predictable ON/OFF state transitions. The off-line approach achieves the optimal performance when the predicted state transitions are correct, while it can result in great performance degradation when the transitions are wrongly predicted. An on-line approach makes decisions of server locations without client ON/OFF state transitions in the future. It can be applied to any NW-Apps and achieves reasonable performance. As the first step for realizing the optimal SMS, this paper proposes an off-line server locations decision approach. Our approach is useful for NW-Apps with predictable client state transitions, and also serves as a benchmark for on-line approaches that cope with NW-Apps with unpredictable client state transitions.

III. Off-line approach for server locations decision

As discussed in Section II, it is important to minimize both the penalty associated with NW-App's QoS violation and the network QoS degradation caused by the additional traffic associated with server migrations. This section first considers the server locations decision as a simple discrete-time model. Then this section formulates an integer-programming model for the discrete-time server locations decision.

A. Discrete-time server locations decision

In order to simplify modeling of server locations decision, we consider a discrete-time model where state-transitions and server locations decision (i.e., when and to which WP servers should migrate) occur at discrete-time instances. This model is referred to as *the discrete-time server locations decision* in the rest of the paper.

In the model, time is slotted, and server and client status changes at the boundary of slots, and the server locations are also determined at the boundary of slots as depicted in Fig. 4. It is assumed that, once the new location (WP) is determined for a server, the server migrates to the WP instantaneously, i.e., there is no network delay associated with moving the server to the new WP. In calculating NW-App's QoS provided by the SMS provider and also the network QoS degradation, it is assumed that those QoSs are static within a slot, and they change only at the boundaries of a slot. These assumptions make it relatively easier to calculate NW-App's QoS that the SMS provider provides and the network QoS degradation that servers create when they migrate.

Because time slots are artificially introduced in the discrete-time model to approximate continuous time, it is important to carefully determine the size of a time-slot. When a slot is large, the deviation in the calculated NW-App's QoS and the network QoS degradation from the actual values will be large. When a slot size is small, frequency of the server locations decision increases, and consequently complexity (e.g., CPU time and memory space) for determining server locations will be high. The optimal length of a time slot is beyond the scope of this paper.

The total penalty refers to the sum of the penalties that arise in all time-slots. In order to calculate the total penalty, the penalty in a single slot is first calculated. The penalty in each slot depends on the NW-Apps' QoSs of communications between servers and their clients, as well as communications between servers. For each communication, *a penalty function* calculates the penalty based on the degree of the NW-App's QoS violation and its duration. In our model, any form of the function may be adopted under the agreement between the SMS provider and the NW-App developer.

The degree of network QoS degradation due to a server migration may be calculated based on factors such as the size of a server and the number of hops that a server takes to move to the new WP. *A network QoS degradation function* calculates the degree of the network QoS degradation. In our model, any form of the function may be adopted by the SMS provider. In order to keep the network QoS degradation below a predetermined level, the sum of the degrees of the network QoS degradations needs to be kept below a predetermined upper bound in a given period of time (i.e., called a network QoS degradation window) that consists of a given number of consecutive time-slots. For example, if the size of a network QoS degradation window is three (slots) and if the upper bound on the sum of the degrees of network QoS degradations in a window is ten in Fig. 4, the sum of the degrees of the network QoS degradations in every window must be less than or equal to ten.

B. Model formulation

Given a sequence of ON/OFF state transitions of all clients, the optimal server locations in every time-slot should be decided so that the total penalty is minimized while keeping the network QoS degradation below a predetermined level. Our model for the server locations decision is based on an integer programming and is described below.

- Parameters (Constants)

T : A set of consecutive slots (= $\{0, 1, 2, \cdots N\}$). Slots 1 to N are included in the time period where the SMS provider migrates servers while slot 0 stands for expressing the initial locations (WPs) of servers.

S : A set of servers (= $\{1, 2, \cdots n\}$).

L : A set of WPs (= $\{1, 2, \cdots r\}$).

R_i : A set of clients that server i supports.

C_i : Capacity of WP i (i.e., the number of servers that WP i can support).

Q_i^t : A binary constant that is equal to 1, if client i is in ON-state in slot t, and 0, otherwise.

P_{ij}^t : Penalty when server i stays at WP j in slot t (the penalty function calculates P_{ij}^t using Q_i^t).

W : A set of network QoS degradation windows (= $\{\{1, 2, \cdots, p\}, \{2, 3, \cdots, p+1\}, \cdots, \{N-p+1, N-p+2, \cdots, N\}\}$ where p is the length (slots) of the window).

U : The upper bound on the sum of the degrees of the network QoS degradations for a given network QoS degradation window.

I_{ij} : The degree of network QoS degradation when a server migrates from WP i to WP j (the network QoS degradation function calculates I_{ij}).

F_i : Initial location (WP) of server i.

- Variables

s_{ij}^t : A binary variable that is equal to 1, if server i stays at WP j in slot t, and 0, otherwise.

o_i^t : A binary variable that is equal to 1, if server i is in ON-state in slot t, and 0, otherwise.

Using the parameters and variables defined above, our model becomes to minimize the objective function (1) subject to (2)–(7) below.

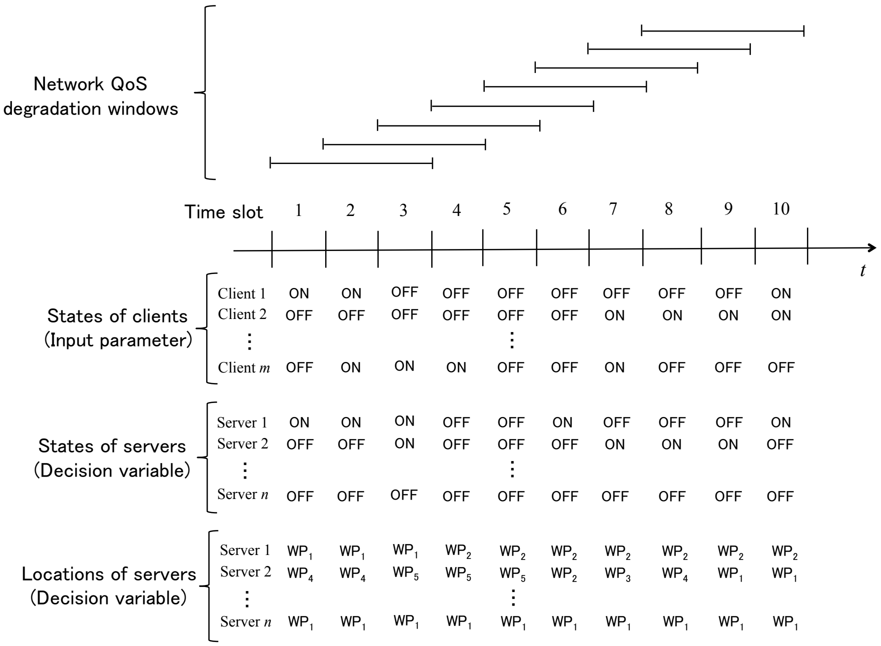
- Objective function: To minimize the total penalty.

$$\text{minimize} \sum_{t\in T}\sum_{i\in S}\sum_{j\in L} P_{ij}^t s_{ij}^t \qquad (1)$$

- Constraints
  - A server's location must fall within all WPs.

$$\sum_{j\in L} s_{ij}^t = 1 \quad \forall i \in S, \forall t \in T \qquad (2)$$

  - The number of servers that reside on a WP must



		1	2	3	4	5	6	7	8	9	10
States of clients (Input parameter)	Client 1	ON	ON	OFF	OFF	OFF	OFF	OFF	OFF	OFF	ON
	Client 2	OFF	OFF	OFF	OFF	OFF	OFF	ON	ON	ON	ON
	⋮					⋮					
	Client *m*	OFF	ON	ON	ON	OFF	OFF	ON	OFF	OFF	OFF
States of servers (Decision variable)	Server 1	ON	ON	ON	OFF	OFF	ON	OFF	OFF	OFF	ON
	Server 2	OFF	OFF	ON	OFF	OFF	OFF	ON	ON	ON	OFF
	⋮					⋮					
	Server *n*	OFF	OFF	OFF	OFF	OFF	OFF	OFF	OFF	OFF	OFF
Locations of servers (Decision variable)	Server 1	WP_1	WP_1	WP_1	WP_2	WP_2	WP_2	WP_2	WP_2	WP_2	WP_2
	Server 2	WP_4	WP_4	WP_5	WP_5	WP_5	WP_2	WP_3	WP_4	WP_1	WP_1
	⋮					⋮					
	Server *n*	WP_1	WP_1	WP_1	WP_1	WP_1	WP_1	WP_1	WP_1	WP_1	WP_1

Figure 4. An example of discrete-time off-line server locations decision.

be less than or equal to the WP's capacity.

$$\sum_{i\in S} o_i^t s_{ij}^t \le C_j \quad \forall j \in L, \forall t \in T \tag{3}$$

- Sum of the degrees of network QoS degradations during an network QoS degradation window must be less than or equal to the predetermined upper bound.

$$\sum_{t\in w}\sum_{i\in S}\sum_{j,k\in L} s_{ij}^{t-1} s_{ik}^t I_{jk} \le U \quad \forall w \in W \tag{4}$$

- A server is in ON-state, if one or more of its clients is in ON-state.

$$o_i^t \ge Q_j^t \quad \forall j \in R_i, \forall i \in S, \forall t \in T \tag{5}$$

- A server's initial WP is given.

$$s_{ij}^0 = 1 \quad j = F_i, \forall i \in S \tag{6}$$

$$s_{ij}^0 = 0 \quad j \ne F_i, \forall i \in S \tag{7}$$

IV. Numerical examples

In this section, we obtain the optimal performance with our integer programming model in section III, and compare it with the performance obtained through a simple greedy on-line algorithm.

A. Parameter settings

With the simple greedy on-line algorithm that we consider in this paper, a server selects the WP with the minimum penalty in the current slot from the candidate WPs that satisfy the network QoS degradation constraint. When there are multiple such candidates with the same minimum penalty, a server migrates to the WP that yields the minimum network QoS degradation when the server migrates to the new WP.

We use 14-node NSFNET (Fig. 5) as the network model. Every router is equipped with one WP with the capacity of one server. The propagation delays of links in this network varies between 1.4 and 11.2 [ms]. In the numerical examples in this section, it is assumed that the delay on each link is dominated by the propagation delay of the link, and the packet transmission time (i.e., packet length divided by channel speed of the link) and the queueing delay at a router are negligible. Negligible packet transmission time is realistic and justified, because channel speed of the link is huge and getting huger. So is negligible queueing delay at a router, because it is reported that a very small buffer (e.g., a buffer for 10–20 packets) is enough for core routers to achieve high TCP throughput [7]. Small buffer yields negligible queueing delay at routers.

In the scenario considered in this numerical result section, there is only one NW-App consisting of one server and 14 clients. 14 clients are uniformly distributed over 14 routers (i.e., one client per router). The SMS provider and the NW-App developer agree that end-to-end delay between the server and each client must be smaller than or equal to 10 ms. It is assumed that clients follow the exponential ON/OFF model where ON-state period and OFF-state period follow the exponential distributions with means μ_{ON} = 2 (slots) and μ_{OFF} = 10 (slots), respectively. The duration of the time period the SMS provider performs server migrations is set to 10 (slots). We consider a total of 100 ON/OFF-state transitions of clients to obtain the average of total penalties. IBM ILOG CPLEX Optimizer [8] is used to solve our integer programming model.

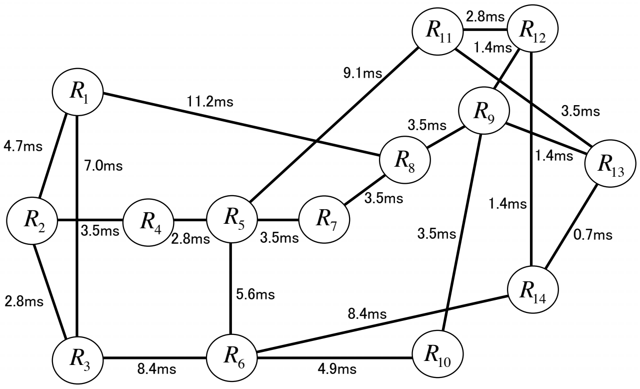


Figure 5. NSFNET

It is assumed that the penalty of the NW-App's QoS violation in one time-slot is proportional to the difference between the end-to-end delay between the server and a client and the predetermined threshold (i.e., 10 ms), i.e., the penalty function is α times the difference, where α is a constant. Note that the difference is regarded as zero when the end-to-end delay is smaller than or equal to the threshold value. In the numerical examples in section IV.B, we set the value of α to ten. As for the network QoS degradation function, we simply consider that the degree of network QoS degradation is equal to the number of hops that a server takes to move to the new WP (e.g., the server's migration on 2-hop counts route causes the network QoS degradation of two). We set the size of the network QoS degradation window to three (slots). It is assumed that a server and a client communicate using the shortest hop path between them. It is also assumed that the server migrates to the new WP using the shortest hop path. In the numerical result examples shown in section IV.B, the upper bound (U) on the sum of the network QoS degradation in a window varies one to ten.

B. Results

Figs. 6 and 7 depict the average total penalty as a function of the upper bound on the sum of the network QoS degradation in a window (U) with 95% confidential interval, when the server's initial locations are set to WP 1 connected to R_1 and WP 5 connected to R_5, respectively.

These figures show that the average total penalties of both our model and the greedy algorithm decrease as U increases. This is because the larger U enables the server to migrate more frequently and/or to the new WP that are further over more number of hops. Consequently, there is a larger possibility of a server finding the new WP that either avoids or reduces the penalty.

Figs. 6 and 7 show that, when U is less than or equal to three (when the degradation of network QoS is little acceptable), our model achieves 36% to 49% lower penalty than the greedy algorithm.

When U is larger than or equal to four, both our model and the greedy algorithm show nearly identical penalty. This is explained as follows. When the value of U is large, namely, when there is no tight upper bound on the network QoS degradation, even if the server migrate to almost any WP, it still meet the network QoS degradation constraint. As a result, with the greedy algorithm, a server often migrates to the WP with the minimum penalty, resulting in nearly identical total penalty as with our model.

We next explore the influence of the server's initial location on the average total penalty. Figs. 8 and 9 depict the average total delays of our model and the greedy algorithm as a function of the server's initial WP. In the figures, the larger the value of U becomes, the smaller the difference of the average total penalties among different server's initial WPs. This is because the larger U leads to extending a server's moving range limited by its initial WP.

V. CONCLUSION AND FUTURE WORK

In this paper, we formulated an integer programming model for a discrete-time off-line server locations decision in the server migration service, and derived the optimal server locations. Numerical examples showed that 1) our model achieves 36% to 49% smaller penalty than the greedy on-line algorithm when the degradation of network QoS due to server migrations is little acceptable and 2) the greedy on-line algorithm can achieve the optimal or near optimal performance when an upper bound on the network QoS degradation is large (i.e., when the network QoS degradation constraint virtually does not exist).

Our future work includes 1) investigation of the optimal length of a time-slot and 2) design of an on-line server locations decision algorithm that achieves a performance close to our integer programming model because the computational complexity of our integer programming model can be large in a practical situation.

REFERENCES

[1] M. Armbrust et al., "A view of cloud computing," Communications of the ACM, vol. 53, Apr. 2010, pp. 50–58.

[2] "Amazon EC2." http://aws.amazon.com/ec2 [retrieved: Jan. 2013].

[3] A. Yamanaka, Y. Fukushima, T. Murase, T. Yokohira, and T. Suda, "Destination selection algorithm in a server migration service," in Proceedings of the 7th International Conference on Future Internet Technologies (CFI), Sept. 2012, pp. 15–20.

[4] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: Research problems in data center networks," in Proceedings of ACM SIGCOMM'08, Jan. 2009, pp. 68–73.

[5] S. Ranjan, J. Rolia, H. Fu, and E. Knightly, "QoS-driven server migration for Internet data centers," in Proceedings of tenth International Workshop on Quality of Service (IWQoS), May 2002, pp. 3–12.

[6] T. Hara, M. Tsukamoto, and S. Nishio, "A scheduling method of database migration for WAN environments," in Proceedings of Brazilian Symposium on Database (SBBD), Oct. 1999, pp. 125–136.

[7] M. Enachescu, Y. Ganjali, A. Goel, N. McKeown, and T. Roughgarden, "Part III: Routers with very small buffers," SIGCOMM Computer Communication review, vol. 35, July 2005, pp. 83–90.

[8] "IBM ILOG CPLEX Optimizer." http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/ [retrieved: Jan. 2013].

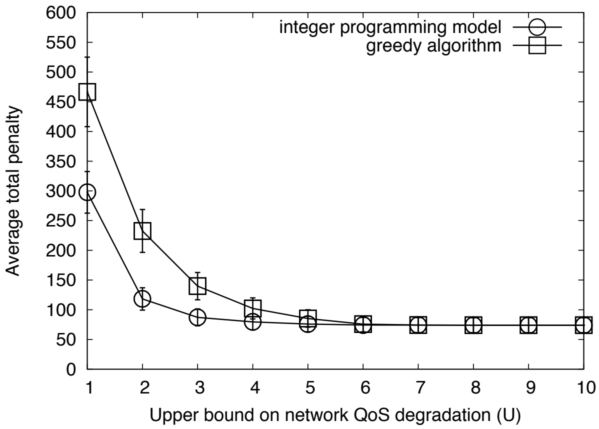


Figure 6. Average total penalty (server's initial WP: WP 1).

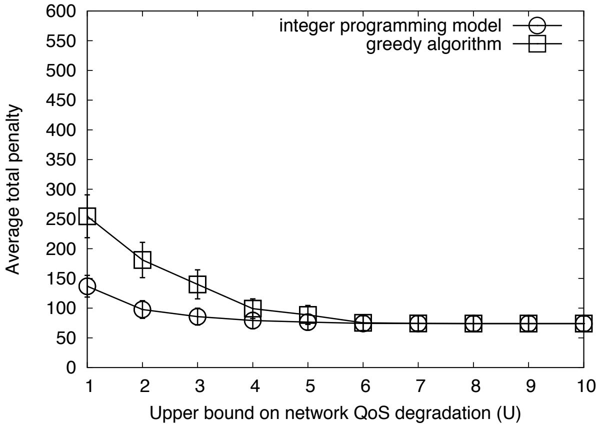


Figure 7. Average total penalty (server's initial WP: WP 5).

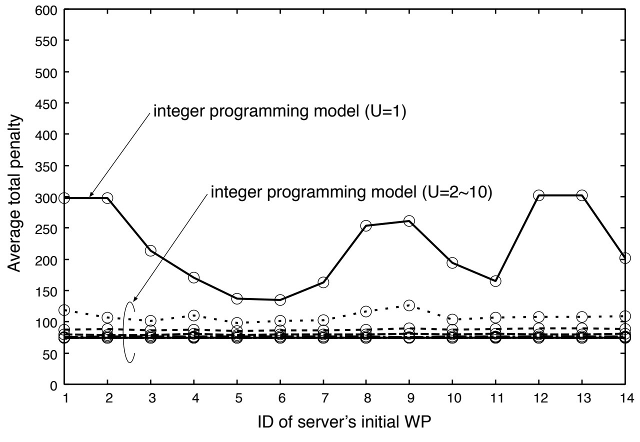


Figure 8. Average total penalty as a function of server's initial WP (integer programming model).

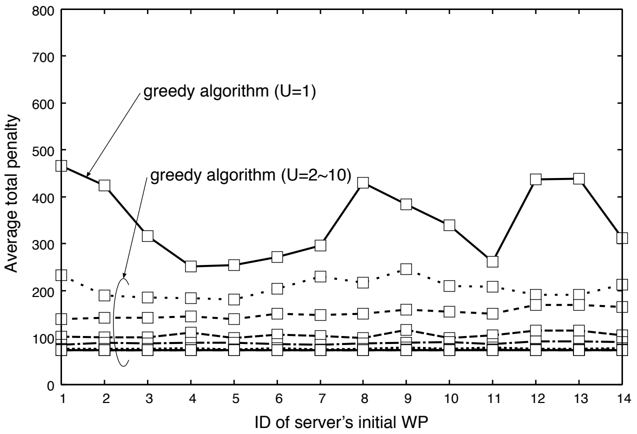


Figure 9. Average total penalty as a function of server's initial WP (greedy algorithm).

SecDEv: Secure Distance Evaluation in Wireless Networks

Gianluca Dini, Francesco Giurlanda, Pericle Perazzo
Dept. of Information Engineering
University of Pisa
Email: [name.surname]@iet.unipi.it

***Abstract*—The problem of measuring the distance between two electronic devices in the presence of an adversary is still open. Existing approaches based on *distance-bounding protocols* are subject to *enlargement attacks* that cause the target to be perceived farther than it actually is. Enlargement attacks represent a new challenge for the research field of secure localization. The contribution of this paper is twofold. First, we propose SecDEv, a secure distance-bounding protocol for wireless channels that withstands enlargement attacks based on jam-and-replay. By leveraging on the characteristics of radio frequency signals, SecDEv establishes a *security horizon* within which a distance is correctly measured and a jam-and-replay attack is detected. Second, we show how SecDEv improves the scalability of secure positioning techniques.**

***Keywords*-secure localization; secure positioning; distance-bounding protocols; distance enlargement attacks**

I. Introduction

The measurement of the distance between two electronic devices is crucial for many practical applications. Many techniques have been proposed over the years [1]. All these techniques fail in the presence of an adversary that wants to disrupt the distance measurement process. Even the well-known and widespread civilian Global Positioning System (GPS) is extremely fragile in adversarial scenarios [2]. Secure location estimation has a plethora of applications including coordination of autonomous guided vehicles [3] and geographical routing [4]. For all these applications, an insecure distance or position estimation could produce security problems such as unauthorized accesses, denial of service, thefts, integrity disruption with possible safety implications and intentional disasters.

Desmedt [5] first introduced the problem of secure location verification and showed that it cannot be solved by solely using cryptography. Brands and Chaum [6] proposed the first *secure distance-bounding* protocol. Since then, many variants have been proposed in the literature [7], [8]. These protocols leverage on both the unforgeability of authenticated messages and the upper bound of the communication speed that is the speed of light. They prevent *distance reduction*, i.e., an adversary cannot make a device appear closer than it really is. The resistance against distance reduction is an important requirement for all the application scenarios involving secure proximity verification [9], [10], [8]. A common example is the problem of proximity-based access control. Let us suppose an RFId card performing an authentication protocol with a reader. If the card correctly performs the protocol, the reader will open a door of a building. An adversary can trick the system by establishing a relay link between the reader and a far away legitimate card, owned by an unaware user. The card correctly performs the authentication protocol via the relay link, and the reader opens the entrance. This attack is known as *mafia fraud*. Along with the correctness of the authentication, the reader has to check even that the card is within a security distance. However, if such a distance measurement is made with insecure methods, the adversary can still break the system. In particular she can perform a distance reduction attack to deceive the reader into believing that the far away card is in the proximity.

The relevance of the secure proximity verification eclipsed the dual problem: the *distance enlargement* attack. By this attack, an adversary makes a device appear farther than it really is. The resistance against both reduction and enlargement attacks is important whenever we want to securely estimate a distance, rather than a proximity. Let us suppose a distributed system that monitors the movement of autonomous guided vehicles. The system relies on distance information to avoid collisions between vehicles. An example of such systems is in [3]. If an adversary is able to make a distance appear larger than it really is, the system could not take collision-avoidance countermeasures in time. This could cause collisions between vehicles, and consequent loss of money and safety threats. Secure distance estimations are extremely useful in trilateration techniques too. These techniques use the distances measurements from at least three anchor nodes, whose positions are known, to estimate the position of a fourth node. If an adversary can enlarge one or more distance measurements, she is able to disrupt the whole positioning process.

In this paper we propose SECure Distance EValuation (SecDEv), a distance-bounding protocol able to resist to enlargement attacks based on jam-and-replay tactics [11], [12], [13]. SecDEv exploits the characteristics of wireless signals to establish a *security horizon* within which a distance can be correctly evaluated (besides measurement errors) and any adversarial attempt to play a jam-and-replay attack is detected. We also show how SecDEv improves the scalability of secure positioning techniques in terms of number of anchor nodes.

The remainder of this paper is organized as follows. In Section II we present related works. In Section III we introduce a reference distance-bounding protocol. In Section IV we define the threat model. In Section V we introduce SecDEv as an improvement of the reference distance bounding. In Section VI we show how SecDEv improves the performance of secure positioning techniques. Finally, we draw our conclusions in Section VII.

II. Related Works

Secure localization has a vast applicability in many technological scenarios, but it has showed to be a nontrivial problem. The silver bullet is yet to be found.

Brands and Chaum [6] proposed distance-bounding protocols, in which a *verifier* node measures the distance of a *prover* node. Distance-bounding protocols do not determine the actual distance, but rather a secure upper bound on it. In this way, the actual distance is assured to be shorter or equal to the measured one, even in presence of an adversary. These protocols were created to assure the physical proximity between two devices, and consequently to contrast *mafia fraud* attack [5].

Hancke and Kuhn [8] fitted distance bounding protocols for RFId tags. Their proposal deals with a variety of practical problems such scarce resources availability, channel noise and untrusted external clock source. Though extensions for RFId's are possible, we focus on more resourceful devices. We assume the clock source is internal and trusted and the channel noise is corrected by FEC techniques.

Clulow et al. [14] focused on a wide variety of low-level attacks, which leverage on packet latencies (e.g. preambles, trailers, etc.) and symbols' modulations. PHY-layer preambles are sent before the cryptographic quantities, in order to permit the receiver to synchronize itself to the sender's clock. The preamble of the response is fixed and does not depend on the content of the challenge. A dishonest prover could thus anticipate the transmission of the response preamble to reduce the measured distance. To deal with this problem, Rasmussen and Čapkun [15] proposed full-duplex distance bounding protocols, in which the challenge and the response are transmitted on separate channels. The prover receives the challenge and meanwhile transmits the response. In this way, a dishonest prover cannot anticipate the transmission of the response, without having to guess the payload. In the present paper, we assume the prover to be honest. This permits us to simplify our reference distance-bounding protocol (cfr. Section III). In particular we use a single channel in a half-duplex fashion.

Flury et al. [10] and, more in depth, Poturalski et al. [16] analyze the PHY-protocol attacks against impulse-radio ultra-wideband ranging protocols (IR-UWB), with particular attention to 802.15.4a [17], which is the *de facto* standard. These studies concentrate only on reduction attacks, and estimate their effectiveness in terms of meters of distance reduction. We instead focus on the opposite problem, distance enlargement, which requires different countermeasures.

Chiang et al. [18] proposed the first technique able to mitigate the enlargement attack in case of dishonest prover. The verifier makes two power measurements of the prover's signal on two collinear antennas. Subsequently, it computes the difference of the two measurements. Given the standard path-loss model, if the difference is low, the signal source will be far away. Otherwise it will be near. The idea is that the adversary cannot modify the way the signal attenuates over the distance, thus the distance estimation is trusted. Obviously such proposal relies on the standard path-loss model, which is poorly reliable. The authors claim that if the path loss exponent varies between 2 and 4, an enlargement of more than twice the measured distance is impossible. In this paper, we focus on external adversaries. The problem of distance enlargement in presence of internal ones is challenging as well, but falls outside our present scope.

III. Reference Distance-Bounding Protocol

A distance-bounding protocol allows a *verifier* (V) to "measure" the distance of a *prover* (P). In its basic form, a distance-bounding protocol consists in a sequence of single-bit challenge-response rounds [6]. In each round, the verifier sends a challenge bit to the prover that replies immediately with a response bit. The round-trip time enables V to compute an upper-bound of the P distance. Then, the distance is averaged on all rounds. Many variants of distance-bounding protocols have been proposed in the literature [7], [8]. Here, we establish a *reference distance-bounding protocol*, similar to those described in [16] for external adversaries. It involves a *request* message (REQ) from the verifier, an *acknowledgment* message (ACK) from the prover, and a final *signature* message (SGN) from the prover. Such a reference protocol is vulnerable to jam-and-replay attacks, as we will show in Section IV, and SecDEv (cfr. Section V) will overcome these vulnerabilities.

The request and the acknowledgement convey, respectively, a and b, which are two independent, random and unpredictable sequences of bits. Note that, differently from the original version of distance-bounding protocol, the request and the acknowledgement are frames, rather than single bits. In fact, it is hard to transmit single bits over an IR-UWB channel. This is due to TLC regulation, which poses strict limits to the transmission power. In 802.15.4a [17], for example, every packet is preceded by a multi-bit synchronization preamble. The signature authenticates the acknowledgement and the request by means of a *shared secret* S. What follows is a formal description of the protocol.

REQ $\mathrm{V} \longrightarrow \mathrm{P} : a$

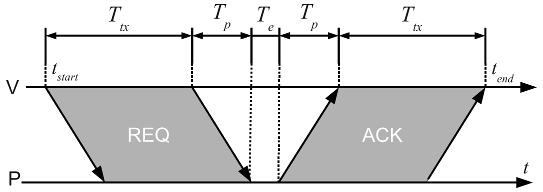


Figure 1. Round-trip time.

ACK P $\longrightarrow$ V : b
SGN P $\longrightarrow$ V : $H_S(a, b)$

The quantities a, b and $H_S(\cdot)$ are k-bit long. Therefore, the probability for an adversary to successfully guess one of these quantities is 2^{-k}. Such a probability gets negligible for a sufficiently large value of k, which we call the *security parameter*.

The verifier measures the distance between itself and the prover, by measuring the round-trip time $\hat{T}$ between the request and the acknowledgement messages. With reference to Fig. 1, we denote by t_{start} the instant when the transmission of REQ begins, and by t_{end} the instant when the reception of ACK ends. We denote by T_e the time interval from the end of REQ reception, to the beginning of ACK transmission. Since ACK does not depend on REQ, T_e does not include any elaboration time. It includes only the time for the antenna to switch from the receive mode to the transmit mode and the necessary hardware delays. We assume T_e to be small and known. Dedicated hardware can fulfill these requirements. We further denote by T_{pkt} the transmission time of the request and acknowledgement messages, and with T_p their propagation time in the medium. The round-trip time will be:

$$\hat{T} = 2T_p = (t_{end} - t_{start}) - 2T_{pkt} - T_e \tag{1}$$

Finally, we obtain a measure of the distance:

$$\hat{d} = \frac{c \cdot \hat{T}}{2} \tag{2}$$

where c is the speed of light.

The distance measurement precision depends on the capability of measuring the time interval with nanosecond precision. Localization systems based on IR-UWB can achieve nanosecond precision of measured time of flight, and consequently a distance estimation with an uncertainty of 30 cm. Also, this feature of time precision are available only with dedicated hardware.

IR-UWB protocols like 802.15.4a provides packets made up of two parts: a preamble and a payload. The preamble permits the receiver to synchronize to the transmitter and to precisely measure the time of arrival of the packet. The payload carries the information bits. In our protocol, a and b are transmitted in the payload part. We suppose the last part of the payload to carry a forward error correction code (FEC), for example some CRC bits.

In a non-adversarial scenario, the *actual distance* d will be equal to the *measured distance* $\hat{d}$. To deceive the measurement process, the adversary has to bring the verifier to measure a fake round-trip time. That is, she must act in a way that the verifier receives the acknowledgement at a different instant of time, while still receiving the correct signature. The basic idea of distance-bounding protocol is that an external adversary cannot deliver a copy of the legitimate acknowledgement *before* than the legitimate one.

On the other hand, she can deliver a copy of the acknowledgement *after* the legitimate one. In other words, she can only *enlarge* the measured distance, not *reduce* it. Thus, we are always sure that $d \leq \hat{d}$, i.e., the measured distance is a secure upper bound for the actual distance.

IV. Threat Model

We assume that the adversary (M) is an external agent, meaning that she does not know the shared secret (S) ant it cannot be stolen. Techniques like trusted hardware and remote attestation can help defending against these possibilities [19]. The objective of M is to deceive the verifier into measuring an enlarged round-trip time:

$$\hat{T} = 2T_p + \Delta T \tag{3}$$

in order to make it infer an enlarged measured distance:

$$\hat{d} = \frac{c \cdot \hat{T}}{2} = d + \frac{c \cdot \Delta T}{2} \tag{4}$$

We do not deal with distance reduction attacks. Since our protocol is an enhancement of the reference distance-bounding protocol of Section III, it offers the same guarantees against distance reduction attacks.

A. Adversary's Capabilities

M can eavesdrop, transmit or jam any signal in the wireless channel. The principle of a jammer is to generate a radio noise at a power comparable or higher than the legitimate one. In case of IR-UWB channels, a jammer could send periodic UWB pulses, in such a way to disrupt the synchronization process [20]. Alternatively, she could simply send random pulses in the payload part, in such a way the receiver discards the packet as corrupted after the FEC test. In both cases, the goal of the jammer is to disrupt the reception of the message.

M can transmit or jam *selectively*, in such a way that only a target node receives. In the meanwhile, M can correctly eavesdrop other signals. To do this, she can place a transmitting device nearby the receiver, and a listening one nearby the transmitter. Alternatively, she can use a single device with two directional antennas. One of them transmits to the receiver, while the other listens to the transmitter.

Another possibility is the *overshadowing* attack. In this attack, M injects a fake signal with higher power than the original one. The original signal becomes entirely overshadowed by the attacker's signal. Ideally, original signal is treated as noise by the receiver. In this paper, we do not deal with this attack, and we focus only with jam-and-replay attacks. The overshadowing attack is indeed interesting and deserves a full analysis, that we are planning to do in future work. Here we only points out that it is not simple to be performed in a real-world IR-UWB protocol. In fact, the verifier does not receive only the fake signal, but the legitimate signal too. Even if the former is much stronger in power, the latter is still a valid IR-UWB signal, which interferes with the packet synchronization and reception. Sending an overshadowing signal is probably not enough. The adversary should also attenuate the legitimate signal with some complementary technique, such as electro-magnetic shields or similar.

We assume that M has no physical access to the prover or the verifier. This has two consequences: (i) she cannot tamper with the nodes and steal their secret material, and (ii) she cannot attenuate the wireless signals with electro-magnetic shields or Faraday cages.

B. Jam-and-Replay Attacks

In the distance-bounding protocol of Section III, the adversary can enlarge the measured round-trip time in the following way (Fig. 2a).

1) M listens to the radio channel, until she hears a REQ signal.
2) M waits for the ACK signal.
3) M jams the ACK signal and eavesdrop it in the meanwhile.
4) After a time ΔT, M replays it.

The adversary must replay the ACK signal selectively, in such a way that only the verifier receives it. Otherwise, the prover will also receive the replayed signal, and could infer that the protocol is under attack.

It is important to highlight that M has to wait for the legitimate ACK to end, before starting the transmission. This is because she must avoid signal collision.

The adversary can perform a similar attack on the REQ signal (Fig. 2b). Even in this case, M has to wait for the end of the legitimate REQ before starting her transmission.

We state the following:

Proposition 1 *In a jam-and-replay attack on REQ/ACK, the adversary must enlarge the round-trip time of a quantity ΔT not smaller than T_{pkt}, i.e., $\Delta T \geq T_{pkt}$.*

Proposition 1 represents the fundamental limitation of the jam-and-replay attacks. SecDEv will leverage on this to withstand them. Note that this limitation comes from the properties of the radio-frequency channel, and does not depend on how many devices the adversary controls. For the sake of simplicity, Figg. 2a and 2b show a single adversary.

V. SecDEv Protocol

SecDEv is a distance-bounding protocol, which measures the correct distance between a verifier V and a prover P in presence of an adversary M performing a jam-and-replay attack. It is similar to the reference distance-bounding protocol (cfr. Section III), except that the length of REQ and ACK do not depend only on the security parameter, but also on a *security horizon*.

Let us consider the Equation 3 for a general enlargement attack and apply the Proposition 1, we obtain the constraint $\hat{T} \geq 2T_p + T_{pkt}$. Hence:

$$\hat{T} \geq T_{pkt} \tag{5}$$

Equation 5 assures us that a measured round-trip time smaller than T_{pkt} has not been affected by any jam-and-replay attack. We can translate T_{pkt} in a distance d_M, that we call *security horizon*:

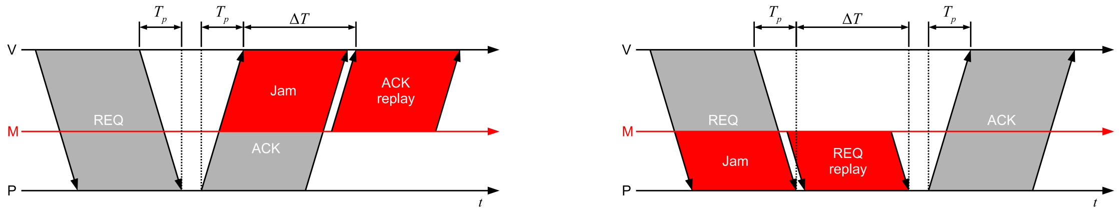
$$d_M \triangleq \frac{cT_{pkt}}{2} \tag{6}$$

In terms of distances, Equation 5 becomes:

$$\hat{d} \geq d_M \tag{7}$$

Equation 7 is our test to distinguish between trusted and untrusted distance measurements. V can extend the packet transmission time to enlarge the security horizon (cfr. Eq. 6), in order to securely measure longer distances. T_{pkt} is enlarged by introducing padding bits after the nounce. Padding bits have not to be unpredictable. They can have a well-known value (e.g. all zeroes), since they serves only to prolong the packet transmission time. V decides on the length of the REQ padding, and P has to respond with the same padding length in the ACK. Therefore, both messages have the same length, to withstand both jam-and-replay on REQ and on ACK.

Let us explain the protocol in detail. We assume that the wireless channel is characterized by the parameter tuple: $\{T_{pre}, R_{pld}, T_e\}$. T_{pre} is the transmission time of the preamble part. R_{pld} is the bit rate of the payload part. T_e is the reaction time of the prover node. In addition, we define the following triplet of protocol parameters: $\{k, S, d_M\}$. k is the security parameter. A higher value for k implies a higher security level, but has an impact on power consumption, as we will see in the following. S is a secret bit sequence shared between V and P. Its length is longer than or equal to k. d_M is the security horizon that distinguishes between trusted and untrusted measured distances. If the actual distance d is longer than d_M, the measured distance cannot be trusted because it may be affected by a jam-and-replay attack. In such a case, the protocol can be executed again with a longer d_M. Alternatively, the distance d can be first estimated in an insecure manner, and then securely confirmed with $d_M > d$.



(a) Jam-and-replay on ACK.

(b) Jam-and-replay on REQ.

Figure 2. Jam-and-replay attack.

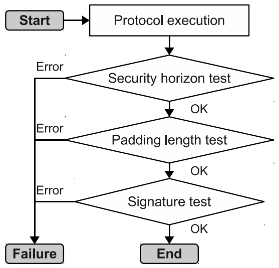


Figure 3. SecDEv algorithm.

A higher value for d_M allows us to measure longer distances, but has an impact on power consumption.

We further define the following quantities. N_{pad} and N_{fec} are respectively the number of bits of the padding and the FEC code. Since the number of bits of a and b is k, the total transmission time will be:

$$T_{pkt} = T_{pre} + (k + N_{pad} + N_{fec})/R_{pld} \tag{8}$$

If with $N_{pad} = 0$, the T_{pkt} identifies the minimum value of d_M. Thus, if the actual distance is smaller than this value, there is not need of padding bits. Otherwise, we determine N_{pad} with the following formula:

$$N_{pad} = \left\lceil \left(\frac{2d_M}{c} - T_{pre} \right) \cdot R_{pld} \right\rceil - k - N_{fec} \tag{9}$$

Using the Equation 9, we can set every value of d_M. Note that T_{pkt} grows with d_M. A larger security horizon causes longer messages, accordingly higher energy consumptions per protocol execution. An implementer must choose d_M as a trade-off between ranging capabilities and power consumption.

Fig. 3 shows the algorithm executed by V. After the protocol execution, V tests whether the measured distance is within the security horizon, that is, if $\hat{d} < d_M$. If this test fails, the measured distance is discarded as untrusted. Then, V tests the length of the ACK padding. If it contains less bits than the REQ one, the measured distance is discarded as untrusted. This is to avoid a jam-and-replay attack on REQ (cfr. Fig. 2b), in which M tries to lower ΔT by replaying REQ with a smaller padding. In such a case, P will respond with an ACK with a smaller padding too, and the attack will not pass the padding length test. Finally, V tests the validity of the cryptographic signature.

VI. Experimental Results

We combined SecDEv with multilateration technique to securely localize the prover. We analyzed the efficiency of this solution in terms of covered area and we compared it with *verifiable multilateration* [12], which is the state-of-the-art technique for secure positioning in wireless networks. Verifiable multilateration involves at least three distance measurements from different verifiers. The distance measurements are performed by means of distance bounding protocols, which are supposed to withstand reduction attacks. Verifiable multilateration deals with possible enlargement attacks by forcing an additional check to the final position estimation. In order to be trusted, the position must be inside the polygon formed by the verifiers, otherwise it is discarded as untrusted. Intuitively, this reduces the coverage area of the positioning technique.

In other words, classic multilateration is more scalable in terms of number of verifiers needed to cover a specific area. To quantify this, we have tested the performance of classic multilateration in terms of number of verifiers needed to cover a working area, and we have compared our results with those of verifiable multilateration, taken from [12]. We supposed that every verifier covers a circular area with radius $250\,\mathrm{m}$.

We neglect planned distributions [12], because in a real deployment, environment may impose constraints on the verifier positioning. Thus, we consider that the verifiers are uniformly distributed over the area of interest.

In order to evaluate the two techniques under the same conditions, our simulation were performed on areas of variable sizes. The verifiers were uniformly distributed in the area and in a boundary region outside the area, whose width was 10% of the area width. We use the boundary region to avoid the boundary effects [12] in the verifiable multilateration.

Fig. 4 shows how many verifiers are required to cover 95% and 90% of the working area. *VM* and *CM* curves are respectively verifiable multilateration with distance bounding and classic multilateration with SecDEv. The number of verifiers is the average of 100 simulations with confidence

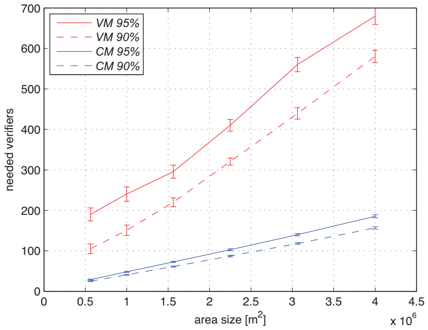


Figure 4. Verifiers required to cover an area.

intervals of 95% calculated for different values of working area from $0.5km^2$ to $4km^2$. The chart shows that classic trilateration needs far less verifiers, because it has not the limitation of the verification triangles. This gives strong motivation to fight distance enlargement attacks.

VII. Conclusions

We proposed SecDEv (SECure Distance EValuation), a distance-bounding protocol able to resist to enlargement attacks based on jam-and-replay tactics. SecDEv exploits the characteristics of wireless signals to establish a security horizon within which any adversarial attempt to play a jam-and-replay attack is detected. We also showed how SecDEv improves the scalability of secure positioning techniques in terms of number of anchor nodes.

References

[1] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems," *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 6, pp. 1067–1080, Nov. 2007.

[2] R. G. Johnston, "Think GPS cargo tracking = high security? think again," Los Alamos National Laboratory, Tech. Rep., 2003.

[3] G. Dini, F. Giurlanda, and L. Pallottino, "Neighbourhood monitoring for decentralised coordination in multi-agent systems: A case-study," in *Computers and Communications (ISCC), 2011 IEEE Symposium on*, 2011, pp. 681–683.

[4] Y. Yu, R. Govindan, and D. Estrin, "Geographical and energy aware routing: a recursive data dissemination protocol for wireless sensor networks," UCLA Computer Science Department, Tech. Rep., 2001.

[5] Y. Desmedt, "Major security problems with the 'unforgeable' (Feige)-Fiat-Shamir proofs of identity and how to overcome them," *SecuriCom*, pp. 15–17, 1988.

[6] S. Brands and D. Chaum, "Distance bounding protocols," in *EUROCRYPT'93*, 1993, pp. 344–359.

[7] L. Bussard and W. Bagga, "Distance-bounding proof of knowledge to avoid real-time attacks," *IFIP/SEC*, pp. 223–238, 2005.

[8] G. P. Hancke and M. G. Kuhn, "An RFId distance bounding protocol," in *Proceedings of IEEE/Create-Net SecureComm 2005*, I. C. S. Press, Ed., 2005, pp. 67–73.

[9] A. Francillon, B. Danev, and S. Capkun, "Relay attacks on passive keyless entry and start systems in modern cars," in *NDSS*, 2011.

[10] M. Flury, M. Poturalski, P. Papadimitrios, J.-P. Hubaux, and J.-Y. Le Boudec, "Effectiveness of distance-decreasing attacks against impulse radio ranging," in *Proceedings of the third ACM conference on Wireless network security (WiSec2010)*, 2010, pp. 117–128.

[11] J. Kong, Z. Ji, W. Wang, M. Gerla, R. Bagrodia, and B. Bhargava, "Low-cost attacks against packet delivery, localization and time synchronization services in under-water sensor networks," in *Proceedings of the 4th ACM workshop on Wireless security*, ser. WiSe '05, 2005, pp. 87–96.

[12] S. Čapkun and J.-P. Hubaux, "Secure positioning in wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 2, pp. 221–232, feb 2006.

[13] N. Tippenhauer and S. Čapkun, "ID-based secure distance bounding and localization," in *Computer Security – ESORICS 2009*, ser. Lecture Notes in Computer Science, M. Backes and P. Ning, Eds. Springer Berlin / Heidelberg, 2009, vol. 5789, pp. 621–636.

[14] J. Clulow, G. P. Hancke, M. G. Kuhn, and T. Moore, "So near and yet so far: Distance-bounding attacks in wireless networks," in *Proceedings of European Workshop on Security and Privacy in Ad-Hoc and Sensor Networks (ESAS)*, 2006, pp. 83–97.

[15] K. B. Rasmussen and S. Čapkun, "Location privacy of distance bounding protocols," in *Proceedings of the 15th ACM conference on Computer and communications security*, ser. CCS '08. ACM, 2008, pp. 149–160.

[16] M. Poturalski, M. Flury, P. Papadimitrios, J.-P. Hubaux, and J.-Y. Le Boudec, "Distance bounding with IEEE 802.15.4a: Attacks and countermeasures," *IEEE Transactions on Wireless Communications*, pp. 1334–1344, 2011.

[17] Z. Sahinoglu and S. Gezici, "Ranging in the IEEE 802.15.4a standard," in *Proceedings of IEEE Wireless and Microwave Technology Conference*, 2006, pp. 1–5.

[18] J. T. Chiang, J. J. Haas, J. Choi, and Y.-c. Hu, "Secure location verification using simultaneous multilateration," *IEEE Transactions on Wireless Communications*, vol. 11, no. 2, pp. 584–591, feb 2012.

[19] W. Hu, H. Tan, P. Corke, W. C. Shih, and S. Jha, "Toward trusted wireless sensor networks," *ACM Transactions on Sensor Networks*, vol. 7, no. 1, pp. 1–25, aug 2010.

[20] M. Poturalski, M. Flury, P. Papadimitrios, J.-P. Hubaux, and J.-Y. Le Boudec, "The cicada attack: degradation and denial of service in ir ranging," in *Proceedings of 2010 IEEE International Conference on Ultra-Wideband*, 2010, pp. 1–4.

A Hybrid Model for Network Traffic Identification Based on Association Rules and Self-Organizing Maps (SOM)

Zuleika Nascimento Djamel Sadok Stênio Fernandes
Informatics Center
Federal University of Pernambuco - UFPE
Recife, Brazil
{ztcn, jamel, sflf}@cin.ufpe.br

***Abstract*—Considerable effort has been made by researchers in the area of network traffic classification, since the Internet grows exponentially in both traffic volume and number of protocols and applications. The task of traffic identification is a complex task due to the constantly changing Internet and an increase in encrypted data. There are several methods for classifying network traffic such as known ports and Deep Packet Inspection (DPI), but they are not effective since many applications constantly randomize their ports and the payload could be encrypted. This paper proposes a hybrid model that makes use of a rule-based model along with a self-organizing map (SOM) model to tackle the problem of traffic classification without making use of the payload or ports. The proposed method also allows the generation of association rules for new unknown applications and further labeling by experts. The proposed hybrid model was superior to a rule-based model only and presented a precision of over 94% except for eMule application. The model was validated against a Measurement and Analysis on the WIDE Internet (MAWI) trace and presented true positive results above 99% and 0% false positives. It was also validated against another model based on computational intelligence, named Realtime, and the hybrid model proposed in this work presented better results when tested in real time network traffic.**

***Keywords*-Association Rules; Self-Organizing Maps; Network Traffic Measurement; Genetic Algorithms.**

I. Introduction

In recent years, the research effort toward network traffics identification has been growing [1] [2] [3] [4] [5]. As the Internet grows exponentially in both traffic volume and number of protocols and applications, it is essential to understand the composition of dynamic traffic characteristics to recognize protocols and applications which are often encrypted.

In this context, identifying traffic that passes over a network is a complex task, since access to the Internet is significantly increasing, bringing with it new users with different goals. Many peer-to-peer (P2P) applications are increasingly popular and accessible, such as eMule, Ares, and BitTorrent. The users behavior is also changing and the growth of streaming video services is notable [5], since Skype, MSN (Messenger), and other instant message services, along with sites that allow their users to upload and share videos in digital format, have become commonplace. To bring to the experts attention what passes through a network is an increasingly important activity.

There are several methods for classifying network traffic as known ports and Deep Packet Inspection (DPI) [6] [7]. The classification method based on ports performs an analysis of port numbers and is employed to identify applications or protocols. This technique proves to be quite ineffective, since most of the applications make use of random ports. The payload inspection technique or DPI, in turn, eliminates the problem of using random port number used for a specific application or protocol. The technique works starting with a classifier that extracts the payload from TCP/UDP packets and scans each packet in search of signatures that can identify the flow type. However, this technique does not work correctly in encrypted traffic data.

Recently, some methodologies have been investigated as network traffic classification tools. The work presented in [8] demonstrates the use of data mining techniques to classify flow and user behavior profiles. In order to classify the network traffic, the clustering k-means algorithm is used and compared to other model-based clustering methods along with rule-based classification models. Associations were found among flow parameters for several protocols and applications, such as Hypertext Transfer Protocol (HTTP), Mail, Simple Mail Transfer Protocol (SMTP), Domain Name System (DNS) and Internet Relay Chat (IRC). However, the variables used were source port, destination port, source IP address and destination IP address, and they may not be efficient when this technique is used for applications that enable obfuscation techniques or which are constantly changing pairs, IP addresses and random generations of ports number (e.g., eMule, BitTorrent, and Gnutella).

Bar-Yanai et al. [3] proposed a methodology based on a hybrid combination of two machine learning algorithms - K-Nearest Neighbor (KNN) [9] and K-Means [10], but this method works only with prior knowledge of the number of analyzed applications, i.e., the number of formed groups. Some works [1] [2] [4] [6] [7] [8] [11] do traffic classification based on port number, payload, or even the use of machine learning algorithms. Some of these works [1] [8] [7] exhibit signatures or association rules, resulting in extracted patterns. However, as already explained, these methods are not efficient for encrypted data and when applications make use of random ports. Furthermore, few studies validate their methods in real time.

Thus, this work presents a methodology to extract patterns automatically, and to identify and classify network traffic in real time without performing payload inspection and without the need for known ports. So, the paper proposes a real time hybrid model that combines the algorithm that automatically generates the association rules with a self-organizing maps (SOM) model to characterize the traffic. The model uses a machine learning algorithm based on Apriori [12] for the automatic extraction of association rules, choosing the most representative rules for each type of traffic. One characteristic of association rules is that these are easy to understand, unlike complex mathematical models. With the generated rules, the expert identifies the patterns more easily, assisting in decision making, e.g., blocking rules in firewalls. The model also uses an algorithm based on SOM [13] to perform grouping of protocols and applications traffic, dividing it by similarity to help with rules generation and as a second classifier. Moreover, the proposed model is capable of grouping unknown traffic for future tagging by experts (e.g., new applications traffic).

This paper is organized as follows. In Section II, we briefly review the techniques used in this paper. Section III shows the proposed model methodology. Section IV presents the experiments and the analysis of the results. Finally, Section V concludes with final considerations.

II. Fundamentals

The machine learning is a very promising approach for traffic classification, since classification using artificial intelligence techniques can be used to identify traffic data without relying on packet payload. To deal with the analysis of huge network traffic data, machine learning techniques have been used as important tools for extracting association rules and performing traffic classification.

A. Self-Organizing Maps (SOM)

Self-Organizing Maps (SOM) or Kohonens Self-Organizing Map is a clustering technique and data visualization technique that uses neural networks. These were based on observations of brain behavior where the unsupervised training is predominant.

The main reason for using SOM networks is to group similar input data into classes or groups called clusters [13]. What distinguishes SOM networks from others is a structure of two layers: one input and one output as shown in Fig. 1. The output layer is formed by a grid of neurons connected only to its immediate neighbors where, in this example, there are j input neurons and 15 output neurons.

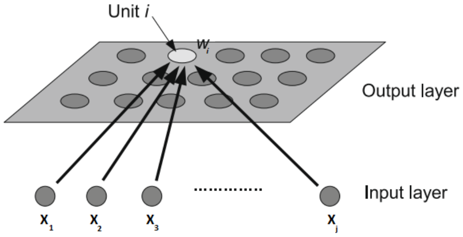


Figure 1. Self-Organizing Maps (SOM). (figure adapted from [14])

SOM networks produce a topological mapping and are based on competitive learning. In competitive learning, an objective function is required to determine the winner neuron, that is, a neuron which weight vector are nearest to the input vector. One of the key metrics used to determine the winner neuron is the Euclidean distance. The Euclidean distance is given by (1):

$$d_{xw} = \sqrt{\sum_{j=1}^{n}(x_j - w_{ij})^2}, \qquad (1)$$

where d_{xw} is the Euclidean distance, x_j is the input sample, w_{ij} is the synaptic weights connecting the input to neurons of the output grid and n is the number of inputs. The Euclidean distance allows calculating the similarity between the input samples and the set of neurons weights.

The competitive training can be done in two ways: The winner-take-all and the winner-take-quota. In winner-take-quota training, used in this study, both winner weight and its neighbors are readjusted, as shown in (2), according to a neighborhood region.

$$w_{ij}(new) = w_{ij}(old) + \alpha.h_i^v.(x_j - w_{ij}^v(old)), \qquad (2)$$

This region is defined by the function which is centered on the winner neuron and is expressed according to (3). There are several formats of neighborhood, including the circular and rectangular. In (3) a Gaussian function is defined. Over the iterations, it is necessary to reduce the size of the neighborhood region for the algorithm to converge.

$$h_i^v(n) = e^{-[\frac{d_i^v}{2r^2}]} \qquad (3)$$

The function $h_i^v(n)$ is the neighborhood value between a neighbor neuron i and the winner neuron v, d_i^v is the Euclidean distance between the winner neuron v and the excited neighboring neuron whose weights will be readjusted and, finally, r is the width of the neighborhood function. A learning rate for the SOM training is defined by α.

B. Association Rules

Association rules show how the occurrence of an item set implies the occurrence of some other distinct item set in records of the same database. The main objective of association rules is to find items that occur simultaneously and often in large transaction databases, facilitating the understanding of data behavior, since the use of association rules techniques make it possible to predict not only the class but any other attributes, like, for example, a network traffic behavioral profile.

An association rule is represented as follows:

$$X(antecedent) \Rightarrow Y(consequent) \tag{4}$$

Agrawal et al. [12] presented a mathematical model in which parameters such as support and confidence are taken into account. The support corresponds to the frequency of the occurrence of patterns throughout the database, while confidence is a measure of the force of rules, which indicates the frequency at which items in Y appear in occurrences containing X.

Analysis by means of association rules using Apriori has been studied and applied in a variety of fields such as web mining [15] and intrusion detection systems [16]. Apriori is able to find all frequent itemsets of any database and subsequently generate association rules. The algorithm, in fact, is based on two main subtasks: The frequent itemsets generation and the generation of rules themselves.

Machine learning algorithms can be used for grouping and extraction of patterns of network traffic, which can be used, after being trained, to classify traffic. Therefore, the advantage of using association rules is that associations between different databases attributes can be explored in the task of network traffic patterns extraction.

C. Information Gain

Before extracting the rules, the stored data must be evaluated according to their relevance by measuring the information gain shown in (5). Information gain is a measure based on the entropy of a system, that is, on the disorder degree of a system. This measure indicates to what extent the whole systems entropy is reduced if we know the value of a specific attribute. Thus, it can show us how the whole system is related to an attribute; in other words, how much information this attribute contributes to the system [17]. The information gain is then used in the process of feature selection.

$$\triangle info = I(parent) - \sum_{j=1}^{k} \frac{N(V_j)}{N} I(V_j) \tag{5}$$

III. Proposed Model

The proposed techniques in the literature cover several models to deal with traffic identification and pattern extraction. However, the number of new Internet applications increases at a high speed, and to extract patterns and identify applications is a complex task, especially when it comes to new applications done without the analysis of the payload and done in real time. To deal with this problem, this paper presents a hybrid model for traffic classification based on Apriori and SOM algorithms.

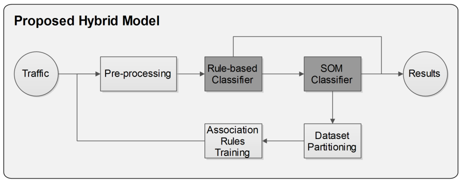


Figure 2. The proposed hybrid model.

A. Architecture

The proposed model is divided into two main modules: a rule-based classifier and a second classifier named SOM Classifier, forming a hybrid model. In Fig. 2, we present an overview of the proposed hybrid model for automatic pattern extraction and real time traffic classification.

The proposed model objective is to generate association rules, i.e., knowledge from network traffic. One of the advantages of a rule-based classifier is that the generated knowledge is easily readable and understood by experts, unlike the case of complex mathematical models. The rules are used to classify traffic with low computational cost, without the need to perform complex calculations. In the proposed hybrid model of Fig. 2, the rule-based classifier was previously trained and is described in Section III-B2. In Fig. 2, after the data is pre-processed (Section III-B1), if traffic is known, the results are presented, and otherwise an SOM classifier is used to aid in traffic classification. The SOM model was previously trained as described in Section Section III-B3.

Unknown traffic can be subjected to the model. In this case, none of the classifiers would identify the traffic. In this scenario, a continuous and automatic process of training is performed. After being subjected to the two classifiers without success in classification, the unknown traffic is accumulated into a database, automatically labeled by the model. This label can be relabeled by an expert. The database is divided into similar statistical distributions, with the aid of a second SOM network which is trained in each cycle. Many clusters are generated with distinct datasets, where each dataset is trained by an Apriori algorithm, thus creating new rules for unknown traffic. At this moment, the expert takes action to label each generated cluster. In a new capture cycle, the traffic that was unknown would be known from an auto-labeling system or a definitive label after the expert analysis.

The analyzed applications and protocols were HTTP, HTTPS, FTP, SSH, Ares, Gnutella, eMule, BitTorrent and Skype. The defined model was implemented in Java using the Jpcap API and some Weka API classes. Furthermore, the Java Kohonen Neural Network Library (JKNNL) was used

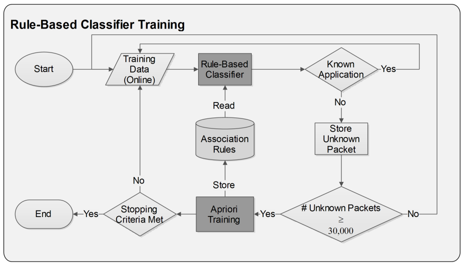


Figure 3. Rule-based classifier training process.

as reference [18].

B. Training Process

The assembly of the proposed model is explained in the next subsections.

1) Pre-processing: The datagram header has several fields that could be used in the network traffic training and classification process. But not all fields are relevant for association rules generation. Thus, an algorithm based on information gain [19] was used to refine the most significant attributes to be used in the proposed hybrid model.

The most relevant attributes determined by the information gain algorithm were: DataLength (referring to the datagram length field), Flag (associated with the flags field), Lay (associated with the upper layer protocol field), TcpLength (relative to the packet payload size (TCP)) and SegmentLength (relative to the packet payload size (UDP)), totaling five attributes. The attributes TcpLength and SegmentLength were calculated since they are not part of the header. Note that neither payload nor port information is used in the process,. The traffic was pre-processed so that only these five attributes were used as input data. In the training of the SOM model (Section III-B3), the duplicated records were removed from the database and then all dataset was normalized within range [0.15,0.85]. The process of database normalization improves the effectiveness and performance of computational intelligence algorithms.

2) Rule-based Classifier: The rule-based classifier used in the proposed hybrid model in Fig. 2 was generated from the training process of Fig. 3. The training was conducted using real time traffic as input (online). The entire process was repeated for each application and protocol, ensuring that only particular application traffic would be trained at any time.

At first, the rules database is empty, and therefore, the application is labeled as unknown. After storing 30,000 unknown packets, new rules are generated to classify this initial unknown traffic. As it is ensured that what is passing through the network traffic belongs to the application that is being trained at the moment, the rules are labeled according to the corresponding application. Each passage through the Apriori training stage corresponds to a training cycle. The training process ends when the result of the classification accuracy rate (see Table I) and the number of rules in the database does not increase after 10 consecutive cycles.

TABLE I. METRICS USED IN THE EXPERIMENTS

Metric	Equation
Accuracy (AC)	$AC = \frac{AD}{D}$ AD = number of packets correctly detected. D = number of packets detected.
Correctness (CR)	$CR = \frac{AD}{A}$ AD = number of packets correctly detected. A = real number of packets of the desired application.
Completeness (CP)	$CP = \frac{D}{A}$ D = number of packets detected. A = real number of packets of the desired application.
False Negative Rate (FN)	$FN = \frac{FN}{A}$ FN = number of packets of a desired application that was not detected. A = real number of packets of the desired application.
False Positive Rate (FP)	$FP = \frac{FP}{D}$ FP = number of packets incorrectly detected as a desired application. D = number of packets detected.

After the training process execution for all applications and protocols used in this work, a single rules database is created by unifying all rules databases found, which is then used as the rules database on the rule-based classifier on the proposed hybrid model (Fig. 2). The Apriori algorithm generates a huge number of rules, so it was necessary to realize a filter process to determine the most relevant rules. Relevant rules were defined as the rules with the greatest number of used attributes in their composition that represent an application or protocol.

The parameters used in the model were support and confidence. Many parameter values were tested during the training process, and the best results were obtained when using 30% for support and 90% for confidence.

3) SOM Classifier: The SOM model used in the proposed hybrid model of Fig. 2 was generated from the SOM network training according to Fig. 4. An offline and stratified dataset with 100,000 packets was used during the training for each application and protocol and was randomly mixed. Several trainings were conducted while varying the SOM network parameters until the largest number of representative clusters formed was obtained, since the SOM network used in this work has the functionality of increasing the grid size dynamically. A cluster is defined as representative when a particular application has predominance over others in 70% of the database. Each cluster represents a particular application or protocol.

As already mentioned, during the training process, diverse parameters of the SOM network were tested. The parameters

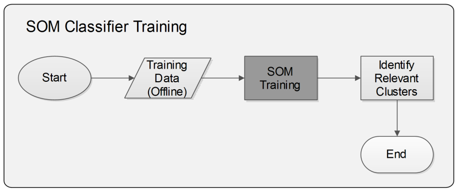


Figure 4. SOM classifier training process.

were the number of iterations and the radius used by the Gaussian neighborhood function. The sets used during the training were [5000, 10000, 15000] and [100, 150, 200] for the iterations and radius parameters respectively. The best results were obtained when using 15000 for the number of iterations and 100 for the radius. The learning rate was fixed in 0.1. The initial weight vector was randomly initialized with values between [0, 0.85].

IV. Experimental Results

This section introduces the experimental results when applying the model against some test datasets and an analysis is performed when the model is compared to another network traffic classifier also based on computational intelligence.

A. Metrics

The described models in Section III-A were measured using the metrics described in Table I.

In the next subsections, we present the obtained results for each metric in order to compare the proposed hybrid model in Section III-A with a rule-based model only described in Section III-B2. We also investigate the results on classifying commonly used public traces like the Measurement and Analysis on the WIDE Internet (MAWI) [20] database with the proposed model. Besides that, a comparative investigation between the model proposed by [3] and the proposed hybrid model described in this work was performed.

B. Dataset

The dataset used in the experiments was generated by exclusively running the protocols and applications investigated in this work, assuring the collection of the ground-truth data. The experiments were performed in real time and it was ensured that only a determined application or protocol was being captured at that time, therefore, the data are submitted without label (unknown data) to the models analyzed in the next subsections. The dataset contains an average of 1,000,000 packets for each application. The packets were pre-processed to extract only the five attributes, serving as input to the classifiers. For the SOM Classifier, the data was normalized. The experiments were performed at the Point of Presence of Pernambuco (PoP-PE) of the National Network on Teaching and Research (RNP).

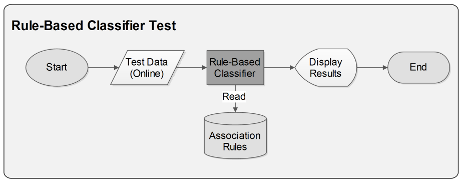


Figure 5. Rule-based classification process.

TABLE II. Rule-based Model Results

Aplications / Protocols	AC %	CR %	CP %	FN %	FP %
BitTorrent	95.98	89.35	93.09	4.07	0.23
eMule	87.86	76.88	87.50	3.95	0.46
Ares	94.01	82.53	87.78	5.25	0.90
Gnutella	99.82	89.10	89.25	0.16	0.00
HTTPS	99.98	98.46	98.48	0.02	0.00
SSH	99.96	98.91	98.95	0.03	0.00
FTP	99.93	99.91	99.97	0.06	0.00
HTTP	99.96	99.17	99.20	0.03	2.14
Skype	53.16	53.02	99.72	46.70	3.73

The proposed hybrid model was also validated against a public database (MAWI). The trace was randomly selected, comprising traces collected in June 8, 2012. The comparative investigation between the model proposed by [3] and the hybrid model proposed in this work was performed using approximately 12GB real time network traffic.

C. Rule-based Model

The rule-based model was first validated as described in Fig. 5. The association rules were previously generated from the training phase (Section III-B2). The dataset was classified by the generated rules and the results were analyzed.

The results are shown in Table II. This shows the obtained results for the evaluated applications and protocols, with their respective rates. For example, for the BitTorrent application, using the found rules, an accuracy of 95.98% was obtained, with a correctness of 89.35%, completeness of 93.09%, false negatives of 4.07% and 0.23% false positives. Thus, with the found rules, it was possible to classify approximately 96% of the traffic generated by the BitTorrent application correctly. False positives presented low values for all applications and protocols, except for Skype and HTTP. The high false positive rate of HTTP was due to the fact that P2P applications such as eMule, Ares and Gnutella accessed web servers during the test. When an application has its datagram size varying in very short intervals, such as Skype, the model does not provide a high accuracy rate, because it does not work with such intervals.

D. Proposed Hybrid Model

Due to the bad results obtained by the Skype application when evaluating the rule-based model, we proposed a hybrid model with the aid of a SOM network model so that the

TABLE III. Proposed Hybrid Model Results

Aplications / Protocols	AC %	CR %	CP %	FN %	FP %
BitTorrent	94.41	91.37	96.77	4.50	0.03
eMule	91.52	76.23	83.29	7.06	0.25
Ares	95.98	92.32	95.83	2.08	0.35
Gnutella	99.91	98.13	98.25	0.02	0.00
HTTPS	99.61	99.59	99.97	0.00	0.00
SSH	99.62	99.58	99.96	0.00	0.00
FTP	99.78	97.74	97.96	0.00	0.00
HTTP	99.70	99.58	99.88	0.22	0.00
Skype	94.97	86.66	90.25	3.58	0.00

TABLE IV. Results Obtained from MAWI Database

Protocols	TP %	FP %
SSH	99.73	0.00
FTP	99.64	0.00
HTTP	100.00	0.00

results could be compared. A SOM network reduces the dimensionality of the data and assists in pattern extraction.

The hybrid model was discussed in Section III-A and the results obtained in validation, i.e., using the dataset described in Section IV-B are displayed in Table III. Notice that the results of accuracy on the Skype application proved to be superior in the hybrid model (94.97%) against a 53.16% (see Table II) rate when the rule-based model is used. Besides that, the results for other applications was superior or equivalent. Rates of FP were better tackled by the proposed hybrid model in all applications and protocols, while FN rates were superior in 67% of all protocols and applications investigated.

The hybrid model performance was also evaluated against public traces from MAWI. The FTP, SSH and FTP protocols were analyzed, since these are the protocols in common with the protocols used in this work. The results were evaluated by two metrics and are shown in Table IV. The first metric is the true positive (TP), which is used to measure the traffic fraction of a certain application that is recognized by the model for this application. Also, the false positive (FP) was used, which measures the traffic fraction that does not belong to a certain evaluated application. The classification rate results exceeded 99% of true positive rate. No false positives were detected during the test.

Besides that, an investigation was performed to determine the effectiveness of the proposed hybrid model (PHM) against the model proposed by [3] (RTM), since it also uses some well-known computational intelligence techniques to classify network traffic data. The experimental results, using the protocols, applications and metric (AC (%)) in common with both works, are shown in TABLE V.

TABLE V. Results for the Realtime Model (RTM) [3] and the Proposed Hybrid Model (PHM)

Applications / Protocols	RTM	PHM
HTTP	93.75	99.7
Skype	90.54	94.97
eMule	94.57	97.4
BitTorrent	84.55	94.41

It can be observed that when both models were validated against a real time traffic scenario, our proposed hybrid model was far superior when compared to the model proposed by [3].

V. Conclusion

This work proposed a hybrid model based on computational intelligence techniques, consisting of a rule-based model and a self-organizing map (SOM) model. The model was proposed to deal with the problem of encrypted data, the constantly changing ports behavior of some applications and the identification of new protocols or applications (unknown network traffic) for future labeling by experts, since the architecture has the ability of being in a constant learning process to extract patterns from new applications or protocols. Besides that, the aim of the method is to extract association rules for a network traffic classification purpose, since rules are easily understood by experts and can be easily applied, for example, in a firewall system for security purposes.

The experimental results showed that the proposed hybrid model is superior to a rule-based model only, with results that exceed 91% of precision, maintaining low rates of false positives and false negatives for most applications and protocols. The hybrid model was also able to better tackle the problem of Skype identification, which presented bad results when classified by a rule-based model only. It was also validated against a known public network traffic database known as Measurement and Analysis on the Wide Internet (MAWI). The proposed model reached levels superior to 99% for true positive rates and 0% for false positive rates for the investigated protocols. The proposed hybrid model was also superior to the Realtime model (RTM) [3] when evaluated against a real time network traffic.

References

[1] G. Szabó, Z. Turányi, L. Toka, S. Molnár, and A. Santos, “Automatic protocol signature generation framework for deep packet inspection,” in Proceedings of the 5th International ICST Conference on Performance Evaluation Methodologies and Tools, Brussels, Belgium, May 2011, pp. 291–299.

[2] V. Carela-Español, P. Barlet-Ros, M. Solé-Simó, A. Dainotti, W. de Donato, and A. Pescapé, “K-dimensional trees for continuous traffic classification,” in Proceedings of the Second international conference on Traffic Monitoring and Analysis, ser. Lecture Notes in Computer Science, F. Ricciato, M. Mellia, and E. Biersack, Eds., vol. 6003. Berlin, Heidelberg: Springer Berlin Heidelberg, Apr. 2010, pp. 141–154.

[3] R. Bar - Yanai, M. Langberg, D. Peleg, and L. Roditty, “Realtime classification for encrypted traffic,” in Proceedings of the 9th international conference on Experimental Algorithms, ser. Lecture Notes in Computer Science, P. Festa, Ed., vol. 6049. Berlin, Heidelberg: Springer Berlin Heidelberg, May 2010, pp. 373–385.

[4] A. Dainotti, A. Pescape, and K. Claffy, “Issues and future directions in traffic classification,” IEEE Network, vol. 26, no. 1, Jan. 2012, pp. 35–40.

[5] J. Summers, T. Brecht, D. Eager, and B. Wong, “Methodologies for generating HTTP streaming video workloads to evaluate web server performance,” in Proceedings of the 5th Annual International Systems and Storage Conference on - SYSTOR ’12. New York, New York, USA: ACM Press, Jun. 2012, pp. 1–12.

[6] G. La Mantia, D. Rossi, A. Finamore, M. Mellia, and M. Meo, “Stochastic Packet Inspection for TCP Traffic,” in 2010 IEEE International Conference on Communications. IEEE, May 2010, pp. 1–6.

[7] M. Ye, K. Xu, J. Wu, and H. Po, “AutoSig-Automatically Generating Signatures for Applications,” in 2009 Ninth IEEE International Conference on Computer and Information Technology, vol. 2. IEEE, 2009, pp. 104–109.

[8] U. K. Chaudhary, I. Papapanagiotou, and M. Devetsikiotis, “Flow classification using clustering and association rule mining,” in 2010 15th IEEE International Workshop on Computer Aided Modeling, Analysis and Design of Communication Links and Networks (CAMAD). IEEE, Dec. 2010, pp. 76–80.

[9] T. Cover and P. Hart, “Nearest neighbor pattern classification,” 1967, pp. 21–27.

[10] J. Hartigan and M. Wong, “A k-means clustering algorithm,” 1 ACM/IEEE-CS Joint Conference, Applied Statistics, vol. 28, 1979, pp. 100–108.

[11] M. Soysal and E. G. Schmidt, “Machine learning algorithms for accurate flow-based network traffic classification: Evaluation and comparison,” Performance Evaluation, vol. 67, no. 6, Jun. 2010, pp. 451–467.

[12] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules in Large Databases,” Sep. 1994, pp. 487–499.

[13] S. Haykin, “Neural Networks: A Comprehensive Foundation,” Jul. 1998.

[14] E. Kita, S. Kan, and Z. Fei, “Investigation of self-organizing map for genetic algorithm,” Advances in Engineering Software, vol. 41, no. 2, Feb. 2010, pp. 148–153.

[15] J. S. Ryu, W. Y. Kim, K. I. Kim, and U. M. Kim, “Mining opinions from messenger,” in Proceedings of the 2nd International Conference on Interaction Sciences Information Technology, Culture and Human - ICIS ’09. New York, New York, USA: ACM Press, Nov. 2009, pp. 287–290.

[16] L. Li, D.-Z. Yang, and F.-C. Shen, “A novel rule-based Intrusion Detection System using data mining,” in 2010 3rd International Conference on Computer Science and Information Technology. IEEE, Jul. 2010, pp. 169–172.

[17] M. Martín-Valdivia, M. Díaz-Galiano, A. Montejo-Raez, and L. Ureña López, “Using information gain to improve multimodal information retrieval systems,” Information Processing & Management, vol. 44, no. 3, May 2008, pp. 1146–1158.

[18] JKNNL, “Java Kohonen Neural Network Library.” [retrieved: March, 2013]. Available: http://jknnl.sourceforge.net/

[19] P.-N. Tan, M. Steinbach, and V. Kumar, “Introduction to Data Mining, (First Edition),” May 2005.

[20] MAWI, “MAWI Working Group Traffic Archive.” [retrieved: March, 2013]. Available: http://mawi.wide.ad.jp/mawi/