ICNS 2014

The Tenth International Conference on Networking and Services

ISBN: 978-1-61208-330-8

April 20 - 24, 2014

Chamonix, France

ICNS 2014 Editors

Eugen Borcoci, University 'Politehnica' Bucharest, Romania

Sathiamoorthy Manoharan, University of Auckland, New Zealand

Tao Zheng, Orange Labs Beijing, China

ICNS 2014

Foreword

The Tenth International Conference on Networking and Services (ICNS 2014), held between April 20 - 24, 2014 in Chamonix, France, 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.

We take here the opportunity to warmly thank all the members of the ICNS 2014 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 2014. 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 2014 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that ICNS 2014 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 Chamonix, France.

ICNS 2014 Chairs:

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
Go Hasegawa, Osaka University, Japan
Abdulrahman Yarali, Murray State University, USA
Emmanuel Bertin, Orange Labs, France
Steffen Fries, Siemens, Germany

Rui L.A. Aguiar, University of Aveiro, Portugal
Iain Murray, Curtin University of Technology, Australia
Khondkar Islam, George Mason University - Fairfax, USA

ICNS Industry/Research Relation Chairs
Eunsoo Shim, Samsung Electronics, Korea
Tao Zheng, Orange Labs Beijing, China
Bruno Chatras, Orange Labs, France
Jun Kyun Choi, KAIST, Korea
Michael Galetzka, Fraunhofer Institute for Integrated Circuits - Dresden, Germany
Mikael Gidlund, ABB, Sweden
Juraj Giertl, T-Systems, Slovakia
Sinan Hanay, NICT, Japan

ICNS 2014

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
Go Hasegawa, Osaka University, Japan
Abdulrahman Yarali, Murray State University, USA
Emmanuel Bertin, Orange Labs, France
Steffen Fries, Siemens, Germany
Rui L.A. Aguiar, University of Aveiro, Portugal
Iain Murray, Curtin University of Technology, Australia
Khondkar Islam, George Mason University - Fairfax, USA

ICNS Industry/Research Relation Chairs

Eunsoo Shim, Samsung Electronics, Korea
Tao Zheng, Orange Labs Beijing, China
Bruno Chatras, Orange Labs, France
Jun Kyun Choi, KAIST, Korea
Michael Galetzka, Fraunhofer Institute for Integrated Circuits - Dresden, Germany
Mikael Gidlund, ABB, Sweden
Juraj Giertl, T-Systems, Slovakia
Sinan Hanay, NICT, Japan

ICNS 2014 Technical Program Committee

Johan Åkerberg, ABB AB - Corporate Research - Västerås, Sweden
Ryma Abassi, Higher School of Communication of Tunis /Sup'Com, Tunisia
Nalin Abeysekera, University of Colombo, Sri Lanka
Ferran Adelantado i Freixer, Universitat Oberta de Catalunya, Spain
Prathima Agrawal, Auburn University, USA
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
Maria Andrade, University of Porto / INESC Porto, Portugal
Annamalai Annamalai, Prairie View A&M University, USA
Mario Anzures-García, Benemérita Universidad Autónoma de Puebla, Mexico
Pedro Andrés Aranda Gutiérrez, Telefónica I+D - Madrid, Spain
Patrick Appiah-Kubi, Indiana State University, USA

Serban Obreja, University Politehnica - Bucharest, Romania
Kazuya Odagiri, Yamaguchi University, Japan
Máirtín O'Droma, University of Limerick, Ireland
Tae (Tom) Oh, Rochester Institute of Technology, USA
Jinwoo Park, Korea University, Korea
Harry Perros, North Carolina State University, USA
Dennis Pfisterer, University of Luebeck, Germany
Zsolt Alfred Polgar, Technical University of Cluj Napoca, Romania
Francisca Aparecida Prado Pinto, Federal University of Ceará, Brazil
Thomas Prescher, TU Kaiserslautern, Germany
Francesco Quaglia, Sapienza Università di Roma, Italy
Ahmad Rahil, University of Burgundy, France
Jelena Revzina, Transport and Telecommunication Institute, Latvia
Karim Mohammed Rezaul, Centre for Applied Internet Research (CAIR), NEWI, University of Wales, UK
Oliviero Riganelli, University of Milano Bicocca, Italy
David Rincon Rivera, Technical University of Catalonia (UPC) - Barcelona, Spain
Diletta Romana Cacciagrano, Università di Camerino, Italia
Paolo Romano, IST/INESC-ID - Lisbon, Portugal
Sattar B. Sadkhan, University of Babylon, Iraq
Alessandra Sala, University of California - Santa Barbara, USA
Francisco Javier Sánchez Bolumar, Centro de Formación Tecnológica - Valencia, Spain
Luz A. Sánchez-Gálvez, Benemérita Universidad Autónoma de Puebla, México
Susana Sargento, University of Aveiro, Portugal
Panagiotis Sarigiannidis, University of Western Macedonia - Kozani, Greece
Reijo Savola, VTT, Finland
Stefan Schmid, TU Berlin & Telekom Innovation Laboratories (T-Labs), Germany
Jeff Sedayao, Intel Corporation IT Labs, USA
René Serral Garcia, Universitat Politècnica de Catalunya, Spain
Xu Shao, Institute for Infocomm Research, Singapore
Fangyang Shen, New York City College of Technology (CUNY), USA
Jian Shen, Chosun University, Gwangju, Republic of Korea
Tsang-Ling Sheu, National Sun Yat-Sen University - Kaohsiung, Taiwan
Eunsoo Shim, Samsung Electronics, Korea
Simone Silvestri, University of Rome "La Sapienza", Italy
Thierry Simonnet, ESIEE-Paris, France
Navjot Singh, Avaya Labs Research, USA
Charalabos Skianis, University of Aegean - Karlovasi, Greece
Vasco Soares, Instituto de Telecomunicações / Polytechnic Institute of Castelo Branco, Portugal
José Soler, Technical University of Denmark, Denmark
Gritzalis Stefanos, University of the Aegean, Greece
Manolis Stratakis, Forthnet Group, 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
Stephan Trahasch, Hochschule Offenburg, Germany
Joseph G. Tront, Virginia Tech, USA

Copyright Information

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

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

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

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

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

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

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

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

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

Table of Contents

Building a House of Cards: On the Intrinsic Challenges of Evolving Communication Standards

Jean-Charles Grégoire
INRS-EMT, CANADA
email:gregoire@emt.inrs.ca

***Abstract*—A critical look at a recent communications standard exposes how matters of feature interaction remain pervasive, not necessarily at a horizontal level, but also through layers of the architectures. This situation arises from the emergence of new standards or popular services based on earlier infrastructures, generic support middleware as well as emerging technologies. Several illustrations of such problems are presented and discussed in this paper, as illustrations of problems to try to avoid in practice. We show how the industrial practices remain lacking but also how some of the difficulties around the emergence of feature interactions are deeply linked to the standardization process.**

Keywords–Feature Interaction; RCS; SIP; OMA CPM; OMA SIMPLE IM; IMS.

I. INTRODUCTION

Much has been written over the years on the problem of feature interaction (FI) and its numerous guises [1], [2], [3], [4], [5], [6]. Different communities, such as requirements engineering and formal methods have come together to present various perspectives on issues pertaining to the specification and implementation of telecommunication services, and further expanding such investigations beyond telephony to other domains, even besides telecommunications. Tools, architectures, methods and insights has been proposed, with varying, yet demonstrated degree of usefulness. At least two questions remain at this stage: what practical impact can we observe from this work and is there still "something" that we have missed.

The fairly recent 5.1 release of the Rich Communications System (RCS) standard [7] has been an opportunity for us to look at these questions. In this paper, we look at a number of issues we perceive as problems related with this specification, and later discuss how a feature interaction perspective may help with such problems, or where more effort is warranted. We begin with giving some background on messaging protocols used in cellular communications. We will then follow with the description of a number of issues of an FI nature raised by the current specification and expand into a more general discussion.

Because of the nature of this special issue, we expect that the reader is familiar with the Session Initiation Protocol (SIP) and the work done on that foundation by different bodies. The introduction only highlights the contributors but not the technology itself. Among many possibilities, references [8], [9] can be used if such information is desired.

The paper is structured as follows. The next Section gives some background on the standards of interest. Section 3 presents and illustrates a number of problems. Section 4 is a general discussion and Section 5 concludes this paper.

II. BACKGROUND

In the age of the Internet, different bodies contribute to the standardization process. While the IETF has created the Session Initiation Protocol and retains the control over its evolution, it has become the foundation of several service infrastructures, notably the IP Multimedia Subsystem (IMS), originally developed by the 3rd Generation Partnership Project (3GPP) for wireless services. Over the years, other standardization bodies have become interested in IMS which built into joint work with 3GPP2, the European Telecommunications Standards Institute (ETSI) and CableLabs. Practically, this has led to the extension of the use of IMS in 3GPP to various network access technologies, and their evolution. While 3GPP was concerned with the middleware, it was not interested in pursuing work on applications beyond audio/video (A/V) services. Other bodies, possibly closer to the ear of operators, have looked into other services, such as the Open Mobile Alliance (OMA) for wireless. Over time OMA focus has moved from being platform neutral to acknowledging IMS as an enabling platform. OMA has produced two standards for personal communications based on SIP, among other protocols: Instant Messaging (IM)[11] and Converged IP Messaging (CPM) [12]. The GSM Alliance (GSMA) has in turn reused and extended some of that work to create the current (5th) release of RCS. As companion to the standard, we find a number of *endorsement documents* for a variety of messaging communication standards which describe to what extent their functions are supported (e.g OMA SIMPLE IM 2.0, OMA CPM 2.0).

Why such a complex picture? While a division of responsibilities may appear clear between IETF and 3GPP, matters become more complicated when competition over forums, markets and cultures emerge. Tensions between different perspectives abound while efforts are constantly made to bind them together in a convergent view–pick your favourite analogy of the One Ring or the Holy Grail–to have as large a market as possible. GSMA, for example, takes upon the mandate to look at standards proposed by major agencies, extracts from complex architectures or a large selection of protocols a set of mechanisms sufficient to support a set of services of interest to its members, in a streamlining process, which in essence does not create anything new except of specific focus for service enabling technology, with different profiles.

This picture explains how the standardization process remains complex, even though quite often the same companies

 ISBN: 978-1-61208-330-8

will be involved with different aspects of the work. Since the devil remains in the details, we see that the traditional issue of going through piles of documents written in prose ornamented with diagrams has not changed. A quick glance at the RCS specification from GSMA reveals a typical document structure with general concepts, feature specifications and use cases. In traditional fashion, we find text describing interactions between the different features of the service, even for the most trivial cases, to. viz.:

> 3.2.2 Interaction with other RCS features
> There are no interactions between the RCS 5.1 Standalone Messaging service and other RCS services.

(Rich Communication Suite 5.1[7] p. 138.)

The questions we raise are the identification of interactions arising from this piecemeal construction of a standard, constrained by the reuse of existing components designed in different fora because of a number of constraints, not the least commercial ones, but also historical, as our last example shall illustrate.

III. Problems

In this section, we illustrate a number of issues of an FI nature with the 5.1 revision of the RCS standard. This is done informally, based on quotes from the document itself and a discussion thereof.

A. Problem 1

Recall that a REGISTER message is a pre-requisite to SIP operations, to bind the user to a proxy (or P-CSCF for IMS) and allow further end-to-end communications; it is essentially the first operation to be performed by the user to allow access to services. An INVITE will initiate a communication. Figure 1 illustrates the process.

In RCS, there are three different ways to send a message: through a SIP MESSAGE message, by initiating a session using the MSRP [13] protocol, or directly within the SIP INVITE message, with a CPIM body. Again, these alternatives stem from the consideration of different communication scenarios in different standards, conversational or standalone messages, and also message size—in the standards, we see different modes of operation for standalone messages: pager mode or large message mode. Furthermore, non-delivered messages are stored in a server and delivered at the next registration. The need to be able to hold messages requires the presence of a server able to temporarily store messages in the path. The server is informed of the registration through a notification and automatically sends the pending messages.

Now let us consider this segment of the RCS specification.

> **2.4.5 Registration frequency optimization**
> *An RCS client shall not send more register requests than what is needed to maintain the registration state in the network. When the IP connectivity is lost and restored with the same IP address, the RCS client shall:*
> - *Only send a register refresh upon retrieval of IP connectivity if the duration for sending a register-refresh since the last REGISTER request has been exceeded, and,*
> - *Only send an initial register upon retrieval of IP connectivity if the registration has expired.*

(RCS 5.1 Advanced Communications Specification [7], p. 55.)

There are several issues here. Again, going back to figure 1, we see that the access link can be of various natures, including WiFi. Depending on the technology used, the loss of the communication link may or may not be detected automatically and notified to IMS. In some way, detection will boil down to the use of a form of timeout but the duration of that timeout is key: it goes from milliseconds in some cases to seconds in others. The matter is worst when timeout detection is done through the loss of the TCP carrier.

We intuitively see that this situation leads to a race condition, where a user can be disconnected and reconnected while a TCP session is still established and this can lead to out-of-order messages delivery if new messages are sent while a TCP connection still holding messages was ended, not unlike what we observe with email when messages cannot be delivered and are retransmitted at a later time. This could occur because of unstable access, typically WiFi, and some messages being stored until they can be delivered.

B. Problem 2

As a corollary of the previous problem, we might wonder what the issue is with the TCP protocol that standard bodies seem reluctant to use it. As we have seen, rather than relying the MSRP protocol for instant communications, we find a number of alternative behaviours.

At stake there can be the compatibility between an Internet-like messaging (SIP-independent) behaviour and an SMS-like behaviour. In a typical Internet-like, IM behaviour, a session is established with a server and we receive notifications when a message is received for us, typically through polling mechanisms. TCP is the underlying transport mechanism and guarantees the stability of the session. Alternatively, an SMS can be received or sent independently of a service session. This latter behaviour has led to the creation of short messages in OMA standards: a message can be carried in an invite and a session—complete with MSRP transport—will be established when a message is sent back.

This behaviour actually alleviates some issues. For one, *if the target of the message is connected through multiple terminals, which one will actually be used for the exchange?* Only when an answer is sent can this be safely assessed, and a unique TCP session established with a full SIP INVITE exchange. Of course, it could be argued that a typical SIP session could instead be used. But, if the user has registered multiple terminals, and one or more are auto-answered enabled, this could lead to the set-up of multiple connections, possibly with the split of the TCP session at a message relay in the call, and increased costs in resource usage and bookkeeping. Practically, different markets have chosen different solutions and a global standard must support them all.

 ISBN: 978-1-61208-330-8

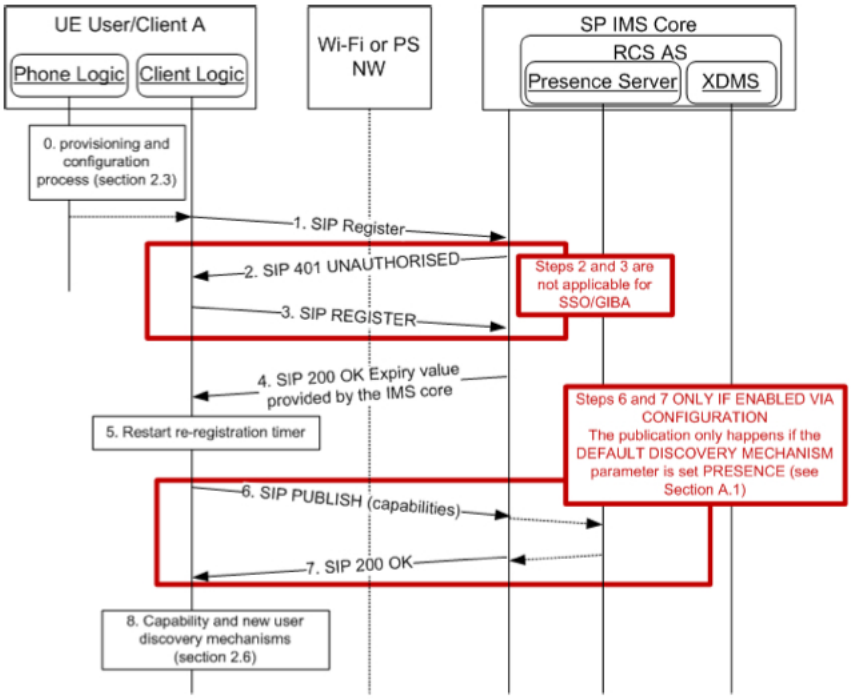


Figure 1. The SIP registration process, from [7].

C. Problem 3

Another corollary of the existence of various forms of messages is the intersection of the signalling path and the media path. Since messages are carried in signalling messages or on MSRP (TCP) sessions and since messages must be stored, if necessary, in a message store and possibly be forwarded to different destinations, there are nodes which must be both on the signalling path and on the media path.

The following quotation illustrates how these matters are acknowledged in standards. In practice, the CPM Feature Tag is used to route the SIP messages to the proper functional components.

> The protocols used for the CPM-PF1 interface are SIP, SDP, MSRP, and RTP/RTCP. SIP is used for CPM Session signalling, for CPM File Transfer signalling and for discrete Pager Mode CPM Standalone Message transfer. SDP is used to describe the set of Media Streams, codecs, and other Media related parameters supported during CPM Session set up and for describing file characteristics during CPM File Transfer initiation. MSRP is used for the transfer of Large Message Mode CPM Standalone Messages, for the exchange of CPM Chat Messages, both small and large, and for the Media transfer of a CPM File Transfer. RTP is used for continuous Media transport and RTCP supports for the exchange of information needed to control RTP sessions.
>
> NOTE: The exact network path used for the actual Media transfers (i.e., MSRP and RTP/RTCP protocols) will be negotiated via the SIP signalling part of this interface. For example, it is possible that direct client-to-client Media transfers are negotiated, or a direct Media transfer between a client and an Interworking Function. The signalling part of the CPM-PF1 interface is dependent on an underlying SIP/IP core infrastructure.
>
> (OMA CPM standard, [12] p. 30.)

For the sake of completeness, let us just add that the choice of MSRP as a support for standalone message transfer (large message mode) depends solely on the size of the message to be sent. Above 1300 bytes, an MSRP session is established and disconnected once the transfer has been accomplished.

D. Problem 4

Another acknowledged challenge has been the proliferation of the means to identify users and the means to access them. The two following quotations from the RCS specification clearly illustrate this. In the first case, we see the link between access technology and the user identity. What is interesting here is that we could assume that such problems are resolved by IMS, which should provide a unique means of identification. With different access technologies and authentication requirements, we end up with a proliferation of such information, and possible inconsistencies. In the second case we see the necessity of being able to use an identity users are still very familiar with, such as a telephone number (tel URI), in a typical sign of legacy constraint.

> Both a SIP and a tel URI may be configured for a user with following clarifications:
>
> - The configured values should not be used in the non-REGISTER transactions; instead the client uses one of the SIP or tel URIs provided in the P-Associated-URI header field

returned in the 200 OK to the SIP REGISTER request as described in [18].
- The user's own tel URI and/or SIP URI identities are configured through the Public_User_Identity parameters defined in [18].
- The public identity used for IMS registration is built according to the procedure defined in [18].
- When the device has either ISIM or USIM present and the RCS client has access to the ISIM or USIM, it does not rely on the SIP URI and tel URI configuration parameters.
- If the device has neither ISIM nor USIM present or is not able to access to it, a SIP URI must be configured. This URI is used for REGISTER transactions.
- Configuration of the tel URI is optional.

(RCS 5.1 Advanced Communications Specification, Version 1.0, [7] p. 332.)

> For device incoming SIP requests, the address(es) of the contact are, depending on the type of request, provided as a URI in the body of a request or contained in the P-Asserted-Identity and/or the From headers. If the P-Asserted-Identity header is present, the From header will be ignored. The only exception to this rule is when a request for Chat or Standalone Messaging includes a Referred-By header (it is initiated by Messaging Server for example in a store and forward use case as described in 3.3.4.1.4), thereby the Referred-By header should be used to retrieve the originating user instead. The receiving client will try to extract the contact's phone number out of the following types of URIs:
>
> - tel URIs (telephone URIs, for example tel:+1234578901, or tel:2345678901;phone-context=*phonecontextvalue*)
> - SIP URIs with a "user=phone" parameter, the contact's phone number will be provided in the user part (for example sip:+1234578901@operator.com;user=phone or sip:1234578901;phone-context=*phonecontextvalue* @operator.com;user=phone)
>
> Once the MSISDN is extracted it will be matched against the phone number of the contacts stored in the Address Book. If the received URI is a SIP URI but does not contain the "user=phone" parameter, the incoming identity should be checked against the SIP and tel URI address of the contacts in the address book instead. If more than one P-Asserted-Identity is received in the message, all identities shall be processed until a matched contact is found.

(RCS 5.1 Advanced Communications Specification, Version 1.0 [7], p. 57.)

E. Problem 5

As terminal technology evolves, traditional, basic assumptions on capabilities and capability negotiation need to be reassessed. It used to be that terminals had simple capabilities and a support service, for example for MMS, could resolve conversion issues simply based on the identification of the terminal and its profile. Nowadays, smartphones will support not only different codecs but also different formats and profiles. Current IETF work [14] reflects this trend and defines extensions to SDP to support new multimedia capabilities but one may wonder if all communication features can be supported in such a mode of negotiation, or what manufacturers do while waiting for the IETF to update its standards accordingly. For example, Apple's FaceTime supports switching video transmission between portrait and landscape orientation, and will notify the receiver side of the proper orientation, which will automatically trigger the appropriate view change for the receiver. This quite useful and rather trivial extension was not reflected in standard communication specifications until RCS 5.1 and has since then pushed into 3GPP.

F. Problem 6

If we consider capability exchange in RCS, that is, which features the terminal or terminal application supports, we see the combination of two approaches, one which assumes that a presence server is used in the network, which follows the OMA Presence specification, and one where terminals will exchange capability end to end. We should note that the use of SIP OPTIONS for such end to end discovery of communication features conforms with the IETF (SIP) RFC 3261, although the purpose is slightly different. The presence server will hold the information of features supported by the multiple terminals a user may have active at some time. A communication attempt can be made based on the features identified based on the query of the server, and will be forked only to the terminals supporting those features, again as per RFC 3841[15].

However, if one resorts to using SIP OPTIONS capability exchange, i.e., in the absence of a Presence server, the target device of the user cannot be selected a priori and the request message has to be forked to all terminals by the terminating CSCF (acting as a SIP proxy server, following the trapeze model.) The first terminal answering will establish itself as the terminating end, but it may not support the features required for the call, and the full set of capabilities available will not be returned, which may result in a communication attempt not being made.

To avoid such a situation, a new application server has to be introduced, in this case the Options AS. ([7], p. 59.) We must note two things about this server. First, it is specified only implicitly in the document, as it is supposed to make sure all the SIP OPTIONS based call flow behave as required. Second, its presence in a network is only required if the Presence server cannot be used. Still we can see why some manufacturers/operators would have preferred one approach vs. the other. Keeping the decision closer to the end terminal allows adaptation to dynamic access conditions, which is harder to achieve when such information must be updated on a remote—here presence—server.

G. Problem 7

Although feature tags are used in accordance with IETF requirements, their interpretation differs slightly as RCS uses them to indicate to the network and other devices the set of

 ISBN: 978-1-61208-330-8

communication methods used by the device, whereas they are meant to help route calls to suitable terminals in RFC 3841.

While the end effect may end up being the same for the user, there are interactions at work here between the information users (and their applications) want, and the needs of the network, specifically in terms of accounting and ultimately billing. Whereas the user application would be happy to indicate that it supports RCS communications, an operator would appreciate having a break down per specific features, e.g., file transfer or video transfer, for content-specific billing. Again, for historical reasons, RCS supports feature tags of both natures, more detailed and more specific.

IV. Discussion

We have presented matters of conflicting and evolving requirements, consensus building, changing context, lack of proper architectural constructs. Some are typical Feature Interaction matters, others might not be considered as such. Overall, such issues are not new, but we see them occurring in new standards and this leads to wonder about the impact of research on feature interaction on standards work.

A. On Standards

An important concern we have attempted to expose here is *evolution*. Evolution of standards, of course, but also evolution of technology. In some cases, we could feel as if the rug had been pulled from under our feet as an issue which appeared resolved is reopened, because new uses are added, or the underlying technology changes. Formally, this can be captured as a matter of machine-closed-ness [8], [9]. For specifications to be machine-closed means that no liveness property imposes restriction on finite behaviours of the system. Once the implementation changes, such a condition may no longer be satisfied and elements of the proof of implementation be broken; if we integrated two bodies of technology into one to support different legacies, we must handle possible assumption mismatch. How we, in practice, detect such conditions in an evolving environment largely remains an open issue, even if this is hardly a new problem, as it can be related to the 1996 Ariane 5 flight 501 failure; see for example [10].

Evolution in standards, context and usages are not coordinated. The terminal changes, and user interfaces will take advantage of it, while application protocols, standardized separately, will evolve separately and introduce limitations, or create issues where none existed before. Different markets will make different decisions on the evolution of service offerings, on matter as diverse as the path to obsolescence for SMS and its transition to IP-based texting. This in turns has an impact on specifications such as RCS. In practice, we have only seen such coordination succeed in situations where a single party controls most of the application and the infrastructure, as is becoming common in cloud-based environments such as offered by Google or Facebook. Interoperability, then, is not a concern. These applications, globally designated as “over the top” by operators, restrict them to a role of carriers without added value and are rather not viewed too fondly by the latter.

B. On Race Conditions

An area, however, where we could do much better is in the explicit capture of the semantics of connectivity and their report, especially failure semantics. It was quite surprising to realize how standards are still missing a clear way to define and report connectivity to applications, consistently across both signalling and media path. That we still find ourselves confronted with matters of race conditions and reordering in these days and age is quite amazing. Sadly, the trend will be to try to patch the problem, and not to go to its core, as we keep on building the house of cards.

This state of affairs also warrants a deeper look. At stake here is the end-to-end conception of communications services, as supported by IMS, and as opposed to a centralized model of control. Most popular communication services we use nowadays over the Internet are based on a model of cloud-based, client server-like centralized control, where is it possible to coordinate sender and receiver through a unique relay. In such circumstances, the effects of temporary disconnection can be easily managed, simply because the client will have the responsibility of querying the server for any new messages. In a straightforward peer-to-peer model, it is also possible to avoid such issues if we operate under a single domain/application (e.g., Skype) and the client can take the responsibility of holding on to messages and queueing them in proper delivery order until their delivery has been confirmed.

Race condition situations arise as domains are split and each domain takes responsibility for its part of the transaction: sending or receiving. On the receiver side, another entity must be introduced to temporarily store messages and therefore adds a further communication path to the receiver. The SIP call model, through its forking mechanism, easily supports placing a server on a call which will accept incoming communications if the user devices cannot and such addition is trivial. The challenge is then to deliver messages which may come directly from the sender and from storage in the right order: the store will need to be informed of the renewed connectivity before it can transmit the message, while the sender can send new messages directly to the receiver and they may arrive out of order. This race condition cannot be resolved unless the client polls the server first after re-establishing its connectivity, but this is not part of the SIP model and puts more demands on the applications. Also, in the IMS model, the server could be systematically put on the signalling path, but to function properly would need to completely intercept the call, i.e., be a back to back user agent (B2BUA), which would break the semantics of the call.

To summarize, we see that race conditions become a side effect of a forcing some features on top of a signalling model which is not fully adequate. Strict adherence to an end-to-end model without intermediate storage would resolve this issue, but then put more complexity in the client. But this leads to other philosophical debates.

We should mention a related approach [19], which we have described in earlier work, that would end the call on a form of user avatar, a client virtualized in a cloud at the edge of the operator’s domain (the edge cloud). It would act as be a stable point for communications while a simple, streamlined GUI protocol would run on the access link, as a form of compromise

 ISBN: 978-1-61208-330-8

between the multi-domain and centralized solutions.

V. CONCLUSION

We have presented a number of problems with a contemporary telecommunications standard, to illustrate how, beyond the progress we have made in requirements engineering and formal analysis we still have work to do as a community to improve the industrial state of the art.

We have illustrated how many of these issues are not really new nor ground breaking in nature. Solutions for them do exist or, in other cases, the nature of the issue can be identified and diagnosed before it makes its way into standards, or worse into implementations. Still, while the cure to the issues may be clear, we still need to better understand their cause and it is our hope that this paper can serve as a salutary lesson in that respect.

REFERENCES

[1] M. Calder, M. Kolberg, E.H. Magill, and S. Reiff-Marganiec, "Feature Interaction: A Critical Review and Considered Forecast", Computer Networks, Vol. 41, Jan. 2003, pp. 115–141.

[2] K.J. Turner and E.H. Magill, Eds. "Feature Interaction in Communications and Software Systems", Computer Networks, Special Issue on Feature Interaction, Vol. 57, Aug. 2013, pp. 2395–2464.

[3] M. Jackson and P. Zave, "Distributed Feature Composition: A Virtual Architecture for Telecommunications Services", Software Engineering, IEEE Transactions on, Vol. 24, Oct. 1998, pp. 831–847.

[4] M. Kolberg and E.H. Magill, "Managing feature interactions between distributed SIP call control services", Computer Networks: Special Issue on Feature Interaction, Vol. 51, Feb. 2007, pp. 536–557.

[5] A. Nhlabatsi, R. Laney, and B. Nuseibeh, "Feature interaction: the security threat from within software systems", Progress in Informatics, Special issue: The future of software engineering for security and privacy, Mar. 2008, pp, 75–89.

[6] A. Gouya and N. Crespi, "Detection and resolution of feature interactions in IP multimedia subsystem", Intl. Journal of Network Management, Vol. 19, Jul/Aug. 2009, pp. 315–337.

[7] GSM Association, Rich Communication Suite 5.1 Advanced Communications Services and Client Specification, Version 1.0, 13 August 2012.

[8] M. Poikselka and G. Mayer, "The IMS : IP Multimedia Concepts and Services", J. Wiley, 3rd ed., 2009.

[9] R. Noldus et al., "IMS Application Developer's Handbook: Creating and Deploying Innovative IMS Applications", Academic Press, 2011.

[10] J. L. LIONS, "ARIANE 5: Flight 501 Failure Report", the Inquiry Board, Paris, 19 July 1996, available from http://www.ima.umn.edu/ arnold/disasters/ariane5rep.html, last accessed February 14th, 2014.

[11] OMA SIMPLE IM V2.0, Open Mobile Alliance, Release date (Candidate Version), 2012-07-31.

[12] Converged IP Messaging Architecture V2.0, Open Mobile Alliance, Release Date (Candidate Version) 2013-06-11.

[13] B. Campbell, R. Mahy, and C. Jennings, The Message Session Relay Protocol (MSRP), RFC 4975, IETF, 2007.

[14] J. Rosenberg, H. Schulzrinne, and P. Kyzivat, Indicating User Agent Capabilities in the Session Initiation Protocol (SIP), RFC 3840, IETF, August 2004.

[15] J. Rosenberg, H. Schulzrinne, and P. Kyzivat, Caller Preferences for the Session Initiation Protocol (SIP), RFC 3841, IETF, August 2004.

[16] L. Lamport, The Temporal Logic of Actions, ACM Transactions on Programming Languages and Systems (TOPLAS), Vol. 16, May 1994, pp. 872–923.

[17] M. Abadi and L. Lamport, The existence of refinement mappings, Theoretical Computer Science, May 1991, pp. 253–284.

[18] Third Generation Partner Project (3GPP), IP multimedia call control protocol based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP), TS 24.229, Release 11, 2012.

[19] S. Islam and J.-Ch. Grégoire, Giving users an edge: A flexible Cloud model and its application for multimedia, Future Generation Computer Systems, June 2012, pp. 823–832.

 ISBN: 978-1-61208-330-8

Secure User Tasks Distribution in Grid Systems

Maxim Kalinin
St. Petersburg State Polytechnical University
St. Petersburg, Russia
maxim.kalinin@ibks.ftk.spbstu.ru

Artem Konoplev
St. Petersburg State Polytechnical University
St. Petersburg, Russia
artem.konoplev@ibks.ftk.spbstu.ru

Dmitry Moskvin
St. Petersburg State Polytechnical University
St. Petersburg, Russia
dmitry.moskvin@ibks.ftk.spbstu.ru

Alexander Pechenkin
St. Petersburg State Polytechnical University
St. Petersburg, Russia
alexander.pechenkin@ibks.ftk.spbstu.ru

Dmitry Zegzhda
St. Petersburg State Polytechnical University
St. Petersburg, Russia
dmitry.zegzhda@ibks.ftk.spbstu.ru

Abstract—**The paper discusses a new approach to provide user tasks distribution in Grid systems using Petri nets unfolding. Branched Petri nets definition is proposed to describe Grid system security. Partial order method is applied to reduce size of Petri net model of Grid. Secure user tasks distribution method and system are suggested for automated security maintenance in the Grid. Solution of state explosion problem in branched Petri net model of Grid is proposed. Implemented access control system and obtained experimental results demonstrate successful solution of data protection against insiders in Grid systems.**

Keywords-grid; information security; Petri net; unfoldings; tasks distribution

I. INTRODUCTION

Implementation of security features in modern distributed computing systems, especially in Grid systems, which process confidential and restricted data, is accompanied by reduction of their scalability and parallelizability. It leads to preventing Grid systems from resource sharing thus turning it into a set of weakly bounded hosts.

Growing number of security violations relative to Grid systems (e.g., CVE-2009-0046 incident in Sun GridEngine, CVE-2013-4039 in IBM WebSphere Extended Deployment Compute Grid) proves high importance of task aimed at protection of data being processed in the Grid with minimal loss of Grid functionality.

Further, in Section 2 we consider existing works dedicated to solve this problem. Application of branching Petri nets to the Grid representation is provided in Section 3. Section 4 discusses the suggested solution of the state explosion problem basing on partial order method. In Section 5, we propose secure user tasks distribution method and system. Section 6 reviews the work results.

II. RELATED WORK

Grid systems provide the availability of increased amounts of valuable computing and information resources. Such information systems heavily depend on the provision of high security level. Naqvi and Riguidel [1] present a survey of the various Grid system threat models.

Special hardware and software components are used to provide protection against denial of service attacks and the spread of malicious software in Grid systems. These components also called security managers include Intrusion Detection Systems (IDS), firewalls and antivirus agents [2]. There are also several works aimed at solving the problem of anomaly detection in distributed computing systems [3][4].

Security managers are integrated with dedicated communication channels, which in the case of intrusion detection alerts are broadcast. After receiving such a notification, each host duplicates it to all resource providers being connected to it. As a result, all hosts isolate the problematic host. It thus prevents the possibility of attacks spreading in Grid systems.

In addition, in some Grid systems, fuzzy trust logic is used [5]. Each host is initially labeled. This label shows the trust level assigned to it by other components of Grid system. If attack from that host is fixed, the trust level is decreased. While search for a suitable host for a user-defined task, the hosts with the highest trust level are chosen for running this task.

In existing products, such as Grid Resource Allocation and Management (GRAM) in Globus Toolkit [6] and Community Authorization Service (CAS) in gLite [7], there are authentication and authorization mechanisms implemented to control user tasks access to Grid resources.

Definition and realization of security policies in these solutions are based on a set of Virtual Organizations (VOs) and fixed states of the Grid [8]. They do not take into account high dynamics and access rights distribution at the level of user tasks. Therefore, it might cause unauthorized access to data being processed in the Grid.

This paper refers to development of Grid system model, taking into account high dynamics of user tasks distribution, and suggests secure user tasks distribution method based on this model. In [9], an algorithmic behavior model of multi-agent distributed system is proposed. This model is based on adaptive random graphs mathematical apparatus and takes into account sufficiently high frequency of the number of nodes and links between nodes changing. There is high frequency of node status and user tasks distribution between nodes in Grid systems that can be observed. Whereas to add or delete a node in the Grid, you must pass the verification procedure which means that the number of nodes in such distributed network changes quite rarely.

In [10][11], an unfolding technique is formally described and applied to colored Petri nets to describe branching processes whose behavior close to the Grid. Branching

 ISBN: 978-1-61208-330-8

process is a Markov process [10] that models a population in which each individual in generation n produces some random number of individuals in generation $n + 1$. They propose a model of branching processes, suitable for describing the behavior of general Petri nets, without any finiteness or safeness assumption [11]. In this paper we extend the results of this work with reference to Grid systems security feature.

III. GRID SYSTEMS MODELING

Implementation of this approach involves the mathematical apparatus of functional colored Petri nets. A Grid system is described with Petri net $N = (RP, T, F, M)$, where $RP = \{rp\}$ is the finite set of vertices of graph which represents Grid nodes (hosts, resource providers, etc.), $T = \{t_i\}$ is a set of transitions between the vertices, $F = RP \times T \cup T \times RP$ is a set of arcs [12]. Markers $\{m\}$ denote user tasks from J (i.e., requests for a particular type of Grid resource).

T-transition of Petri net (N, M_0) or a simple transition is defined as a transition $t_{ij} \in T$, where mark M' directly accessible from mark M has the following form: $M' = (m_1, \ldots, m_i - 1, \ldots, m_j + 1, \ldots, m_n)$. Graphical model of T-transition is presented in Fig. 1.

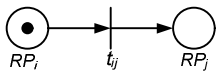


Figure 1. T-transition.

F-transition of Petri net (N, M_0) or branching is defined as a transition $t_{ijk} \in T$, where mark M' directly accessible from marking M has the following form: $M' = (m_1, \ldots, m_i - 1, \ldots, m_j + 1, \ldots, m_k + 1, \ldots, m_n)$. Graph representation of F-transition is provided in Fig. 2.

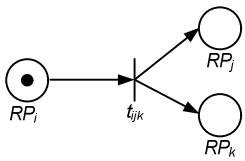


Figure 2. The graph representation of F-transition.

Let us define the branching Petri net as a subclass of colored functional Petri net.

Definition 1. The branching Petri net is a colored functional unlimited Petri net which has only T- and F-transitions.

Any Grid system can be represented as a graph of branching Petri net. T-transition means the simple migration of user request from one node to another. There are no new markers appear in that case, which means that the summary marker counts taken before and after transition activation are the same.

F-transition means situation when the computing power of multiple Grid nodes is required to cope with user task. The total count of markers is increased according to the number of nodes involved into user task processing.

Any user task migration in the Grid keeps arcs multiplicity at the level of one. Taking into account specified definition, there are only T- and F-transitions can exist in the Grid.

IV. SOLUTION OF STATE EXPLOSION PROBLEM

While modeling such a high distributed systems as Grid systems number of states grows exponentially with an increase in the number of nodes. For example, an initial marking of any Petri net representing the Grid has the following form: $M_0 = (m_1, \ldots, m_n)$. Each marker of this marking set assumes a value in the range of 0 to n_A, where n_A is a number of active nodes of Grid. Assume that there is no user task can be produced on any node before previous one would have been finished ($n_A \leq n$). Then the total number of states describing such Grid is n_A^n (e.g., $n = 1000$, then power of states set is 10^{3000}). These problem is known as a state explosion problem.

There is a partial order method implemented to solve this problem. Define the partial order on the set of markings of Petri net. Let M_1 and M_2 be the markings of Petri net (N, M_0). Assume that $M_1 \leq M_2$ being in a partial order relationship, if for every marker m of marking M_1 situated in p position, there is a marker m' of marking M_2 in the same position p, where m' is equal to or greater than m.

M_1 is less than M_2 relatively to order $\leq$, if marking M_1 can be obtained from marking M_2 by sequential markers removing from Petri net vertices. Using this fact as a basis, an inductive definition of minimal partial order can be done.

Definition 2. Minimal partial order of marking M of Petri net is a natural number D such that for any marking M' being in partial order with M: $M' \leq M$, there is no marker m' of marking M', where $m' < D$.

Lemma 1. Minimal partial order of the branching Petri net marking is equal to 1.

Proof. According to definition 1, the branching Petri net is unlimited. It means that it could be any number of markers (user tasks) at any position. At least one marker at the position is enough for transition to be triggered because of the multiplicity of branched Petri net. Therefore, 1 is a minimal natural number to which it is possible to decrease the amount of markers at every position of Petri net.

Theorem 1. Any marking of the branched Petri net N reachable from marking M is also reachable from marking $M' = (m'_1, \ldots, m'_n)$, $m_i' = \{0, 1\}$ which is obtained from N by applying to it the partial order equal to 1.

Proof. Consider the simple case when the branched Petri net is 2-limited. Common case can be proved in induction. For every $m \in M : m \leq 2$. By the definition the branching Petri net consists of aggregate of T- and F-transitions. Consecutively, consider all possible ways of such Petri net fragment aggregation.

 ISBN: 978-1-61208-330-8

- T-transition—T-transition. Branched Petri net fragment of such kind has 1 of 2 forms, as shown in Fig. 3.

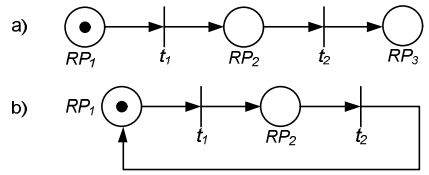


Figure 3. The branched Petri net fragment of T-transition—T-transition type.

In that case, the amount of markers at each position can be arbitrarily large. Find the set of reachable states for each form of branched Petri net fragment, taking into account deterministic form of transition function which means that *ceteris paribus* transitions are triggered simultaneously. For simplicity we also agree that all markers have the same type.

The fragment of Petri net illustrated in Fig. 3 generally has marking M_1 = (a, b, c), where a, b, c — natural numbers, and has the following set of the states: $R(M_1)$ = {(k, l, m), (0, l, c), (k, 0, m), (k, l, 0), (k, 0, 0), (0, l, 0), (0, 0, m), (0, 0, 0)}, where k, l, m — natural numbers and max(k, l, m) ≤ max(a, b, c).

Apply net partial order relationship, where D = max(a, b, c) - 1. Then the resulting Petri net has marking M_1' = (x, y, z), where x, y, z — natural numbers and max(x, y, z) = max(a, b, c) – 1. Therefore $R(M_1')$ = {(k', l', m'), (0, l', c'), (k', 0, m'), (k', l, 0), (k', 0, 0), (0, l, 0), (0, 0, m'), (0, 0, 0)}, where k', l', m' — natural numbers and max(k', l', m') ≤ max(x, y, z) ≤ max(a, b, c), i.e., $R(M_1) = R(M_1')$.

Thereby in induction: $R(M_1) = R(M_1')$ for $\forall D \in N : M_1' \leq M_1$. According to the Lemma 1 the minimal partial order of branching Petri net marking $D_{min} = 1$. In addition $\forall m' \in M' : m' = \{0, 1\}$. Thus, we have $M' = (m'_1, \ldots, m'_n)$, $m_i' = \{0, 1\}$.

- F-transition—F-transition. Branched Petri net fragment of such kind has 1 of 2 forms, as shown in Fig. 4.

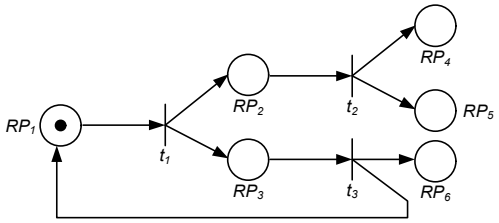


Figure 4. The branched Petri net fragment of F-transition—F-transition type.

The fragment of Petri net illustrated in Fig. 4, generally, has marking M_2 = (a, b, c, d, e, f), where a, b, c, d, e, f — natural numbers and has the following set of reachable states $R(M_2)$ = {(k, l, m, n, o, p), (0, 0, 0, n, o, 0), (0, 0, 0, n, 0, 0), (0, 0, 0, 0, o, 0), (0, 0, 0, 0, 0, p), (0, 0, 0, 0, 0, 0)}, where k, l, m, n, o, p — natural numbers and min(k, l, m) ≥ min(a, b, c, d, e, f).

Apply partial order relationship, where D = min(a, b, c, d, e, f) + 1. Then, the resulting Petri net has marking M_2' = (x, y, z, α, β, χ), where x, y, z, α, β, χ — natural numbers and min(x, y, z, α, β, χ) = min(a, b, c, d, e, f) + 1. Hence $R(M_2')$ = {(k', l', m', n', o', p'), (0, 0, 0, n', o', 0), (0, 0, 0, n', 0, 0), (0, 0, 0, 0, o', 0), (0, 0, 0, 0, 0, p'), (0, 0, 0, 0, 0, 0)}, where k', l', m', n', o', p' — natural numbers and min(k', l', m', n', o', p') ≥ min(x, y, z, α, β, χ) ≥ min(a, b, c, d, e, f), i.e., $R(M_2) = R(M_2')$.

Thereby in induction: $R(M_2) = R(M_2')$ for $\forall D \in N : M_2' \leq M_2$. According to the Lemma 1 the minimal partial order of branching Petri net marking $D_{min} = 1$. Thus, we have $\forall m' \in M' : M' = (m'_1, \ldots, m'_n)$, $m_i' = \{0, 1\}$.

- F-transition—T-transition. Branched Petri net fragment of such kind has 1 of 2 forms, as shown in Fig. 5.

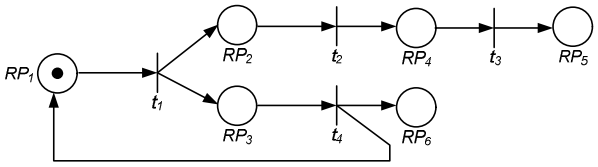


Figure 5. The branched Petri net fragment of F-transition—T-transition type.

The fragment of Petri net illustrated in Fig. 5 is a composition of fragments illustrated in Figs. 3-4. Following the induction, we have $R(M_3) = R(M_3')$ for $\forall D \in N : M_3' \leq M_3$, where M_3 — marking of specified Petri net fragment. Thus, we get $\forall m' \in M' : M' = (m'_1, \ldots, m'_n)$, $m_i' = \{0, 1\}$.

The provisions of this theorem allow to reduce a set of the states which describe the Grid from to n_A^n to 2^n.

V. SECURE USER TASKS DISTRIBUTION

Current security-based solutions referenced to describing and enforcement of security policies operate with a set of virtual organizations and fixed Grid states. These solutions do not take into account real access rights distribution at the level of user tasks running on Grid system nodes. They also miss the fact of high intensity of user tasks migration between such nodes. Thus, it leads to the possibility of unauthorized access to processed data.

The proposed method of secure user tasks distribution is based on the solution of a reachability problem in terms of Petri net describing the Grid. There is a reachability graph suggested to create for the specified Petri net N= (RP, T, F, M) with initial marking $M_0 = (m_1, \ldots, m_n)$, where vertices are the marking with minimum partial order equal to 1. According to the Theorem 1, that tree must be a finite one. Vertices of this graph form a set of states which the system may reach. For every state, a formalized transition function

 ISBN: 978-1-61208-330-8

is used to determine compliance with security policy constraints. Verification is performed by comparison of security policy requirements with the current state.

As a result, there is the set of Grid nodes user tasks, transmission to which is permitted by security policy rules. After that, the rules of user tasks distribution are transmitted to such nodes (Fig. 6).

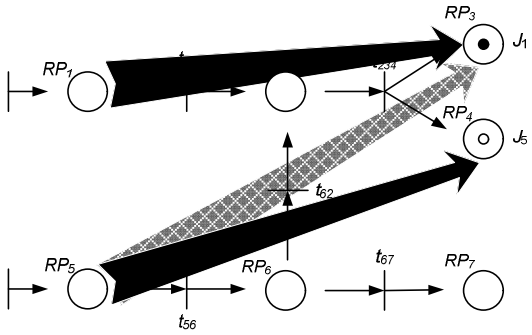


Figure 6. Secure user tasks distribution scheme

The secure user tasks distribution system is proposed and its effectiveness estimation shown. Total time required to process some user task in the Grid includes a time of user task processing by n nodes, time of data transmission between n nodes and time required to define permitted task distribution according to the discussed method.

$$T = t_p + t_m + t_s, \quad (1)(1)$$

According to (1), t_p is a time of user task processing by n nodes; t_m is a time of data transmission between n nodes; t_s is a time required to define permitted task distribution according to the discussed method.

$$t_p = \frac{k_1}{s(n)}, s(n) = \frac{1}{k_2 + \frac{1-k_2}{n}}, \quad (2)$$

According to (2), k_1 is a total time of user task processing by single node; k_2 is a portion of the task that cannot be parallelized (it remains serial).

$$t_m = n\frac{k_3}{k_4}, t_s = \frac{k_5 \cdot 2^n}{k_6}, \quad (3)$$

According to (3), k_3 is an amount of data required for user task processing being transported to each Grid node; k_4 is a data rate; k_5 is a number of operations required for Grid state verification that is being represented by the reachability tree of branched Petri net; k_6 is a computing power of node.

Relative reduction in time costs required for processing of restricted information in Grid system is calculated as a division of time required to process user task with proposed method and without it.

$$Q = \frac{\sum_{i=1}^{d}(k_{1i} \cdot (k_2 + \frac{(1-k_2)\cdot c}{n}) + \frac{n}{c} \cdot \frac{k_3}{k_{4i}})}{\sum_{i=1}^{d}(k_{1i} \cdot (k_2 + \frac{(1-k_2)}{n}) + n \cdot \frac{k_3}{k_{4i}} + \frac{k_5 \cdot 2^n}{k_{6i}})}, \quad (4)$$

In (4), c is a number of restricted data classification categories being processed in the Grid; d is a number of iterations used to perform the specified task.

Typical user tasks in complex distributed analytical systems which use Grid software to perform processing of huge amount of data (e.g., CERN [8]) have the following parameters: $k_1 \approx 1$ hour, $k_2 \approx 20\%$, $k_3 \approx 512$KB, $k_4 \approx 100$ Mb/sec, $k_5 \approx 102$ oper, $k_6 \approx 4 * 1010$ oper/sec, $n \approx 2000$, $c = 5$ (sample). Experimental results for Grid sample with such characteristics are shown in Fig. 7. As one can see, Grid system with integrated proposed access security subsystem requires significantly less time to process user tasks than Grid system with organizational measures aimed to divide the Grid in isolated segments to prevent restricted data leaks. Experiments have been performed on a laboratory stand included of 300 compute nodes, which are virtual machines of multiprocessor computer running on Xen Cloud Platform [13]. Software infrastructure is based on Globus Toolkit 5 [6].

Implementation of secure user tasks distribution system in the Grid allows us to protect data from user privilege escalation, simultaneously reduce the time expenses associated with the security assessment and increase the productivity of Grids which processes classified data.

VI. CONCLUSION AND FUTURE WORK

The new approach of Grid systems modeling based on mathematical apparatus of Petri nets is proposed. Subclass of colored Petri nets called 'branched Petri nets' is used to represent behavior of the Grid. Extremely large size of models representing real high distributed systems causes problem known as state explosion. Partial order method is applied to branched Petri net to solve it.

Proposed secure user tasks distribution method based on technique of reachability tree construction and subsequent security verification of obtained tree nodes. Implemented access control system provides successful solution of data protection against attacks based on user privilege escalation technique.

Future work of proposed solution involves access control system integration to most popular Grid software infrastructures. Different types of authorization techniques also must be taken into account.

REFERENCES

[1] S. Naqvi and M. Riguidel, "Threat model for grid security services", Lecture Notes in Computer Science, vol. 3470, 2005, pp. 1048-1055, doi:10.1007/11508380_107.

 ISBN: 978-1-61208-330-8

[2] H. Lohr, H. V. Ramasamy, A. Sadeghi, S. Schulz, M. Schunter, and C. Stuble, “Enhancing grid security using trusted virtualization”, Lecture Notes in Computer Science, vol. 4610, 2007, pp. 372-384, doi: 10.1007/978-3-540-73547-2_39.

[3] T. Stepanova, D. Zegzhda, M. Kalinin, and P. Baranov, “Mobile anomaly detector module based on power consumption analysis”, Proc. International Conference on Information Security and Privacy (ISP-10), Jul. 2010, pp. 85-89.

[4] M. Burgess, “Probabilistic anomaly detection in distributed computer networks”, Science of Computer Programming, vol. 60, Mar. 2006, pp. 1-26, doi:10.1016/j.scico.2005.06.001.

[5] S. Song, K. Hwang and M. Macwan, “Fuzzy trust integration for security enforcement in grid computing”, Lecture Notes in Computer Science, vol. 3222, 2004, pp. 9-21, doi: 10.1007/978-3-540-30141-7_6.

[6] D. Gomez, “Secure collaborative grid computing”, Johns Hopkins Whiting School of Engineering, Spring 2008.

[7] L. Pearlman, C. Kesselman, V. Welch, I. Foster, and S. Tuecke, “The community authorization service: status and future”, Proc. of Computing in High Energy Physics 03 (CHEP '03), 2003, pp. 44-52.

[8] A. Chakrabarti, “Grid computing security”, Berlin: Springer, 2007.

[9] T. Stepanova and D. Zegzhda, “Stochastic model of interaction between botnets and distributed computer defense systems”, Computer Network Security, Lecture Notes in Computer Science, vol. 7531, 2012, pp. 218-225, doi: 10.1007/978-3-642-33704-8_19.

[10] V. Kozura, “Unfoldings of coloured Petri nets”, Lecture Notes in Computer Science, vol. 2244, 2001, pp. 268-278, doi:10.1007/3-540-45575-2_27.

[11] J. Couvreur, D. Poitrenaud and P. Weil, “Branching processes of general Petri nets”, Lecture Notes in Computer Science, vol. 6709, 2011, pp. 129-148, doi: 10.1007/978-3-642-21834-7_8.

[12] P. Zegzhda, D. Zegzhda, M. Kalinin, and A. Konoplev, “Security modeling of grid systems using Petri nets”, Computer Network Security, Lecture Notes in Computer Science, vol. 7531, 2012, pp. 299-308, doi: 10.1007/978-3-642-33704-8_25.

[13] “XCP Overview”, web site: *wiki.xenproject.org/wiki/XCP_Overview*.

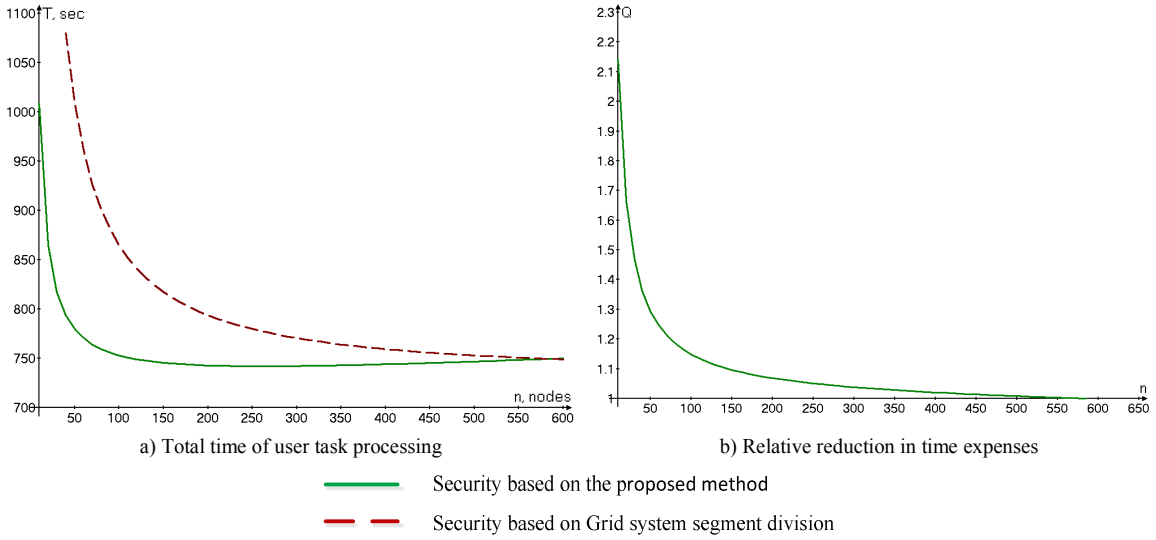


Figure 7. Secure user tasks distribution system experimental results.

Analysis of Scheduling Algorithms with Migration Strategies in Distributed Systems

Francisca Aparecida P. Pinto*, Chesley B. Chaves†, Lucas G. Leite‡, Francisco Herbert L. Vasconcelos§ and Giovanni C. Barroso¶

Federal University of Ceará (UFC)
Department of Teleinformatics Engineering *§¶, Department of Computer Science‡ and UFC Virtual Institute†§, Ceará, Brazil
{aparecida.prado, herbert}@virtual.ufc.br, chesleybraga@gmail.com, lucasgml@alu.ufc.br, gcb@fisica.ufc.br

***Abstract*—Task scheduling is a problem which seeks to allocate, over time, various tasks from different resources. In this paper, we consider group task scheduling upon a heterogeneous multi-cluster system. Two types of job tasking are considered, parallel and sequential. In order to reduce fragmentation caused by the scheduler group, migration mechanisms were implemented. Moreover, the dispatchers global and local use distribution of jobs in order to minimize delays in the task queues, as well as in response time. To analyze the different situations, performance metrics were applied, aiming to compare schedulers in different situations.**

Keywords-parallel job; scheduling; distributed Systems;

I. INTRODUCTION

Traditionally, the main focus of the industry has been to improve the performance of computing systems through more efficient projects, hereby increasing the density of its components. Associated with the exponential growth of data size in simulation/scientific instrumentation, storage and internet publications, the increased computational power of such systems boosted investment by big providers, government research laboratories and computing environments, thus enhancing robustness in order to host applications ranging from social networks to scientific workflows [1].

In this context, distributed systems emerge as an interesting solution for the provision of physical resources upon demand, as they allow additional computational power of several nodes interconnected by a computer network, in order to perform tasks. Distributed computing systems have been used due to their important attributes, such as: cost efficiency, scalability, performance and reliability. In grid computing, there are three important aspects that must be treated: task management, task scheduling and resource management [2]. In particular, Grid Task Scheduling (GTS) plays an important role in the system as a whole, and its algorithms have a direct effect on the grid system. Task scheduling in a heterogeneous computing environment proved to be a NP-complete problem [3].

To solve this problem, various scheduling algorithms have been proposed for distributed environments, whereby they have been classified in several different ways. For example, as shown in [4], we propose a hierarchical tree classification which splits at the highest hierarchy level into the local and global algorithms. With reference to [5], the authors classify the algorithms according to the types of applications found in the grids: meta-task and Directed Acyclic Graph (DAG) algorithms. Actions can be executed simultaneously and independently through meta-task algorithms, whereas the type DAG algorithm contains precedence constraints. Moreover, they are classified into traditional algorithms, deterministic and heuristic intelligence, for the use of different optimization technologies.

Existing scheduling techniques, highlight scheduling groups or co-schedulers [6], are considered efficient algorithms for the purpose of scheduling parallel jobs that consist of tasks that must be allocated and executed simultaneously on different processors [7]. These types of scheduling algorithms provide interactive response times for tasks with low execution time by means of preemption, with the disadvantage of causing fragmentation and reducing system performance [8]. The fragmentation of resources has been a common subject of research in the last two decades [16]. Various approaches to the fragmentation of resources have been developed, whereby better fit and task migration are the two most common approaches.

Based upon the above, this paper aims to reduce the fragmentation caused by the scheduler group and response time. Therefore, the main contributions of this work are: i) application of a migration mechanism task schedulers group, in order to minimize the fragmentation and response time. ii) implementation of strategies inside dispatcher managers, aiming the distribution of tasks to the clusters, to avoid unnecessary migrations, improving system efficiency; and iii) implementation of a heterogeneous multi-cluster system with the objective of analyzing the performance of schedulers in different situations, as well as the system behaviour in different contexts.

This paper is organized as follows: Section II presents related work; Section III presents the model of the proposed system; Section IV describes the operation of the system; Section V illustrates the group schedulers and mechanisms of migration; Section VI presents the performance metrics; Section VII presents the results of simulations, and finally, Section VIII covers the completion of the work, along with the prospects for future development.

II. RELATED WORK

The group schedulers, Adapted First Come First Served (AFCFS) and Largest Gang First Served (LGFS), used in

 ISBN: 978-1-61208-330-8

this work, have been studied in a distributed environment, [7][9][10][11]. The aim of this study is to make them more efficient in scheduling of parallel tasks, since these algorithms cause fragmentation in the system. In [7][10][11][12], there are proposed migration mechanisms, which are utilized in order to minimize the fragmentation caused by the schedulers group in the distributed environments. In this study, besides the technical migration, other strategies are used (Section V-A), in order to avoid unnecessary migrations, as well as overloading of the system. In [13][14], the authors propose a two levels system model within a grid environment. In the first level, containing the global scheduler, there is an overview of the application of the job task, and subsequently, the second level scheduler has knowledge of all resource details. Moreover, they consider load balancing through the global manager scheduler only. In our proposal, we implemented a model that uses heterogeneous multi-cluster system managers, Grid Dispatcher (GD) and Local Dispatcher (LD), for the purpose of allocating jobs on resources. Therefore, unlike the authors mentioned above, we introduce the GD before sending jobs to clusters, which is information feedback regarding the load of the clusters, so that a more efficient load balance is achieved. Moreover, the distribution of tasks for the processor queue, the LD, requests information regarding the load of each processor, namely the number of running tasks in a long processor queue. Such information is required in order to reduce the time in the task queue and the response time of a job.

In some work studies concerning the group schedulers above, migration mechanisms are used to reduce the fragmentation of the system, whereby a metric used to evaluate the scheduler in relation to fragmentation is not applied. Therefore, in this paper, in order to analyze the different applied situations in addition to other metrics, we use the Loss of Capacity (LoC) function, so that we are able to analyze the performance of schedulers in different situations, as well as the behaviour of the system in different contexts. The metric LoC is relevant to measure both the use of the system as the fragmentation [15][16][17]. These authors applied the metric in different contexts.

III. Environment Description

The simulation environment consists of a grid multi-cluster system which uses the hierarchical structure of two layers. Such an environment was developed by the Research Group in Applied Computational Modeling of the UFC, in Java. This system is composed of managers, Grid Dispatcher (GD) and Local Dispatcher (LD).

The GD is responsible for the sending of jobs, both parallel and non-parallel for clusters, and LD is responsible for sending the task to the jobs belonging to the ranks of processors based on the algorithm Join The Shortest Queue (JSQ). Each cluster is comprised of a LD and a set of processors. The system is heterogeneous in terms of clock rate *r* and the number of processors *p* per cluster. The clock on machines can vary between $1500 \leq r \leq 3000$ (megahertz), which is randomly generated upon creation of the resources in the simulation environment. More importantly, the larger the value of *r*, the time between each processing cycle will be shorter and therefore tasks are executed in less time. In the implemented system, the distribution processor *p* is made in two clusters of 32 and 64 processors respectively, whereby each processor has its own queue. The model system is illustrated in Fig. 1.

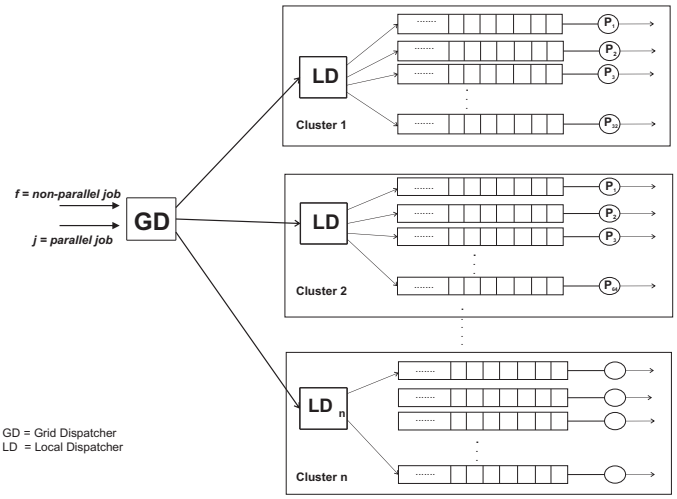


Figure 1. Multi-cluster system based on queues

In this environment, we assume that the two clusters belong to the administrative domain, such that they can communicate with GD. Moreover, communication between processors is contention free. Thus, we consider the communication latencies are implicitly included on the service time of the job. The workload applied in this system was extracted from a real distributed environment [18]. This workload is composed of two different kinds of jobs that are competing for the same resources: non-parallel job *f* and parallel job *j*. In the workload, each job is described by a tuple (*id*, *at*, s_j, *pt*): identification of job *id*, arrival time *at* where $at > 0$; s_j size of a job, in which $1 \leq s_j \leq 64$, and the processing time *pt*, whereby $pt > 0$.

In this paper, we assume that the value of a job *pt* workload will be applied in a machine, whereby $r \doteq 2000$ megahertz, as it currently has a median frequency processor. Otherwise, the *pt* undergoes change and may vary proportionally according to the clock of the machine, that is, when the value of $r \neq 2000$ megahertz (standard). Therefore, the calculation of the new processing time P_t is defined by (1).

$$P_t = pt \times \frac{C_d}{C_m} \tag{1}$$

whereby *pt* is the processing time of the job on the workload, C_d is the standard clock where $C_d \doteq 2000$ (megahertz), and C_m represents the value of the *r* of the machine, which relates to $1500 \leq r \leq 3000$.

A job, *f*, consists of a single task, whose execution begins immediately upon its arrival on the grid. The system is not capable of responding to more than 64 jobs *f* per time unit. At present, a job *j* consists of t_j tasks, where $1 < t_j \leq 64$, hence, the number of tasks in a job *j* cannot exceed the number of processors in a cluster. Thus, the risk that a job may never be answered is null. Moreover, mapping between tasks and processors must be one to one. Thus, tasks from the same job cannot be attributed to the same processor queue. A job *f* is a high priority task, requiring only a processor *p* for its execution. Therefore, a processor that receives a priority task must immediately stop the execution of any other task type

 ISBN: 978-1-61208-330-8

j, in order to serve the task type *f*. If a job *j* has one of its tasks interrupted by a job *f*, then each sibling task of *j* has to stop its execution, and then be rescheduled. Stopping the task t_j that belongs to a job *j*, can affect even more response time of *j*. As soon as *f* ends its execution, interrupted tasks can begin their execution again, as well as the entire process. It is noteworthy that disruption only occurs when all processors are busy. Moreover, jobs *f* can not interrupt one another.

IV. Operation of System Managers

This section will describe in detail, the operation of system managers, in other words, the Grid and Local Dispatchers, as illustrated in Fig. 1.

A. Grid Dispatcher

As stated earlier, the GD is responsible for sending jobs to the clusters. This submission is made based upon feedback information concerning the total load of each cluster, i.e., the total number of jobs in the queues, plus the number of tasks that are running on the processors. This cluster load information will be sent only at the request of GD, since excessive feedbacks may cause overload on the system. The average load between clusters is very important for more efficient load balancing. If the clusters are randomly balanced, then dispatch occurs. The calculation of the total load of the cluster is defined by (2).

$$C_i = \frac{1}{n} \times \sum_{p=1}^{n} [t(p) + t_s(p)] \quad (2)$$

whereby C_i is the total load per cluster, *n* is the total number of processors per cluster, *t(p)* is the total number of tasks in the queue of each processor *p*, and $t_s(p)$ represents the existence or non existence of a task running on processor *p*, $t_s(p) = 1$, if there is a running task, otherwise $t_s(p) \doteq 0$.

B. Local Dispatcher

After the parallel job *j* has been sent to a cluster *c*, according to the load of *c*, LD assigns the tasks to the available queues based on algorithm JSQ. JSQ is responsible for sending the tasks that belong to a job queue for processors that have fewer tasks in their own queues. It is important to emphasize that when a job *f* arrives at LD, it is forwarded to the cluster by JSQ.

In this work, an adaptation has been implemented in LD as a new criterion in the selection of the processor. This adaptation works as follows: LD receives information about the cluster load of each processor, i.e., the number of tasks in the processor queue plus the running task t_s. The information feedback only occurs when LD asks, thus avoiding system overload. The knowledge of the load of each processor is to minimize the delay of tasks to the processor queues and the response time of the job. The calculation of the adaptive LD, the adaptive Local Dispatcher (aLD) is defined by (3).

$$N_t(q) = n_t + t_s \quad (3)$$

whereby $N_t(q)$ is the total number of jobs per queue, *q* is the size of the processor queue, n_t represents the number of jobs in the queue and t_s represents the existence or non existence of a task running on processor, $t_s \doteq 1$ if there are any tasks running on the processor, otherwise, $t_s \doteq 0$.

With the aLD, JSQ sends tasks to the processors more efficiently, since this algorithm now has information of the effective value of the load of each processor. In Section VII, we present the impact that the adaption on LD causes on the results of the response time of the jobs. For this analysis, the LD will be applied with and without the adaptation in both group scheduling algorithms. In the next section, we present the group schedulers that were used for scheduling jobs in the queues of the processors.

V. Group Schedulers

In the simulation model, the following policies have been applied to the analysis of queues: AFCFS and LGFS [7][9] [10][11]. These schedulers were modified and implemented in each cluster, separately.

The algorithm AFCFS tends to favour jobs that consist of a number of smaller tasks, so that jobs that require a smaller number of processors, result in an increased response time for larger jobs. But the LGFS tends to favour the performance of bigger jobs instead of the smaller ones, i.e., bigger jobs have their jobs put on queues of processors before any others belonging to a job with a smaller size, resulting in an increase of response time of smaller jobs. In addition, LGFS involves a considerable amount of overhead in the system. Therefore, such scheduling algorithms cause fragmentation in the system, which happens in two stages: i) the schedulers cannot always meet the requirements of a job *j*, since the latter requires a number of available processors equal to the number of tasks, in order to execute; and ii) when there are idle nodes and tasks waiting in the queue to be executed, they are not able to schedule these tasks.

A. Migration

Assuming that the group scheduling causes fragmentation in the environment, we look at the use of migration to reduce fragmentation. In this work, we have studied different migration schemes for heterogeneous systems, in order to minimize such problems. Therefore, we assume two types of migration: local migration m_l and external migration m_e.

The difference between migration m_l and m_e is that the latter causes a higher overhead on the system, since it involves the transfer of tasks from one cluster to another. Therefore, the following strategies have been proposed in order to reduce fragmentation, as well as unnecessary migration, and consequent overloading of the the system:

1) checks all clusters that have processors available;
2) analyzes which jobs has its tasks at the beginning of the queue of idle processors;
3) based upon the above analysis, it checks which of the jobs has the least or equal number of tasks to that of the idle processors;
4) and finally transfers the tasks of the job that has fewer number of tasks to migrate.

During the migration tasks, the destination nodes are reserved, in order to prevent other tasks using them. When the target

 ISBN: 978-1-61208-330-8

processor is reserved, we ensure the immediate start up of their performances after the migration of the tasks are chosen. The only way you can prevent the execution of these migrated tasks, is the arrival of a job *f*. If this problem occurs, migrated tasks are reserved, and when job *f* liberates the processor, they return and begin execution immediately. The reserved node can only be occupied by another job if it is of type *f*.

The reservation mechanism prevents elected tasks returning to the queue, because the scheduler AFCFS or LGFS would not be able to schedule such tasks. These schedulers are not suitable for parallel scheduling of tasks in different clusters. Eventually, the m_e was applied in an attempt to use more resources as to avoid resources remaining idle. In addition, we have the use of aging, in order to regulate the number of migrations that occur in the system, as well as reducing the starvation of tasks that came before.

The migration strategies were employed in the algorithms of AFCFS schedules and LGFS (Section V), which were used in each cluster separately. Therefore, these algorithms with migration are defined as AFCFSm and LGFSm, respectively. First, the scheduling hierarchy requires to run the scheduling algorithms AFCFS and LGFS, and then the migration m_l tries to schedule the jobs that were not able to be allocated by the scheduling algorithm. The migration m_e can only be used in an attempt to make use of more resources. The reason for this hierarchy is the overhead imposed by each of these steps. Unlike migration techniques, the schedulers (AFCFS and LGFS) without migration does not cause any additional overhead.

In the next section, we will describe the performance metrics applied to analyze the system model behaviour in different situations.

VI. Performance Metrics

In this work, we applied the following performance metrics: Average Response Time (ART), Utilization (U) and Loss of Capacity (LoC), which constitutes everything required in order to analyze the performance of schedulers in different situations, and the system behaviour in different contexts.

A. Average Response Time

The metric response time *rt* (in time units) measures the time interval between the arrival of the job in the system until the end of its execution [17], thus, the average response time or ART is given by (4).

$$ART = \frac{1}{m} \times \sum_{j=1}^{m} rt(j) \tag{4}$$

whereby ART is the average response time of jobs, *rt(j)* represents the response time of a job, and *m* is the total number of jobs executed.

B. Utilization

In simulation studies, the metric used is simply an indirect measure of makespan [20], with the workload constant for all schedulers. The calculation of the metric used is given by (5):

$$U = \frac{\sum_{j=1}^{m} p_j \times te_j}{makespan \times N} \tag{5}$$

whereby U is utilization of the clusters, p_j represents the number of processors that each job needs for its implementation, te_j represents the execution time of each job, N represents the size of the system and *m* is the total number of jobs executed.

C. Loss of Capacity

This metric is important for measuring both uses of the system as fragmentation. In a system, fragmentation occurs when: (i) there are tasks waiting in the queue to run; and ii), there are idle nodes, but still unable to run the waiting tasks. LoC reflects the costs of fragmentation. These metrics have been used in some of the works as detailed in [15][17][19], respectively. To use the LoC, we assume two factors: 1) the number of tasks of a job, *j*, can not exceed the number of processors in the system, avoiding the starvation, in other words, the job would never be serviced; 2) the variable δ (delta), (6), represents the state of the processors and jobs in the system. For example, $\delta = 1$ indicates the existence of enough available processors to perform at least one job in the queue at the moment a new job is scaled. Likewise, when $\delta = 0$, if the queues are empty or do not exist in the same job size less than or equal to the number of idle processors. The metric LoC is calculated as follows:

$$LoC = \frac{\sum_{j=1}^{q-1} n_i(t_{i+1} - t_i)\delta}{N(t_q - t_1)} \tag{6}$$

whereby *q* represents the number of jobs in the staggered moment when a new job is scheduled or when a job terminates execution. This is indicated by the time t_i, for $i = 1 \cdots q$ and n_i represents the number of idle nodes between *i* and $i + 1$.

VII. Analysis of Results

A. Input Parameters

The simulations were performed using a simulation application implemented in Java, which was developed by the Research Group in Applied Computational Modeling, allowing the simulation of entities in systems of parallel and distributed computing. Therefore, this application was developed for the following purposes: analysis of the mechanisms used in dispatchers, responsibility for the distribution of jobs in the clusters, and evaluation of different scheduling algorithms covered in this work.

In this environment, we assume that two clusters (32 and 64 processors) belong to the administrative domain, so that they are able to communicate with the GD. For analysis of the environment, we use various traces extracted from a real distributed environment. However, the workload used for the experiment consists of 1,500 jobs in total and 24,152 tasks, which are described by the tuple (*id*, *at*, s_j, *pt*), Section III. Furthermore, the workload used in the simulation of job, *j*, has different characteristics, such as size of each job, processing time, among others. The mapping between tasks and processors is one to one, with a total of 96 processors in the system. For the simulation, we proposed three scenarios: i) in the S1, the schedulers AFCFS and LGFS (without migration) were used; ii) in the S2, the schedulers AFCFS and LGFS with migration mechanisms (AFCFSm) and (LGFSm) were applied; iii) and in the S3, the schedulers AFCFSm and LGFSm with

 ISBN: 978-1-61208-330-8

the adaptation strategy on the Local Dispatcher (aLD) were used.

It is noteworthy that in all three scenarios, we applied the strategy of GD as decision making in the distribution of jobs among clusters. For each scenario, ten simulations were performed, from which we calculated the average values of waiting times, response times and LoC, with a confidence interval of 95%. In the next section, we present the results of simulations performed, using the metrics as described in Section VI.

B. Simulation Results

The results that follow describe the impact on system performance mentioned above in relation to migration in the applied schedulers group AFCFS and LGFS. Furthermore, the impact of adaptive place order is examined.

- Average response time versus number of jobs executed

In Fig. 2, the ART is illustrated in three scheduler scenarios, AFCFS and LGFS without migration (S1), AFCFSm and LGFSm with migration (S2), and AFCFSm and LGFSm with migration and aLD (S3), respectively, where the *x-axis* represents the number of jobs executed. Note that we take into account the response time increase of jobs rescheduled.

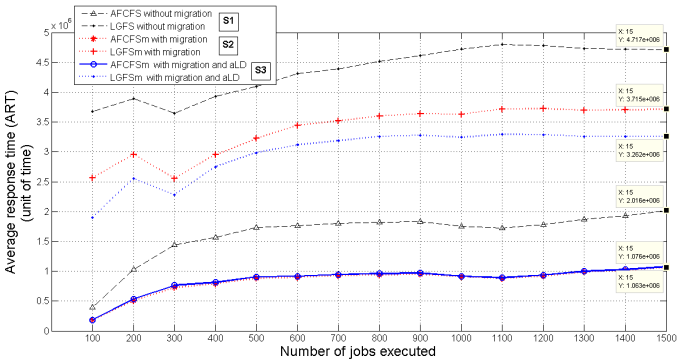


Figure 2. ART versus No. of executed Jobs - S1, S2 and S3

In the three scenarios, as illustrated in Fig. 2, it can be seen that the AFCFS had the lowest ART of all amounts of jobs performed with respect to LGFS. Furthermore, the ART showed an increase that conformed to the number of jobs carried out. This is justified for two reasons: firstly, an increase of executed jobs, *j* and *f*, and secondly, the processing time of the jobs are different. The scenario S2 shows a fairly significant reduction in the ART in relation to the scenario S1. This shows that the use of migration causes great impact on reducing the response time. Therefore, the suggested method could use the available processors more efficiently, thus reducing the fragmentation, and subsequently, the response time.

At the S3 stage (with the migration and aLD), Fig. 2, the ART was reduced in comparison to the S1 and S2 scenarios. This occurred for three reasons: firstly, the use of the migration strategy was implemented, which presented satisfactory impact on the results, as mentioned above; secondly, the migration with aLD, the JSQ algorithm distributes the tasks in the queue more fairly; and thirdly, the aLD acts before the schedulers begin to queue, thus minimizing the limitations of these at the time of allocation of tasks to resources. Furthermore, it can be seen that the scheduler LGFS in the scenario S3 ($ART = 1.076e{+}006$) showed a small reduction in the average response time with respect to S2 ($ART = 1.063e + 006$). The case scheduler AFCFS now visibly presented the best result in the three scenarios. The information concerning the total task, i.e., the number of jobs in the queue over existence or non existence of a running task on the processor, implies a reduction of task waiting time and response time. Furthermore, we observed a reduction in the number of migrations, causing direct impact on improvement of the system.

- Utilization of clusters

In this section, the performance analysis based on the use of resources using the three scenarios S1, S2 and S3 is presented. Fig. 3 illustrates the percentage utilization of the clusters in each scenario based on the number of interactions. In Fig. 3 (S1), on the intervals $400 - 2600$ (aFCFS) and $400 - 2700$ (LGFS), the average utilization of clusters is 50% and $42,5\%$, respectively. The LGFS (S1) showed an increase in the range from $1600 - 2400$. This happens because this algorithm takes care of larger jobs. Therefore, LGFS tend to offer greater fragmentation in the system.

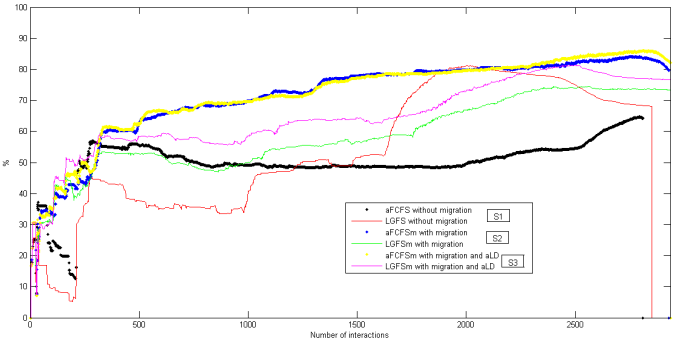


Figure 3. Utilization (%) versus number of interactions - S1, S2 and S3

In Fig. 3 AFCFSm and LGFSm (S2), the average utilization of the clusters is 60% and 50%, respectively. These results show an increase compared to the results of the scenario (S1), even with the arrival of high priority tasks in the environment. Furthermore, it can be seen that algorithms with the migration strategy could more efficiently use the resources. The results in Fig. 3 algorithms AFCFSm and LGFSm aLD (S3) show an increase of 10% over S2. That is, the average resource utilization is 75%. The scenario S3 distributes the tasks in the fairest way to the processors, causing direct impact on improving resource utilization.

- LoC versus Scenarios

Fig. 4 illustrates the loss of system capacity by fragmentation in three scenarios S1, S2 and S3. In the scenario S1, the scheduling policy AFCFS presents the lowest of LoC ($30,4\%$) compared to LGFS ($LoC = 31,1\%$). This result confirms that the LGFS tends to offer greater fragmentation in the system, since this algorithm favours regarding jobs with a larger number of tasks. The scenario S2 shows a considerable decrease in fragmentation compared to S1, confirming once again that the migration minimizes fragmentation in the system. Furthermore, it can be seen that the AFCFS presents a lower percentage relative to LGFS. This implies that the AFCFS with migration is able to schedule jobs more effectively and

 ISBN: 978-1-61208-330-8

subsequently reduces fragmentation. The S3 scenario shows the best results compared to the others. Therefore, the strategy implemented in LD offers an efficient scheduler.

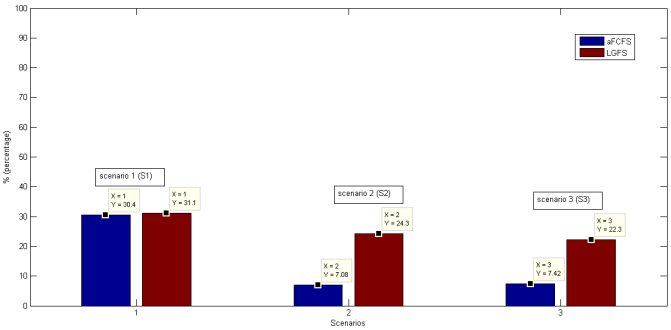


Figure 4. LoC (in %) - Fragmentation in three scenarios

VIII. Conclusion and Future Work

In this work, experiments were carried out using a heterogeneous environment based multi-cluster, a structure of two layers which were applied to different scenarios. For analysis of these experiments, we used performance metrics to evaluate the performance of schedulers in different situations.

The ART analysis results indicated a reduction of response times by implementing the migration mechanism, (Section V-A), in the schedulers (AFCFS and LGFS). This implied that the suggested method of migration could use the idle processors more efficiently and therefore reducing the fragmentation. Comparing results of the scenarios S2 and S3, we conclude that the strategy implemented in LD has a better response time. This shows that the adaptive in order site (aLD), the algorithm distributes JSQ queues of tasks more efficiently by minimizing the waiting time of the task, as well as response time. From the results of the simulations, one can observe the reduction of migration, causing a direct impact on efficiency. The algorithm AFCFS in ART metric shows the best results when compared with LGFS in the three scenarios. Regarding the utilization of metrics clusters, it was confirmed that the migration technique minimizes idle processors in the system, as well as fragmentation, with the most significant results obtained with the further scenario (S3). The latter was even more efficient, reducing the overhead on the system caused by excessive migration. The LoC metric measures the impact that the schedulers bring to the system in relation to fragmentation. The results obtained in Fig. 4, AFCFS without migration algorithm (30,4%), were achieved through less fragmentation with respect to LGFS $(31,1\%)$. With the suggested method of migration, fragmentation was considerably reduced in AFCFS $(7,08\%)$ and LGFS $(24,3\%)$, and with the implementation aLD, the results were more than satisfactory. This still showed that AFCFSm caused less fragmentation with the aLD system in relation to schedulers (LGFSm with migration and LGFSm with aLD). From the results, we aim to reduce the fragmentation through the controlled use of task migration between the rows of multi-cluster processors in a heterogeneous environment, as well as better use of them, implying a reduction in operating costs by the service providers in QoS expectations of the users.

As a future perspective, we must examine the proposal in other heuristic algorithms, comparing it with them schedulers used in the different approaches of this work. Furthermore, there is a proposal to implement the proposal in a real environment.

References

[1] J. M. U. de Alencar, R. Andrade, W. Viana and B. Schulze, P2P scheme: a P2P scheduling mechanism for workflows in grid computing, Concurrency and computation: Practice and experience, John Wiley Sons, Ltd, 2011.

[2] H. Luo, D. Mu, Z. Deng and X. Wang, A review of job scheduling for grid computing, Research of computer, vol. 22, 2005, pp. 16-19.

[3] H. Topcuoglu, S. Hariri and M. S. Wu, Performance-effective and low complexity task scheduling for heterogeneous computing, IEEE Transactions parallel distributed systems, vol. 13, 2002, pp. 260-274.

[4] T. L. Casavant and J. G. Kuhl, A taxonomy of scheduling in general-purpose distributed computing systems, Transactions on software engineering, vol. 14, Feb. 1988, pp. 141-154.

[5] T. Ma, Q. Yian, W. Lu, D. Guan and S. Lee, Grid Task Scheduling: Algorithm review, IETE Technical, vol. 28, Apr. 2012, pp. 158-167.

[6] J. Ousterhout, Scheduling techniques for concurrent systems, Proc. of the 3rd ed. Intl. Conference on distributed computing systems, 1982, pp. 22-30.

[7] Z. Papazachos and H. D. Karatza, Performance evaluation of bag of gangs scheduling in a heterogeneous distributed system, Journal of systems and software, vol. 83, Jan. 2010, pp. 1346-1354.

[8] X. Wang, Z. Zhu, Z. Du and S. Li, Multi-cluster load balancing based on process migration, Lecture notes in computer science, Springer, Berlin, vol. 4847, 2007, pp. 100-110.

[9] H. D. Karatza, Performance of gang scheduling strategies in a parallel system, Simulation modeling practice and theory, Elsevier, vol. 17, Feb. 2009, pp. 430-441.

[10] Z. Papazachos and H. D. Karatza, Gang scheduling in multi-core clusters implementing migrations, Future generation computer systems, vol. 27, Feb. 2011, pp. 1153-1165.

[11] I. Moschakis and H. D. Karatza, Evaluation of gang scheduling performance and cost in a cloud computing system, The journal of supercomputing, vol. 59, Feb. 2012, pp. 975-992.

[12] Z. Papazachos and H. D. Karatza, Gang Scheduling in a two-cluster system implementing migrations and periodic feedback, Transactions of the society of modeling and simulation international, vol. 87, Dec. 2011, pp. 1021-1031.

[13] S. K. Garg, S. Venugopal, J. Broberg and R. Buyya, Double auction-inspired meta-scheduling of parallel applications on global grids, Journal parallel distrib. Comput, vol. 73, Apr. 2013, pp. 450-464.

[14] Z. K. Gkoutioudi and H. D. Karatza, Multi-criteria job scheduling in grid using an accelerated genetic algorithm, Journal grid computing, vol. 10, Mar. 2012, pp. 311-323.

[15] V. J. Leung, G. Sabin and P. Sadayappan, Parallel job scheduling policies to improve fairness: A case study, ICPP Workshops, 2010, pp. 346-353.

[16] W. Tang, Z. Lan, N. Desai, D. Buettner and Y. Yu, Reducing fragmentation on torus-connected supercomputers, Parallel & distributed processing symposium, 2011, pp. 107-115.

[17] W. Tang, D. Ren, Z. Lan and N. Desai, Adaptive metric-aware job scheduling for production supercomputers, 41st International conference on parallel processing workshops, 2012, pp. 107-115.

[18] D. Feitelson (2014, Mar.) The Standard Workload Format. [Online]. Available:http://www.cs.huji.ac.il/labs/parallel/workload/swf.html

[19] Y. Zhang, H. Franke, J. Moreira and A. Sivasubramaniam, The impact of migration on parallel job scheduling for distributed systems, Proceedings of europar, 2000, pp. 242-251.

[20] A. Burkimsher, I. Bate and L. S. Indrusiak, Survey of scheduling metrics and an improved ordering policy for list schedulers operating on workloads with dependencies and a wide variation in execution times, Future Generation Computer Systems, vol. 29, Oct. 2012, pp. 2009-2025.

 ISBN: 978-1-61208-330-8

A Cloud-based Multimedia Function

J-Ch. Grégoire and M. Amziab
INRS-EMT, CANADA
{gregoire,amziab}@emt.inrs.ca

***Abstract*—Two dominant trends of the Internet are the increasing importance of multimedia traffic, not only in the form of streaming videos but also for interactive communications, and the use of cloud technology to deploy services. In this paper, we look at the intersection of these trends and expose a number of considerations to help with the deployment of multimedia functions for interactive, mobile-adaptive and time-constrained applications in the cloud. We show how virtual servers can be CPU or bandwidth constrained and how to use them effectively.**

Keywords–Multimedia processing, Edge Cloud, IMS, mobile media services.

I. INTRODUCTION

Deployments of multimedia functions in a cloud are of interest for a number of reasons. First, it is a way to optimize services offered by operators, through economies of scale. Second, for many applications, it is a way to avoid end-to-end connectivity issues posed by middleboxes (e.g., network address translation boxes). More generally, it is a way to leverage third party service offerings, through outsourcing.

On the other hand, there are multiple ways to implement and exploit cloud technology, designated through different variants of platform, software or infrastructure as a service (PaaS, SaaS, IaaS, resp.), and we can wonder what is the best way to do such deployments to take full advantage of the scalability and flexibility offered by the cloud.

In this paper, we look at media from the perspective of a service infrastructure such as the IP Multimedia Subsystem (IMS), that is, a mobile-supporting media environment where control and processing are split, and control will be related to some signalling infrastructure. In the IMS world, one talks of a media controller and a media processor [1], [2], but we must point out that, although we adopt this decomposition and terminology, this work is in no way specifically tied to IMS. Doing such a split is interesting for a number of reasons but scalability comes naturally to mind, as the demands of media processing, especially video, will dwarf those of control processing.

The main challenge for the cloud deployment problem of a media function then becomes the dual issue of server placement, so as to avoid problems related with latency and general control of quality of service (QoS), especially in the presence of mobility, and of spreading the processing load across a number of processors, which is essentially a scheduling problem. This must be repeated for multiple instances of controllers, possibly for different customers, each requiring the services of multiple processors. However, before we can address the design of such a scheduler, we need to study the requirements of the processors themselves. Such a problem has been widely investigated in the community, but not in such a case as propose here.

In this paper, we study the performance constraints of a number of video codecs used for interactive communications (e.g., video conference), a particularly demanding application, and present a control architecture to support their efficient operations in an Edge cloud environment. Our focus on video processing is meant to expose the needs of the most demanding services, but our long term goal is to support a general set of media types, including voice and audio.

This paper is structured as follows: Section II sets the background on this work. Section III describes the experimental context and Section IV presents the results of our evaluation of the performance constraints. In Section V, we present a sketch for a scheduling function for media operations within the cloud. Section VI has a discussion of our work and we conclude in Section VII.

II. BACKGROUND

Moving a service to a cloud presents a number of benefits, including lower infrastructure costs and scalability of offer through on-demand activation of servers. Indeed, adapting to demand has been an important sales point for cloud-based services. Offers have typically been confined to computation and storage, and media restricted to streaming.

Much has been written on the various guises of cloud infrastructures and service offerings [3]. More recently, there has been a specific interest in the use of clouds for multimedia services [4], [5], typically Video on Demand (VoD), a growing commercial segment with commercial offerings such as Netflix, which incidentally largely relies on a combination of Amazon's Elastic Compute Cloud (EC2) service and Content Delivery Networks (CDN) providers such as Akamai for content delivery. Companies with large cloud infrastructures, such as Google, Apple, and Amazon itself, offer competing services. Gaming has been another topic of interest.

Media streaming, such as VoD, has to be responsive but is non-interactive and supports a large amount of buffering. The constraints on the server side are of a storage and bandwidth nature: deliver content from storage—or memory caches if the content is in high demand—through a network interface. Furthermore, CDNs can be used in conjunction with cloud storage to scale delivery to a large number of customers.

Media services expand beyond VoD, however, and many are interactive, which implies reaction times in the couple of hundreds of milliseconds in the worst case, and very little margin for buffering. Streaming of real time programmes (i.e., live TV) is another example. There have been several studies of the use of cloud to support mobile services, also in the context

 ISBN: 978-1-61208-330-8

of IMS [6], [7], which presents a clear distinction between control and processing supporting distribution, including our own work on the Edge Cloud [8].

IMS also illustrates the need for different varieties of media processing. Beyond the streaming services already described, we find services strongly related to telephony, such as DMTF (tone) decoding, voice mail or also interactive voice response (IVR), but also more generic services such as transcoding or conference bridges. Beyond IMS, IP/TV is another example of media processing, especially in contexts, such as mobility, where a uniform multicast model can be difficult to deploy at the network level and needs to be provided as an adaptive application. There have been so far little effort reported to study the effects of the deployment of interactive media services in the cloud. We can note that some commercial offerings, in the form of virtual media servers, are required to be run alone on a hardware platform and have strong limitations (e.g., Microsoft Media Platform).

Unlike streaming services, which can be accessed and controlled from a web page, interactive services tend to be related to a signalling protocol, typically SIP. Also, for scalability reasons, the media function is separated into control and processing. From this perspective, we argue that it makes sense to study how the media processing function can be deployed in the cloud, to take advantage of the latter's flexibility and scalability. In the following sections, we study first the cost of hosting a media function in the cloud and second, how can it be properly orchestrated.

III. Media Function Performance Characterization

We look here at the characterization of the performance cost of running a media processing function on a generic processor. We have created a simple testbed to isolate the contribution of video flows on computer resources along three parameters: CPU, memory and bandwidth consumption; we have also measured latency. The purpose of the characterization is to identify the key parameters to be used by a scheduling function, which we will explore in the next section.

The goal, quite straightforward, is to study how it is possible to multiplex different functions onto single processors. To illustrate our purpose we consider only one example–transcoding–a function that is however quite certainly CPU demanding and subject to latency.

Video processing: Our experimental environment is based on the use of the GStreamer framework. Such a framework, extensible through plugins and composition is an ideal vehicle for custom tests. It is also quite efficient, in spite of its flexibility, as has been demonstrated in performance evaluations [9].

A GStreamer application is a pipeline of different modules which contribute one specific element of the audio/video transmission and processing chain, including coding/decoding, mixing, filtering, scaling, effects, etc.

To illustrate how processing pipelines can be specified in succinct term, we present in Fig. 1 an example of a simple GStreamer pipeline, and the matching code is presented below. The pipeline starts with a live video capture from a camera (Microsoft LifeCam Studio, 1080p), although a network or

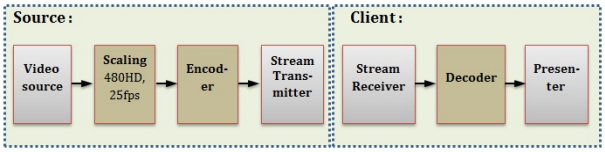


Figure 1: Simple GStreamer Pipeline

disk stream is also a possibility. This stream is scaled to a smaller video size, coded, transmitted, received, decoded and presented on a screen.

Latency measures: We use GStreamer extensively to generate our media streams, with different codecs. But also quite important for our study is the possibility of using it to insert data in a media stream, which can be identified at the receiver and used for latency measurements. For this purpose, we use the Zbar module, a generic part of the GStreamer framework, which allows the detection of the presence of a barcode in a picture. The Zbar module reads frames from a video stream, detects barcodes and sends them as element messages to the GStreamer bus, from where we can retrieve the detected barcode data and the timestamp of the frame that triggered the message.

The use of the Zbar module is key to measuring latency. A barcode picture is integrated in the stream and marked with a timestamp as part of normal processing for transmission with RTP transport. The module detects that barcode at the receiving end and retrieves the corresponding timestamp. The latency is the difference between the timestamps retrieved from the stream and the current time at the receiver. We note also that this technique is robust in the presence of video transcoding.

For such a measure to be useful, the time on both machines must be synchronized. In our studies, all computers run the NTP protocol with a server polling interval of 10 seconds. The latency reported is the average of 10 samples.

Codecs: We have used three codecs suitable for use for video conferencing over the Internet. They were meant to be representative of the most popular techniques, but not necessarily to present an exhaustive choice of possibilities. Most specifically, we have used motion JPEG (MJPEG), MPEG2, et MPEG4-AVC (H.264), all available within the GStreamer framework. The key rate was fixed at a standard 25fps, and the video format was 480p, which, with the arrival of a new generation of mobile terminals, is slowly becoming the standard low end of video resolution. In the case of H.264, we have configured the implementation we used (x264) for an interactive application and not for its default streaming operation, which uses a large amount of buffering to achieve high video compression rates.

IV. Experimental Results

All tests were performed on a computer with an AMD Phenom(tm) II processor at 2.7 GHz, with 16 GiB of memory, running a XEN-enabled bare-bone Debian linux distribution. All media processing instances were running single-threaded, to avoid conflicts between different levels of scheduling, in separate processes but without virtual machines. The communication link speed was limited to 100 Mbps, a realistic

 ISBN: 978-1-61208-330-8

perspective if we consider that this machine would be part of a cloud and have many siblings performing the same operations.

We begin with the performance limits imposed by the computing platform itself, and then consider the impact on latency.

Physical setup: The end-to-end view of our basic system is presented in Fig. 2. More specifically, it shows the sender side, relay and receivers side. On the sender side, a GStreamer pipeline encodes the video stream and transmits it over UDP. The streams received at the relay are transcoded and retransmitted to the receivers. Each receiver decodes and renders the stream to the screen.

The number of receivers is limited by the number of instances executed in the relay. Once the sender transmits the stream, on the relay side, we continue adding instances, all the while measuring the resource consumed, until we reach the point where the relay has exhausted its capacity to do further useful work. After the generation of each instance we wait 10 seconds before making a measurement to avoid transient effects. Indeed, during the evaluation of the CPU metric we found that, after the generation of an instance, the results were not correlated until the processor had stabilized, which required 5 seconds on the average. Also, to reduce interference, all background tasks were disabled during measurements.

On the relay side, the CPU and memory utilization were measured based on the standard process statistics report that contain information related to overall system status. We have used the libpcap library to capture the network traffic into a file, which was later analysed using the Wireshark graphic analysis tool to extract the bandwidth information. This analysis was performed offline to avoid interfering with the experiments. The latency between the sender and the receiver side was measured by the barcodes frames as explained before.

Platform limits: Fig. 3 presents the results of our performance tests for two types of videos: a talking head-type video stream, typical of video-conference applications, and an action video stream with many changes of background. For all figures, the x-axis presents the number of instances of a video operation and the y-axis the percentage of a CPU or the amount of bandwidth used for each type of video (memory is not presented because of space considerations). We are considering only homogeneous instances: only one type of video for all instances. The results are presented for three codecs denoted as jpeg (MJEPG), Mpeg2 (MPEG2, which gives equivalent results as H.263 and MPEG4-part 2) and x264 (H.264).

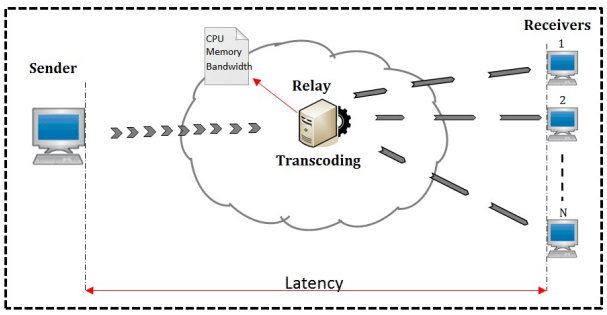


Figure 2: Experimental setup

We see that, as could be predicted, H.264 coding is the most demanding in terms of CPU, with the best compression results for dynamic video content. At the other extreme, MJPEG is low in CPU demand, while requiring higher bandwidth. Mpeg2 trails x264 closely and presents little noticeable advantage over it. In the case of x264, the limiting factor will be the CPU consumed—between 25 and 30 instances—while for jpeg, it is the bandwidth—about 18 for a 100 Mbps link.

These results clearly establish the soundness of using a standard computational platform to perform video processing, as the number of simultaneous instances that can be supported is rather large and lends itself to a mix of operations.

Latency: We next consider the impact of running multiple instances on latency. As explained above, these measures were done by the insertion of barcodes with a timestamp in the video stream and their extraction at the reception. A null operation was performed to eliminate the delay due to this technique and we consider only the increase in latency in our results.

Fig. 4 again shows the results for two different kinds of content, static and dynamic. At first glance we see that we do not have the linear behaviour that we had observed with the performance indicators. At about 10 instances, in both cases and for two codecs, we see that we lose the linear behaviour, and latency increases dramatically for the MJPEG codec. A correlation with the use of bandwidth points to a likely answer for this behaviour, which would denote a greater level of contention at the network interface. The x264 module, on the other hand, behaves rather linearly, with similar results for both types of video. Furthermore, the results are quite acceptable for interactive communications, remaining inferior to 100 ms.

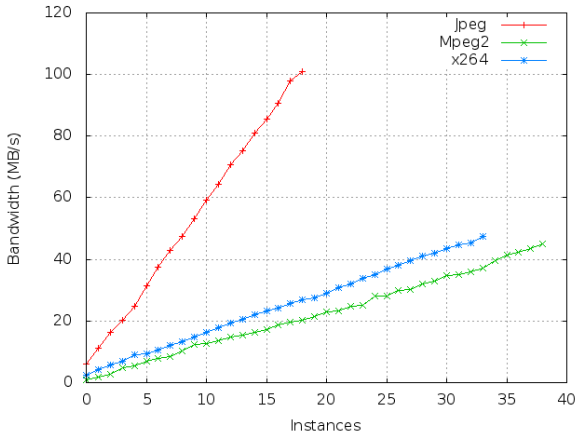
Other factors: We have also analysed jitter and video quality degradation. Jitter does increase with the number of instances but we could not measure it with sufficient precision to have statistically significant results. Similarly, no statistically significant degradation has been observed in the video content, on the basis of a PSNR-based comparison of original vs. received content.

V. Edge Cloud Deployment

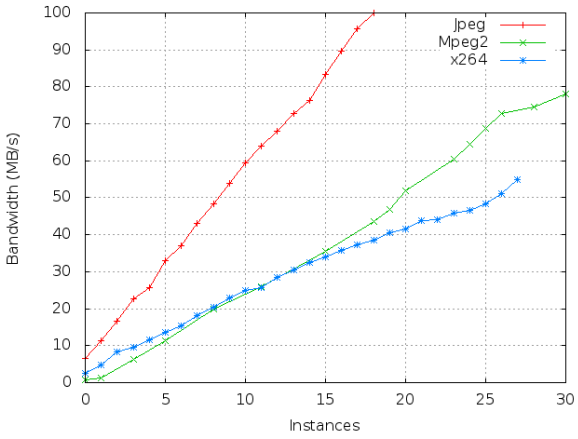
Through the assessment of the performance of a media function, our experiments have established that it is possible to run several instances of demanding media functions on a single computer. We now analyse how such a deployment could be orchestrated under the supervision of cloud management, and on demand by a control function.

Edge Cloud: First, we consider that processing is deployed on Edge Clouds, that is, broadly speaking, a Cloud within an access provider's domain, close to users. There are several benefits to this model as it provides more flexibility to integrate user feedback based on the nature of the access link, be it for cellular phone services or over-the-top services. It also supports scalability through the availability of such infrastructures across many sites. While not acknowledged as such, the Edge Cloud is a reality as most ISPs have taken to deploy cloud infrastructures of their own, to take advantage of this booming market. Furthermore, it can be closely related to the way CDN is deployed, although CDN is not meant to deliver computing power and is traditionally quite limited in that respect, and typically restricted to the support of dynamic web content.

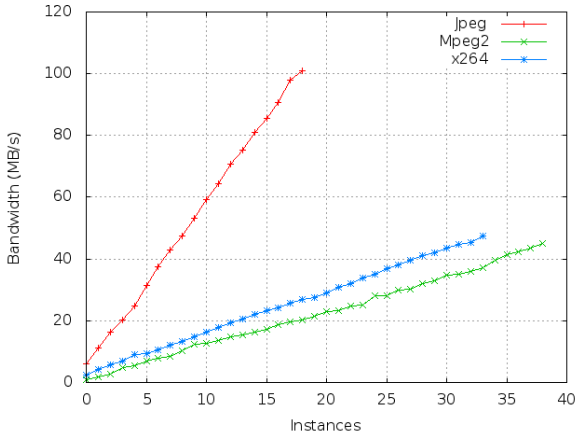
 ISBN: 978-1-61208-330-8



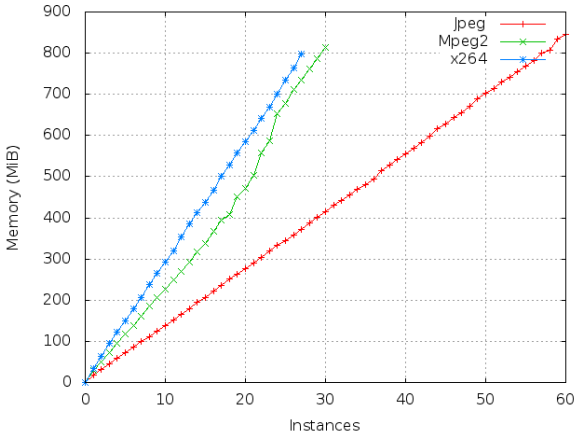
(a) Talking Head Bandwidth



(b) Dynamic Content Bandwidth

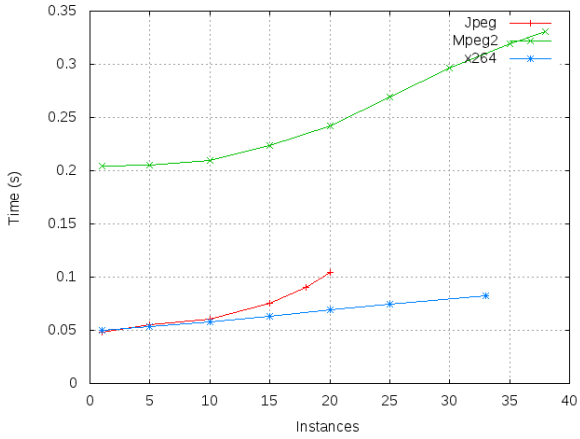


(c) Talking Head Memory

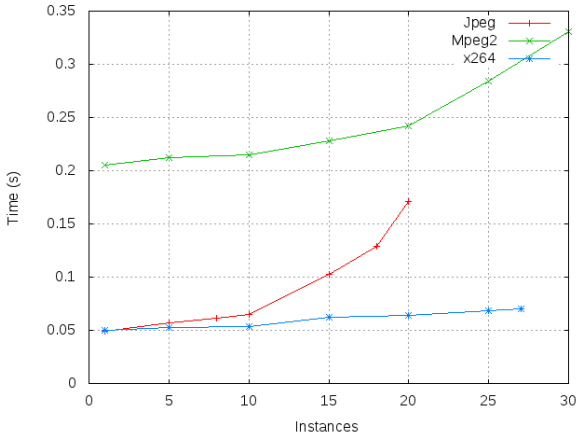


(d) Dynamic Content Memory

Figure 3: Performance of static and dynamic video content



(a) Talking Head



(b) Dynamic Content

Figure 4: Latency

Media Processing: As we have already said, we postulate that media control can be done remotely while processing will be closer to the access network–what we call the network edge. We have already discussed that we consider media

functions with tight time constraints. This includes media relay (to bypass firewalls or middleboxes), transcoding, live broadcasting, IVR interactions, etc.

In the line of the experiment described above, we also characterize each form of processing in terms of the resources it requires (CPU, Memory, Bandwidth, Storage).

Computing structure: Remember that virtualization is the tool of choice for deployment in the cloud, from small, language-specific virtual machines (e.g., Python, Java) to a full virtualized computer. This is however not necessary in our case: we only need minimal support from a generic OS to be able to run multiple instances of GStreamer pipelines, encapsulated in their own process, and virtualization is irrelevant. Of course the observation we make about these pipelines generalizes to other implementations of media processes.

We propose then that media processing be organized along the lines of a grid; that is, a pool of machines dedicated to media processing, running on a software-tuned platform. To harmonize with normal cloud operations, this software platform could run on top of the hypervisor used for PaaS operations. This also implies that the size of the pool could be elastic, with more machines assigned to its operations as required. Each software platform has a control module to accept new instances of media processes.

Monitoring is organized in similar fashion, keeping in mind that the ratio of monitoring to media functions can be quite small. Monitoring is in turn connected with media control, which can run on a different cloud and be more centralized.

Scheduling: Scheduling, in this environment, takes several dimensions:

- *static* adjustment of the vocation of machines in the pool based on the number of control functions activated;
- *dynamic* dispatching of activation of a media processor;
- *meta* management of elastic behaviour (size of the pool) based on runtime demands.

We concern ourselves only briefly with meta management here, and do not discuss the static dimension, as they largely depend on contractual terms balanced with a history of the behaviour and consequent demand prediction model [10], [11]. Still, the adjustment of the size of the pool requires some degree of concern to make sure that the resources we need are available when we need them, but are also not kept around beyond the time they are required. Dynamic scheduling, on the other hand, is directly relevant to our study as we must make sure that machines running media functions are well used. In that spirit, observe that, unlike streaming content, we do not know a priori what the duration of the service will be, especially for communications where transcoding/conferencing is involved. Under such conditions, it is difficult to hope to find an optimal scheduler.

Remember that we characterize the different media processes in terms of their CPU, memory and bandwidth needs. Since we have seen CPU load is the dominant factor for video processing, we propose, as a first approximation, a straightforward scheduler where machines are sorted from least loaded to most heavily loaded and a suitable machine is chosen based on that order, in a greedy fashion, also matching the latency constraint of the application. For meta-management, When the least loaded machine's load increases beyond a high water mark, the active pool size is increased; similarly, when the most heavily loaded machine's load falls below a low water mark, the active pool size is decreased. The high/low water marks would be adjusted with the rate of subscription and departure of media functions, to give enough reaction time to allocate resources. Such considerations, however would depend on the nature of the service(s) offered in practice.

Orchestration: The connection between monitoring and processing is easily achieved through a management function. The control requests a processing resource for a specific operation; the management function will schedule the activity on a suitable machine, and return to the control the characteristics required to integrate it in an end-to-end media flow, e.g., IP address and port numbers.

The operation would have to be pre-registered with the management function, in the form of an execution script to establish its performance characteristic, to be used by the scheduler, with a test and calibration protocol. This model provides strict resource confinement and acts as a contract for the execution of a specific function, valid over all its instances.

The control also notifies the management of the end of the execution of a function, so that the resource can be terminated and recycled.

A variant: To illustrate the flexibility of our architecture and its scheduling model, we present here another context, which is suitable for videoconference, where quality can be degraded within reason if resources are saturated. We have imagined four quality scenarios, characterized by the profiles presented in Table I.

TABLE I: DIFFERENT SERVICE PROFILES

Profile	Best	Default	Fast	Ultrafast
CPU(%)	13	7.2	5.4	3.6
PSNR	40.52	38.22	37.00	35.28
MOS	Excellent	Excellent	Good	Good

The idea, in this case, it to work with a fixed pool of resources, but to degrade the quality of the communication, within the confines of quality constraints characterized by a satisfactory MOS. As the load increases beyond a limit set a percentage of CPU load, the quality of the flows will be lowered, and similarly increased, with suitable hysteresis to avoid oscillations, when the load decreases. Fig. 5 shows the behaviour of the load as the number of instances increases beyond the best quality and flow quality is slowly downgraded; the algorithm itself is based on thresholds and straightforward. The threshold maximum load is set at 75 % to allow temporary overruns. Note that GStreamer allows a transition between quality profiles without interruption of the video transmission.

VI. Discussion and Related Work

Complementarity between grid and cloud has been discussed by Foster in [12], and [13]. Taking a subset of a cloud

 ISBN: 978-1-61208-330-8

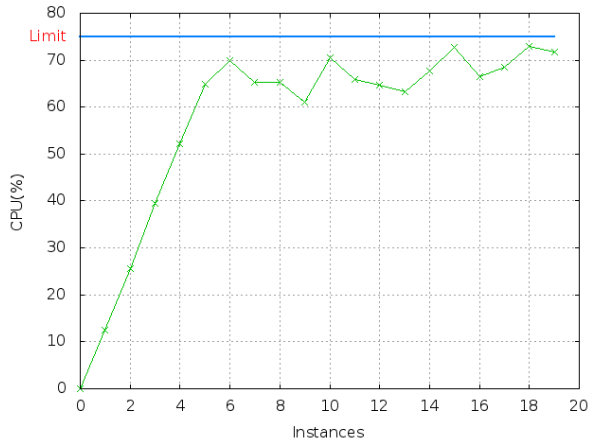


Figure 5: CPU Load

and using it as a computational grid, as we describe here, is not a new concept. It is also consistent with the use of a cloud as a streaming farm, which has itself been the focus of some research [5]. Our work differs in its concentration on interactive media processing.

It also complements the large body of work on the management of QoS in clouds and the establishment of SLAs [14], [15], as it is meant to provide a predictable model of performance requirements for management. Our approach is simpler as we show that we do not have to worry about the effects of virtualisation, which is unnecessary for our purposes. We also differ from streaming applications since we do not have to worry about load balancing or disk access, which can lead to other scheduling issues.

Most important, media distribution is closely related to the research done on gaming clouds [16]. The similarities in time constraints and computing load between both types of problem are clear although some elements are clearly different. Taking into consideration single user or group video games only, they will be implemented in a single server, with no need for transcoding. Furthermore, specialized hardware, typically GPUs, will be used to achieve better performance results. Finally, some latency in game set-up is acceptable, which leads to more flexibility in scheduling. Nevertheless, it is clear that the same infrastructure can benefit both types of application and such convergence will be the focus of future work.

VII. Conclusion

We have presented a performance analysis for a media function and shown how these results can be used for their scheduling in a grid environment. We have chosen this demanding function to assess the practical limits and the results show that it is quite reasonable to not only mix such functions on the same processor with a simple containment, but to also mix them with other functions, also interactive, but possibly less demanding.

This work overall establishes the suitability of the edge cloud, as opposed to dedicated hardware or boxes, as a host and support for media processing. The latency of the operations is quite acceptable for such applications as IMS provided the cloud be deployed closer to the edge, e.g., for mobile service providers.

In future work we plan to develop and refine our management function and design a scheduling toolbox to support a variety of operations along the line of those we have presented. More specifically, we plan to integrate learning mechanisms to assess the nature of the processing.

Acknowledgment Our technique to measure latency was inspired by [17], but adapted to the tools offered by GStreamer.

References

[1] Multimedia Resource Function Controller (MRFC) - Mp interface: Procedures Descriptions, 3rd Generation Partnership Project TS 23.333, Rev. V11.0.0, Jun. 2012.

[2] Media server control using the IP Multimedia (IM) Core Network (CN) subsystem, 3rd Generation Partnership Project TS 24.880, Rev. V8.2.0, Jun. 2008.

[3] M. Armbrust et al. “A view of cloud computing,” Commun. ACM, vol. 53, no. 4, Apr. 2010, pp. 50–58.

[4] S. Dey, “Cloud mobile media: Opportunities, challenges, and directions,” in Computing, Networking and Communications (ICNC), 2012 Int’l Conference on, Feb. 2012, pp. 929–933.

[5] Y. Wu, C. Wu, B. Li, X. Qiu, and F. Lau, “Cloudmedia: When cloud on demand meets video on demand,” in Distributed Computing Systems (ICDCS), 31st Int’l Conference on, June 2011, pp. 268–277.

[6] J.-L. Chen, S.-L. Wuy, Y. Larosa, P.-J. Yang, and Y.-F. Li, “IMS cloud computing architecture for high-quality multimedia applications,” in Wireless Communications and Mobile Computing Conference (IWCMC), July 2011, pp. 1463–1468.

[7] P. Bellavista, G. Carella, L. Foschini, T. Magedanz, F. Schreiner, and K. Campowsky, “QoS-aware elastic cloud brokering for IMS infrastructures,” in Computers and Communications (ISCC), IEEE Symposium on, July 2012, pp. 157–160.

[8] S. Islam and J.-C. Grégoire, “Giving users an edge: A flexible cloud model and its application for multimedia,” Future Generation Computer Systems, vol. 28, no. 6, 2012, pp. 823–832.

[9] V. Sentongo, K. Ferguson, and M. Dlodlo, “Real-time performance evaluation of media pipeline plug-in architectures,” in Southern Africa Telecommunication Networks and Applications Conference (SATNAC 2011), East London, South Africa, Sep. 2011.

[10] M. Peixoto et al., “A metascheduler architecture to provide QoS on the cloud computing,” in Telecommunications (ICT), IEEE 17th Int’l Conference on, Apr. 2010, pp. 650–657.

[11] I. N. Goiri, F. Julià, J. Fitó, M. Macías, and J. Guitart, “Resource-level QoS metric for cpu-based guarantees in cloud providers,” in Economics of Grids, Clouds, Systems, and Services, Lecture Notes in Computer Science Series, J. Altmann and O. Rana, Eds. Springer Berlin / Heidelberg, vol. 6296, 2010, pp. 34–47, 10.1007/978-3-642-15681-6_3.

[12] I. Foster, Y. Zhao, I. Raicu, and S. Lu, “Cloud computing and grid computing 360-degree compared,” in Grid Computing Environments Workshop, GCE ’08, Nov. 2008, pp. 1–10.

[13] I. Foster, “There’s grid in them thar clouds,” http://ianfoster.typepad.com/blog/2008/01/theres-grid-in.html, 2008, last access on February 14th, 2014.

[14] S. Ferretti, V. Ghini, F. Panzieri, M. Pellegrini, and E. Turrini, “QoS-aware clouds,” in Cloud Computing (CLOUD), IEEE 3rd Int’l Conference on, July 2010, pp. 321–328.

[15] J. Pedersen et al., “Assessing measurements of QoS for global cloud computing services,” in Dependable, Autonomic and Secure Computing (DASC), IEEE Ninth Int’l Conference on, Dec. 2011, pp. 682 –689.

[16] R. Shea, J. Liu, E. C.-H. Ngai, and Y. Cui, “Cloud gaming: Architecture and performance,” IEEE Network, July/August 2013, pp. 16–21.

[17] O. Boyaci, A. Forte, S. Baset, and H. Schulzrinne, “vDelay: A tool to measure capture-to-display latency and frame rate,” in Multimedia, 2009. ISM ’09. 11th IEEE Int’l Symposium on, Dec. 2009, pp. 194–200.

 ISBN: 978-1-61208-330-8

Secure Heterogeneous Cloud Platform for Scientific Computing

To ensure the dynamism of modern business requires scalable and reconfigurable systems, however, the transformation of isolated and static corporate IT- resources is problematic.

Vladimir Zaborovsky, Alexey Lukashin
Department of Telematics
Saint-Petersburg State Polytechnical University
Saint-Petersburg, Russia
vlad@neva.ru lukash@neva.ru

***Abstract*— New technical systems and facilities are now using more accurate models that require high performance and large amounts of data to be processed. All these add new constraints on the effectiveness of configuration, scalability, and reliability of services. Data protection computation is used at various stages of the life cycle. In response to such demands, it is necessary to develop applications that can achieve high performance in heterogeneous computing infrastructure. Its components can function as in the modes of virtualization, and in the form of clusters, optimized for parallel calculations. Supercomputer Center «Polytechnic» is being created within the national research university, especially designed for high performance, scalability, heterogeneity and security resources for industrial applications and research. This paper proposes the way to use cloud services for hybrid high performance computing resources management and describes the implementation of hybrid cloud using heterogeneous computing resources, OpenStack platform, and stealth firewalls.**

Keywords-Cloud computing; security; heterogeneous platforms; firewalls; OpenStack

I. INTRODUCTION

Cloud providers, such as Amazon, Rackspace, Heroku, and Google may provide different services on the models of Infrastructure as a Service (IaaS), Platform as a Service (PaaS) or Software as a Service (SaaS), whose integration into a specific environment of industrial development is carried out by highly qualified engineers and IT- specialists. So far, actual challenge is to develop cloud services for scientific computing and computer aided engineering. These services have to provide human resources as well as computing environment. Existing engineering centers are being built today on a specially designed software and hardware platforms, which limits their performance and flexibility, or on the IaaS model that also does not allow you to efficiently solve a variety of engineering problems.

The center for Supercomputing Applications Platform "Polytechnic" was designed to solve a wide range of engineering tasks. It uses four types of systems combined to take into account the characteristics of different types of applications: system with globally addressable memory; hybrid cluster based on CPU and gpGPU; reconfigurable flow-computers; and cloud that span the computing systems in a single environment for shared storage of data and services to control access to resources. The use of such a heterogeneous computing environment has the following advantages:

- computing environment allows expanding the range of information services that allows to quickly and cost-effectively implement multidisciplinary projects;
- virtualization and heterogeneity provide scaling resources to ensure high performance of computation at all stages of the implementation of engineering projects;
- cloud architecture implements automatic configuration of hardware and software components, versioning of applications and monitoring the integrity of the computing environment;
- network services provide benefits of network centric approach in the implementation of complex engineering projects by geographically and logically distributed development teams and specialists;
- stealth security system implements a common policy in the field of information security.

There are different proposals and implementations for organizing scientific computing services using cloud services [1] [2]. In this paper we propose using IaaS OpenStack platform and security services for organizing heterogeneous engineering center.

Heterogeneous cloud platform is the basis of computing infrastructure for engineering centers; the principal difference from the classic data centers is to provide remote access not only for computing resources or applications, but also intelligent services that are implemented by teams of specialists in different areas working in outsourcing within the chosen corporate information security policy. Using the resources of modern cloud-based engineering centers it is possible to create equivalent social networks that bring together professionals and experts to perform multi-disciplinary engineering project including computations, verification of test results based on the use of different materials, virtual prototyping and data visualization of computation. The above-listed problems from the point of view at the computational algorithms can be combined into technological chains, which form a network of operations. Their implementation is provided within a heterogeneous

 ISBN: 978-1-61208-330-8

cloud. The components of the platform (Figure 1), based on the OpenStack, include: IaaS cloud class segment, computing infrastructure within the cluster, the specialized high-performance hybrid system based on reconfigurable computing nodes.

resources configuration in real time. These approaches do not accurately identify the change in level of risk and take steps to block dynamically emerging threats.

To solve the problem of controlling access in the cloud, it is required to continuously monitor resources that cannot be

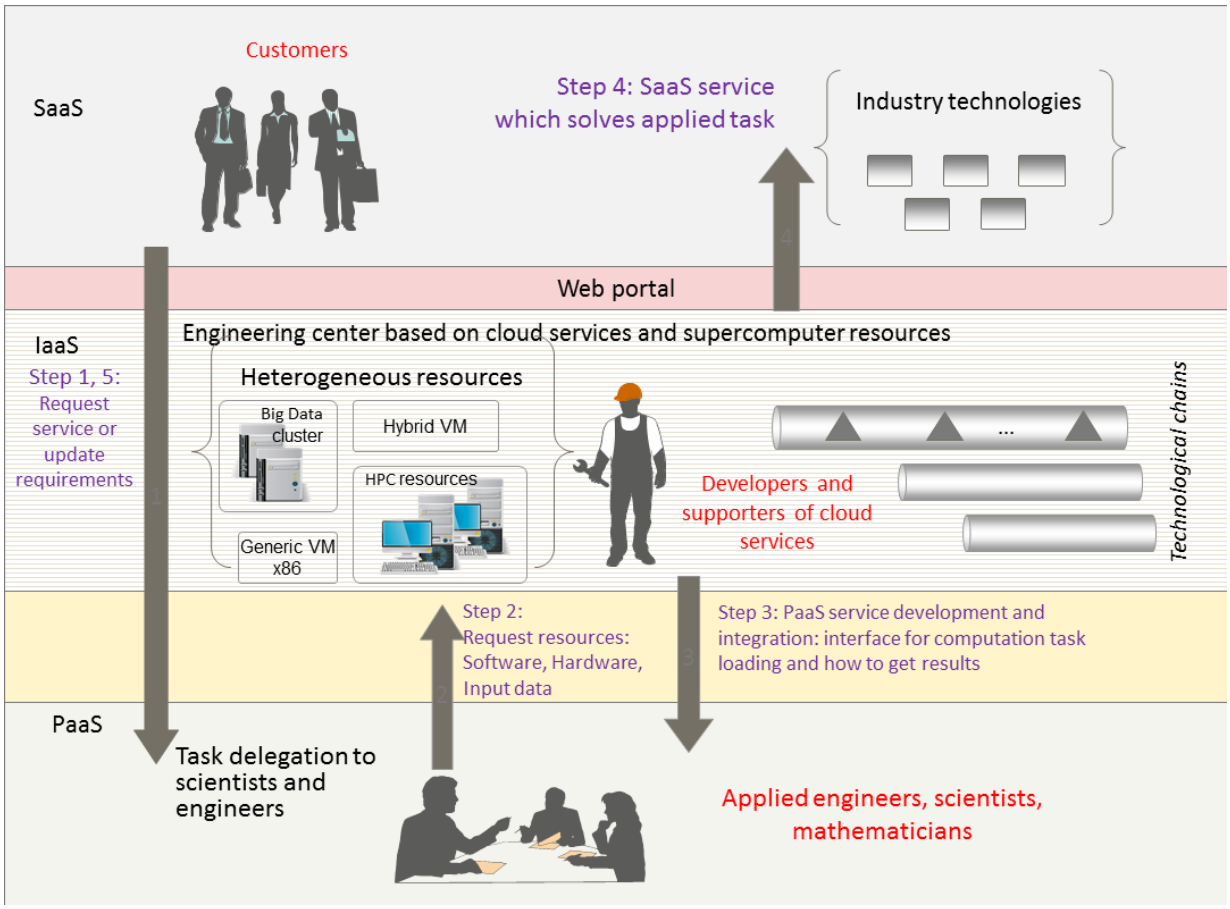


Figure 1. Functional scheme of a cloud platform with heterogeneous computing resources.

This paper is organized as follows. Section II covers security aspects of cloud platform. Section III covers methodology and technology overview of creating computation segments in cloud environment. The paper concludes with Section IV and presents future work on Section V.

II. ENSURING SECURITY

Virtualization has changed the approach of deploying, managing, and using enterprise resources by providing new opportunities for consolidation and scalability of computational resources available to applications; however, this led to the emergence of new threats posed by the complexity and dynamic nature of the process of providing resources. These threats can lead to the formation of a cascade process of security violations, which are powerless to traditional data protection systems. The existing approaches like “Scan and Patch” do not work in a cloud environment — network scanners cannot track changes of provided without the automatically generating rules for filtering and firewall log files analysis. Information security management products in a dynamic cloud environment should include mechanisms that provide: total control over processes for deploying virtual machines; proactive scanning virtual machines for the presence of vulnerabilities and configuration errors; tracking the migration of virtual machines and system configuration to control access to resources. Therefore, within the center of the "Polytechnic" a series of measures are set out to improve information security resources, namely:

- Enhanced Control of virtual machines. Virtual machines as active components of the service are activated in the cloud application random moments, and Administrator cannot enter and exit virtual machine out of operation, until the security scanner checks the configuration and evaluate security risks.
- Automatic detection and scanning. Information security services are based on discovery of

 ISBN: 978-1-61208-330-8

vulnerabilities in the computing environment. This discovery is based on the current virtual machine configurations and on reports of potential threats that come from trusted sources, such as antivirus update servers.

- Migration of virtual machines. Proactive application migration is an effective method to control security.

users and services running at any given time. On the platforms of this type, there are rare situations when user needs a single virtual machine. Therefore, cloud services support the dynamic creation of secure networks with a set of preconfigured virtual machines. Secured networks are connected to the firewalls which are integrated with the distributed SDN switch Open vSwitch and OpenStack

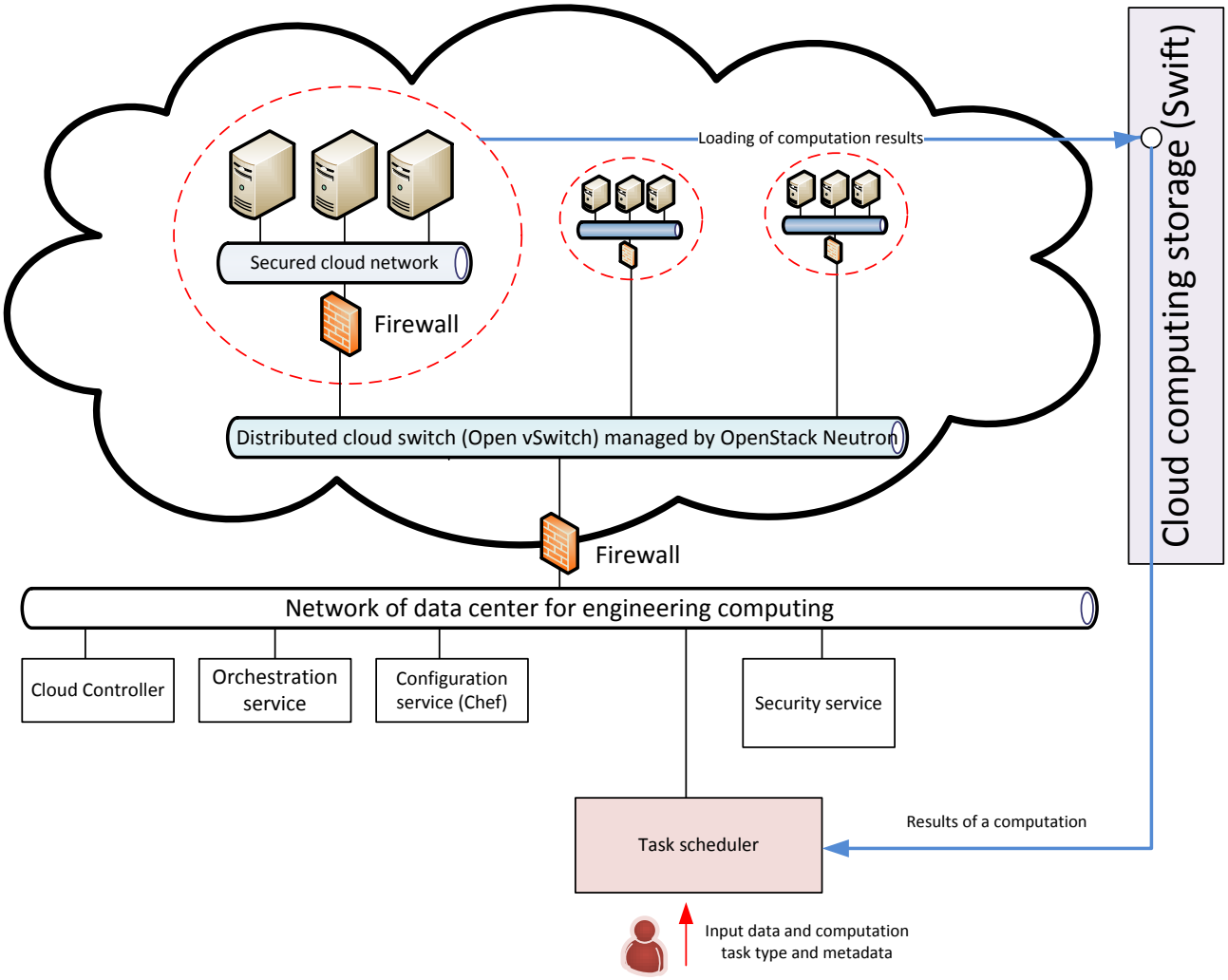


Figure 2. Components of cloud platform and general workflow.

In addition, computing center has the access control service for the cloud services protection. Its main features are support of dynamic infrastructure, scalability, and the ability to support security policies without reference to the composition of resources. The service is built on the technology of stealth traffic filtering and software defined networks (SDN) and provides the reconfiguration of access isolation system in accordance with the current state of the environment (in detail, the question of access control in cloud environments considered in [3] and [4]). Static platform segments applied the principle of "rent" under which the filter rules to access segments are formed only by Neutron networking service.

Firewalls which protect the dynamically generated cloud networks are created during computing segment initialization. An important feature of the firewall is its ability to function in the address less mode [5]. It allows implementing invisible protection of a cloud, and security system integration will not require the reconfiguration of a cloud network subsystem. The firewall acts as a virtual machine. The firewall of a network segment filters network traffic based on the rules that created the access policy service. Access policy in a cloud computing environment is based on the Role Based Access Control (RBAC) model of

 ISBN: 978-1-61208-330-8

access control. This policy can be represented as a set of following attributes:

- user IDs, that are involved in the management of virtual machines and information services;
- privileges that are described in the form of permitted information services (privileges set rules for user access to services, it is possible to change the privileges for the user in the specified virtual machine filtering rules for your firewall, which allow access to a network service);
- set of roles that can be assigned to users;
- user sessions in a computing environments based on the network connections between subjects and objects.

Access policy is translated to firewall filtering rules according to computing environment state. This state can be represented by a set of IP addresses of computing resources, with assigned user labels. Label represents user which is responsible for computing resource. When a state of a computing environment changes, then it is necessary to generate a new set of filtering rules and reconfigure firewalls. For that, a method of the dynamic configuration rules is developed, which consist of substitution of the network address lookup in user-owners privileges for each virtual machine. This approach formed the rules of access to the services of the computational resource and of computing resource to services of other users.

It is required at least one virtual firewall in each virtualization server and one general bare-metal firewall for protecting cloud services from external threats. In a cloud-based system there is a dedicated management network separated from virtual machines, so this network is used for the information exchange between the components of the access control system and cloud services (Figure 2). OpenStack cloud platform is implemented by using service bus for communication between its components. Service bus is based on Advanced Message Queuing Protocol (AMQP) technology and RabbitMQ service [6]. Access control security service was integrated with OpenStack bus by subscribing its software components to events of OpenStack Compute service, which is managing the lifecycle of virtual machines, and OpenStack Neutron service, which is managing the lifecycle of cloud networks. When security service receives an event that a new virtual machine is starting, it generates and distributes filtering rules for the firewalls and generates rules for the virtual switch using OpenFlow technology which redirects traffic from virtual machine to the firewall. Firewall-based approach allows controlling traffic between instances which are connected to one virtual switch but belong to different users or security groups.

The proposed security service requires additional resources in the cloud. Traffic filtering costs make up about 10% of the virtualization server's resources [4].

III. PROTECTED SEGMENTS FOR ENGINEERING APPLICATIONS

For tasks which require heterogeneous computing resources, it is necessary to automate creation of the protected segments. We describe heterogeneous computing system as a set of logical computing resources. Such a segment must be applied to the specified security policy to permit the possibility of access to computing resources for the owner, but forbid access to these resources to other users. When the task is complete, the results must be loaded into the data warehouse, and the computing resources are freed. At the same time, it is essential to guarantee access to computing resources in simultaneous execution of multiple tasks.

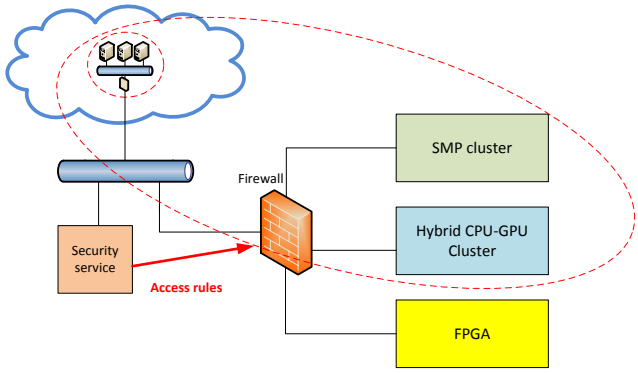


Figure 3. Reconfiguration of hybrid supercomputer center using firewalls.

We used OpenStack for creating groups of virtual machines in a cloud environment service. This service supports description of configurations in an Amazon Cloud Formation format that ensures compatibility with public services such as Amazon AWS. This service allows creating groups of virtual machines according to pattern, virtual networks, cloud-based routers and other components. The images of virtual machines contain a basic set of services. Any other application specific packages are installed using the automation services provided by Opscode Chef tool [7] that provides automated deployment of software configurations in virtual machines and bare-metal servers. When new computation segment is being created the security system spawns and configures virtual firewall which is filtering access to newly created network which serves computation. Dynamic network creation is supported by OpenStack Neutron services and distributed virtual switch Open vSwitch. After computing and receiving the results, the segment is removed, the cloud resources are released, and the results are uploaded to cloud storage and become available to the other consumers of the service. Every operation is automated: there are not any steps which need to involve human operations.

Reconfigurable segments of the cloud allow solving a wide range of scientific and technical tasks, among them: tasks that operate on large data sets based on the MapReduce technology; Bioinformatics tasks, including processing of genetic information in distributed systems; tasks of class CAD/CAE; calculation jobs not requiring high-speed networks. Tasks that cannot be solved in the cloud virtual machines (for example, requiring quick access to globally

 ISBN: 978-1-61208-330-8

addressable memory and massively-parallel or streaming computations) are transferred to the dedicated hybrid clusters for high performance computing, computing infrastructure platform and equipped with an internal high speed communication bus, nodes-accelerators based on FPGA and GPU. Firewalls provide protection from unauthorized access to computing resources in a time of challenge and consolidation of heterogeneous segments (cloud and high-performance) computing resources into a single computation network, which components can communicate with each other, using the allowed protocols.

Built this way, infrastructure allows dynamically creating secure computing segments and thus provides an opportunity to organize a simultaneous solution of various tasks on a single set of hardware resources (Figure 3). The proposed solution implements reconfigurable federated cloud with one interface and multiple computation segments. A similar approach was used for organizing mobile cloud for intelligent transport systems and presented in [8].

IV. CONCLUSION

The proposed approach of organizing engineering center which is based on cloud services enables ability to reconfigure computing resources for different computation tasks. Integrated security services allow sharing computing resources between different users and clients. Reconfiguration of computing resources by using cloud firewalls is not a standard approach. It requires additional resources and makes platform more complex. From other side, it provides opportunity of reconfiguration of resources on network level. Stealth technology allows leaving applied software without modification. Dynamic computation segments creation service allows to effectively using IaaS resources on demand.

V. FUTURE WORK

The next step in our project is to integrate heterogeneous virtual machines in cloud platform. We are working on adding GPGPU devices like NVidia k20x and FPGA devices to virtual machines by using PCI pass-through capabilities in modern hypervisors like XEN and KVM. Recently released OpenStack Havana has support of PCI pass-through to virtual machines so we started evaluating how it works with heterogeneous compute devices.

REFERENCES

[1] D. Horat, E. Quevedo, A. Quesada-Arencibia, ”A hybrid cloud computing approach for intelligent processing and storage of scientific data”, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 8111 LNCS (PART 1), 2013, pp. 182 – 188

[2] J. Yimu, K. Zizhuo, P. Q. Yu, S. Yanpeng, K. Jiangbang, H. Wei, “A cloud computing service architecture of a parallel algorithm oriented to scientific computing with CUDA and monte carlo”, Cybernetics and Information Technologies 13 (SPECIALISSUE), 2013, pp. 153 – 166

[3] V. S. Zaborovsky, A. A. Lukashin, S. V. Kupreenko, and V. A. Mulukha, “Dynamic Access Control in Cloud Services. International Transactions on Systems Science and Applications”, ISSN 1751-1461, Vol. 7, No. 3/4, December 2011, pp. 264-277

[4] A. A. Lukashin and V. S. Zaborovsky. “Dynamic Access Control Using Virtual Multicore Firewalls”, The Fourth International Conference on Evolving Internet INTERNET 2012, ISBN: 978-1-61208-204-2, June 24-29 2012, Venice, Italy, pp. 37-43.

[5] A. Lukashin, V. Zaborovsky, S. Kupreenko, “Access isolation mechanism based on virtual connection management in cloud systems: How to secure cloud system using high perfomance virtual firewalls”, ICEIS 2011 - Proceedings of the 13th International Conference on Enterprise Information Systems 3 ISAS, 2011, pp. 371 – 375.

[6] S. Vinoski, “Advanced Message Queuing Protocol”, Internet Computing, IEEE, Volume: 10 , Issue: 6, 2006, pp. 87-89, doi 10.1109/MIC.2006.116.

[7] D. Spinellis, “Don't Install Software by Hand”, Software, IEEE, Volume: 29 , Issue: 4, 2012, pp. 86-87, doi 10.1109/MS.2012.85.

[8] V. .S. Zaborovskiy, A. A. Lukashin, S. G. Popov, A. V. Vostrov, “Adage mobile services for ITS infrastructure”, 2013 13th International Conference on ITS Telecommunications, ITST 2013, 2013, pp. 127 - 132

 ISBN: 978-1-61208-330-8

An Architecture for Risk Analysis in Cloud

Paulo F. Silva[1,2], Carlos B. Westphall[1], Carla M. Westphall[1], Mauro M. Mattos[2], Daniel Ricardo dos Santos[1]

[1]Networks and Management Laboratory
Post-Graduate Program in Computer Science
Federal University of Santa Catarina, Florianópolis, Brazil
{westphal, carlamw, danielricardo.santos}@inf.ufsc.br,

[2]Development and Transfer Technology Laboratory
Regional University of Blumenau, Blumenau, Brazil
{paulofernando, mattos}@furb.br

Abstract - **Cloud computing offers benefits in terms of availability and cost, but it transfers the responsibility of information security management to the cloud service provider. Thus, the consumer looses control over the security of their information and services. This factor has prevented the migration to cloud computing in many businesses. This paper proposes a model where the cloud consumer can perform risk analysis on providers before and after contracting the service. The proposed model establishes the responsibilities of three actors: Consumer, Provider and Security Labs. The inclusion of the Security Labs provides more credibility to risk analysis making the results more consistent for the consumer.**

Keywords-cloud computing; information security; risk analysis.

I. INTRODUCTION

Cloud computing brings several challenges for the scientific community of information security. The major challenges are data privacy of users, protection against external and internal threats, identity management, virtualization management, governance and regulatory compliance, Service Level Agreement (SLA) management, and trust gaps [1]-[4].

A strategy to meet the challenges of information security in cloud computing is based on risk analysis [5]. Several papers have worked on risk analysis on cloud computing [6]-[12], focusing on specific techniques for identifying and assessing risks.

Current solutions for risk analysis in cloud computing do not specify the agents involved and their responsibilities during the implementation of risk analysis. This uncertainty creates deficiencies in risk analysis, as:

- Deficiency in scope occurs when the selection of security requirements is performed by the Cloud Service Provider (CSP) or an agent without sufficient knowledge. Detrimental to their own environment, thus skewing the results of the risk analysis. An agent that is not knowledgeable enough may specify wrong or insufficient requirements, thus creating an incorrect risk analysis;
- Deficiency in adhesion to Cloud Consumer (CC) occurs when the agent responsible for defining impacts ignores the technological environment and business nature of the CC. In this case, the specification can disregard the impact scenarios relevant to the CC or overestimate scenarios that are not relevant, thus creating an incorrect risk assessment;
- Deficiency of reliable results occurs when the quantification of the probabilities and impacts is performed by an agent who is interested in minimizing the results of the risk analysis. For example, if the analysis is performed solely by CSP, he can soften the requirements and evaluation of impacts, thus generating a satisfactory result for the CC. However, such results are incorrect.

The deficiencies outlined above can generate a lack of trust on the part of CCs in relation to risk assessments, as in current models where CSPs are performing their own risk analysis, without the participation of CCs or any other external agent.

This paper proposes a model of shared responsibilities for risk analysis in cloud computing environments. The proposed model aims to define the agents involved in the risk analysis, their responsibilities, language for specifying risks and a protocol for communication among agents.

The rest of this paper is organised as follows. Section 2 discusses related works. The proposed model is presented in Section 3. Section 4 discusses the results. The conclusion and future works are presented on Section 5.

II. RELATED WORK

Architectures for risk analysis in cloud computing are presented in many solutions.

Hale and Gamble [7] show an architecture called SecAgreement which enables the management of security metrics between CSPs and CCs. A SLA for risk management in the cloud is presented by Morin et al. [8]. Ristov et al. [9] discusses the analysis of risk in cloud computing environments based on ISO 27001 and proposes a model for assessing security in cloud computing.

Chen et al. [10] present an architecture that defines levels of security from the risk of each service offered by CSP. Zech

 ISBN: 978-1-61208-330-8

et al. [11] portray a model for security testing in cloud computing environments based on a risk analysis of these environments. Wang et al. [12] explore the risk analysis in the cloud using techniques based on intrusion attack-defense trees and graphs.

The related works presented above discuss risk analysis of requirements or specific scenarios on cloud computing, but they do not address the definition of the agents involved and their interactions during the risk analysis.

III. THE PROPOSED ARCHITECTURE

The proposed architecture defines the sharing of responsibilities between three agents during the risk analysis. Information Security Labs (ISL) is an agent that represents a public or private entity, which specializes in information security, e.g., an academic or private laboratory. The CC is an agent that represents the entity that is hosting their information assets in the cloud. The CSP is an agent that represents the entity being analyzed.

The three agents defined by the proposed architecture divide the responsibilities of running a risk analysis, according to the concepts defined by ISO 27005 [5]. In this context, threats exploit vulnerabilities to generate impacts on information assets.

A risk analysis works with many variables. The variables used in the proposed architecture are: (i) DE – Degree of Exposure, defines how the cloud environment is exposed to certain external or internal threat, (ii) DD – Degree of Disability, defines the extent to which the cloud environment is vulnerable to a particular security requirement, (iii) P – Probability, defines the probability of an incident occurrence, i.e., a threat exploiting a vulnerability (iv) I – Impact, defines the potential loss in the event of a security incident, (v) DR – Degree of Risk, defines the degree of risk for a given scenario of a security incident.

The risk analysis of the proposed model is organized in two well-defined phases: risk specification and risk assessment.

The risk specification phase defines threats, vulnerabilities and information assets that will compose the risk analysis. At this stage it is also defined how to quantify the threats, vulnerabilities and assets specified.

The risk assessment stage comprises the quantification of the variables DE, DD and I, for threats, vulnerabilities and information assets, respectively. In this phase the quantification of variables of P and DR for each incident scenario is also performed (a combination of threat, vulnerability and asset information).

Figure 1 illustrates the flow of interactions between components of the architecture and the ISL, CSP and CC agents in the risk specification phase. Initially each agent must register with their respective registry component (Fig. 1 a, b, c). After their registration, the ISL is responsible for identifying threats and vulnerabilities in cloud computing environments. Then, the ISL specifies how to quantify threats and vulnerabilities.

The architecture provides a language for the specification of risk, the RDL – Risk Definition Language. This language is used by ISL to specify threats and vulnerabilities. The RDL is specified in XML and contains information such as: risk ID; ISL ID; threat and \ or vulnerability ID and reference to a WSRA – Web Service Risk Analyzer. The WSRA is a Web Service specified by ISL to quantify the Degree of Disability (DD) and Degree of Exposure (DE).

After developing its RDLs and WSRAs, the ISL exports the records for the RDLs repository (Fig. 1-d) and publishes WSRAs.

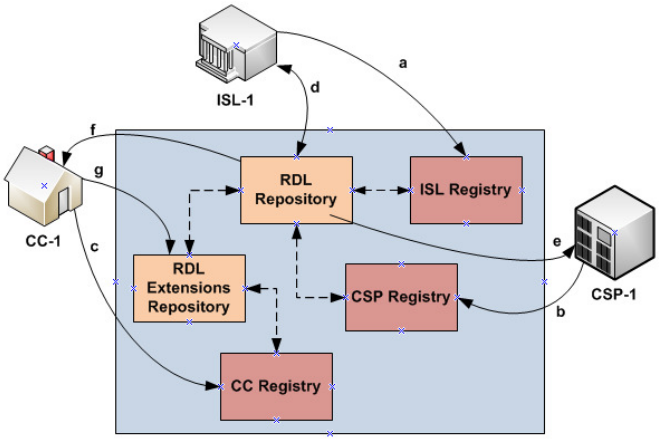


Figure 1. Risk specification phase.

The responsibility of the CSP on the specification phase of risk consists in importing RDLs and implementation of calls to WSRAs (Fig. 1-e).

ISL is responsible for the correct identification of threats and vulnerabilities. CSP is responsible for the correct execution of the quantification of threats and vulnerability. The CC agent is responsible for the identification of information assets and the quantification of impact, as this is the most fitting agent to express the cost of an information security incident.

In order to perform the identification of an information asset and quantifying an impact on this asset a CC must import the RDLs (Fig. 1f) and extend them including information on information assets and their impacts.

The method of quantification of impacts may be static or dynamic. In the static method the CC determines a fixed value for the impact and in the dynamic method the CC specifies a Web Service to quantify the impact. After specifying their information assets and their impacts, the CC exports the extension to the RDL Extensions Repository (Fig. 1g).

Figure 2 illustrates the flow of interactions between the components of the proposed architecture and the ISL, CSP and CC agents during risk assessment.

The Risk Analysis component coordinates the interaction between external agents and other internal components of the proposed architecture. The RDL Repository and RDL Extensions Repository components store records of threats and vulnerabilities of ISLs and information assets of the CC, respectively. The RA Processor component is responsible for establishing the relationships between information assets,

 ISBN: 978-1-61208-330-8

threats and vulnerabilities, as well as performing the calculation of risk.

The CC, ISL and CSP agents present the components Impacts Evaluation, Evaluation WSRA and CSP Proxy, respectively. Impacts Evaluation is a component that contains the Web Services for dynamic definition of impacts or tables for static impacts. Evaluation WSRA is a component that contains the Web Services assessment of threats or vulnerabilities identified by an ISL. CSP is a proxy component deployed in CSP to perform the call of the WSRAs.

The risk assessment begins with the CC informing the CSP to be analyzed (Fig. 2a). Then the Risk Analysis component queries the RDL repository (Fig. 2b) and performs a call of the CSP Proxy component passing the information about each risk (Fig. 2c).

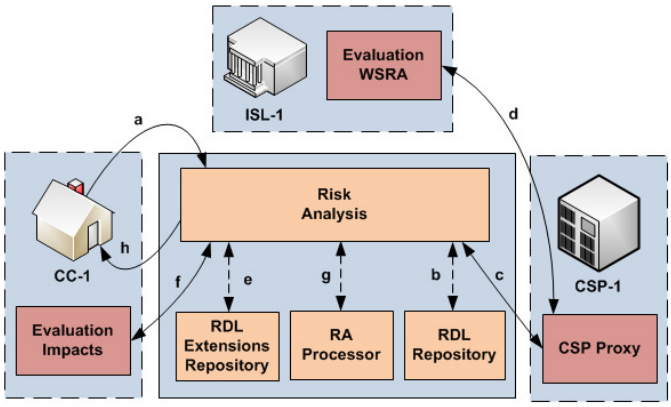


Figure 2. Risk assessment phase.

Based on each RDL received, the CSP performs a call of the WSRA (Fig. 2d). The WSRA is run by ISL and returns the quantification of the threat (DE - Degree of Exposure) or vulnerability (DD - Degree of Disability). Then, the quantification of the threat or vulnerability is returned to the Risk Analysis component (Fig. 2c) and stored. The steps "b", "c" and "d" in Fig. 2 are executed for each RDL in RDL Repository.

The quantification of impacts as defined by the CC starts after the quantification of all threats and vulnerabilities. The Risk Analysis component queries the RDL Extensions Repository (Fig. 2e) and performs a call of the Evaluation Impacts component for the quantification of the impact (I - Impact) (Fig. 2f).

After obtaining the quantification of all impacts the Risk Analysis component is able to perform the calculation of the probability and risk. Therefore, all records showing the quantification of threats, vulnerabilities and impacts are sent to the RA Processor component (Fig. 2g).

The RA Processor component sets the valid relationships between information assets, threats and vulnerabilities, and performs the calculation of the probability (P - Probability) and of the risk (DR – Degree of Risk) through the variables DD, DE and I previously quantified.

After calculation of risk analysis the result is returned in XML for Risk Analysis component (Fig. 2g), and transferred to CC (Fig. 2h).

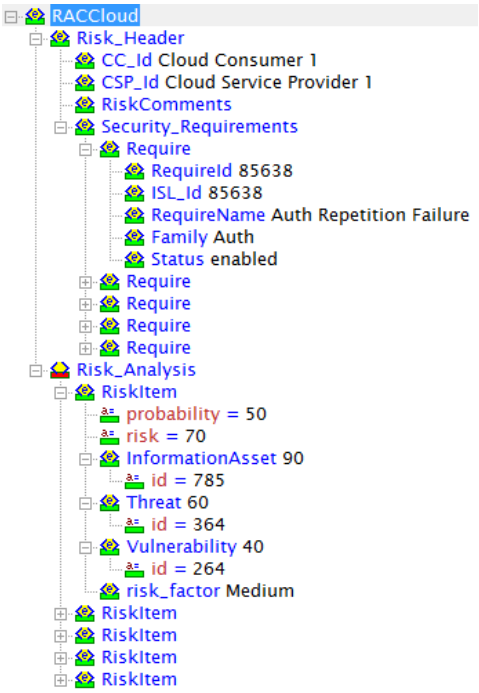


Figure 3. Risk Analysis result.

The XML resulting from the risk analysis (Fig. 3) contains the ID of the CC and CSP agents, and a list of security requirements that were defined by ISLs. Each requirement contains its ID and the ID of the ISL that created it. The resulting XML still contains the probability (P) and the degree of risk (DR) which was calculated for each requirement and the results of variables, DE, DD and I.

IV. Results and Discussion

With the information from the risk analysis, the CC may decide to allocate or not their information assets in a particular CSP.

The proposed model aims to reduce the three main deficiencies presented by current models of risk analysis in the cloud: deficiency in scope, deficiency in adhesion and deficiency of reliable results.

The reduction in adhesion deficiency occurs when the proposed model includes the CC as a key agent in the process of risk analysis. The CC agent has an important role in risk analysis, defining information assets and quantifying impacts on these assets.

The CC is the most suitable agent for the definition of impacts. It is the agent which best understands the relevance of each information asset within its area of expertise. CSP and ISL agents are not able to identify or quantify the impacts on information assets. They are not experts in the business of CC.

The proposed model acts to reduce the deficiency in scope by adding the ISL agent. ISL is an agent specializing in information security. It is the entity best suited to define

 ISBN: 978-1-61208-330-8

security requirements, threats and vulnerabilities (specification of the RDLs), as well as to define how to qualify such threats and vulnerabilities (specification of the WSRAs).

The proposed model acts on the deficiency of reliable results because in our model the CSP has more restricted responsibilities than in traditionally models presented by related works.

Traditionally, the CSP is responsible for defining security requirements and the tests that are applied to evaluate the risk of their own environment. In this scenario, the risk assessment can be smoothed by CSP. The inclusion of the ISL agent removes responsibilities which are traditionally assigned to the CSP, such as the identification and quantification of threats and vulnerabilities, thus making the result of the risk analysis more reliable.

The proposed model allows multiple ISLs defining RDLs and WSRAs jointly (Fig. 1). Thus, the definitions of risk can come from different sources and can be constantly updated in a dynamic and collaborative way, forming a large and independent base of risk definition for cloud.

The way WSRAs are specified is also a feature that impacts the improvement of scope. The use of Web Services to specify safety requirements allows them to be platform independent. It also allows the use of a wide variety of techniques for quantifying the threats and vulnerabilities because the only limit is set by the programming language chosen for implementation of WSRA.

Related works of risk analysis in the cloud do not consider the role of the CC agent on risk analysis. These works usually focus on the vulnerability assessment by the CSP, without considering the impact it will have on the vulnerability of the different information assets of the CC. The proposed model assigns the responsibilities of the identification and quantification of impact to the CC. Thus, the performing of risk analysis is shared among different agents, so the responsibility for quantifying the variables of risk analysis is not centered on a specific agent.

The CSP is the agent that will be analyzed; therefore it is not able to set any of the variables of the risk analysis, as this could make the results of risk analysis incorrect. The role of CSP is only to inform the data requested by ISL, so ISL itself performs the quantification of each requirement of information security.

A CC can perform analysis on multiple CSPs before deciding to purchase a cloud service. It is also possible to perform periodic reviews of its current provider and compare them with other providers in the market, choosing to change CSP or not.

V. CONCLUSION AND FUTURE WORK

This paper presented a model of shared responsibilities for risk analysis in cloud computing environments. In addition to the traditional CC and CSP agents the model adds the ISL agent, which is responsible for identifying and specifying the security requirements.

The model presented in this paper is an initiative to allow the CC to perform the risk analysis on its current or future CSP. Also, this risk analysis is broad, current, unbiased and reliable.

The characteristics presented in this article aim at generating a more reliable risk analysis for CC, so that it can choose its CSP based on more solid information.

Several papers on cloud computing indicate the lack of trust from CC to CSP as a relevant factor in avoiding the purchase of cloud computing services. A risk analysis can act to reduce or eliminate this suspicion and boost the acquisition of cloud computing services.

The presented model performs a free and reliable risk analysis because the analysis is not centered in the CSP. The identification and quantification of threats and vulnerabilities are carried out collaboratively by several laboratories. Safety and impact on information assets are quantified by the CC.

The risk analysis of the proposed model is broad because the security requirements are defined by specialized laboratories and the CC itself defines and quantifies their information assets. It is dynamic because the various ISLs can modify their security requirements for considering new vulnerabilities in future risk analysis.

This work opens possibilities for the development of future research. There is a need for research on the reliability of the data reported between CSP and ISL during risk analysis. The RDL - Risk Definition Language can be further explored in specific jobs. Further research should be done on the inferences on the results of risk analysis. These inferences can help all stakeholders in understanding the causes of incidents and their solutions. Finally, there is the need to extend this work so that the proposed model can also suggest the controls or countermeasures to the CSPs.

REFERENCES

[1] M. K. Srinivasan, K. Sarukesi, P. Rodrigues, M. S. Manoj, and P. Revathy, “State-of-the-art cloud computing security taxonomies: a classification of security challenges in the present cloud computing environment”. ICACCI '12: Proceedings of the International Conference on Advances in Computing, Communications and Informatics. August 2012.

[2] H. Yu, N. Powell, D. Stembridge, and X. Yuan, “Cloud computing and security challenges”. ACM-SE '12: Proceedings of the 50th Annual Southeast Regional Conference. March 2012.

[3] K. Ren, C. Wang, and Q. Wang, "Security Challenges for the Public Cloud," *Internet Computing, IEEE*, vol. 16, no. 1, pp. 69-73, Jan.-Feb. 2012 doi: 10.1109/MIC.2012.14.

[4] B. Grobauer, T. Walloschek, and E. Stocker, "Understanding Cloud Computing Vulnerabilities," *Security & Privacy, IEEE* , vol. 9, no. 2, pp. 50-57, March-April 2011 doi: 10.1109/ MSP.2010.115.

[5] ISO/IEC 27005:2011, Information Security Risk Management. [Online]. Available: http://www.iso.org.

[6] J. Zhang, D. Sun, and D. Zhai, "A research on the indicator system of Cloud Computing Security Risk Assessment," *Quality, Reliability, Risk, Maintenance, and Safety Engineering*

 ISBN: 978-1-61208-330-8

(ICQR2MSE), 2012 International Conference on , pp. 121-123, 15-18 June 2012 doi: 10.1109/ ICQR2MSE. 2012.6246200.

[7] M. L. Hale, and R. Gamble, "SecAgreement: Advancing Security Risk Calculations in Cloud Services," *Services (SERVICES), 2012 IEEE Eighth World Congress on* , pp. 133-140, 24-29 June 2012 doi: 10.1109/ SERVICES.2012.31.

[8] J. Morin, J. Aubert, and B. Gateau, "Towards Cloud Computing SLA Risk Management: Issues and Challenges," *System Science (HICSS), 2012 45th Hawaii International Conference on* , pp. 5509-5514, 4-7 Jan. 2012 doi: 10.1109/ HICSS.2012.602.

[9] S. Ristov, M. Gusev, and M. Kostoska, "A new methodology for security evaluation in cloud computing," *MIPRO, 2012 Proceedings of the 35th International Convention* , pp. 1484-1489, 21-25 May 2012.

[10] J. Chen, Y, Wang, and X. Wang, "On-Demand Security Architecture for Cloud Computing," *Computer*, IEEE, vol. 45, no. 7, pp. 73-78, July 2012 doi: 10.1109/MC.2012.120.

[11] P. Zech, M. Felderer, and R. Breu, "Towards a Model Based Security Testing Approach of Cloud Computing Environments," *Software Security and Reliability Companion (SERE-C), 2012 IEEE Sixth International Conference on* , pp. 47-56, 20-22 June 2012 doi: 10.1109/SERE-C.2012.11.

[12] P. Wang, W. Lin, P. Kuo, H. Lin, and T. C. Wang, "Threat risk analysis for cloud security based on Attack-Defense Trees", *Computing Technology and Information Management (ICCM), 2012 8th International Conference on*, vol. 1, pp. 106-111, 24-26 April 2012.

 ISBN: 978-1-61208-330-8

MEDIACTIF

A dynamic, centralized and real-time digital signage system for smooth pedestrian flow control with arbitrary topologies

Thierry Simonnet, Samuel Ben Hamou
IT Department.
ESIEE-Paris
Noisy le Grand, France
{thierry.simonnet, samuel.benhamou}@esiee.fr

Jacques Angelé
inExtendis
Malakoff, France
j.angele@inextendis.fr

***Abstract*—Digital signage systems distribute all kinds of information to specific locations, including retail stores, public areas and transportations. Thanks to the massive adoption of displays and wide application of wireless networks, digital signage can deliver targeted messages designed to accurately reach the passing audience and eventually influence customers. Digital signage, including Digital Out Of Home, is a natural evolution of the old sign painting business. The aim of the MEDIACTIF project is to develop a new digital signage concept, using both human traffic modelling and compartmental behavior based on interests. This is the major issue we are facing as of now: crowd motion has extensively been studied on peculiar test cases, on both macroscopic and microscopic levels, but it is lacking considerable social and personal inputs, which are significant for each site and scenario.**

Keywords-digital signage; real-time; network; mesh.

I. INTRODUCTION

The MEDIACTIF Project [1] has been started a few months ago to propose solutions to different problems:

- Reducing the stress of fair visitors or airports clients by reducing crowd congestions and increasing the pedestrian flows.
- Giving relevant information to users and the operations team in real time.
- Offering a centralized system that may adapt signage to any situation.
- Increase security levels with adapting signage (redirect pedestrians easily in case of emergency situations)
- Reducing global waste of both printed signage materials, but also aiming at a massive drop in the consumption of high toxicity materials like inks.

Partners on this project have different fields of expertise and some of them are facing day to day issues regarding their fixed or dynamic signage impact over people behaviors. They would like to efficiently inform clients or simply redirect them to a specific location.

As the literature is quite extensive on traffic management systems, and also because of a gathered experience on this field, it will serve as a basis to this project.

We will first talk about the general purpose of the system, followed by an overview of current traffic systems. Then we will focus on standards for four identified areas of interest before briefly going over the current architecture we are considering. We will also slightly talk about additional services considered for the system.

II. GATHERED DESIDERATA

Presently, digital signage systems handle different media sources and dispatch them to a pool of selected screens. Information can be static (map, pictures, static adverts [2]), dynamic (flight departures and arrivals), animated (film trailers, animated adverts). Obviously, any combination of these previous elements can be displayed on the same screen using defined scenarios. Different applications are currently being considered and worked on:

- Public information: area map, dynamic pathway, exit roads.
- Communication channels: static and dynamic signs, mobile apps, kiosks, social/collaborative interactions.
- User experience and environment enhancements: use of mobile/apps as a natural extension of signage, dynamic path finding, and profile adapted suggestions, interaction with the digital signage information system.
- Behavior influence: events propagation and advertising, customized information, lunch/dinner indications, etc., based on activities, objectives, profiles and customer interests.

The MEDIACTIF Project aims not only to display different contents on screens, but also to integrate multiple sensors in the processing loop. This would allow for signage adaptation depending, for example, on:

- crowd densities
- personal preferences
- commercial factors of interest
- specific or recurrent events
- previously computed models and statistics

 ISBN: 978-1-61208-330-8

These models should be driven by the analysis of previous sessions (when available) and will be enhanced with the sensors variations/data constantly.

The needs considered can be so far classified in the following categories: users, operations, security, and environment.

A. Users

In a fair or in an airport, a good way to reduce stress is to reduce the global walk time and crowding. "Way finding" digital signage can be used to allow for optimal pathways for everybody. MEDIACTIF would like to provide a comfortable walk by dynamically adapting path suggestions with adaptive signage closely linked to affluence sensors. It will not only allow for "navigation" but also provide real time information (a potential stress factor for some).

Users will be provided with different kinds of extra information which may influence their behaviors. An example of this would be to push restaurant information to some visitors before lunch time to eventually allow them to eat before the rush hour. These data can be fully public and basic (restaurant opening hours or placement), more practical (remaining available seats), but could also influence the subject (inducing the need of a meal with restaurant discounts, tastings at specific stands, etc). Advertising has also been considered as a part of a future remunerated service.

B. Exploitation

As usual for such a complex system, the first approach is to centralize, at least in a first iteration, all information gathering. Doing so it will be easier to process all sensor values gathered in real time and compute relevant actions in order to update displays.

All computed information will be attached to different scenarios or strategies. Basic information would consider of fixed maps and schedules which will cycle with other somewhat basic type of data. Each "screen" will be displayed for a specific amount of time and cycle according to the crowd estimation also in concordance with past statistics, or specific times for lunch or particular events.

This basic mode will be later on enhanced by an adaptive algorithm. Depending on unexpected events, pathways congestions, specific commercial deals, maps will be adapted to drive people efficiently and ensure a general fluidity.

During each fair or over a period of time for airports or different clients, every bit of data (crowd movements, computed from cameras and sensors, positioning of opened desks, active or inactive zones, which will pose problems to a global fluidity), will be collected in order to generate new model revisions.

For corporate users, the benefits of such a system are not negligible:

- Fair organizers will be provided a tangible prediction of the flow of persons at their exhibition based on to the simulations and past data characterizations.
- Possibility to quantify the impact of any layout modification on the traffic flows.
- Commercial trade centers or security offices will be able to anticipate influence of any departure or arrival of a flight, depending on destination or origin. They would also be able to manually pilot path redirections to offer a better service.

They ideally should have access to tools allowing them to qualify and quantify any kind of influence on people traffic interaction.

C. Security

Security management takes a great part in MEDIACTIF. In case of fire or emergency, digital signage can be used to redirect people to the nearest exits or safe routes. Obviously all displays must embed a security mode, with an autonomous mode and power supply.

The second security mode we are considering is to allow for easier and faster interventions for dedicated services: in case of an emergency with a person (malaise for example), we would want to adapt the signage around an area to drain the crowd and offer a clean pathway for emergency.

Once again, here, the main advantage regarding the security sector, is to allow for a traffic fluidity in order to avoid stress and nervous feeling.

D. Environment

VIPARIS, one of MEDIACTIF's partners, is in charge of the ten most important congress and exposition centers in the Paris area. It accounts for around 330 fairs, more than 100 spectacle representations, 150 conventions and 620 enterprise events driving more than 11Milion visitors each year. In Paris, one fair produces a mean of 200m^3 of waste, especially printed signs and posters. Not only this number is quite high, but these prints also use highly toxic inks which are difficult to recycle. Considering even only 1000 events per year for our area, we could suppress at least 200 000m^3 of waste per year using fixed or mobile reusable displays, etc.

III. EXISTING URBAN TRAFFIC CONTROL (UTC) SYSTEMS

MEDIACTIF will base its first iterations on the experience and results gathered with a urban traffic lights control system. Many systems are deployed all around the world, handling specific drivers' behaviors. So far, traffic control systems can either be distributed or centralized, using plans, local adaptation or be traffic responsive [3].

So far, we have settled on static and scheduled models as entry level rules, which will be influenced by real-time adaptive algorithms later on.

 ISBN: 978-1-61208-330-8

A. Fixed time systems

The method used to calculate timings defines the objective that the system will seek to minimize. This is often used to reduce network vehicle delays. The designer can optimize different parameters of the network to attain different objectives. Although the timing can be biased against the main traffic movements, it can generally be restrained by adjusting the splits at critical junctions.

Fixed time systems cannot respond dynamically because they use pre-calculated timing plans. They also are unable to respond automatically to incidents. This implies that they are unfortunately not suitable for situations with any variability pattern. In fact they cannot adapt to any change in traffic patterns. The Traffic Network Study Tool (TRANSYT) in UK was a system that used this technology [4].

B. Plan selection systems

Plan selection systems use fixed time plans, but select which plan to use according to sensors data, rather than by timetables. However, this type of system has not proven to be any better than simple time-of-day implementations with fixed time plans. Plan selection systems have the same advantages and drawbacks as fixed time systems.

C. Plan generation systems

Plan generation systems generate their own fixed time plans from sensors information and implement them. These systems are way less under direct control from exploitation people. In theory, this kind of system could respond to unexpected incidents, but in reality, their degree of freedom is too small to allow them large enough changes in order to respond effectively. The Sydney Co-ordinated Adaptive Traffic System (SCAT) uses this technology [5].

D. Traffic responsive centralized systems

Traffic responsive systems are fully dynamic. The system is based on a central server and communication to controllers. The advantages of responsive systems are that they can respond to traffic demand, day to day variation, unexpected events and traffic evolution. A responsive system adjusts its control depending on sensors inputs. A centralized system has the advantage that all the relevant information, from sensors but also from system is fully centralized and available. Centralized traffic management systems offer also a better reaction to events and a better efficiency [6].

SCOOT (Split, Cycle and Offset Optimisation Technique) [7][8] in UK and GERTRUDE (Gestion Electronique de Régulation de Trafic Routier Urbain Défiant les Embouteillages) in France are traffic responsive centralized systems.

E. Traffic responsive systems with distributed processing

The features and advantages of distributed responsive systems are basically the same as those of centralized responsive systems. A major difference is the communication system used. A centralized responsive system has continuous communication between each controller and the central server. A distributive system has a router module and each controller is connected to neighboring controllers. Messages can be passed between any machine or controller to others, routing the message by intermediate controllers when necessary. A distributed responsive system should be able to work with a route guidance system, but interaction is much more complicated than for a centralized responsive system. PRODYN in France has been implemented and tested on the Zone Experimentale et Laboratoire de Trafic de Toulouse (ZELT) [9][10]. OPAC (USA) and UTOPIA/SPOT (Italy) implements this architecture. It was a mesh like implementation, years before mesh network was fully formalized.

MEDIACTIF will provide a prototype implementation for a “people responsive centralized system”, and intends to enhance the system using a distributed architecture.

The major evolution compared to a UTC system is people, more particularly individual behaviors. Cars drive on separate ways, each line going in a specific direction, whereas crowds are more hectic. Another difference is the response time of the system. A car traffic (with a max speed of 50 km/h) real-time system uses a 1 second internal cycle. It is at the moment not easy to establish a correct value for crowd behaviors.

The type and positioning of sensors, as well as a crowd dispersion model will take a big part in a sub module evaluation during the project. Some partners already have an extensive review of different products and technologies which will help us get a good grasp of what to avoid at least.

IV. STANDARDS AND TOOLS

A. Digital signage standards

The ITU published a whitepaper [11] in which Synchronized Multimedia Integration Language is cited as "a key standard for the digital sign industry," and that it "is increasingly supported by leading digital sign solution providers." It is reported in [12] that SMIL players are deployed for nearly 100,000 screens in year 2011, and a single software provider has won three major projects, deploying more than 1,000 SMIL players over the same period of time for each one.

POPAI has released several digital sign standards [13] to promote "interoperability between different providers". The objective of these standard documents is to establish a foundation of performance and behavior that all digital sign systems can follow. The current sets of standards published by POPAI are:

- "Screen-Media Formats" to specify compatible and supported file formats.
- "POPAI Digital Sign Device RS-232 Standards"
- "POPAI Digital Sign Playlog Standards V 1.1"

 ISBN: 978-1-61208-330-8

- "Digital Control Commands"
- "Industry Standards of Digital Sign Terms"

MEDIACTIF will try to use existing digital signage appliances, if already installed. Some airport terminals are already equipped for arrival/departure schedule information. These systems will be operated using communication standards (HTML5, SMIL, SOAP, REST). For new contracts or area, the MEDIACTIF Project may have to define its own digital signage system that will:

- be standard compliant
- use different communication channels: IP (wired and Wi-Fi), wired and radio RS232, etc.

B. *Crowd motion modelling*

The major issue in handling pedestrians is to define accurate models valid for a wide range of topologies and flow densities. Models should also be reasonable realistic, robust against incomplete data, and of course computationally manageable.

Pedestrian dynamics share some similarities with fluids, and it is not surprising that the first models of crowds were inspired by hydrodynamics or kinetics of gases. Henderson found as early as 1971, from measurements of motion in crowds, a good agreement of the velocity distribution functions with Maxwell-Boltzmann distribution [14].

In microscopic models, each individual is represented separately. In contrast, in macroscopic models, different individuals are not distinguished. Instead, crowd densities are at play, usually a mass density derived from cumulative locations of persons and also a corresponding locally averaged velocity of this density.

Social forces have been introduced by Helbing in a microscopic model [15] based on the idea that pedestrians have different perceptions about intimate/personal and social space, which leads to repulsive forces between persons.

Cellular automata [16][17][18] are another important class of models that are discrete in space and time. Most of these models represent pedestrians by particles that can move to one of the neighboring cells based on transition probabilities which are determined by the desired direction of motion, interactions with other pedestrians, and interactions with the infrastructure (walls, doors, etc.).

However, unlike Newtonian particles, persons have a free will and may want to avoid jams by changing their preferred path when approaching a crowded area, find a new path, even if it is not the preferred one. To take into account such strategies, microscopic models are to be extended well above the present state of the art. Stochastic behavioral rules may lead to potential realistic representations of complex systems like pedestrian crowds, however, parameterization and calibration of such models may remain elusive.

Following Maury [19], models can also be classified according to their stiffness. Soft congestion models are applicable when the distance between individuals becomes smaller, while hard congestion models propose solutions in the case of physical contacts between individuals (when people are packed, the overall motion is perturbed by the fact that two persons may not occupy the same place at the same time).

Mainly, we have to focus on congestion in crowds in motion [19][20][21].

This poses a non-trivial modeling problem as we not only have to characterize a general walking behavior for users, but also have to take into account the fact that contacts are usually avoided and mostly not anticipated when they occur. We will probably limit our model in the first iteration to basic constraints raised by the environment and goals derived from a specific location: the expected behavior will be surely different between a person in an airport terminal and the same person at a fair. Thus we have to consider that each individual will move according to its current desires though we cannot exclude any spontaneous irregularity. Also the model has to be considered on both a microscopic and macroscopic level: the initial goal of the user will drive its original direction and velocity on a macroscopic scale while the interaction with its environment (other users or the physical structures defining the area map itself) will be evaluated at a microscopic level.

The other phenomenon to take into account is the fact that not only does a small congestion affect the behavior of a single individual but a wider one tends to have an effect on a group motion at a global scale. At the same time we can also consider that an individual, with common sense, will tend to avoid high density areas when possible.

We currently are reviewing the literature on this particular problem to find a suitable modeling possibility for different scenarios and are considering multiple options, such as neural networks, pure statistical models but also a quite interesting representation of crowds as fluids to which fluid dynamics theories could be applied.

C. *Positionning terminals in mobile computing*

To offer precise services using smartphones, it is necessary to be able to pinpoint a position with a good precision. There are different techniques available using wireless positioning [22]. The smartphone collects the received signal and compares the computed vector to the vectors previously recorded along the walk. It is also possible to measure the distance between mobile terminals [23] or fixed access points.

Presently, some companies are testing such techniques to position their staff inside the different airports areas using either Wifi or Bluetooth low energy depending on the smartphone (Android or iOS). We are also looking at emerging solutions such as IEEE® 802.15.4/ZigBee® technology positioning options and also some recent announcements made by STMicroelectronics with the

 ISBN: 978-1-61208-330-8

LPS331AP, an ultra-compact, absolute piezoresistive pressure sensor which advertises for *"3D indoor positioning and enhanced GPS in portable devices"*.

D. Crowd sensors

In order to minimize costs at first, we will mostly try to use the infrastructure already in place. For some places, this includes a camera network covering almost all areas of interest. The video feeds will be analyzed in real time using OpenCV, an open source computer vision library. Once integrated with a digital representation of the concerned areas it will allow us to characterize at least individuals movements and crowds formations, or as mentioned earlier fluid mechanics of each area. Though it will not specifically generate any particular quantitative data, it will serve as our main source for the crowd dispersion management models and redirection protocols in case of emergencies.

For the numerical part (initially mostly for fairs), we have already considered a deeper integration with the organizers to be able to monitor the ticketing desks streams.

For both sectors we will also surely use counting sensors based on different technologies (pressure, video, radio) to be able to monitor efficiently all concerned area and be able to generate appropriate reactions to any event.

V. ARCHITECTURE

MEDIACTIF will be an adaptive, centralized system. All relevant data will first be collected by a specific module and will serve to compute crowd models. These will be used for primal crowd traffic prediction. They will act as the base module for all corporate models like airports and event organizers. It should also allow mobile services (free or not) for visitors via apps or web services. According to the state of the art regulations, it has to follow network standards to pilot existing digital signage systems and to communicate with its own devices.

As an integrated platform, we must offer different tools for different users and/or operation modes:

A. Front office tools

At the early stage, conception tools, like a computer assisted design software, are needed for designing sensors placements and optimize screens positioning to the site topology. Sensors values should be able to be controlled from the interface and a synoptical backplane view must be able to display a synthetic survey of the whole physical implementation. It will allow for enabling/disabling sensors, forcing states or displaying specific messages.

Basic behaviours will be held by computed plans using models but adaptive rules will mostly be managed by the use of programming languages. Scilab [26] will be the original programming tool suggested, handling each controller separately and also each area globally. As an interpreted language, it proves more convenient to program specific rules and test them using the included web interface as a frontend. Once the tuning phase is finished, files will probably be compiled and integrated as executable application inside the backend system for performance and confidentiality reason.

B. Backoffice tools

The back-office architecture is organized around communication, events and sensors management.

For a fair, or for an airport terminal, it is possible to define basic events or rules (holidays, week-end, plane arrivals and departures) and associate, to each of them, fixed plans to anticipate crowd variations.

Using key sensors values, it is possible to manage adaptive modes (last minute gate changes, system failure, human weaknesses, etc.). One of the key rules will be to take a crowd increase immediately into account, but awaiting stabilization for a decrease.

C. Data management

During an event like a fair or even on a 24/7 schedule for the airport case, the system will store key data, time stamped in a big data server, which will later on be used to compute a new model version adapted to a specific location.

The first system model will surely be an a priori model, based on knowledge and extensive statistics. During operations, real time data and predictions will be compared and analyzed, with a set of specific events and conditions. For the next fair, a new refined model will be setup and tested with all past data. The objective is to minimize differences between model prediction and real situation. Unexpected events will be removed to compute a new model.

Based on physical and goals differences between each particular location, models will evidently be different in the long run but can be tuned for specific events based on similarities.

D. Tools for operations users

MEDIACTIF has to stand as a complete system for corporate users.

Taking the airport scenario, such a system must firstly reduce walk time and crowd generation/increase but it could also have a more commercial aim, especially for shops. Today, it is quite impossible to evaluate the influence of an arrival or a departure over shop businesses. With this system, it will be possible to anticipate these events and adjust an adequate commercial response.

For fair organizers, the situation is quite different. It is necessary to ensure a global fluidity in the alleys but also to maximize visitors just before a congestion level is attained.

We also want to be able to help organizers in designing their fairs and anticipate the temporal and spatial distributions of global and local traffic flows. Consequences are evident: scaling a fair with an optimized topology and pathways, identifying probable critical spots and times for congestions, lowering the stress and improving the security, while enabling new business models based on deliveries of

 ISBN: 978-1-61208-330-8

quantitative predictions and actual traffic flow measurements.

E. End users/passerby benefits

For the end user, benefits are much more concrete. First, a reduced walk time and/or a better walk fluidity decreases the perceived fatigue. This not only avoids the greater part of stress but can lead for a global increase in the positive perception of the fair and/or a longer stay, which could translate in more business opportunities. Another direct benefit is relevant and up to date information for any unexpected events.

For commercial or enhanced services, end users can also have individual and exclusive information or vouchers on their smartphone depending on each and everyone's profiles. Signage for a business man with luggage will not be the same as signage for a family.

Another example would also be to handle access points like car parks or cash desk locations. Of course this implies a refined management of every parameter of a particular site and a better anticipation, avoiding manual transit time measures.

F. Commercial tools

MEDIACTIF is not only a Digital Signage System. It can also be considered as an engineering commercial methodology.

It can be used as a conception tool for commercial purposes design, crowd modelling, fair design, etc.

Localization, path finder, thematic itineraries, specific profile handling (disables people, families, etc.) can be managed with the right plug-in on the system.

VI. SERVICES

MEDIACTIF is a dynamic Digital Signage management system. But it can also be considered as a bartering platform. It can establish relationships between different operators like restaurants, advertising and media companies. The system offers data and capabilities to address users' needs which can be driven through services by external operators.

We plan to offer different levels of services, from low level like security to high level like data exchange.

The first level to handle is security. Even if the system is down, or if there is no communication to controllers, signage must work fully independently to indicate at least a fixed map and fire exits during a defined time.

The first enhanced mode would be to offer a basic plan to allow users pathway indications.

In case of a failure from the central server or in the transmission of data, controllers can work with fixed time scheduled plan, displaying basic signage information for each programmed schedule. Controllers are then considered as autonomous.

In the "centralized plan selection mode", the central server and data transmissions are up. Plans are computed and controllers are synchronized to provide a basic signage based only on the model.

The "centralized traffic responsive mode" will have computed plans with synchronized controllers and sensor data handled to provide a full adaptive system. It also includes an enhanced security mode to manage security intervention of paramedics, police or fire men in case of serious unexpected events (illness, fire, robbery, etc.).

The last level is the real bartering platform. It includes commercial services like advertising, external high level information (number or remaining seats in a restaurant, extra waiting room in an airport terminal, etc.) and paid services for end users.

Digital signage systems can also work together to contribute to a large network. Specific time slots on the displays can be sold off to different partners, for example via auctions. This concept is known as a digital sign exchange [24][25].

It is also interesting to integrate smartphones as part of digital signage. Information displayed on these devices should be more specific depending on user profiles. Some information like pathway or localization could be free of charge, but some others must remain profitable like restaurant reservation or VIP services.

The whole system perimeter can be different depending on corporate users' wishes. The basic localization is the fair itself or an airport terminal. But it also can include peripheral areas that generate all incoming flows of people like cash register, passport control desks, car parks, all points of arrival.

Here, the key point will be about the zoning itself: each zone must have a complete consistence for human behavior. Car parks never generate the same crowd congestion as a passport control desk, for example.

We perceive MEDIACTIF as a crowd management system, however it must mainly be an engineering method: by understanding crowd phenomena, human behaviors and interactions, we can propose to equip an area with sensors and displays to generate the relevant information at the right place and time.

VII. CONCLUSION

The MEDIACTIF Project is at its very early stages of development. Surely some critical points must be solved first. Models must be setup to take care of the different usage scenarios considered so far (fairs, airports, etc.). We must handle people behaviors depending on tangible parameters (time, events, etc.) but also on unquantifiable parameters like personal interests.

Another point to solve is indoor localization. A multimodal approach is planned using physical sensors and networking (Wi-Fi) though we are also looking at other options based on mesh networks.

A first basic prototype with real experiment is planned at the beginning of year 2015.

 ISBN: 978-1-61208-330-8

ACKNOWLEDGMENT

ESIEE-Paris works in close cooperation with !nstantané (project leader), Aéroports de Paris, VIPARIS, Scilab Enterprises, INNES, PERTIMM, B2B EN-TRADE.

Some research leading to these results has received funds from the French Governement (DGCIS), Conseil Régional d'Ile de France and Conseil Général de Seine Saint-Denis for the FUI 16 program.

REFERENCES

[1] Mediactif, http://mediactif.esiee.fr/, [retrieved: March, 2014].

[2] J. V. Harrison and A. Andrusiewicz, "The digital signage exchange: a virtual marketplace for out-of-home digital advertising", Proceedings 4th ACM Conference on Electronic Commerce, June 2003, p. 274

[3] K. Wood, "urban traffic control : systems review", Transport Research Laboratory, Crowthorne, Berkshire, 1993

[4] B. M. Chard and C. J. Lines, "TRANSYT: The latest developments", Traffic Engineering and Control, vol. 28, 1987, pp. 387-390

[5] A. M. N. Mocofan, R. Ghiţă, V. R. T. López and F. C. Nemţanu, "Comparative Assessment of Traffic Control Parameters within an UTC-Distributed System", J. Transp. Eng., 10.1061/(ASCE)TE.1943-5436.0000634, vol. 140, March 2014.

[6] E. Franceries, "Centralized traffic management system as response to the effective realization of urban traffic fluency", Archives of Transport System Telematics, vol. 4, 2011, ISSN 1899-8208

[7] K. Wood and R. T. Baker, "Using SCOOT weightings to benefit strategic routes", Traffic Engineering and Control, 1992

[8] R. D. Bretherton and G. T. Bowen, "Recent enhancements to SCOOT - SCOOT Version 2.4", Proceedings 3rd International Conference on Road Traffic Control, IEE, May 1990

[9] J. J. Henry, J. L. Farges and J. Tuffal, "The PRODYN real time traffic algorithm", IFAC/IFIP/IFORS 4th Conference on Control in Transportation Systems, 1984

[10] S. P. Shepherd, "A review of traffic signal control", Institude of Transport Studies, University of Leeds, Working paper 349, January 1992.

[11] "Digital signage: the right information in all the right places", http://www.itu.int/oth/T2301000015/en, [retrieved: March, 2014].

[12] "The Popularity of SMIL in Digital Signage", http://www.iadea.com/technology/smil, [retrieved: March, 2014].

[13] "POPAI Digital Signage Content Standards", http://popai.com/docs/DS/ScreenFormat%20Standards%20Draft%20rev097.pdf, [retrieved: March, 2014].

[14] L. F. Henderson, "The statistics of crowd fluids", Nature, vol. 229, 1971, pp. 381-383.

[15] D. Helbing and P. Molnár, "Social force model for pedestrian dynamics", Phys. Rev. E, vol. 51, 1995, pp. 4282-4286.

[16] V.J. Blue and J. L. Adler, "Cellular automata microsimulation of bi-directional pedestrian flows", Transportation Research Board, vol. 1678, 1999, pp. 135-141.

[17] M. Fukui and Y. Ishibashi, "Jamming transition in cellular automaton models for pedestrians on passageway", Journal of the Physical Society of Japan, vol. 68, 1999, p. 3738.

[18] P. G. Gipps and B. Marksjö, "A micro-simulation model for pedestrian flows", Mathematics and Computers in Simulation, vol. 27, 1985, pp. 95-105.

[19] B. Maury, A. Roudneff-Chupin, F. Santambrogio and J. Venel, "Handling congestion in crowd motion modeling", Networks and heterogeneous media, American Institute of Mathematical Science, vol. 6, no. 3, September 2011

[20] B. Maury and J. Venel, "A discrete contact model for crowd motion", ESAIM: Mathematical Modelling and Numerical Analysis, vol. 45, January 2011, pp. 145-168.

[21] B. Maury, A. Roudneff-Chupin and F. Santambrogio, "A macroscopic crowd motion model of gradient flow type", Mathematical Models and Methods in Applied Sciences, vol. 20, 2010, pp. 1787-1821.

[22] Y. Gu, A. Lo and I. Niemegeers, "A survey of indoor positioning systems for wireless personal networks", IEEE Commun. Surveys & Tutorials, vol. 11, no. 1, pp. 13-32, 1Q 2009

[23] F. Barcelo-Arroyo, M. Ciurana and I. Martin-Escalona, "Process and system for calculating distances between wireless nodes", U.S. Patent 8 289 963, October 2012.

[24] J. V. Harrison and A. Andrusiewicz, "The digital signage exchange: a virtual marketplace for out-of-home digital advertising", Proceedings 4th ACM Conference on Electronic Commerce, June 2003, p. 274, ACM 2003, ISBN 1-58113-679-X.

[25] J. V. Harrison and A. Andrusiewicz, "An Emerging Marketplace for Digital Advertising Based on Amalgamated Digital Signage Networks", IEEE International Conference on Electronic Commerce (CEC 2003), June 2003, ISBN 0-7695-1969-5.

[26] Scilab, http://www.scilab.org/, [retrieved: February, 2014].

 ISBN: 978-1-61208-330-8

System Design Artifacts for Resilient Identification and Authentication Infrastructures

Diego Kreutz, Oleksandr Malichevskyy
LaSIGE/FCUL, Portugal
{kreutz, olexmal}@lasige.di.fc.ul.pt

Eduardo Feitosa, Kaio R. S. Barbosa and Hugo Cunha
IComp/UFAM, Manaus, Brazil
{efeitosa, kaiorafael, hugo.cunha}@icomp.ufam.edu.br

***Abstract*—The correct and continuous operation of identity providers and access control services is critical for new generations of networks and online systems, such as virtualized networks and on-demand services of large-scale distributed systems. In this paper, we propose and describe a functional architecture and system design artifacts for prototyping fault- and intrusion-tolerant identification and authentication services. The feasibility and applicability of the proposed elements are evaluated by using two distinct prototypes. Our results and analysis show that building and deploying resilient and reliable infrastructure services is an achievable goal through a set of system design artifacts based on well-established concepts from security and dependability. We also provide a performance evaluation of our resilient RADIUS service compared with the long standing FreeRADIUS.**

Keywords*—*System design; fault and intrusion tolerance; identification and authentication services; network access control.

I. Introduction

The growth of Authentication and Authorization Infrastructure (AAI) services is motivated by the fact that users are allowed to transparently access different services (e.g., Facebook, Google, Twitter, and Amazon) with a single credential or authentication session. These services rely on Identity Providers (IdPs) or Authentication, Authorization, and Accounting (AAA) protocols to identify and authenticate the user before granting him access to the requested resources or services. OpenID [1] and RADIUS [2] are examples of such services.

Despite the importance of AAIs for service infrastructures such as clouds and virtual networks, there are still open questions regarding their availability and reliability. This can be supported by recent work showing that digital attacks and data breach incidents are growing [3]. Additionally, advanced persistent threats [4] are becoming one of the top priorities of security specialists. Therefore, security and dependability properties should be the top priority of future AAIs.

Most of the existing RADIUS-based services and OpenID-based IdPs do not completely address security and dependability properties such as confidentiality, integrity, and availability. This can be observed on the services' online documentation and deployment recommendations [5][6][7][8][9]. Some implementations and deployments provide basic mechanisms to improve the service's reliability and robustness, such as SSL communications and simple replication schemes to avoid eavesdropping and tolerate stop failures, respectively. Hence, there are opportunities for further research with the ultimate goal of designing more resilient solutions which are able to deal with new threats and cyber attacks.

To the best of our knowledge, this paper proposes the first set of system design artifacts and functional architecture to design and deploy fault- and intrusion-tolerant identification and authentication services. Two distinct prototypes, RADIUS and OpenID, are used as proof of concept to demonstrate the applicability of the functional architecture and system design artifacts. We also briefly, we briefly discuss components and essential building blocks for implementing more robust and secure services. We conclude with an analysis of the approaches and requirements for deploying resilient services.

The next section introduces the motivation of our work. In addition, the functional elements and system design artifacts to develop robust and reliable AAI services are described in Section IV. Thereafter, the results are analyzed and discussed in section V. Lastly, Sections III and VI comprise of the related work and final remarks.

II. Motivation

AAI solutions are often based on protocols like OpenID and RADIUS. However, both of them are not designed with features for robust security (e.g., strong confidentiality) and dependability (e.g., high availability), being frequent targets of attacks and data theft attempts (e.g., user credentials).

OpenID is a framework to build identity providers [1]. It is based on open Hypertext Transfer Protocol (HTTP) standards, which are used to describe how users can authenticate on third party services through their own IdP. There are two main advantages of this approach. First, it allows identification and authentication protocols to be transported over standard Web protocols. Second, users need only one single credential to access different services provided that service providers accept external IdPs.

In spite of allowing the user to have a single credential to access multiple domains, there are different security issues on the OpenID identification scheme and service availability. Recent research has shown that the discovery and authentication steps are vulnerable to cross-site request forgery attacks, phishing and man-in-the-middle [10]. Also, its availability is vulnerable to denial of service attacks and protocol handling parameters [11][12].

RADIUS is an AAA protocol. The authentication verifies user's identity prior to granting access to the network or service. Authorization is used to determine which actions a user can perform after a successful authentication. Accounting provide methods for collecting data about the network or service usage. Collected data can be used for billing, reporting and traffic accounting. Therefore, RADIUS is commonly used to provide AAA features for infrastructures such as corporate networks and carrier grade provider networks.

The main security issues of RADIUS are in the protocol

 ISBN: 978-1-61208-330-8

specification and poor implementations [13]. Regarding flaws in the protocol, RADIUS does not validate the integrity of some packages (`Access-Request`) and does not provides mechanisms against reflection attacks. As well as this, existing implementations of RADIUS are also susceptible to dictionary, man-in-the-middle and spoofing attacks.

Fault and intrusion tolerance. We can choose two different approaches when designing secure and resilient systems. First, one can assume that it is possible to build robust and secure enough systems. However, as it is well known, a system is as secure as its weakest link. Moreover, it can be considered as secure until it gets compromised. Hence, the second approach is to assume that eventually the system will fail or be intruded. With this in mind, one can design highly available and reliable systems by leveraging mechanisms and techniques capable of enabling them to operate under adversary circumstances, such as non-intentional failures and attacks.

Our system design artifacts and functional architecture support the second approach, i.e, we do not intend to solve all security and dependability problems of AAI's services. Yet, by taking advantage of advanced techniques and resources from different domains we can build fault- and intrusion-tolerant systems capable of ensuring essential properties such as integrity, confidentiality, availability and reliability.

III. Related Work

Despite the existence of different solutions and components that can be used to improve the security and dependability of AAI services, such as advanced replication techniques, virtualization, proactive and reactive recovery techniques and secure components, there are no methodologies, functional architectures or a set of system design artifacts that are capable of demonstrating how different elements can be orchestrated to build highly available and reliable systems. Existing approaches and solutions are designed for specific scenarios or a particular system. One example is to use TPMs to create trustworthy identity management systems [14]. While the solution allows one to create trustworthy mechanisms for issuing tickets, it is not designed for high availability or fault and intrusion tolerance. Another example is a cooperative coordination infrastructure for Web services [15], which proposes security mechanisms that allow reliable coordination of services even in the presence of malicious components. It works as an integration infrastructure for Web services, being supported by a set of service gateways and one resilient tuple space. This solution offers fault and intrusion tolerance capabilities for the coordination infrastructure. However, it does not represent a generic or adaptable architecture that can be applied to improve the robustness and security of different systems. Furthermore, it only protects the data confidentiality of a particular service through authentication and encryption mechanisms. Therefore, such scenarios indicate the need of more general architectures and system design artifacts that can be combined in a more systematic way to develop and deploy services with higher security and dependability properties.

IV. System Design Artifacts

A. Overview of functional elements

Figure 1 shows a simplified representation of the four main functional elements: (a) client; (b) service; (c) gateway; and (d) Critical Infrastructure Service (CIS). This is the typical functional architecture of computing environments where identification and authentication solutions are separated services. In addition, the fifth element is a secure component, which can be used in conjunction with any of the previously mentioned elements. Its purpose is to provide additional support for ensuring properties such as confidentiality, integrity, and timing, when ever required.

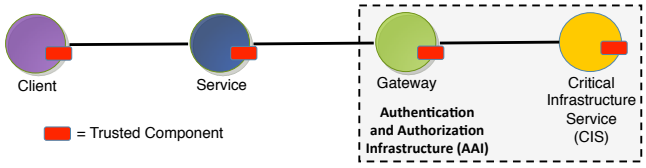


Fig. 1. Main functional elements.

A client can be a user trying to access the network or a networking element. In other words, it represents a generic element whose definition depends on the target scenario.

In a typical network, the service may represent elements such as wireless routers or Ethernet access switches. For Web applications based on OpenID, a service can be a relying party or an access control subsystem of online shopping Web sites.

A gateway provides connection between the service and the critical infrastructure service, a.k.a., back-end service. It has two basic functions. First, it handles multiple protocols from both sides, acting as a protocol gateway. The second function of this element is to mask the replication protocols and mechanisms used to deploy resilient back-end services, providing transparent backward compatibility with AAI protocols such as RADIUS and OpenID.

The back-end service is the critical element of the infrastructure, which is assumed to require higher levels of security and dependability assurances. A CIS can be part of the local domain or provided by third parties as an on-demand service, for instance. It is assumed that these services must tolerate different types of faults such as those caused by unexpected behavior or attacks, and work correctly in case of intrusions. OpenID providers and RADIUS services are examples of critical AAI systems for networked infrastructures and online services. Therefore, failures on these services can potentially impact a corporation's systems and business.

Lastly, secure components can help to ensure properties such as data confidentiality, integrity checks, and timing assurances to specific parts of the system. As an example, user keys can be stored on a smart card. Similarly, server keys and authentication tokens can also be securely stored in secure components. Furthermore, all critical cryptographic operations can be safely executed by these trusted components, without compromising sensitive data in the case of intrusions. Currently, server and self-signed Certificate Authority (CA) keys are stored in the server's file systems. Hence, any system administrator has easy access to them, representing potential security threats.

B. Design artifacts for resiliency

Figure 2 illustrates architectural elements with added system design artifacts for augmented security and dependability properties. The architecture allows different fault thresholds or

 ISBN: 978-1-61208-330-8

replication techniques in a per element basis. For instance, the service, gateway and CIS elements can have distinct characteristics when ever they need to ensure specific reliability and availability requirements, such as crash faults, arbitrary faults, or resist to resource depletion attacks.

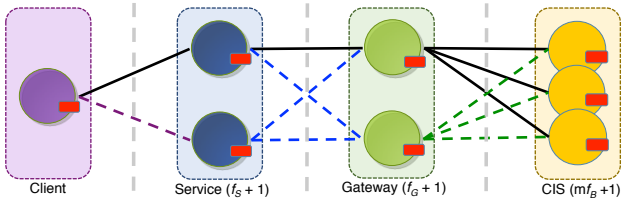


Fig. 2. Architectural components overview.

We assume that the service and gateway elements are designed to tolerate crash faults and detect some specific problems such as message integrity or authenticity violations while tolerating up to f_S and f_G simultaneous failures, respectively. Secure components can be used to verity message integrity and authenticity if sensitive data and procedures are required to accomplish the task. Otherwise, simple software-based verification methods can be sufficient to do the job.

Clients connect to any service replica, while a service can connect to any gateway. The connection to the replicas can be controlled using simple lists, which happens in AAA protocols, or round-robin for load balancing, for instance. However, in functional architecture there is no strict need for load balancing since the main goal is to provide fault tolerance. Hence, it is reasonable to assume that components are configured with at least a simple circular list of replicas.

On the other hand, the CIS does not support distinct methods to choose which replica to connect to. Gateways have to know all replicas required to support the number of faults assumed in the system. Using as an example a system that requires $2f + 1$ replicas to tolerate f faults, gateways need to known at least $2f + 1$ replicas to ensure that arbitrary failures on the CIS are going to be masked as long as $f + 1$ replicas are correct.

As a fault- and intrusion-tolerant infrastructure, the CIS is implemented with protocols to tolerate arbitrary faults. Gateways receive the responses from all (or at least enough to ensure a safe voting) back-end replicas and decide which one is the correct response that should be forwarded to the service or client. To achieve this goal the back-end service requires $mf + 1$ replicas, where m refers to the specific BFT algorithm [16] in use (e.g., $m = 2$, $m = 3$).

C. Essential building blocks

The main building blocks are technologies and components that make it possible to conceive resilient and more secure services based on the proposed functional architecture.

Virtual machines provide flexibility and agility to deploy systems and services. In highly resilient systems, mechanisms like proactive-reactive recovery and diversity can leverage functionalities provided by hypervisors, such as fast start, stop, suspend, resume and migration of virtual machines.

Replication protocols represent one of the major building blocks of resilient services. State machine replication and quorum protocols are common approaches to mask arbitrary faults on dependable systems. Replicas allow the system to tolerate up to f simultaneous faults without compromising its operation. Additionally, complementary ingredients for high availability and robustness are proactive and reactive recovery mechanisms and techniques to increase the diversity of the system [17].

Secure components are small and reliable pieces of software, and/or hardware, capable of ensuring critical properties or functions of the system, such as integrity control, timing and data confidentiality. They can be used in different parts of the functional architecture. For instance, in an OpenID-based authentication solution both end user and the server can trust their sensitive data (e.g., certificate, keys) and crypto functions to a trusted component. Hence, a compromised server will not leak confidential data like private keys or user's tokens.

Secure end-to-end communication. It is necessary to achieve confidentiality and privacy of user data. Protocols such as Transport Layer Security (TLS) and Tunneled Transport Layer Security (TTLS) can be leveraged to provide reliable channels, mutual authentication and server authenticity verification. These functions can be helpful to avoid attacks like man-in-the-middle and eavesdropping.

D. Requirements and components

Table I shows the properties and requirements for designing and deploying services with different levels of resiliency and trustworthiness. We use the notion of trust to indicate whether the system is capable of ensuring data confidentiality of sensitive data such as private keys. Most of the existing identification and authentication services belong to the first three classifications, where "−−" means only primary-backup replication to tolerate crashes of the master server. In other words, those services are less secure and not highly resilient. For instance, an attacker can sequentially compromise all servers since there are no advanced recovery mechanisms in place, such as proactive recovery, to ensure the system's liveness and reliability.

TABLE I
SERVICE PROPERTIES AND REQUIREMENTS/COMPONENTS.

Properties	Secure comp.	Replication protocol	Recovery mechanism	Wormhole model	Intrusion-tolerant
1. Untrusted	no	no	no	no	no
2. Trusted but not resilient	yes	no	no	no	no
3. Resilient (−−) but not trusted	no	yes	no	no	no
4. Resilient but not trusted	no	yes	yes	no	no
5. Resilient but not trusted	no	yes	yes	yes	no
6. Resilient and trusted	yes	yes	yes	yes	yes
7. Resilient and trusted (++)	yes++	yes	yes	yes	yes

Our system design artifacts and functional architecture are expected to contribute to the development of services with properties of classes 4 to 7, where "++" indicates multiple verification points (e.g., the correctness and authenticity of a message could be verified within any element of the functional architecture, potentially reducing the request-response time by taking further actions as soon as possible). Property 4 does not use the wormhole model [18], while 5 does. This model

 ISBN: 978-1-61208-330-8

proposed hybrid distributed systems comprised of two parts: the payload (main parts and functions of the system) and a tiny subsystem with stronger properties for ensuring minimal timing requirements of the system, such the synchrony those needed to ensure the finalization of consensus protocols. It means that the former can not support an asynchronous system since there is no way to assure that consensus protocols, which are one of the basic building blocks of replication protocols, will finish their execution.

A system of class 5 is not intrusion-tolerant because it does not use secure elements to ensure data confidentiality. This is precisely one of the state machine replication and BFT limitations. These protocols are designed to ensure integrity and availability, but not data confidentiality. Therefore, additional mechanisms, such as secure components, are required to ensure the system's sensitive data confidentiality. Furthermore, the worm hole model is required to ensure minimal synchrony requirements of consensus protocols if the system if assumed to be asynchronous.

Lastly, systems of type 6 and 7 need a wormhole to rely on specially designed trusted components to ensure the minimal and critical system properties, which are required if the system is assumed to be asynchronous. However, if the system is synchronous or partially synchronous, then a wormhole is not required for timing purposes, for instance.

Another difference between systems of type 6 and 7 lies in the more extensive use of secure components. While in type 6, trusted components are used only in the client and back-end service, a system of type 7 requires secure components in other elements as well, such as gateways and services. One use of these additional secure components, on different architectural elements, is to safely earlier detect corrupted messages.

E. Deployment scenarios

Figure 3 shows the main trade-offs of service deployments. Despite the performance gains when using shared memory replication solutions such as Intrusion Tolerance based on Virtual Machines (ITVM), which uses a single physical machine and shared memory for communication purposes between virtual machines [19], these services can suffer from resource depletion attacks and be affected by infrastructure incidents [17]. Nevertheless, depending on the needs and requirements of the target environment, an ITVM-based system can be the most adequate solution. On the other hand, resilient systems using techniques provided by replication frameworks, such as BFT-SMaRt [20], allows us to create more robust services which are capable of achieving higher degrees of availability through multiple physical machines and/or multiple domains. A physically distributed system can leverage the resources and defense mechanisms of multiple domains, but with the burden of lower overall performance. Therefore, one can conclude that there is no unique solution to all problems. The mechanisms and protocols to be employed have to be analyzed and chosen based on requirements and guarantees needed by the target environment. Only with these inputs can one decide which are the best system artifacts for building a particular solution.

Resilient services using distributed machines across multiple domains (e.g., distinct clouds) are capable of tolerating physical hazards of machines and domains (e.g., network connection problems, energy outages and disk failures) as well as logical problems (e.g., misconfigurations of systems/networks and resource depletion attacks) by leveraging the support offered by each infrastructure. In practical terms, it has already been shown that cloud providers can tolerate DDoS attacks of big proportions without incurring losses for customers [21]. One of the resources against this kind of attack are the geographically distributed data centers, which together can form a robust and diversified infrastructure.

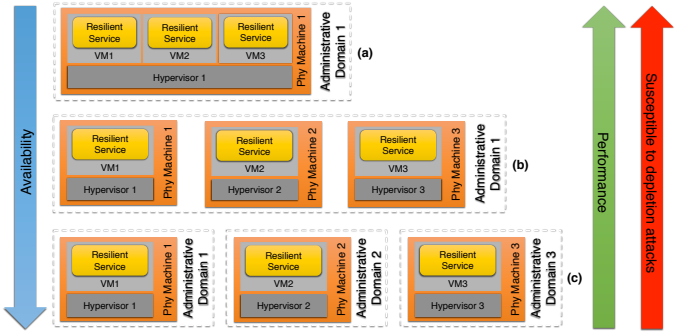


Fig. 3. Service deployment configurations.

V. Results and Discussion

A. RADIUS and OpenID prototypes

To evaluate the functional architecture and system design artifacts, two fault- and intrusion-tolerant prototypes were developed, one an OpenID provider and another, a RADIUS server. Our prototypes use the BFT-SMaRt [20], an open source and free Java-based library that provides high performance Byzantine Fault-Tolerant (BFT).

BFT-SMaRt can be leveraged to deliver a resilient service architecture that requires higher levels of availability and a lower probability of faults due to depletion attacks (e.g., performance or functionality degradation caused by a resource consumption racing or exhaustion). In addition, the state machine replication framework allows us to deploy replicas in a single physical machine, in multiple physical machines, or in multiple physical machines spread throughout different domains (e.g., multiple clouds). Consequently, replicas need to send and receive data over reliable and authentication channels, as is the case in BFT-SMaRt, to avoid attacks such as eavesdropping and man-in-the-middle.

The two prototypes follow the current standards of OpenID and RADIUS protocols, respectively. This means that any application based on these protocols will work with the resilient and more secure version of the service without requiring any modification. Hence, OpenID providers can offer to their users more reliable and secure services, which are able to support faults and intrusions in a smooth and transparent way. Due to space constraints, we only briefly introduce the OpenID BFT prototype in the following section. It is worth mentioning that most of the system design and implementation aspects apply to both RADIUS BFT and OpenID BFT.

B. OpenID BFT implementation

Figure 4 gives an overview of the OpenID BFT, which is based on the proposed functional architecture and system

 ISBN: 978-1-61208-330-8

design artifacts. Dashed lines indicate alternative paths used in case of failures of the default paths (solid lines). The timeouts are used to detect faults. While timeouts A and B and default properties of the functional architecture, timeout C is a specific requirement of OpenID.

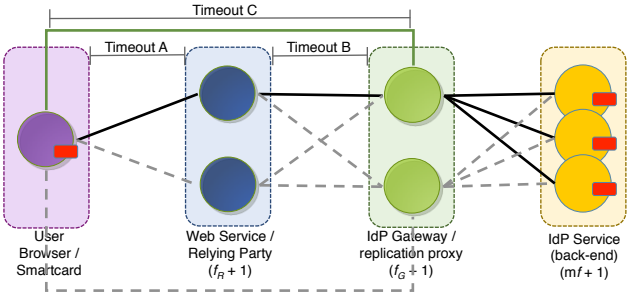


Fig. 4. Overview of the OpenID-BFT functional architecture.

In the OpenID BFT architecture, a client uses a Web service (relying party) that redirects him to his respective OpenID provider (IdP gateway). The user's credentials are required in the identification process, done by the IdP service replicas.

We have implemented the OpenID prototype with the *openid4java* library [22] (version 0.9.8), which supports OpenID 1.0 and 2.0. Our implementation defines OpenID 2.0 as the default authentication scheme. It is a fully fledged identity provider which is capable of tolerating up to f arbitrary replica failures. Moreover, secure elements are employed on the client side and authentication server to ensure the confidentiality of sensitive data such as user and server keys, self-signed CA's private key and session tokens. More information regarding the system's building blocks, design and implementation choices (e.g., state machine replication and secure elements) can be found in [17].

C. Fault thresholds and detection mechanisms

RADIUS BFT and OpenID BFT use the BFT-SMaRt framework, requiring $3f + 1$ replicas in the CIS to tolerate up to f faults or intrusion. We have introduced a third, yet fictitious, prototype called SM Service, where SM stands for shared memory. Its purpose is to describe the requirements of an ITVM like solution, which uses virtual machines and shared memory to tolerate faults and intrusions [17]. SM Service requires $2f + 1$ replicas to tolerate up to f arbitrary faults. However, it works with the best case, i.e., only $1f + 1$ active replicas. When ever the consensus protocol cannot finish due to differences in replicas' responses, the remaining replicas (f) are awakened and used to reach an agreement. One of the main disadvantages of this approach is the fact that it runs only on a single machine, i.e., its availability and operation can be affected by resource exhaustion attacks and infrastructure breakdowns. When high availability is required, or the system is subject to depletion attacks, solutions such as OpenID BFT and RADIUS BFT are needed to address those challenges.

Table II summarizes the fault model and fault threshold of the main component of the architecture. It also identifies example of components in real environments, such as AAA and OpenID infrastructures. As can be observed, it is assumed that services and gateways have a fail-stop (crash) behavior. Nevertheless, they can also detect some arbitrary behavior, such as malformed packets and corrupted messages, as is the case of the gateway element of our resilient RADIUS service. Another interesting potential use case are software defined networks, which still lack security and dependability mechanisms and protocols [24].

Services and gateways support fail-stop and a sub-set of arbitrary faults. Figure 5 shows the fault detection mechanisms. Only abnormal behavior like message corruption or forgering can be detected by clients and services. Yet, gateways are capable of detecting and masking arbitrary faults of the CIS element. Lastly, the CIS tolerates intrusions in up to f replicas once those events can be treated as arbitrary faults.

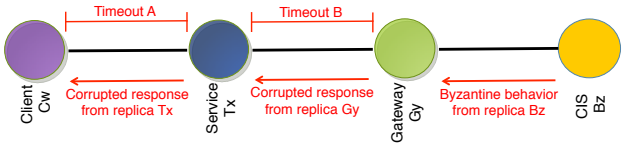


Fig. 5. Fault detection mechanisms.

D. A secure TLS component

Table III summarizes the methods required to implement a secure TLS component. To ensure mutual end-to-end authentication in RADIUS and OpenID, secure components can be used both at the client and server side [25]. These components are key design pieces to ensure properties such as confidentiality, storing sensitive data (e.g., secret key and attributes) and executing critical operations on it in a safe way.

TABLE III
SECURE TLS COMPONENT INTERFACE.

Method	Input	Output
`generate-Random`	Random number `size` (in bytes).	`Random number` with a specific size.
`extract-PreMaster`	Client's `premaster secret`.	`True` if *premaster* is correctly deciphered, `false` otherwise.
`generate-Master`	Random numbers from client and server.	`True` if the *master* secret was generated, `false` otherwise.
`getServer-FinishMsg`	`Hash` of the *record stream*.	Finalization message of the server.

A secure TLS component needs to be designed to provide the required methods to execute a TLS handshake, in case, four methods are needed to accomplish this task. Any outside software component can invoke those methods to execute a handshake between a client and the authentication server. No sensitive data leaves the secure TLS component. For instance, the server's private key, which is required to decipher the premaster key sent by the client and to generate a master key, can only be used inside the secure component. Therefore, an intruder cannot compromise the confidentiality of the server.

We implemented a secure TLS component based on the methods specified in Table III. It is used by RADIUS BFT's CIS to ensure a secure and reliable mutual EAP-TLS authentication between the user (using a certificate) and the AAA authentication server. In the case of OpenID BFT, the secure element ensures the confidentiality of user credentials and server keys and session tokens.

 ISBN: 978-1-61208-330-8

TABLE II
FAULT MODEL, FAULT THRESHOLD AND EXAMPLE OF COMPONENTS.

Component	Fault model	Replicas	Faults	Example of real use case components
Client	—	—	—	End user, network device
Service	Crash	$f_S + 1$	f_S	WiFi router, network management services, OpenFlow switch
Gateway	Crash	$f_G + 1$	f_G	New element interfacing with the target service and CIS
CIS	Byzantine	$mf_B + 1$	f_B	RADIUS AAA server, OpenID server, NIB of an OpenFlow controller [23]
Secure component	Crash	1	0	TLS keys and cryptographic methods

E. Properties, characteristics and performance

In Table IV, we sum up the main properties and characteristics of our prototypes. As can be observed, they are similar for OpenID BFT and RADIUS BFT because both of them leverage the same kind of architectural elements, replication mechanisms and secure components. Again, we have the SM Service for simple comparison purposes.

TABLE IV
PROTOTYPES' PROPERTIES AND CHARACTERISTICS.

Property/support	OpenID BFT	RADIUS BFT	SM Service
1. Multiple physical machines	yes	yes	no
2. Trusted components	yes	yes	yes
3. Hypervisor is trusted and secure	no	no	yes
4. Depletion attacks susceptibility	low	low	moderate
5. Performance (operations/s)	moderate	moderate	high
6. Availability guarantees	high	high	moderate
7. Arbitrary faults tolerance	yes	yes	yes
8. Intrusion tolerance	yes	yes	yes

Susceptibility to depletion attacks is intrinsically related to virtual machines using the same hypervisor. It is moderate to high on solutions based on a single hypervisor because a depletion attack can compromise the performance of non-malicious virtual machines in more than 50% of cases, depending on the specific attack. Examples of such resource exhaustion attacks and their impact on virtual machines of the Xen hypervisor can be found in [17].

Virtual machines on a single hardware platform can use shared memory spaces to execute protocols such as consensus [19], while frameworks such as BFT-SMaRt rely on message communication systems, whose performance depends on the specific algorithms being used and the corresponding implementation details. In OpenID BFT and RADIUS BFT, we use the BFT-SMaRt framework, which implements a set of optimization for state machine replication [16]. Therefore, we can consider its performance as moderate when compared to an ITVM-based solution, which is the best known performance setup for executing a state machine replication protocol. Moreover, other state machine replication implementations using non-optimized protocols would lead the system to lower performance measurements when compared to BFT-SMaRt.

It is well known that fault- and intrusion-tolerant mechanisms introduce some overhead in the system. To give an idea of the overhead, we measured our RADIUS BFT implementation and compared it with FreeRADIUS, which is a well-known and widely deployed implementation of RADIUS. The authentication latency increases by approximately an order of magnitude. Even though, it keeps below 200ms, which is an acceptable (non-perceptible by normal users) value for an authentication system. We also observed a drop in the throughput, i.e., number of authentications per second. In this case the difference narrows down to an overhead of about 5% to 40%, depending on the system's specific configurations and optimization. One of the reasons for the lower impact on the system's throughput is regarding the fact that BFT-SMaRt has a highly optimized batching subsystem for state machine replication, which is able to process a high volume of requests without impairing the system latency.

In summary, system design and development decisions should take into consideration the specific requirements of the target environment. While a solution like SM Service would be more suitable when depletion attacks are unlikely to happen, the hypervisor can be considered a trustworthy element and where high availability (e.g., support operation even under network disruption or other infrastructure disasters) is not a requirement. Yet, services such as OpenID BFT and RADIUS BFT are more indicated when high availability is required, performance is not the most critical issue, the hypervisor cannot be trusted, or when resource exhaustion attacks can happen. These sorts of solutions can resist to different kinds of threats or infrastructure incidents once replicas can be deployed in different physical machines and domains.

Resilient RADIUS versus FreeRADIUS

The throughput of the resilient RADIUS service and the FreeRADIUS server was measured using 2 to 20 simultaneous supplicants. Each client was configured to execute 10.000 sequential authentications using the same credentials. Furthermore, each authentication requires exactly ten packets, which needs to be considered when calculating the number of authentications per seconds. Therefore, we used a C program to measure the number of packets cached by `tcpdump`.

As shown in Figure 6, the throughput of the resilient RADIUS remains almost stable, while it varies for `FreeRADIUS`. This variation is due to the dynamic pool of threads, which automatically increases the number of working threads based on the number of authentication requests. Thus, it causes a slightly decrease in performance from 4 to 10 simultaneous clients. Afterwards, by activating new threads, the system's performance goes up again. The FreeRADIUS was configured with a minimum of 3 active threads and a maximum of 30 threads.

 ISBN: 978-1-61208-330-8

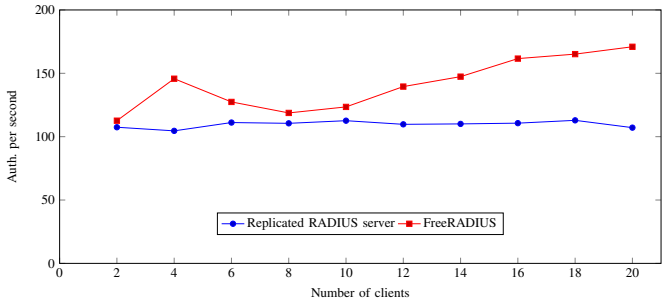


Fig. 6. Resilient RADIUS and FreeRADIUS throughput.

The stable throughput of the replicated RADIUS can be attributed to the gateway's throughput and the AAA replication. It sequentializes the network access servers' requests, sends each request to the replicas and clocks waiting for an answer. A thread pool, similarly to FreeRADIUS, could be used to increase the gateway's throughput. Moreover, the replicated RADIUS server also poses some limits to the system performance since requests must be acknowledged and ordered among all replicas. Nevertheless, the BFT-SMaRt system has a highly optimized batching and parallelization sub-system capable of sustaining high throughputs with higher number of clients (e.g., thousands) [20].

On the other hand, the latency almost doubles, going from nearly 100ms for FreeRADIUS to almost 200ms for the resilient RADIUS. The main problem of the latency lies on the gateway and the trusted component, due to their implementations. Both sub-systems could be re-designed to better explore concurrency and/or parallelism.

VI. Conclusion

This paper presented the first functional architecture and system design artifacts for designing and deploying more robust and reliable identification and authentication services such as OpenID and RADIUS. We believe that this is an important step for developing more systemic countermeasures against new security threats. Our results and evaluations indicate that we are able to build resilient and more secure identification and authentication infrastructures by combining important mechanisms and techniques from security and dependability.

We discussed how a functional architecture can be combined with system design artifacts to build different fault- and intrusion-tolerant services. The same components were successfully applied to build two distinct prototypes.

Interestingly, we believe that the proposed functional architecture and system design artifacts can be also extended to different scenarios. Based on our previous work and observations [17], we could apply the same concepts and components to create secure and dependable control platforms of software defined networks [24], where the CIS is a consistent and fault tolerant distributed data store [23], and resilient event brokers for monitoring cloud infrastructures [26], for instance.

Acknowledgments

This work is supported by EU's 7th Framework Program (FP7), through the project SecFuNet (FP7-ICT-STREP-288349), and by CNPq/Brazil, through grants 590047/2011-6 and 202104/2012-5.

References

[1] D. Recordon and D. Reed, "OpenID 2.0: a platform for user-centric identity management," in 2nd Workshop on Digital IDM, 2006, pp. 11–16.

[2] C. Rigney, S. Willens, A. Rubens, and W. Simpson, "RFC 2865 - Remote Authentication Dial In User Service (RADIUS)," 2000.

[3] Verizon RISK Team, "Data breach investigations report," Verizon, Tech. Rep., 2013, goo.gl/7mIBy, [retrieved: March, 2014].

[4] C. Tankard, "Advanced persistent threats and how to monitor and deter them," Network Security, vol. 2011, no. 8, 2011.

[5] FreeRADIUS, "Documentation," 2012, goo.gl/6g8Qy, [retrieved: March, 2014].

[6] RADIUS Partnerships, "Deploying RADIUS: Practices and Principles for AAA solutions," 2008, goo.gl/fslu7, [retrieved: March, 2014].

[7] Juniper Networks, "Steel belted RADIUS carrier 7.0 administration and configuration guide," 2010, goo.gl/Y5b9k.

[8] OpenID, "OpenID community wiki," 2010, goo.gl/PCASy, [retrieved: March, 2014].

[9] Clamshell, "Clamshell: An OpenID Server," 2013, goo.gl/09pYF, [retrieved: March, 2014].

[10] S.-T. Sun, K. Hawkey, and K. Beznosov, "Systematically breaking and fixing openid security: Formal analysis, semi-automated empirical evaluation, and practical countermeasures," Computers & Security, pp. 465–483, 2012.

[11] B. van Delft and M. Oostdijk, "A security analysis of OpenID," in IFIP Advances in Information and Comm. Tech., vol. 343, 2010, pp. 73–84.

[12] M. Uruena, A. Munoz, and D. Larrabeiti, "Analysis of privacy vulnerabilities in single sign-on mechanisms for multimedia websites," Multimedia Tools and Applications, pp. 1–18, 2012.

[13] J. Feng, "Analysis, Implementation and Extensions of RADIUS Protocol," in Conference on Networking and Digital Society, 2009.

[14] A. Leicher, A. Schmidt, Y. Shah, and I. Cha, "Trusted computing enhanced OpenID," in ICITST, 2010, pp. 1–8.

[15] E. Alchieri, A. Bessani, and J. Fraga, "A dependable infrastructure for cooperative web services coordination," in ICWS, 2008, pp. 21–28.

[16] J. Sousa and A. Bessani, "From byzantine consensus to BFT state machine replication: A latency-optimal transformation," in EDCC, 2012, pp. 37–48.

[17] D. Kreutz, H. Niedermayer, E. Feitosa, J. da Silva Fraga, and O. Malichevskyy, "Architecture components for resilient networks," SecFuNet Consortium, Tech. Rep., 2013, http://goo.gl/xBHCNb, [retrieved: March, 2014].

[18] P. E. Veríssimo, "Travelling through wormholes: a new look at distributed systems models," SIGACT News, vol. 37, Mar. 2006.

[19] J. Lau, L. barreto, and J. da Silva Fraga, "An infrastructure based in virtualization for intrusion tolerant services," in ICWS, june 2012.

[20] A. Bessani, J. ao Sousa, and E. Alchieri, "State Machine Replication for the Masses with BFT-SMaRt," DI/FCUL, Tech. Rep., Dec. 2013, http://hdl.handle.net/10455/6897, [retrieved: March, 2014].

[21] M. Prince, "The DDoS That Almost Broke the Internet," 2013, goo.gl/g5Qs1, [retrieved: March, 2014].

[22] OpenID4Java, "OpenID 2.0 Java libraries," 2013, goo.gl/c3kFV, [retrieved: March, 2014].

[23] F. Botelho, F. M. V. Ramos, D. Kreutz, and A. Bessani, "On the feasibility of a consistent and fault-tolerant data store for SDNs," in Second European Workshop on Software Defined Networking, 2013.

[24] D. Kreutz, F. M. Ramos, and P. Verissimo, "Towards secure and dependable software-defined networks," in SIGCOMM HotSDN, 2013, pp. 55–60.

[25] P. Urien, E. Marie, and C. Kiennert, "An innovative solution for cloud computing authentication: Grids of EAP-TLS smart cards," in Fifth International Conference on Digital Telecommunications (ICDT), 2010, pp. 22–27.

[26] D. Kreutz, A. Casimiro, and M. Pasin, "A trustworthy and resilient event broker for monitoring cloud infrastructures," in IFIP DAIS, 2012, pp. 87–95.

 ISBN: 978-1-61208-330-8

Design and Implementation of an Interaopreable and Extednable Smart Home Semantic Architecture using Smart-M3 and SOA

Haitham S. Hamza, Enas Ashraf, Azza K. Nabih, Mahmoud M. Abdallah, Ahmed M. Gamaleldin
SECC-ITIDA, MCIT
Giza, Egypt
{hhamza, inas, azkamal, mmabdallah, agamal}@itida.gov.eg

Alfredo D'Elia
ARCES, Alma Mater Studiorum, Università di Bologna
Bologna, Italy
adelia@arces.unibo.it

Hadeal Ismail, Shourok Alaa, Kamilia Hosny, Aya Khattab, Ahmed Attallah
ANSR Lab, Cairo University
Giza, Egypt
{hismail, salaa, khosny, akhattab, aattallah}@ansr.cu.edu.eg

***Abstract* — Smart homes attempt to automate interaction with the environment and existing appliances to optimize resources while maintaining user convenience. *Interoperability* is a key feature, as well as a main challenge in developing smart systems. This is due to the wide interactions among appliances, users, and the surrounding environment, which are heterogeneous by nature. Moreover, smart systems may evolve over time by integrating new actors, devices, and applications, making *extendibility* another key challenge in designing such systems. Accordingly, developing a framework that that provides both interoperability and extendibility for smart systems is of a great interest in practice. Accordingly, this paper proposes a framework based on the semantic and service oriented architecture (SOA) technologies for smart homes. The proposed architecture makes use of open source SOA and Smart-M3 framework that provide the core technologies to enable interoperability and extendibility. Ontology is designed and used to enable semantic middleware for integration. The proposed framework is demonstrated using a simple smart home scenario.**

Keywords-Smart home; SOA; Semantic middleware; Smart-M3; SIB; Smart space; Ontology; Interoperability; Extendibility

I. INTRODUCTION

A smart environment is a context-aware environment that is able to interact with its inhabitants through autonomous devices embedded all around the physical world [1]. Three main keywords define smart environments: *Context Awareness*, *Interoperability*, and *Extendibility*.

Context Awareness is the ability of devices to be aware of the context and changes that take place within the environment [2]-[6]. Being aware is one aspect; however, being reactive to the expected changes is the main goal of context awareness.

Interoperability refers to the ability of the system ability to interact based on a common information language. The distributed and heterogeneous nature of smart environments requires interoperability support in many levels. The following three interoperability levels are identified in [7] and [8]: (1) device interoperability to enable communication among devices, (2) service interoperability to exploit the information originating from heterogeneous devices as services for various end-user applications, and (3) information interoperability across application domains. However, not all smart environments are able to ensure interoperability at these three levels.

Extendibility refers to the ability of the smart space to adapt to configuration changes such as adding, removing, and updating hardware or software parts of the environment.

Achieving interoperability and extendibility in smart home solutions have been demonstrated in the literature using Service Oriented Architecture (SOA) or semantic exist, but rarely both [9][10][21][22]. However, we argue that, to achieve better interoperability and extendibility at both the middleware and application layers, an integrated SOA-semantic architecture is needed. Accordingly, this paper proposes a solution that exploits both SOA (using open source SOA technology) and semantics (using Smart-M3 framework, to be elaborated later). The proposed solution makes use of ontologies to enable interoperability among the various devices to allow interoperability as well as extendibility at both the services and application levels. It is worth noting that the proposed architecture is demonstrated in the context of smart homes; however, its structure and operational concept can be adopted to develop services in various domains, such as healthcare, which have received an increasing interest in adapting smart environment solutions in recent years [16]-[20].

The rest of this paper is organized as follows. Section II summarizes related work. Section III presents the proposed architecture. The use of the proposed architecture in a simple smart home prototype is presented in section IV. Section V discussed how the proposed architecture provides extendibility and interoperability; Conclusions are given in Section VI.

II. RELATED WORK

Smart environments refer to any type of user environment (homes, offices, universities, etc.), where technology is utilized to provide the user with the maximum possible automation, convenience, safety, and comfort. There are common challenges exist in all smart environments including the ability to integrate a wide variety of different objects, with each of which having its own operating standards, programmability, sensing ability, etc. In addition, smart systems do need to evolve over time to meet

 ISBN: 978-1-61208-330-8

changing and increasing demands of their users and environments, and hence, extendibility is another key challenge that needs to be addressed in designing such systems.

In addition to the above, smart homes, and environments in general, need to provide the best services to its users while maintaining low deployment and operational costs.

Integrated SOA and semantic architectures/platforms provide good approach to tackle the above challenges. SOA technologies support interoperability and extendibility at the service and application layers; whereas semantics enables interoperability and extensibility at the devices (physical) layer. Research efforts that demonstrate the integration of both semantic and SOA technologies have been reported in various domains (e.g., [16][17]). However, most of these solutions have limited usability as they are tuned to tackle domain-specific issues and challenges.

In [18] and [19], a framework for wearable devices is proposed and demonstrated for use in monitoring sportsman/sportswoman. The framework is shown to be effective; however, it is optimized for use in this particular domain with limited applicability. For instance, services in this framework (sensor agents) are running on the sensor nodes themselves, which makes sensor powerful; but at the cost of an increased complexity. Such complexity (and hence, cost) makes this framework expensive to deploy in smart home applications, where several sensors with much lower complexity and cost are needed.

The framework proposed in the Chiron ARTEMIS project [20] makes use of the SOFIA smart space architecture [23]. A key feature in this framework is the registration of the Semantic Information Broker (SIB) as a service on the SOA. However, this design approach may result in an increased coupling that can affect the extendibility of the framework in dynamic environments with several changes.

Our proposed architecture, on the other hand, suggests an information and service interoperability platform that is highly de-coupled. This is achieved by adopting the concept of Enterprise Service Bus (ESB) to compose and publish services, having a single software component representing the low level services as a single integration point. This design approach would allow for the use of multiple semantic interoperability platforms (middleware), where each middleware is used for implementing and publishing the low level services keeping the high level ones on the ESB intact. Primitive services will be the ones accessing the data in the semantic repository; Semantic Information Broker (SIB) .

III. SYSTEM ARCHITECTURE

The proposed architecture aims at introducing a middleware that is flexible to be used with a wide range of scenarios and devices. To achieve this, interoperability and extendibility are the key design features that the developed middleware should address at both devices and application levels. Interoperability and extendibility are achieved at the devices and application layers by adopting, respectively, Smart-M3 and SOA technologies.

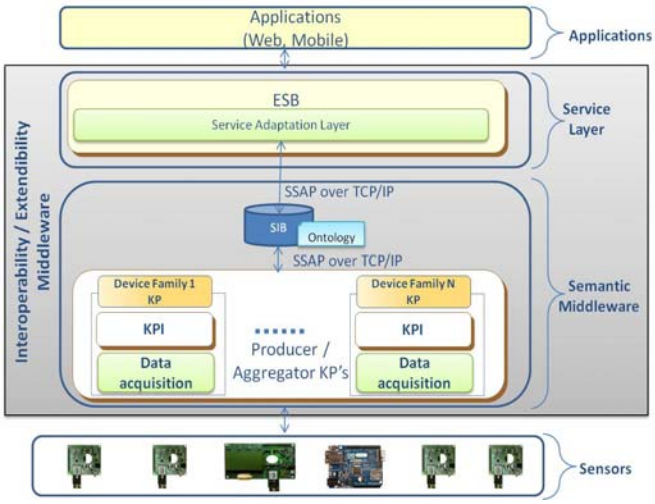


Figure 1. System Layered Architecture

Figure 1 depicts the proposed architecture. As show in the figure, the key components in the proposed middleware are: the SIB, the context ontology, and the SOA/ESB. The following subsections briefly explains these components.

A. *Smart-M3 Interoperability Platform*

Smart-M3 is an open-source software project that provides an infrastructure for sharing semantic information among software entities and devices. Smart-M3 deploys an information interoperability approach for devices to easily share and access local semantic information. In Smart-M3 the information is represented by using same mechanisms as in semantic web, thus allowing easy exchange of global and local information. Two main components make up the Smart-M3 infrastructure (Figure 2): SIB and Knowledge Processors (KPs).

1) Semantic Information Broker (SIB): The SIB is the access point for receiving information to be stored in the smart space or retrieving such stored information. The information is stored in the form of RDF graph (according to some defined ontology) in order to link data between different domains causing cross-domain interoperability to be much easier.

2) Knowledge Processors (KPs): KPs are the active parts of a Smart-M3 system. It provides the information (producer KP), modify and query it (consumer KP) and take actions based on the information seen in the SIB. This is done via a simple XML-based communication protocol called Smart Space Access Protocol (SSAP) that provides KPs with primitives to access SIB data (insert/add/remove/query/etc.).

B. *Context Ontology*

Ontologies are used to represent knowledge in machine understandable format. Ontologies represent the knowledge for describing certain domain. The context is all the information used to define the situation of an entity. As

 ISBN: 978-1-61208-330-8

detailed in [11], the context can be physical, computational or user context. To completely describe a context, the ontology used should be simple, complete, expressive, and evolvable. As a result, defining an ontology that can be extended with new concepts is a key factor. For example, if the ontology includes a concept for device that is associated with some data, it is a good practice to define a separate class for data that has many types that can be extended later (sensor data, identification data, etc.). However, to achieve better extendibility and interoperability, it is required not only to make ontology concepts extendible, but also to add the extendibility concepts to the ontology itself.

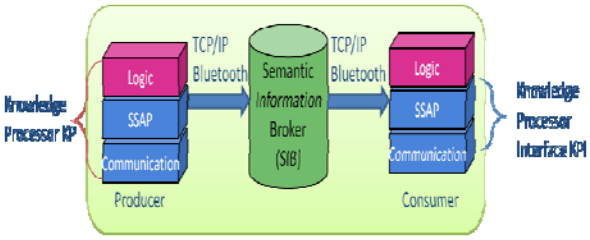


Figure 2. Smart-M3 Component Interaction

C. Service-Oriented Architecture

Service-oriented architecture (SOA) is a set of principles and methodologies for designing and developing software in the form of interoperable services. These services are usually representing the high level business functionalities that are built as software components, which can be reused for different purposes. Many techniques and approaches can be used to achieve this, e.g., OSGi framework [12], even SSAP protocol used in Smart-M3 can be considered as a SOA approach that provides low level services that can guarantee interoperability between devices compliant with the same approach [13]. In the proposed architecture, the ESB technology is used to achieve this goal. ESB helps achieving SOA goals through providing a flexible connectivity infrastructure for integrating applications and services.

IV. Demonstration of Smart Home Scenario

This section demonstrates the implementation of the proposed architecture (Figure 1) in a simple smart home scenario. The key features of the demonstrated smart home are: (1) providing instruments to edit and apply user preferences related to installed devices, (2) using low cost sensors to detect user presence and identity in order to configure the installed devices, (3) implementing a simple and complete web based interface to manage system behavior and to provide remote monitor and control, (4) detecting and reacting to anomalous situations such as faults and alarms, and notifying the user on their mobile devices, and (5) enabling maintenance companies to register in order to offer their services to fix faults and maintain devices installed in the smart home.

In this scenario, the system should consider the following key aspects: users should be identified; user preferences should be applied upon user identification, and faults are automatically detected and communicated to users via mobile and web applications. The main actors of this scenario are: (1) the administrator who defines the smart home and localize devices into smart home rooms, (2) the owner of the smart home, who is registered to the system using security credentials beside the RFID identification tags, and (3) the maintenance companies who will register to offer their maintenance services.

This scenario is realized using the proposed architecture, where Redland SIB and Java based KPs are used for the semantic middleware [14], and Mule standalone ESB v3.3 [15] installed on an Ubuntu LTS v12.04 are used to implement the SOA technology. Mule is chosen for its simplicity, and because it different web services protocols including Simple Object Access Protocol (SOAP) and Representational State Transfer (REST).

A. Hardware and Wireless Sensor Network

The WSN hardware layer at the lower level is realized through a network of ZigBee nodes. The ZigBee nodes can operate stand alone like the ones reporting the environmental status or attached to devices. In addition to the environmental node that should report information on ambient temperature, humidity, and light intensity level, a set of devices is selected to be demonstrated within this scenario in order to demonstrate the value of smart environments. Specifically, ZigBee nodes are attached to AC, refrigerator, multimedia board, and light. The ZigBee network has a single coordinator that is responsible for collecting data from the distributed nodes and sending via serial interface to the device hosting the dedicated KP. User identification is done via RFID module so that system can apply specific actions on the home devices as per a pre-set user preferences. The RFID information is sent via Ethernet connectivity to the device hosting the dedicated KP.

For extendibility on the level of devices, a self-exposure approach for device features is adopted. Each device sends an identification packet upon switch on. By doing so, new devices can be easily plugged into the home network allowing self-configurability of the system.

B. Smart Home Ontology

Several ontologies have been developed for smart home applications. For simplicity, and order to avoid any complexity induced by full-scale smart home ontology, we decided to develop our own ontology (See Figure 3). The used ontology is developed with extendibility in mind.

Figure 4 5 illustrate how a new device is defined in the knowledge base. Each device will send an identification packet that defines the device type, id, capabilities, capability parameters, measurements, and data ranges for measures and possible commands. Once received by the dedicated KP, the data is interpreted to the related ontology concepts.

C. KPs Design

The composition and distribution of KPs is one of the main considerations to promote interoperability and extendibility in the smart space. A proper design of information exchange mechanisms in the smart space is crucial. Modularity, de-coupling of information producers

 ISBN: 978-1-61208-330-8

and consumers, and providing minimal points of integration with hardware and SOA layers can even take extendibility to another level. Figure 6 illustrates how this design issue was handled in our architecture. The `Adapter` KP is responsible for receiving data from the hardware layer, interpreting the data, and inserting the interpreted semantic information into the smart space. This data includes device identification as well as context information. Inserted information can be then consumed by the `Fault Detector` KP and the `Brain` KP. The `Fault Detector` KP consumes and analyzes the smart space information to produce fault related information and insert it to the smart space. The `Brain` KP is triggered by the user presence and identification information from the smart space and uses these data to produce new commands to the smart space. The commands will be then consumed by the `Driver` KP to actuate the devices. In order to expose the semantic information to the service and application layer a Service Adaptation Layer is designed. This is a special KP that will reside on the ESB exposing the required information as low level web services that will be used later to compose higher level services based on application needs.

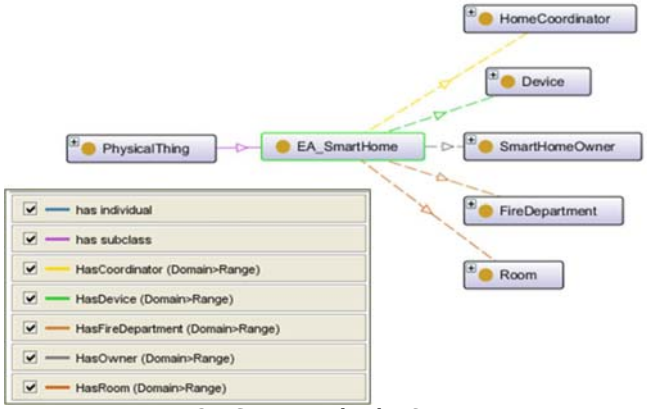


Figure 3. Concepts Related to Smart Home

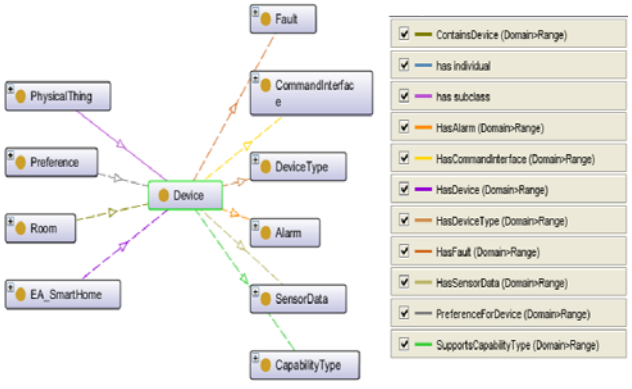


Figure 4. Device Ontological Definition

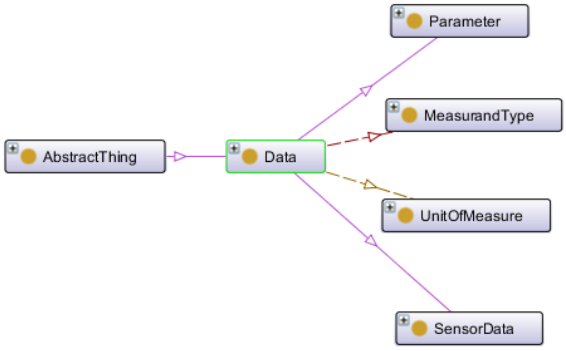


Figure 5. Types of Device Data

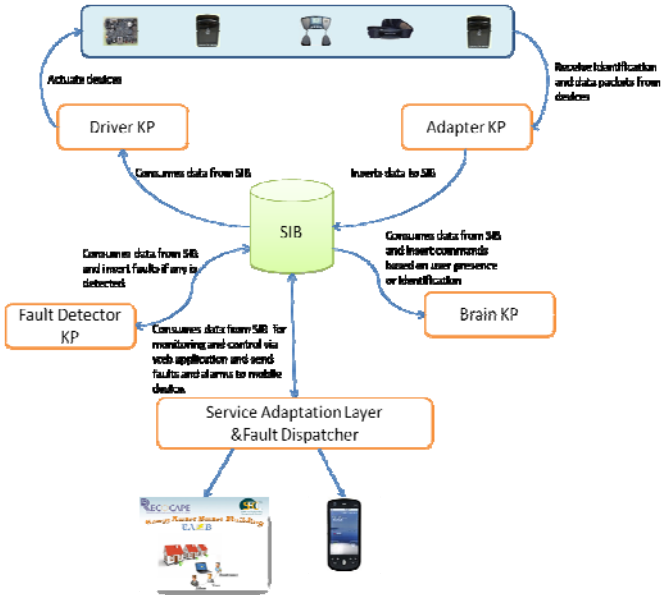


Figure 6. System Data Flow and Component Interaction

D. *Service and Application Layer*

At the higher level, services and user applications exist. Web and Android applications are designed hiding the lower details of the architecture and presenting their smart functionality through compositions of several services deployed on the ESB. This guarantees a complete isolation as the application can use a service to get ambient temperature in a smart home regardless of the underlying semantic middleware. Dynamic service composition techniques can even allow creating a new service at run time to extend the application features. By deploying the services in the ESB, our system can seamlessly integrate with external applications or services regardless of the communication protocol, message structure and implementation technology they are using. Furthermore, utilizing this service layer makes applications platform independent with applications can run on any user platform, without need to do this exhausting application porting task needed in traditional systems.

 ISBN: 978-1-61208-330-8

V. DISCUSSION ON INTEROPERABILITY AND EXTENDIBILITY

This section discusses how the proposed architecture enables interoperability and extendibility via the demonstration of how a new device can be seamlessly plugged into the smart home scenario discussed above.

The proposed architecture provides a systematic and easy approach to add new devices and services and to create various applications that can seamlessly interoperate with the existing environment. Clearly, no modifications are needed when a new instance of an existing device type is added to the system, as in such a case, simply the added instance will be plugged and operated in the system. Thus, the question is how the proposed architecture will seamlessly accommodate the addition of a new type of devices and their corresponding services? We will investigate this question by demonstrating the steps needed to allow the user to add a new service (mobile application) that allow her to monitor the energy consumption profile, and receive notifications for energy misuse based on a set of pre-defined preferences for energy control.

To realize this new service, the system, our smart home scenario needs to be extended with a new device, example, a smart meter to allow for the reading of the energy consumption and report these data to the system. Using our system, the sought smart meter will interoperate with the rest of the devices (already plugged and operating) by simply sending its identification packet and extending the ontology with the new concepts of the smart meters domain. In addition to extending the physical space, the application layer needs to be extended by adding a new set of services to the service pool. The proposed architecture enables this in a straightforward way as the use of ESB makes it easy to create this service from scratch or the consumption a ready-made services available already from other vendors. In the following, we summarize the steps needed to incorporate the smart meter into our system.

- Update the data set to be used in the device identification and data packets. As a result of having a generic packet format (Figure 7) the smart meter can be interoperated with the system by adding a new device type, types and ranges of the measurements read by the smart meter, and commands that the smart meter can receive (if any).
- Extend the ontology with a new device type, MeasurandType, and units. As shown in Figures 8 and 9. This can be achieved by simply adding a new set of individuals to the ontology without changing the main graph.
- For the KPs, we need to modify the data interpreter (message receiver) in the `Adapter` KP as well as the command dispatcher in the `Driver` KP. The message receiver and command dispatcher are two java modules independent of the semantic middleware. The two modules are only dependent on the used ontology. This implies that in case the Smart-M3 interoperability platform is changed, the same modules can be reused.

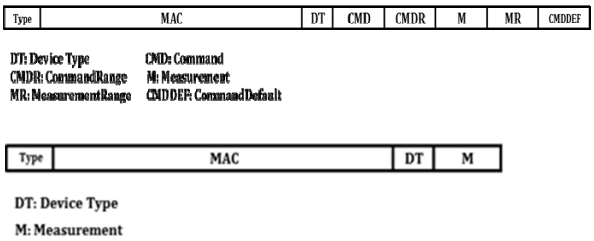


Figure 7. Top: Identification packet format (Type = 'I'), Bottom: Data packet format (Type = 'D')

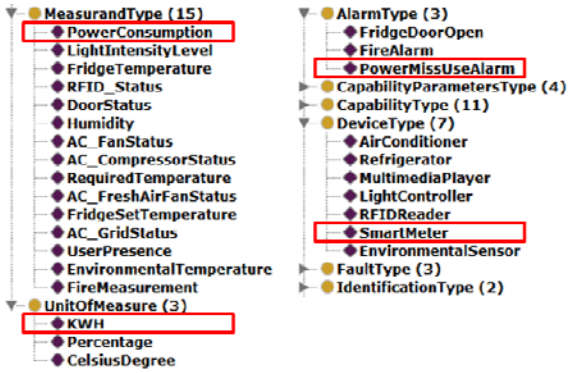


Figure 8. Adding a new device type

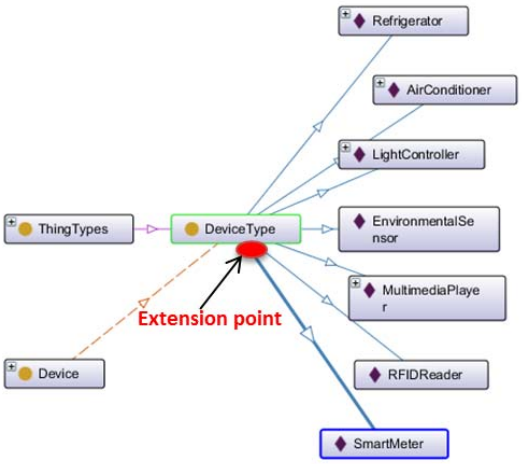


Figure 9. Adding a new device type

- Finally, we will expose the new features as services and deploy it to the ESB upon which new application features are built for both web and mobile applications. In this example two services are added. The first is to enable the user to remotely monitor power consumption, and the second is to notify the user of unexpected load patterns.

As illustrated above, deploying the proposed integration architecture can support system extendibility and interoperability by providing a modular design with reduced points of modification in the existing smart system.

Device interoperability and extendibility are achieved by allowing each device to self-expose its features and

 ISBN: 978-1-61208-330-8

capabilities through generic data packets to be interpreted into semantic information according to the defined ontology via the dedicated KP to feed the smart space. Using ESB as a deploymrnt environment for numourous services of different technologies eases the creation of new services and permits extending the application layer to new domains.

Moreover, the use of SOA/ESB supports *modularity* of services and applications by nature due to the loose coupling of services. Added to this, the proper design of data flow and component interaction in the semantic space is crucial. The used approach in the smart home scenario separates components that feed the smart space with information from those that consume the information. Moreover, data consumers that act on devices are de-coupled from data consumers exposing semantic information to the SOA layer.

VI. CONCLUSION

An effective design of smart homes (or environments, for this matter) needs to deal with various practical challenges in order to enable real value to the uses. Among these challenges are the *interoperability and extendibility* that arise due to the need for heterogeneous devices to be integrated and interoperate to deliver the required business value of the system. This paper proposes and demonstrated, via the real implementation of a simple smart home scenario, the use of the known semantic Smart-M3 and the well-adopted SOA/ESB technologies to develop an architecture that can enable interoperability and extendibility at both devices/information and services/applications layers. Smart-M3 supports interoperability at the information layer by utilizing semantic repository at its core, where information are stored in a standard representation based on domain specific ontology. Smart-M3 also is an enabler for system extendibility, with any components can be easily involved into the system under the condition that it uses same ontology. SOA/ESB provides the required support to realize interoperability at the service level, where applications running on any devices or based on any platforms or operating systems can access system features by the aid of these loosely coupled, platform independent services deployed in the service layer. It can extend the system functionality through composition of these services to support more complex scenarios and through deployment of new set of system services. With the ESB added, not only devices using the same semantic middleware will interoperate, but also those with different semantic middleware can also be integrated, which allows legacy systems to coexist and operate with new ones.

ACKNOWLEDGMENT

This work supported by the RECOCAPE project funded by the EU Commission under call FP7-INCO-2011-6.

REFERENCES

[1] C. A. N. Guerra, and F. S. C. da Silva, “A Middleware for Smart Environments,” Proc. of AISB Symposium on Intelligent Agents and Services for Smart Environments, 1-4 April 2008, vol. 8, pp. 22-24.

[2] J. Krumm, Ubiquitous Computing Fundamentals, Taylor and Francis Group, LLC, 2010.

[3] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of Wireless Indoor Positioning Techniqus and Systems," IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews, vol. 37, No. 6, November 2007, pp. 1067-1080.

[4] A. Flammini, P. Ferrari, D. Marioli, E. Sisinni, and A. Taroni, "Wired and wireless sensor networks for industrial applications," Microelectronics J.l, vol. 40, Issue 9, Sept 2009, pp. 1322-1336.

[5] S. Farahani, ZigBee Wireless Networks and Transceivers, Elsevier Ltd., 2008.

[6] R. Verdone, D. Dardari, G. Mazzini, and A. Conti, Wireless Sensor and Actuator Networks – Technologies, Analysis, and Design, Academic Press, 2008.

[7] L. Roffia, T. S. Cinotti, P. Azzoni, E. Ovaska, V. Nannini, and S. Mattarozzi, “ Smart-M3 and OSGi: The Interoperability Platform,” Computers and Communications (ISCC), 2010 IEEE Symposium, June 2010, pp. 1053-1058.

[8] J. Honkola, H. Laine, R. Brown, and O. Tyrkkö, “Smart-M3 Information Sharing Platform,” Computers and Communications (ISCC), 2010 IEEE Symposium, June 2010, pp. 1-6.

[9] L. Ngo , “Service-oriented architecture for home networks,” TKK T-110.5190 Seminar on Internetworking, 2007, pp. 1-6.

[10] T. Perumal, A. Ramli, C. Y. Leong, S. Mansor, and K. Samsudin, “Interoperability among Heterogeneous Systems in Smart Home Environment,“ SITIS '08. IEEE Int. Conf., Dec. 2008, pp. 177 - 186.

[11] E. Ovaska, L. Dobrica, A. Purhonen, and M. Jaakola, “Technologies for Autonomic Dependable Services Platform: Achievements and Future Challenges,” Software and Data Technologies Comm. in Computer and Information Science, vol. 303, 2013, pp. 199–214.

[12] OSGi™-The Dynamic Module System for Java™, http://www.osgi.org [retrieved: March 2014].

[13] Requirements for SOA Protocols, DIEM – Devices and Interoperability Ecosystem, BA – Building Automation, Aalto University – Media Technology .

[14] Smart-M3 Interoperability Platform, http://sourceforge.net/projects/smart-m3/ [retrieved: March 2014].

[15] Mule ESB, http://www.mulesoft.org/ [retrieved: March 2014].

[16] I. Corredor, J. Iglesias, A. M. Bernardos, and J. R. Casar, "A development methodology to facilitate the integration of Smart Spaces into the Web of Things," 4th Int. Workshop on Sensor Networks and Ambient Intelligence, 23 March 2012, pp. 829 – 834.

[17] V. Luukkala, D. Binnema, M. Börzsei, A. Corongiu, and P. Hyttinen, "Experiences in implementing a cross-domain use case by combining semantic and service level platforms," IEEE Symposium on Computers and Comm. (ISCC), 22-25 June 2010, pp. 1071 – 1076.

[18] J. Rodríguez-Molina, J. Martínez, P. Castillejo, and L. López, "Combining Wireless Sensor Networks and Semantic Middleware for an Internet of Things-Based Sportsman/Woman Monitoring Application," Sensors Jjournal, doi:10.3390/s130201787, 2013, pp. 1787-1835.

[19] P. Castillejo, J. Martínez, L. López, and G. Rubio, "An Internet of Things Approach for Managing Smart Services Provided by Wearable Devices," International Journal of Distributed Sensor Networks, vol. 2013, Article ID 190813, 2013, pp. 1-9.

[20] CHIRON EU project, ARTEMIS, http://www.chiron-project.eu/ [retrieved: March 2014].

[21] M. Kasza, V. Sz˝ucs, V. Bilicki, G. Antal, and A. B´anhalmi, "An Implementation Model for Managing Data and Service Semantics in Systems Integration," The Fifth International Conf. on Advances in Databases, Knowledge, and Data Applications, 2013, pp. 182-189.

[22] G. Pan, L. Zhang, Z. Wu, S. Li, L. T. Yang, M. Lin, Y. Shi, "Pervasive Service Bus: Smart SOA Infrastructure for Ambient Intelligence," IEEE Intelligent Systems, 20 Dec. 2012.

[23] http://www.sofia-community.org/ [retrieved: March 2014].

[24] http://www.secc.org.eg/RECOCAPE/ [retrieved: March 2014]

 ISBN: 978-1-61208-330-8

Theoretical Suggestion of Policy-based Wide Area Network Management System

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

Shogo Shimizu
Gakushuin Women's College
Tokyo, Japan
shimizu-syogo@aiit.ac.jp

Naohiro Ishii
Aichi Institute of Technology
Aichi, Japan
ishii@aitech.ac.jp

***Abstract*— In the current Internet-based systems, there are many problems using anonymity of the network communication such as personal information leak and crimes using the Internet systems. This is because the TCP/IP protocol used in Internet systems does not have the user identification information on the communication data, and it is difficult to supervise the user performing the above acts immediately. As a solution for solving the above problem, there is the approach of Policy-based Network Management (PBNM). This is the scheme for managing a whole Local Area Network (LAN) through communication control of every user. In this PBNM, two types of schemes exist. The first is the scheme for managing the whole LAN by locating the communication control mechanisms on the course between network servers and clients. The second is the scheme of managing the whole LAN by locating the communication control mechanisms on clients. As the second scheme, we have been studied theoretically about the Destination Addressing Control System (DACS) Scheme. By applying this DACS Scheme to Internet system management, we realize the policy-based Internet system management. Though the Wide Area DACS system (wDACS system) part-I, it was shown that it has problems because the processes of data transmission and reception are not encrypted. In this paper, we show the DACS system part-II with the encrypting mechanism theoretically.**

Keywords-policy-based netwok management; DACS Scheme; NAPT.

I. INTRODUCTION

In the current Internet systems, there are many problems using anonymity of the network communication, such as personal information leak and crimes using the Internet systems. The news of the information leak in the big company is sometimes reported through the mass media. Because TCP/IP protocol used in Internet systems does not have the user identification information on the communication data, it is difficult to supervise the user performing the above acts immediately. Many solutions and technologies for managing Internet systems based on TCP/IP protocol have been emerged, namely, Domain Name System (DNS) [3], Routing protocols, firewall (F/W) [7], and Network Address Port Translation (NAPT) [8] / Network Address Translation (NAT) [9]. However, they are for managing the specific part of the Internet systems, and have no purpose of solving our target problems.

PBNM might be a solution for solving these problems. However, it is a scheme for managing a whole LAN through communication control of every user, and cannot be applied to the Internet systems. It is often used in a scene of campus network management. In a campus network, network management is quite complicated. Because a network administrative department manages only a small portion of the wide needs 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 server machines, it is necessary for some users to update a client machine's setups. Most of computer network users in a campus are students. Because they do not check frequently their e-mail, it is hard work to make them aware of the settings update. This administrative operation is executed by means of web pages and/or posters. For the system administrator, it is difficult to support every student in terms of time and workload. Because the PBNM manages a whole LAN, it is easy to solve this kind of problem. In addition, for the problem such as personal information leak, the PBNM can manage a whole LAN by making anonymous communication non-anonymous. As the result, it becomes possible to identify the user who steals personal information and commits a crime swiftly and easily. Therefore, by applying the PBNM, we study about the policy-based Internet system management.

In the existing PBNM, there are two types of schemes. The first is the scheme of managing the whole LAN by locating the communication control mechanisms on the course between network servers and clients. The second is the scheme of managing the whole LAN by locating the communication control mechanisms on clients. It is difficult to practically apply the first scheme to Internet system management, because the communication control mechanism needs to be located on the course between network servers and clients necessarily. Because the second

 ISBN: 978-1-61208-330-8

scheme locates the communication control mechanisms as the software on each client, it becomes possible to apply the second scheme to Internet system management by devising the installing mechanism so that users can install the software to the client easily.

As the second scheme, we have been studied, theoretically, about the Destination Addressing Control System (DACS) Scheme. As the works on the DACS Scheme, we showed the basic principle of the DACS Scheme [28], and security function [29]. After that, we implemented a DACS system to realize a concept of the DACS Scheme [30]. By applying this DACS Scheme to Internet systems, we realize the policy-based Internet system management. As a first step, we studied wide area DACS system part-I [31] realized by applying the DACS Scheme to a wide area network. It has problems that processes of the data transmission and reception are not encrypted. In this paper, we show the encrypting mechanism, which is suitable for the wDACS system.

In Section II, motivation and related research are described. Existing DACS Scheme are described in Section III. Then, in Section IV, after describing the wDACS system part- I, wDACS system part- II are suggested theoretically.

II. MOTIVATION AND RELATED RESEARCH

In the current Internet systems, problems using anonymity of the network communication, such as personal information leak and crimes using the Internet systems occur. Because the TCP/IP protocol used in Internet systems does not have the user identification information on the communication data, it is difficult to supervise the user performing the above acts immediately.

Many solutions and technologies for Internet systems management using TCP/IP [1][2] have been proposed and are in use:

(1) DNS [3]
(2) Routing protocols:
  (2-a) Interior Gateway Protocols (IGP), such as Routing Information Protocol (RIP) [4] and Open Shortest Path First (OSPF) [5]
  (2-b) Exterior Gateway Protocols (EGP), such as Border Gateway Protocol (BGP) [6]
(3) F/W [7]
(4) NAT [8] / NAPT [9]
(5) Load balancing [10][11]
(6) Virtual Private Network (VPN) [12][13]
(7) Public Key Infrastructure (PKI) [14]
(8) Server virtualization [15]

However, they are for managing the specific aspect of the Internet systems, but have no purpose of solving our target problems.

In the following, we are focusing on policy-based thinking, to study the policy-based Internet system management.

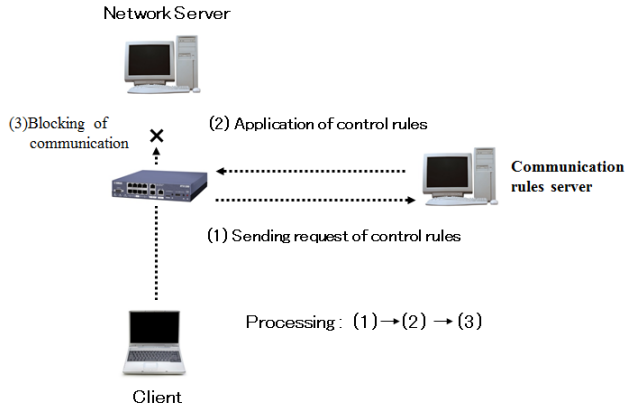


Figure 1. Principle in the first scheme.

In PBNM, there are two types of schemes. The first scheme is described in Figure 1. This scheme is standardized in various organizations. In IETF, a framework of PBNM [16] was established. Standards about each element constituting this framework are as follows. As a model of control information stored in the server called Policy Repository, Policy Core Information model (PCIM) [17] was established. After it, PCMIe [18] was established by extending the PCIM. To describe them in the form of Lightweight Directory Access Protocol (LDAP), Policy Core LDAP Schema (PCLS) [19] was established. As a protocol to distribute the control information stored in Policy Repository or decision result from the Policy Decision Point (PDP) to the Policy Enforcement Point (PEP), Common Open Policy Service (COPS) [20] was established. PDP is the point which performs the judgment about the communication control, and PEP is the point which performs the communication control based on the judgment. Based on the difference in distribution method, COPS usage for RSVP (COPS-RSVP) [21] and COPS usage for Provisioning (COPS-PR) [22] were established. RSVP is an abbreviation for Resource Reservation Protocol. The COPS-RSVP is the method as follows. After the PEP 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 the Distributed Management Task Force (DMTF), a framework of PBNM called Directory-enabled Network (DEN) was established. Like the IETF framework, control information is stored in the server called Policy Server which is built by using the directory service, such as LDAP [23], and is distributed to network servers and networking equipment such as switch and router. As the result, the

 ISBN: 978-1-61208-330-8

whole LAN is managed. The model of control information used in DEN is called Common Information Model (CIM); the schema of CIM (CIM Schema Version 2.30.0) [25] was published. CIM was extended to support DEN [24], and was incorporated in the framework of DEN.

In addition, Resource and Admission Control Subsystem (RACS) [26] 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) [27] 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. Essential principle is described in Figure 2. To be concrete, in the PDP, 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. Based on that judgment, the control is added for the communication that is going to pass by.

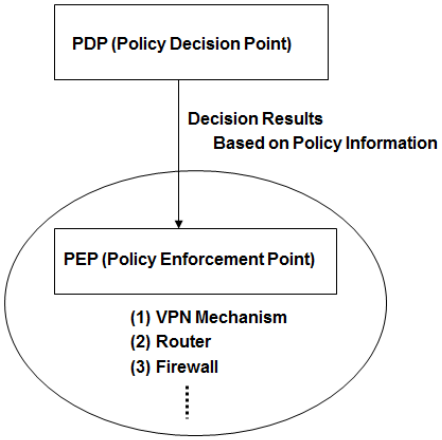


Figure 2. Essential Principle.

The principle of the second scheme is described in Figure 3 [28][29][30][31]. By locating the communication control mechanisms on the clients, the whole LAN is managed. Because this scheme controls the network communications on each client, the processing load is low. However, because the communication control mechanisms need to be located on each client, the workload becomes heavy.

We aim at realizing the PBNM management of an Internet system by applying these two schemes. However, it was difficult to apply the first scheme to Internet system management practically. In the first scheme, the communication control mechanism needs to be located on the course between network servers and clients, necessarily. As the result, the mechanism is operated from outside. It is more likely to violate the network and security policy of each organization.

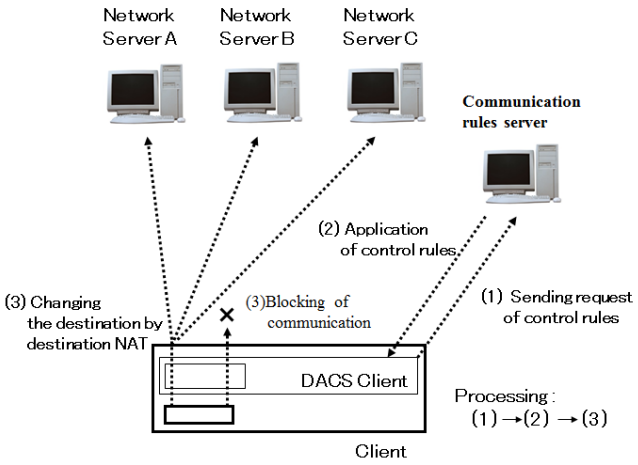


Figure 3. Principle in second scheme.

On the other hand, the second scheme locates the communication controls mechanisms on each client. The software for communication control is installed on each client. Therefore, by devising the installing mechanism letting users install software to the client easily, it becomes possible to apply the second scheme to Internet system management. As a first step, we showed the wDACS system part-I [31]. This system manages a wide area network which one organization manages. However, this system has a problem, namely, user authentication processes and transmission and reception processes of control information called DACS rules are not encrypted.

III. EXISTING DACS SCHEME

In this section, the content of the DACS Scheme is described.

A Basic Principle of the DACS Scheme

Figure 4 shows the basic principle of the network services by the DACS Scheme. At the processing 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) Processing of a user logging in the client.
(b) Processing of a delivery indication from the system administrator.

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

(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 4. In Figure 4, the system administrator can distribute a communication of the user used for login to the specified server among servers A, B or C. Moreover, the case (2) is described. For example, when the system administrator

 ISBN: 978-1-61208-330-8

wants to forbid an user to use Mail User Agent (MUA), it is performed by blocking IP Packet with the specific destination information.

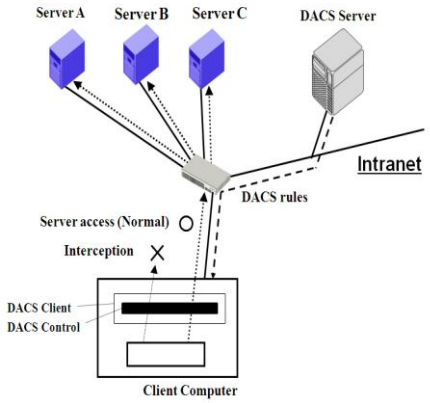


Figure 4. Basic Principle of the DACS Scheme.

In order to realize the DACS Scheme, the operation is done by a DACS Protocol, as shown in Figure 5. As shown by (1) in Figure 5, 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 5.

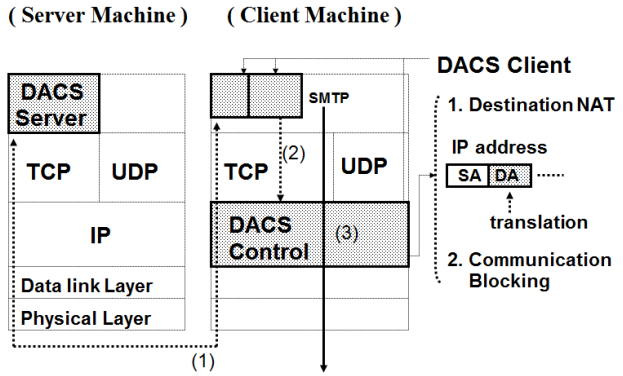


Figure 5. Operation of the DACS Protocol.

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 5.

B Communication Control on Client

The communication control of every user was given. However, it may be better to perform communication control 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 every client is described, and the coexistence method with the communication control of every user is considered.

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, 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.

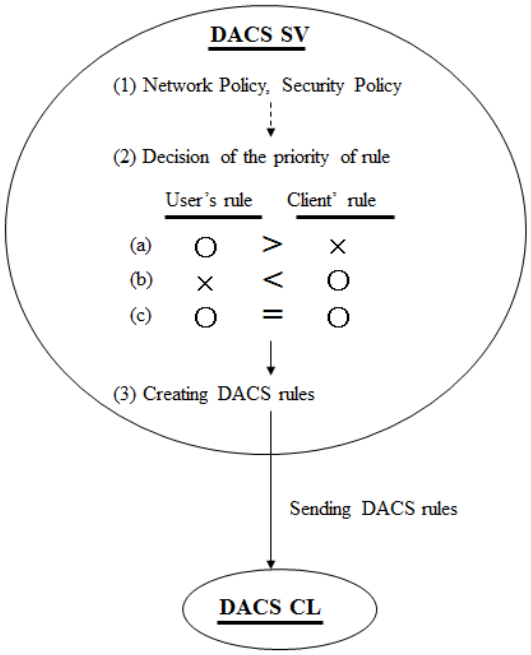


Figure 6. Creating the DACS rules on the DACS Server.

When using the communication control of 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 6. 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

In this section, the security function of the DACS Scheme is described. The communication is tunneled and encrypted by use of Secure Shell (SSH) [32]. By using the

 ISBN: 978-1-61208-330-8

function of port forwarding of SSH, it is realized to tunnel and encrypt the communication between the network server and the DACS Client, which the DACS Client is installed in. Normally, to communicate from a client application to a network server by using the function of port forwarding of SSH, the local host (127.0.0.1) needs to be indicated on that client application as a communicating server. The transparent use of a client as the virtue of the DACS Scheme is lost. The transparent use of a client means that a client can be used continuously without changing setups when the network system is updated. The function that does not fail the transparent use of a client is needed. The mechanism of that function is shown in Figure 7.

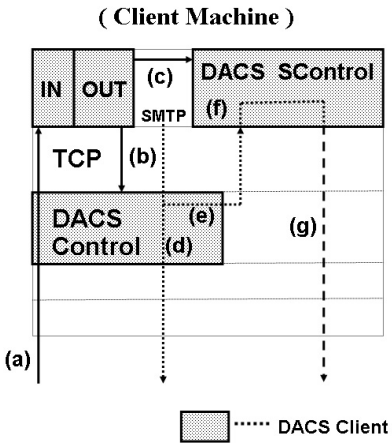


Figure 7. Extend Security Function.

The changed point on network server side is shown as follows, in comparison with the existing DACS Scheme. SSH Server is located and activated, and communication, except, SSH is blocked. In Figure 7, the DACS rules are sent from the DACS Server to the DACS Client (a). On the DACS Client that accepts the DACS rules, the DACS rules are applied to the DACS Control in the DACS Client (b). These processes are same as the existing DACS Scheme. After functional extension, as shown in (c) of Figure 7, 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 7. When communication is tunneled and encrypted, destination of the communication is changed by the DACS Control to localhost, as shown in Figure 7. In Figure 7, the communication to localhost is shown with the arrows from (e) to the direction of (f). After that, by the DACS SControl which is used for the VPN communication, the communicating server is changed to the network server and tunneled and encrypted communication is sent as, shown in (g) of Figure 7, 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. As the functional extension explained in the above, the function of tunneling and encrypting communication is realized in the state of being suitable for the DACS Scheme, that is, with the transparent use of a client. Distinguishing the control in the case of tunneling and encrypting or not tunneling and encrypting by a user unit is realized by changing the content of the DACS rules applied to the DACS Control and the DACS SControl. By tunneling and encrypting the communication for one network service from all users, and blocking the not tunneled and decrypted communication for that network service, the function of preventing the communication for one network service from the client, which DACS Client is not installed in, is realized. Moreover, the communication to the network server from the client on which DACS Client is not installed in is permitted; each user can select whether the communication is tunneled and encrypted or not.

D Technical Points in Implementaion of DACS System

(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 8, Winsock2 SPI is a new layer which is created between the existing Winsock API and the layer under it.

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().

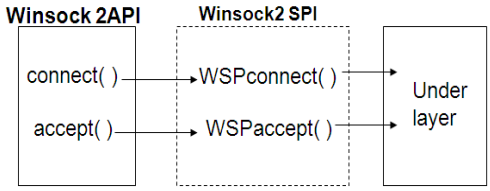


Figure 8. Winsock2 SPI.

(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,

 ISBN: 978-1-61208-330-8

and sends the VPN communication by using the port forward function.

IV. wDACS SYSTEM

In this section, the content of wDACS system is explained. First, we show contents of the wDACS system part-I, and describe solving method of technical problem in the wDACS system part-I. Therefore, the necessary mechanism of the wDACS system part-II is described.

A System Configuration of wDACS system part-I

The system configuration of the wDACS system part-I is described in Figure 9.

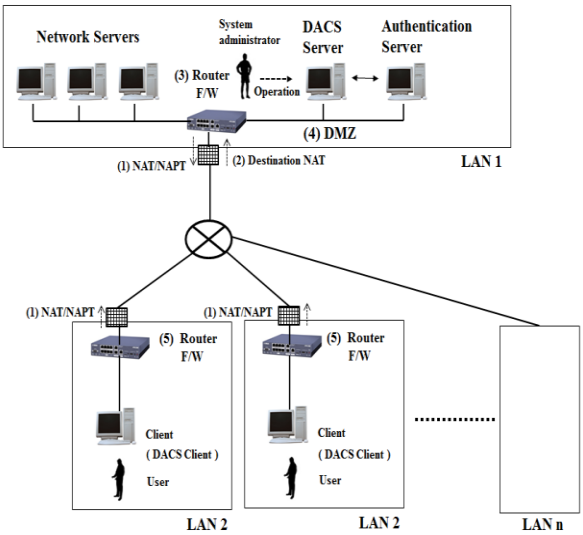


Figure 9. Basic System Configuration of wDACS system part-I.

First, as preconditions, because private IP addresses are assigned to all servers and clients existing in from LAN1 to LAN n, mechanisms of NAT/NAPT are necessary for the communication from each LAN to the outside. In this case, NAT/NAPT is located on the entrance of the LAN such as (1), and the private IP address is converted to the global IP address towards the direction of the arrow.

Next, because the private IP addresses are set on the servers and clients in the LAN, other communications except those converted by Destination NAT cannot enter into the LAN. But, responses for the communications sent form the inside of the LAN can enter into the inside of the LAN because of the reverse conversion process by the NAT/NAPT.

In addition, communications from the outside of the LAN1 to the inside are performed through the conversion of the destination IP address by Destination NAT. To be concrete, the global IP address at the same of the outside interface of the router is changed to the private IP address of each server.

From here, system configuration of each LAN is described. First, the DACS Server and the authentication server are located on the DMZ on the LAN1 such as (4). On the entrance of the LAN1, NAT/NAPT and destination NAT exists such as (1) and (2). Because only the DACS Server and network servers are set as the target destination, the authentication server cannot be accessed from the outside of the LAN1. In the LANs from LAN 2 to LAN n, clients managed by the wDACS system exist, and NAT/NAPT is located on the entrance of each LAN such as (1). Then, F/W, such as (3) or (5), exists behind or with NAT/NAPT in all LANs.

B Key ExchangeMechanism for wDACS system part-II

As the additional mechanism for the wDACS system part-II, the mechanism in Figure 10 is newly introduced.

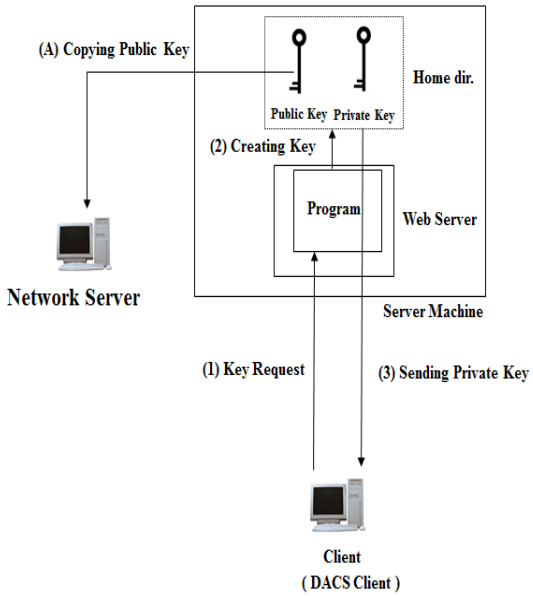


Figure 10. Mechanism of Key Exchange.

This is a periodical key exchange mechanism which is necessary for encrypted communications between the network servers and the client computers. This mechanism is incorporated at the last part of the initialization process of the DACS Client. The preconditions are as follows.

(a) The communications between the DACS Client and the Web Server are encrypted by the https.
(b) The communications between the Server Machine moving the Web Server and network servers are encrypted by SSH.
(y) This mechanism is located on the Server Machine which is separated physically with DACS Sever for the management of a large-scale network with many clients.

Next, the processing of this mechanism is described. First, the key request is performed from the DACS Client (1). The program on the Web Server receives the request, and creates two kinds of keys which are a public key and a private key (2). Then, the program sends the private key to the client (3).

 ISBN: 978-1-61208-330-8

The public key stored in the home directory on the Server Machine is copied and stored on the network server by mirroring through SSH. To be concrete, network commands such as rsync and rdiff-backup are used. The mirroring process is performed just before the transmission of the private key.

C *Encrypted Communication Mechanism for wDACS system part-II*

In this section, two functions to realize the encrypted communication are described.

(1) Function of encrypted communications in user authentication processes

In this section, the function of the encrypted communications in user authentication processes, which is suitable for the wDACS system, is described. The content of the function is shown in Figure 11.

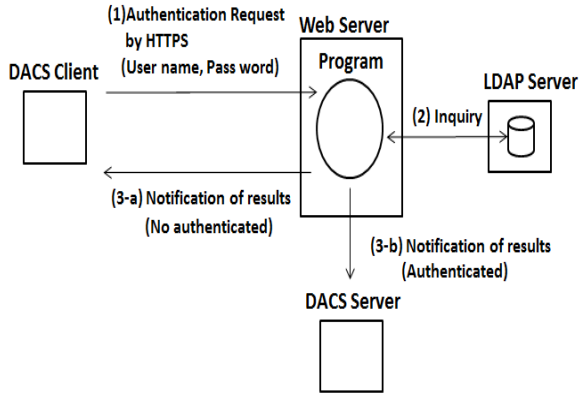


Figure 11. Function of user authentication processes.

First, authentication request is performed form the DACS Client to the program on the Web Server (1). The Program performs inquiry to the LDAP Server which stores user accounts (user name, pass word) (2). As the result, if authentication is not permitted, the results are notified to the DACS Client (3-a). The DACS Client stops performing subsequent processing. If authentication is permitted, the results are notified to the DACS Server (3-b). The DACS Server performs the processing described in next section.

(2) Function of encrypted communications

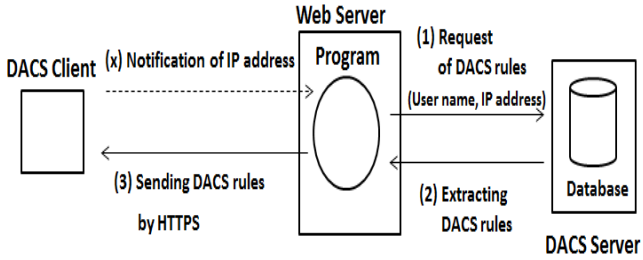


Figure 12. Function of transmission and reception processes for control information.

In this section, the function of the encrypted communications in the DACS rule's transmission and reception processes, which is suitable for the wDACS system, is described.

First, as a part of process (1) in Figure 11, the IP address of the client where the DACS Client is installed is notified with user name and password to the program on the Web Server. This process is described as process (x) in Figure 12, which is shown by a dotted arrow.

Next, based on them, the program performs a request of the DACS rules to the DACS Server (1). The DACS rules are extracted from the database of the DACS Server, and sent to the program on the Web Server (2). The program receives them, and sends to the DACS Client.

Specific to these two functions is the use of HTTPS. Because this wDACS system needs to be extended for Internet management, we chose HTTPS used widely in the world of the Internet.

V. CONCLUSION AND FUTURE WORK

In this paper, we showed the policy-based wide area network management system called wDACS system part-II. This system was realized by the extension of the policy-based network management system called wDACS system part-I which manages the WAN for one organization. The wDACS system part-I had two problems. First problem was unencrypted communication of user authentication processes. The second problem is unencrypted communication of transmission and reception processes of the DACS rules. To solve these two problems, additional two mechanisms were described in this paper. By these two mechanisms, it became possible to perform the encrypted communications using HTTPS in anticipation of expansion to Internet management. As a future study, the wDACS system part-II will be implemented to manage the WAN and evaluations will be performed.

REFERENCES

[1] V. Cerf and E. Kahn, "A Protocol for Packet Network Interconnection," IEEE Trans. on Commn, vol. COM-22, pp. 637-648, May 1974.

[2] B. M. Leiner, R. Core, J. Postel, and D. Milis, "The DARPA Internet Protocol Suite," IEEE Commun.Magazine, vol. 23 pp. 29-34 March 1985.

[3] P. Mockapetris and K. J. Dunlap. "Development of the domain name system," SIGCOMM'88, 1988.

[4] http://tools.ietf.org/html/rfc2453 [retrieved: 2, 2014]

[5] http://www.ietf.org/rfc/rfc2328.txt [retrieved: 2, 2014]

[6] http://tools.ietf.org/html/rfc4271 [retrieved: 2, 2014]

 ISBN: 978-1-61208-330-8

[7] A. X. Liu and M. G. Gouda, "Diverse Firewall Design," IEEE Trans. on Parallel and Distributed Systems, vol. 19, Issue. 9, pp. 1237-1251, Sept. 2008.

[8] http://tools.ietf.org/html/rfc1631 [retrieved: 2, 2014]

[9] M. S. Ferdous, F. Chowdhury, and J. C. Acharjee, "An Extended Algorithm to Enhance the Performance of the Current NAPT," Int. Conf. on Information and Communication Technology (ICICT '07), pp. 315-318, March 2007.

[10] 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.

[11] 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.

[12] C. Metz, "The latest in virtual private networks: part I," IEEE Internet Computing, vol. 7, no. 1, pp. 87-91, 2003.

[13] C. Metz, "The latest in VPNs: part II," IEEE Internet Computing, vol. 8, no. 3, pp. 60-65, 2004.

[14] R. Perlman, "An overview of PKI trust models," IEEE Network, vol. 13, issue 6, pp. 38-43, Nov/Dec 1999.

[15] A. Singh, M. Korupolu, and D. Mohapatra, "Server-storage virtualization: Integration and load balancing in data centers," Int. Conf. for High Performance Computing, Networking, Storage and Analysis, pp. 1-12, Nov. 2008.

[16] R. Yavatkar et al., "A Framework for Policy-based Admission Control," IETF RFC 2753, 2000.

[17] B. Moore et al., "Policy Core Information Model -- Version 1 Specification," IETF RFC 3060, 2001.

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

[19] J. Strassner et al., " Policy Core Lightweight Directory Access Protocol (LDAP) Schema," IETF RFC 3703, 2004.

[20] D. Durham et al., "The COPS (Common Open Policy Service) Protocol, " IETF RFC 2748, 2000.

[21] S. Herzog et al., "COPS usage for RSVP", IETF RFC 2749, 2000.

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

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

[24] M. Wahl et al., "Lightweight Directory Access Protocol (v3)," IETF RFC 2251, 1997.

[25] CIM Schema: Version 2.30.0, 2011.

[26] 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.

[27] 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 specifica-tion", April 2006.

[28] 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.

[29] 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, November 2008.

[30] 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 (AINA 2010), Perth, Australia, Japan, IEEE Computer Society, pp. 348-354, May 2010.

[31] K. Odagiri, S. Shimizu, M. Takizawa, N. Ishii, "Theoretical Suggestion of Policy-Based Wide Area Network Management System (wDACS system part-I)," International Journal of Networked and Distributed Computing (IJNDC), vol. 1, no.4, pp. 260-269, 2013.

[32] http://tools.ietf.org/html/rfc4251 [retrieved: 2, 2014]

 ISBN: 978-1-61208-330-8

Multi-controller Scalability in Multi-domain Content Aware Networks Management

Eugen Borcoci, Mihai Constantinescu, Marius Vochin
University POLITEHNICA of Bucharest
Bucharest, Romania
emails: eugen.borcoci@elcom.pub.ro
mihai.constantinescu@elcom.pub.ro, marius.vochin@elcom.pub.ro

***Abstract* — This paper studies the management system scalability properties of a networked media eco-system. The system offers guaranteed Quality of Services (QoS) multimedia delivery, over multiple domain networks, based on creation of data plane slices named Virtual Content Aware Networks (VCAN). The VCANs are realized under control of a management plane, by centralized per network domain "controllers" - cooperating in order to span the VCANs over multiple IP domains. Scalability is an issue- similar to the multi-controller problem, in emerging Software Defined Networking (SDN) technologies. The management system architecture considered in this paper has been previously defined. This work provides a simulation model and results concerning the multi-controller communications. It is shown that SDN-like control approach is conveniently feasible in a multi-domain environment.**

Keywords — Content-Aware Networking; Software Defined Networking, Multi-domain; Management; Resource provisioning; Future Internet.

I. INTRODUCTION

The architectural solutions for the Future Internet constitute hot research topics today. It is recognized that traditional Internet has fundamental architectural limitations, and also ossification if compared to the current needs and considering its global extension [1], [2].

A significant trend in Internet is the information/content-centric orientation; consequently, significant changes in services and networking have been recently proposed, including modifications of the basic architectural principles. The revolutionary approaches are often referred to as *Information/Content-Centric Networking (ICN/CCN)*, [3], [4]. In parallel, evolutionary (or incremental) solution emerged, as Content-Awareness at Network layer (CAN) and Network-Awareness at Applications layers (NAA). This approach can create a powerful cross-layer optimisation loop between the transport and applications and services.

A new trend, targeting to achieve more flexibility in networking is *Software Defined Networking* (SDN) architecture and its associate OpenFlow protocol [5], [6], [7] where the control plane and data planes are decoupled and the network intelligence is more centralized, thus offering a better and flexible/programmable control of the resources.

The European FP7 ICT research project, *"Media Ecosystem Deployment Through Ubiquitous Content-Aware Network Environments"*, ALICANTE, [12], adopted the NAA/CAN approach, to define, design, and implement a Media Ecosystem spanning multiple network domains. This work considers as input the ALICANTE management architecture [14], which is similar to SDN with respect to the distribution of the main management and control functions among several controllers. Communication between controllers is necessary in order to accomplish multi-domain tasks.

This paper studies by simulation, based on Extended Finite State Machines (EFSM) model, some scalability aspects of the signalling protocol for multi-controller communication.

Section II presents some related work. Section III describes the management architecture. Section IV defines the inter-controller communication. Section V introduces the simulation model and Section VI describes the simulation results. Conclusion, open issues, and future work are shortly outlined in the Section VII.

II. RELATED WORK

The ICN approaches are very promising, but raise some research and deployment challenges like the degree of preservation of the classic transport (TCP/IP) layering principles, naming and addressing, content-based routing and forwarding, management and control framework, in-network caching, etc.

In SDN, the network intelligence is more centralized, thus offering a better and flexible control of the resources, quality of services, etc., due to the possibility to have an overall image of the system in the control plane and by allowing programmability of the network resources. The operators will get more freedom and speed in developing their services, without waiting long time for new releases of vendors' networking equipment. Although it seems to be very attractive, e.g., for data centers but not only, SDN exposes also many research challenges and open issues, both from architectural and from deployment point of view. Degree of centralization and relationship with scalability and reliability are examples. An extension of the SDN concepts is proposed in so-called *Software Defined (Internet) Architecture* [9], where the idea is also to decouple the architecture from infrastructure as to lower the barriers to architectural evolution. The SDIA approach tries to exploit SDN concepts but also traditional technologies (e.g., MPLS, software forwarding, etc.) in order to obtain evolvable architectures. SDN and SDIA are still evolutionary in contrast with "clean slate" ones, which are disruptive.

 ISBN: 978-1-61208-330-8

Currently there are concerns about SDN performance, scalability, and resiliency [8], [11], the main source for these problems being the centralization concept. It is clear that a central controller will have a limited processing capacity and the solution will not scale as the network grows (increase the number of switches, flows, bandwidth, etc.). Apart from the obvious solution to increase the controller performance, the second idea is to define a SDN multi-controller architecture. However, SDN still wants a consistent centralized logical view upon the network; this creates the need for controllers to cooperate and synchronize their data bases in order to provide together a consistent view at network level. Work in progress is developed at IETF towards constructing an inter-controller, [12]. While the vertical protocols between Control and Data Plane have seen significant progress by specifying Open Flow versions, [7] and implementing several types of controllers, [6], [10], the inter-controller cooperation and scalability issues are still open research issue.

ALICANTE architecture has considered, from the beginning, the scalability requirements in its management and control specification. These aspects will be more developed in the next section.

III. MANAGEMENT SYSTEM ARCHITECTURE

In ALICANTE architecture, [12], [13], [14], several cooperating environments are defined containing business actors: User Environment (UE), containing the End-Users; Service Environment (SE), containing SPs and Content Providers (CP); Network Environment (NE), where we find the new CAN Provider and the traditional Network Providers (NP) managing the network elements, in the traditional way at IP level. The "environment", is a generic grouping of functions working for a common goal and which possibly vertically span one or more several architectural (sub-) layers. Between actors, dynamic Service Level Agreements (SLA) can be established. A novel business entity, CAN Provider was defined, to which several SPs can independently ask customizable Virtual Content Aware networks, and then use them. Network Providers can cooperate to VCAN construction but preserving their independency in resource allocation. Flexible connectivity services have been achieved offering: Fully/partially/un-managed services.

The Internet is sliced by creation on demand logically isolated VCANs, realised as parallel logical planes based on light virtualisation (in the Data Plane only), and optimising inter and intra-domain mapping of VCANs, onto several domain network resources.

The architecture supports both V/H integration SLAs for several level of guarantees. Distributed M&C (each domain has its Intra-NRM and an associated CAN Manager) assures large scale provisioning. The VCANs are flexible supporting: unicast, multicast, broadcast, P2P and combinations with different levels of QoS/QoE, availability, etc. End Users and their Home-Boxes can ultimately benefit from CAN/NAA features by using VCANs.

Services Providers only ask VCANs and use them; they are not burdened with tasks to construct them. The architecture assures QoS/QoE optimization based on: CAN/NAA interaction; Cooperation between resource provisioning (SLA) and media flow adaptation; Hierarchical monitoring at CAN and network layers cooperating with the upper layers.

The ALICANTE architecture is conceptually similar to recently proposed SDN, although not following full SDN specifications (Fig. 1).

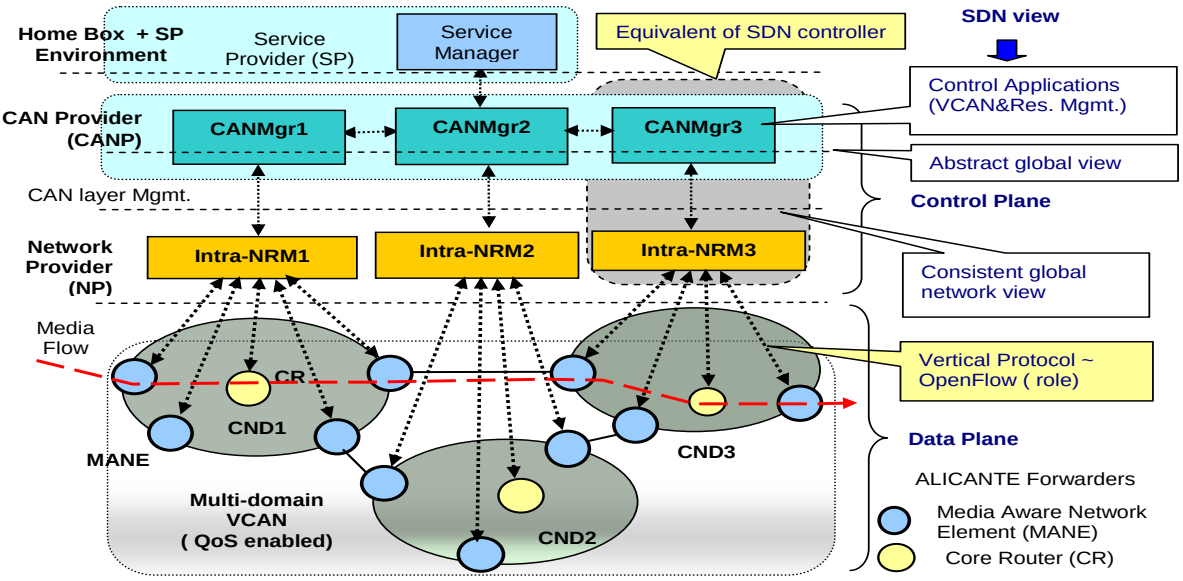


Figure 1. ALICANTE partially centralized management architecture and equivalence with SDN

Notations: SP – Service Provider; CANP - CAN Provider; NP – Network Provider; CND - Core Network Domain; CANMgr - Content Aware Network Manager; Intra-NRM – Intra-domain Network Resource Manager;

 ISBN: 978-1-61208-330-8

Both architectures are evolutionary and can be seamlessly developed. The Control Plane and Data Plane are separated. Note that Control Plane in SDN terminology is here actually Management and Control Plane. The QoS constrained routing, resource allocation, admission control and VCAN mapping are included in the CAN Manager. The "virtualization" of the network is performed by Intra-NRM, which hides the characteristics of MPLS technology by delivering to the CAN Managers an image of abstract matrix of connectivity logical pipes.

In SDN the network intelligence is (logically) centralized in SW -based SDN controllers, which maintain a global view of the network: maintain, control and program Data Plane state from a central entity. In our case [*CAN Manager + Intra-domain Network Resource Manager*] play together the role of an SDN controller for a network domain, controlling the MANE edge routers and interior core routers. Actually, we have a multi-domain logical network governed by several "SDN controllers" – which cooperates for resource management and routing. However, the degree of centralization is configurable in ALICANTE by defining the placement of CAN Managers and the sets of routers to be controlled.

In both SDN and this architecture, the Control Plane SW is executed on general purpose HW. The decoupling of the control with respect to specific networking HW is realized**:** the MANE and core routers are viewed by the upper CAN layer in abstract way.

Data Plane is programmable: all configurations for MANE and Core routers are determined in CAN and Network M&C and downloaded in the routers. ALICANTE architecture defines the control for a whole network (and not for single network devices): at CAN Manager level there exists an overall image on the static and dynamic characteristics of all VCANs; at Intra-NRM level there is a full control on the network domain associated with that Intra-NRM.

In SDN and our case also, the network appears to the applications and policy engines as a single logical switch. In our case, the network appears at higher layers as a set of parallel planes VCANs. This simplified network abstraction can be efficiently programmed, given that the VCANs are seen at abstract way; they can be planned and provisioned independently of the network technology.

IV. INTER-DOMAIN MANAGEMENT COMMUNICATIONS

One CANMgr (belonging to CANP) is the initiator of VCAN construction, at request of an SP. The VCANs asked should be mapped onto real multi-domain network topology, while respecting some QoS constraints. This provisioning is done through negotiations [14], performed between CAN Managers associated to each network domain. 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 entity inside each CANMGr 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. Previous papers, of the [13], developed and implemented the combined VCAN mapping algorithm.

The overall system flexibility and scalability essentially depends on its Management and Control. For VCAN planning, provisioning and exploitation: it was adopted per-domain partially centralized solution; this avoids full-centralized VCAN management (non-scalable), but allowing a coherent per-domain management. However, the initiator CAN Manager, like in SDN approach, has the overall consistent image of a multi-domain VCAN.

There is *no per-flow signaling* between CAN Managers. The VCAN SP–CANP negotiation is performed per VCAN, described in terms of aggregated traffic trunks. The SP negotiates its VCAN(s) with a single CAN Manager irrespective, if it wants a single or a multi-domain spanned VCAN.

A hierarchical overlay solution is applied for inter-domain peering and routing where each CAN Manager knows its inter-domain connections. The CAN Manager initiating a multi-domain VCAN is the coordinator of this hierarchy, without having to know details on each domain VCAN resources. The monitoring at CAN layer and network layer is performed at an aggregated level.

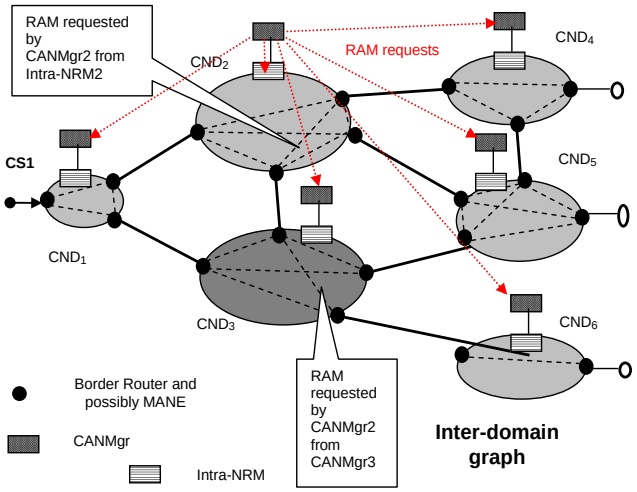


Figure 2. CAN Manager 2 issues RAM requests from each CANMgr involved in a VCAN

Fig. 2 shows an example of inter CAN Managers inter-controller signaling (i.e., inter-controllers in SDN terminology) in which the initiator CAN Manager 2 asked (in hub style) the other involved CAN Mangers (3, 4, 5, 6) to deliver to it their Resource Availability Matrices. Based on the received information, the initiator performs the VCAN mapping.

V. SIMULATION MODEL

The simulation objective of this paper is the evaluation of the M&C signaling overhead related to the negotiation

 ISBN: 978-1-61208-330-8

activities between SP, CAN Managers, Intra_NRMs when the number of network domains and CAN Managers is a variable parameter.

Given the complexity of the M&C subsystem a simulation study has been developed. Real Time Developer Studio is a Specification and Description Language simulator, developed by PRAGMADEV. It comes in two versions: SDL and Specification and Description Language - Real Time (SDL-RT) [15].

SDL-RT is ITU standard SDL (based on Extended Finite State Machine Model - EFSM) extended with real time concepts. It is object oriented, has graphical language, allows modeling real-time features, combining dynamic and static representations, supporting classical real time concepts, extended to distributed systems, based on standard languages. It retains the graphical abstraction brought by SDL while keeping the precision of traditional techniques in real-time and embedded software development and making simpler the re-use of legacy code by using natively the C language. In SDL-RT, the C language is used to define and manipulate data. The ALICANTE management simulation model consists in: one Service Provider; N x CAN Managers, N x Intra_NRMs, where N is variable (1, ...16).

The specific target is to evaluate the time spent from the instant when an SP issues a VCAN request to an initiator CAN Manager, until the final confirmation of the VCAN installation is obtained by SP. The SP can choose any CAN_Mgr, based on their proximity and involvement in the requested VCAN. The chosen CAN_Mgr, named afterward the initiating CAN_Mgr, will interrogate the inter domain database about the network capability of the others domains, performs the inter-domain mapping algorithm, and will communicate with each CAN_Mgr identified by the inter-domain mapping algorithm as being involved in the requested VCAN. Note that the simulation model assumes parallelism in communication process from the initiator CAN Manager to different others (in “Hub” style). This is an important feature and design decision, assuring the scalability of negotiation process.

Fig. 3 describes the system processes. It contains the global variables, the instance of each class and the other blocks involved.

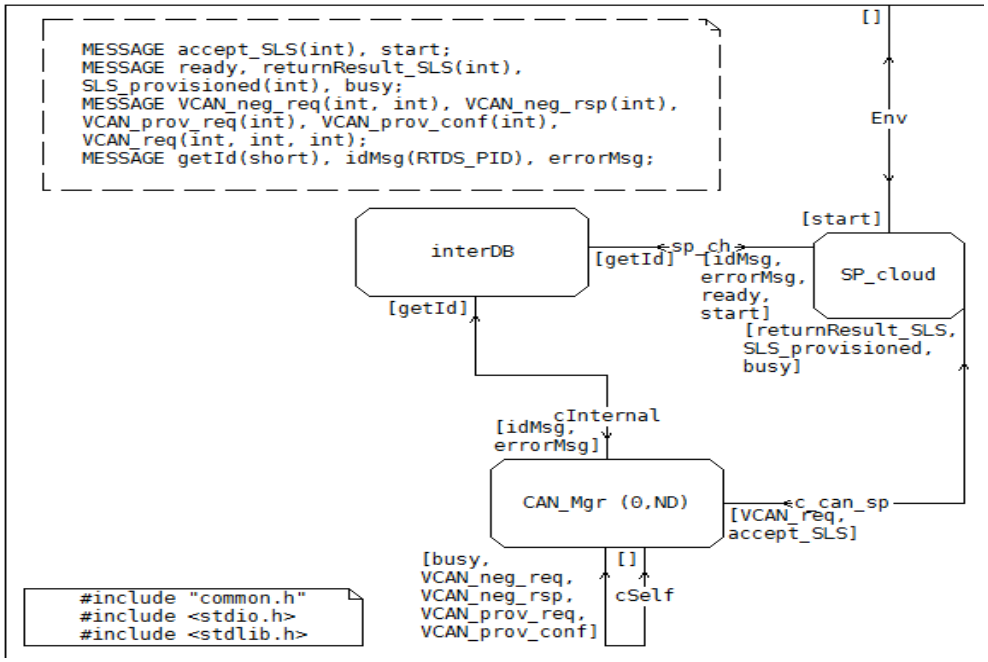


Figure 3. The system model used in RTDS simulations

The system consists of an *interDB* block, (used in simulation only, corresponding to an interdomain database that contains inter-domain network topology), a SP_cloud block, associated with the SP/CP requestors, and ND (Number of Domains) CAN_Mgr(s).

The work [14] describes in Message Sequence Chart (Fig. 4) forms the details of the signaling process between an initiator CAN Manager, and other CAN Managers involved in constructing a multiple domain VCAN. Here a simplified description is done.

The initiating CAN_Mgr send a *VCAN_neg_req* to each of corresponding CANMgr and enter into *"negociating"* state. Each corresponding CANMgr check its own capabilities (intra-domain mapping algorithm), respond to initiating CANMgr with a *VCAN_neg_rsp* message, and transits to *"waiting_for_acceptance_ext"* state. The initiating CAN_Mgr waits for all corresponding CAN_Mgr to respond, integrate the response, send that integrate response by a return_result_SLS message to SP_cloud and wait for a decision. That message indicates to SP that all requested resources are available and could be provisioned.

The SP analyzes the response and sends a provision request to the initiating CAN_Mgr, using the message accept_SLS. The initiating CAN_Mgr sends a provision request message, VCAN_prov_req, to each corresponding CAN_Mgr and waits for their response.

 ISBN: 978-1-61208-330-8

VI. SIMULATION RESULTS

The simulations are focused on identifying the system behavior, and to determine a quantitative and qualitative estimation of the signaling time.

Being a real time simulator, the RTDS SDL-RT uses the internal PC clock to estimate the time for each task/process from the system. Therefore, the results are defined in "ticks", which are relative time units.

The simulation model just simulates the time consumed by the inter-domain and intra-domain mapping algorithms, but it does not actually compute that algorithms. However, the result of the mapping algorithm, the chosen CAN_Mgr and the Round Trip Time (RTT) delay between two communicating CAN_Mgr are introduced in simulator using a configuration file; its data is shown in Table 1.

TABLE 1. VARIATION IN RESPONSE TIME FROM DIFFERENT CAN_MANAGERS TO INITIATOR

	Average RTT=100	Average RTT=200	Average RTT=300	Average RTT=400	Average RTT=500
CAN_Mgr	RTT var	RTT var	RTT var	RTT var	RTT var
1	50	50	50	50	150
2	120	220	420	420	420
3	100	300	500	500	600
4	130	230	230	630	630
5	63	63	163	163	363
6	137	337	537	537	837
7	82	182	382	382	482
8	118	218	218	518	518
9	96	96	96	96	96
10	104	304	504	504	704
11	51	251	351	551	551
12	149	149	149	649	749
13	70	70	270	270	470
14	130	330	330	430	630
15	81	81	81	181	281
16	119	319	519	519	519
17	100	200	300	400	500

Each simulation uses one column from that table. Random variations have been introduced to emulate a real situation where the CAN Mangers are placed in different network domains and communicates via Internet.

While the simulation model uses an abstract time clock, in the experiments done, we can evaluate a time unit comparable to 1ms.

Two sets of data are used: one with a fixed (constant) RTT value for each corresponding CAN_Mgr, and one with the same average value as the constant RTT, but with a big dispersion. Both sets of data are shown in the Table 1.

The simulations are performed on two different machines, (i.e. named "Processor-1" - low power, and "Processor_2" - high power (Table 2).

TABLE 2 PC CONFIGURATION

PC configuration	Windows Experience Index	
	Processor-1	Processor-2
Processor	6	7.5
Memory(RAM)	5.9	7.8
Primary hard disk	5.9	7.9

The computing difference between the two PCs are just a qualitative criteria on evaluating the performance of a real CAN_Mgr machine when computing VCAN requests in ALICANTE environment.

Most of simulations are performed on a powerful PC, named "Processor_2". However, some of simulations are performed also on a slower PC, named "Procesor_1". These simulations allow a validation of results obtained from Processor_2". As expected, the results from "Procesor_1" have bigger relative time values compared with the values from "Processor_2", due processing time is shorter on powerful machine, as "Processor_2" (Table 2).

The range of RTT varies from 100 to 500 relative time units. Figures 4, 5 and 6 present the signaling time dependency (in relative time units) on the domains number implied in the requested VCAN, both from constant RTT and variable RTT.

The important result obtained and shown by the above diagrams is that the system is scalable versus the number of network domains. The simulation results shows that, even the start values are different, the signaling time has a tendency of convergence to 2500 value (~ms).

That convergence is explained by the fact that the signaling is made in parallel with all the CAN_Mgr involved, and the signaling time is depending on the domains numbers, the computation time spent on CAN_Mgr, and the RTT delay between each two communicating CAN_Mgr.

The simulator time unit can represent approximately one ms. This means that a single initiator CAN Manager could perform ~ 1200 VCAN requests per hour which is considered completely satisfactory into ALICANTE environment.

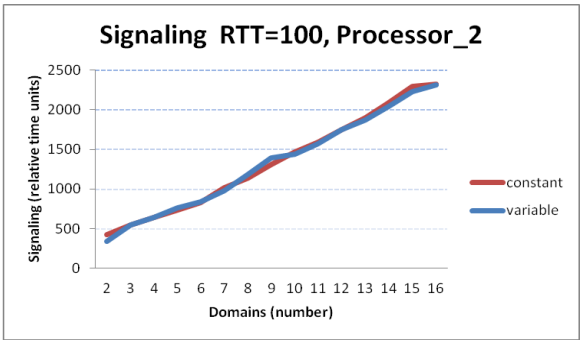


Figure 4. Signaling RTT=100, Processor 2

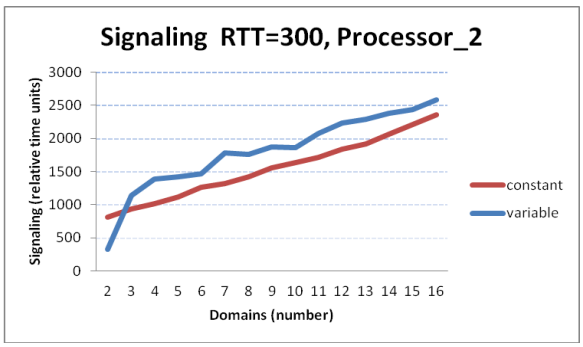


Figure 5. Signaling RTT=300, Processor_2

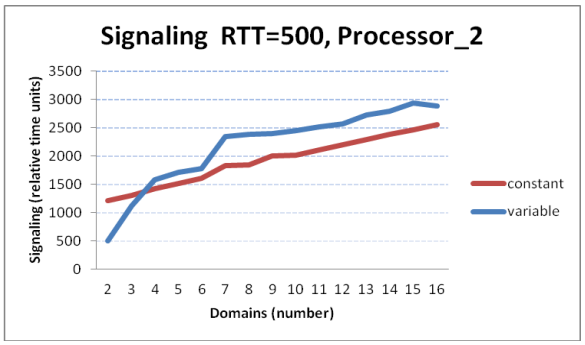


Figure 6. Signaling RTT=500, Processor_2

Fig. 7 shows a comparison between simulations on "Processor-1" and "Processor-2". It is clear that the system behavior is similar, only the convergence value is different (4500 for "Processor-1" and 2500 for "Processor-2"). Again, the convergence is present, and the difference on convergence value is the result of different computing time inside CAN_Mgr.

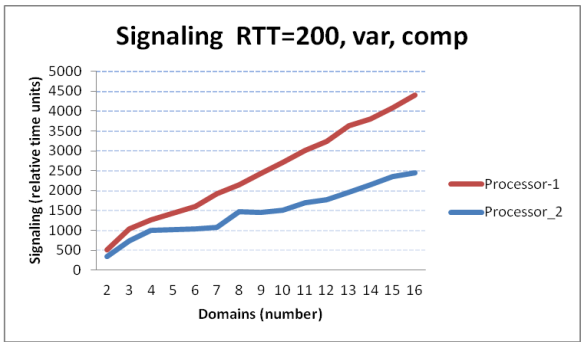


Figure 7. Signaling RTT=200, var, comparison

In order to have a validation from statistical point of view, a set of simulations were performed twice on "Processor-2", using different seeds. The simulation results are shown in Fig. 8. The convergence is present again, the small difference of the convergence value occurs due the seed influence on simulator internal algorithm, shown on the relative time units obtained on each simulation. That difference has a minor importance, as demonstrated by the several comparisons made based on the whole set of simulation results (Fig. 9, 10, and 11).

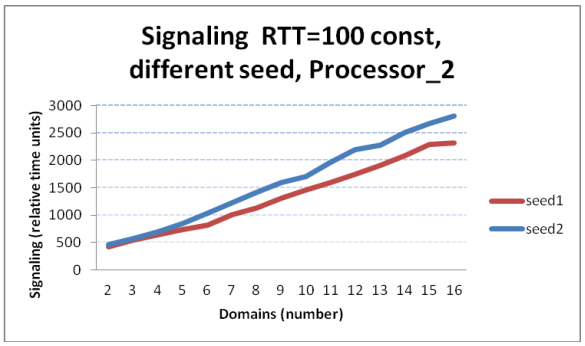


Figure 8. Signaling RTT=100 const, different seed, Processor_2

A convergence analysis is presented on Fig. 9, 10, and 11. The convergence is proved on both machines, for both fixed (constant RTT), and variable RTT.

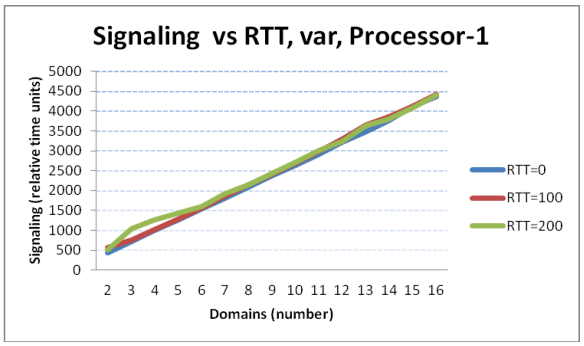


Figure 9. Signaling vs RTT, var, Processor_1

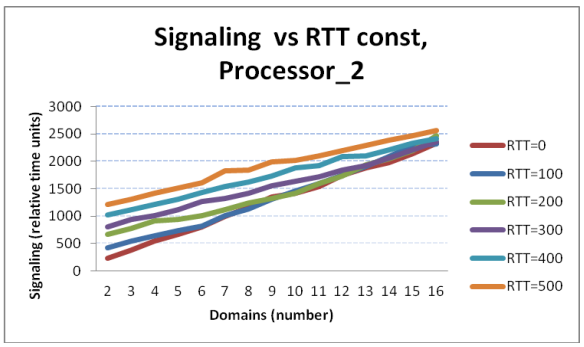


Figure 10. Signaling vs RTT const, Processor_2

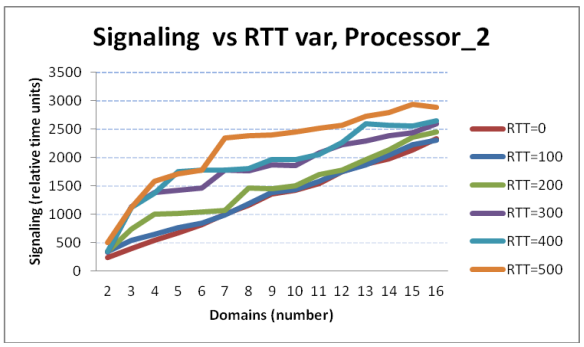


Figure 11. Signaling vs RTT var, Processor_2

Fig. 11 is the most interesting graph, showing that, despite big dispersion of RTT used, the signaling time is converging on the same value (2500) as for constant RTT. On that graph, the simulations used the same seed, but a range of average RTT from 0 to 500ms, with a high dispersion on each subset (same average RTT, different value for each RTT). Each RTT subset used in simulations is described in Table 1.

Moreover, the results from Fig.11 highlight that the RTT delay between each two communicating CAN_Mgr has a minor influence on the convergence value. According with the overall simulation results and comparing with the computing capability of CAN_Mgr and number of domains involved, that influence could be ignored.

 ISBN: 978-1-61208-330-8

VII. CONCLUSIONS

The management architecture of the ALICANTE system has been described, showing the similarity with SDN approach. Equivalence between an SDN controller and the pair {CAN Manager and Intra-domain Network Resource Manager} has been analyzed. Scalability problems appear in this multi-controller environment.

A simulation model based on EFSM model has been constructed.

It was found that signaling time dependency on the domains number implied in the requested VCAN is converging to an approximately constant value, being determined by the speed of the processor equivalent to the CAN_Mgr + IntraNRM used for simulation and by the number of domains involved.

That convergence is not influenced by the communication delay between the communicating CAN_Mgr.

The signaling in ALICANTE system is growing slowly and linearly with the number of domains involved, validating ALICANTE management system scalability.

Further work should evaluate the capacity of one controller to command a given number of network elements (routers) by using a vertical protocol (similar to OpenFlow).

ACKNOWLEDGMENT

This work has been partially supported by the FP7 European research project ALICANTE (FP7-ICT-248652) and partially by the projects POSDRU/107/1.5/S/76909.

REFERENCES

[1] J. Schönwälder, M. Fouquet, G. D. Rodosek, and I. Hochstatter (2009), "Future Internet = Content + Services + Management", IEEE Communications Magazine, vol. 47, no. 7, Jul. 2009, pp. 27-33.

[2] 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.

[3] J. Choi, J. Han, E. Cho, T. Kwon, and Y. Choi, "A Survey on Content-Oriented Networking for Efficient Content Delivery", IEEE Communications Magazine, March 2011.

[4] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard, "Networking Named Content," CoNEXT '09, New York, NY, 2009, pp. 1–12.

[5] N. McKeown, et. al., "OpenFlow: Enabling Innovation în Campus Networks", March 2008, - http://www.openflow.org/documents/openflow-wp-latest.pdf.

[6] B. A. A. Nunes, M. Mendonca, X. N. Nguyen, K. Obraczka, and Thierry Turletti, "A Survey of Software-Defined Networking: Past, Present, and Future of Programmable Networks", ian. 2014, http://hal.inria.fr/hal-00825087

[7] OpenFlow Switch Specification, V 1.3.0 (Wire Protocol 0x04), June 25, 2012, https://www.opennetworking.org/images/stories/downloads/sdn-resources/onf-specifications/openflow/openflow-spec-v1.3.0.pdf

[8] S. H. Yeganeh, A. Tootoonchian, and Y. Ganjali, "On Scalability of Software-Defined Networking", IEEE Communications Magazine, February 2013, pp.136-141.

[9] B. Raghavan, T. Koponen, A. Ghodsi, M. Casado, S. Ratnasamy, and S. Shenker, "Software-Defined Internet Architecture: Decoupling Architecture from Infrastructure", HotNets-XI Proceedings of the 11th ACM Workshop on Hot Topics in Networks, 2012, pp. 43-48, doi: 10.1145/2390231.2390239.

[10] A. Tavakkoli, M. Casado, and S. Shenker, "Applying NOX to the Datacenter," Proc. ACM HotNets-VIII Wksp., 2009.

[11] H. Yin, H. Xie, T. Tsou, D. Lopez, P. Aranda, and R. Sidi, draft-yin-sdn-sdni-00.txt, "SDNi: A Message Exchange Protocol for Software Defined Networks (SDNS) across Multiple Domains", June, 2012

[12] FP7 ICT project, "MediA Ecosystem Deployment Through Ubiquitous Content-Aware Network Environments", ALICANTE, No248652, "D2.1: ALICANTE Overall System and Components Definition and Specifications", http://www.ict-alicante.eu/.

[13] E. Borcoci, R. Miruta, and S. Obreja, "Multi-domain Virtual Content-Aware Networks Mapping on Network Resources", EUSIPCO Conference, 27-31 August 2012, Bucharest, http://www.eusipco2012.org/home.php

[14] C. Cernat, E. Borcoci, and V. Poenaru, "SLA Framework Development for Content Aware Networks Resource Provisioning", AICT 2013 Conference, http://www.thinkmind.org/index.php?view=article&articleid=aict_2013_8_30_10084

[15] http://www.sdl-rt.org/, Apr 23, 2013

 ISBN: 978-1-61208-330-8

Analytical Evaluation of Call Admission Control for SFR-Based LTE Systems

Seung-Yeon Kim and Hyong-Woo Lee
Department of Electronics and Information Engineering,
College of Science and Technology,
Korea University, Sejong, Korea, 339-700
Email: (kimsy8011, hwlee)@korea.ac.kr

Choong-Ho Cho
Department of Computer and Information Science,
College of Science and Technology,
Korea University, Sejong, Korea, 339-700
Email: chcho@korea.ac.kr

***Abstract*—Inter-cell interference coordination (ICIC) is considered to be a promising technique for Long Term Evolution (LTE) to increase spectral efficiency. One efficient approach for ICIC is Soft Frequency Reuse (SFR). In this paper, we consider a call admission control in SFR-based systems and derive the call blocking and forced termination probabilities using a Markov chain analysis. In addition, SFR with spectrum handoff, called S-SFR, is proposed. Numerical results show that the analytic derivations are valid, and the proposed scheme outperforms the SFR scheme without spectrum handoff for the forced termination probability.**

Keywords- call admission control; inter-cell interference; soft frequency reuse; Markov chain; spectrum handoff.

I. INTRODUCTION

The Third Generation Partnership Project-Long Term Evolution (3GPP-LTE) system has focused on several interference management schemes for improving system performance. These schemes include an optimized frequency allocation policy based on semi-static radio resource management approaches, optimal power assignment and control schemes, and smart antenna techniques to null interference from other cells [2]. In particular, a Fractional Frequency Reuse (FFR) scheme has been proposed for 3GPP-LTE systems as an inter-cell interference coordination (ICIC) technique [3].

There are two major variants of FFR, which are static FFR and adaptive FFR. Static FFR includes pre-planned Frequency Reuse factor 1 (FR1) scheme, or FR3 scheme, or a combination of these schemes, such as Fractional Reuse Partitioning (FRP). Further improvements can be achieved by dynamically adapting FFR with techniques such as Soft Frequency Reuse (SFR).

In FRP and SFR, a cell is divided into two regions, namely the *cell-center zone* and the *cell-edge zone*. The *cell-center users* arriving in the cell-center zone utilizes the entire frequency band, whereas the *cell-edge users* arriving in the cell-edge zone operate in a sub-band using an FR 3 scheme, as shown in figure 1. Thus, the effective overall frequency reuse factor is still close to ensuring a high spectral efficiency [4]. SFR differs from FRP as follows. Because the cell-center users share bandwidth with neighboring cells, they typically transmit at lower power levels than the cell-edge users in SFR,

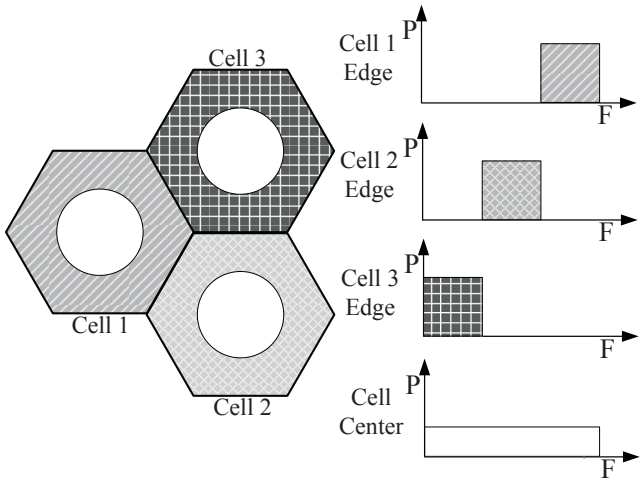


Fig. 1: Frequency-Power arrangement of SFR scheme. P and F denote the power and frequency, respectively.

as shown in figure 1. It is possible that when no resources for the cell-edge users are available, the eNodeB (base station) forcibly terminates (preemption) one of the cell-center users that occupies resources of a sub-band with an FR 3 if any such user exists[1].

In [1], a Markov chain model with 3-dimensional state variables was proposed to describe the call admission control (CAC) in SFR-based LTE systems. Throughout this model, the authors evaluated the system performance in terms of the call blocking probability and the forced termination probability. When comparing SFR with static FFR using FR 3, the call blocking probability of SFR is lower than that of static FFR. However, SFR suffers from a non-zero forced termination probability. From the user's point of view, the forced termination of an ongoing call is significantly less desirable than blocking of a new call attempt [5].

In this paper, we propose a SFR with the spectrum handoff technique, called S-SFR. For the spectrum handoff technique, when cell-center users using cell-center resources are released,

 ISBN: 978-1-61208-330-8

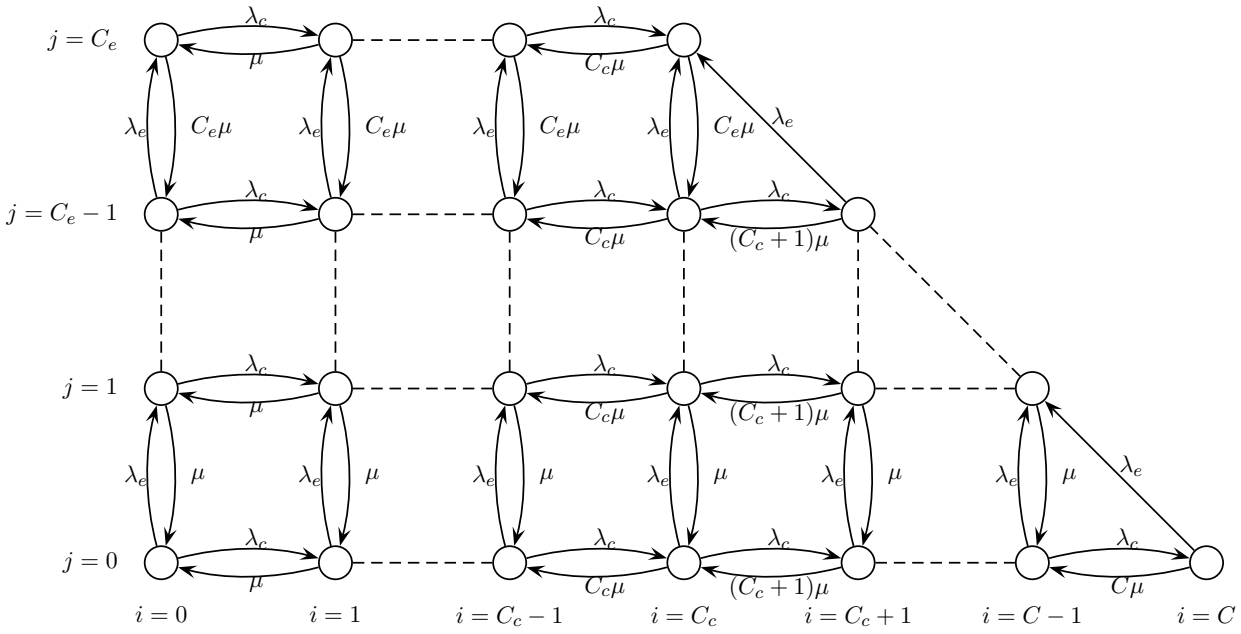


Fig. 2: 2-dimensional Markov chain model for S-SFR.

TABLE I: SUMMARY FOR VARIOUS SCHEMES.

Technique	S-SFR	SFR	FRP
Preemption	Yes	Yes	No
Spectrum handoff	Yes	No	No

a cell-center ongoing call occupying the cell-edge resource can be reconnected using the cell-center resource. Thus, by reducing the number of the cell-center users utilizing cell-edge resources, we can reduce the amount of forced termination of the cell-center users. Specifically, we describe the CAC for S-SFR and provide a 2- dimensional Markov chain analysis, where the state variables are the number of cell-center users and cell-edge users. This scheme is similar to channel reservation used in circuit-switched networks [8].

Additionally, we compare the performances of S-SFR, SFR, and FRP schemes, where each scheme is summarized as shown in Table 1.

This paper is organized as follows. In section 2, we present the system model. In section 3, we introduce an analytic model for the evaluation of performance of SFR based LTE systems. Section 4 presents numerical results, and concluding remarks are given in section 5.

II. SYSTEM MODEL

A. Model Description

An SFR-based LTE system is considered. There is one eNodeB in the center of the cell. In this paper, user equipments (UEs) located in the cell-center zone or the cell-edge zones are called *cell-center UEs* or *cell-edge UEs*, respectively. A number of UEs can initiate multiple calls trying to occupy radio resources, where the basic unit of radio resources is referred to as Physical Resource Block (PRB).

We assume that there are C PRBs that consists of C_c *cell-center PRBs* and C_e *cell-edge PRBs*, where $C_c = C - C_e$. For the sake of simplicity, we assume that the eNodeB can only assign one PRB to the UE at the time it initiates a call and, via an appropriate power allocation, the data rate of each call with one PRB is fixed regardless of its location within the cell.

As mentioned in [1], the data rate of the UE can be maintained using two approaches. First, more than one PRB can be allocated to the UE. Second, appropriate powers are allocated to the PRBs. Thus, the co-channel interference from adjacent cells can be minimized. Hence, the data rate of the UE can be guaranteed. In our model, it is assumed that, via an appropriate power allocation scheme, the data rate of each UE with one PRB is fixed regardless of its location within the cell.

B. New Call Arrival Processes and Call Duration Time

Traditionally, since CAC schemes have been based on call-level QoS measures, such as call blocking and dropping probabilities for voice or data, we assume that the eNodeB serves voice traffic call [6,7]. We also assume that the new call arrival process within a cell is a Poisson process with a mean request rate of λ calls/sec, and the UEs are uniformly distributed over the cell. Let ω be the ratio of the area of the cell-edge zone to the total area of the cell. The new call arrival rates of the cell-center UE and cell-edge UE are assumed to be $\lambda_c = (1-\omega)\lambda$ and $\lambda_e = \omega\lambda$, respectively. The call duration time is assumed to be exponentially distributed with mean μ^{-1} sec.

 ISBN: 978-1-61208-330-8

III. SOFT FREQUENCY REUSE WITH SPECTRUM HANDOFF

A. Call Admission Control

When a new call from the cell-center UE (*cell-center call*) is initiated, it will attempt to occupy a cell-center PRB. If one or more cell-center PRBs are available, it will be admitted. If no cell-center PRB is available, then it will further attempt to occupy a cell-edge PRB. It will be blocked once no cell-edge PRB is available. When a cell-center call occupying a cell-center PRB is released, a cell-center ongoing call occupying a cell-edge PRB is reconnected using a released PRB. This is called *spectrum handoff*.

When a new call from the cell-edge UE (*cell-edge call*) is initiated, it will attempt to occupy a cell-edge PRB. At this time, if all cell-edge PRBs are in use by cell-edge ongoing calls, the call will be blocked. If at least one cell-edge PRB is available, it will be admitted. If no cell-edge PRB is available, but one or more cell-edge PRBs are occupied by cell-center ongoing calls, one of the cell-center calls is randomly chosen and forced to release the cell-edge PRB it is occupying. The released PRB is then assigned to the newly arriving cell-edge call. The probability that the cell-center call is forcibly terminated is referred to as the *forced termination probability*.

B. Traffic Analysis

Assuming the characteristics of traffic, the process of PRB occupation can be modeled as a continuous time Markov chain. For CAC of S-SFR, the state transition diagram is described by an integer pair $(i,\ j)$, as shown in figure 2, where i and j denote the *number of cell-center calls* and *cell-edge calls*, respectively. As the cell-edge UEs have priority to utilize the cell-edge PRBs, the cell-center UEs utilizing the cell-edge PRBs can be preempted by the cell-edge UEs. Depending on the existence of the cell-center UE occupying the cell-edge PRB, a forced termination in state $(i,\ j)$ can move the state to $(i-1,\ j+1)$, where i is greater than C_c. This is because there are only cell-center UEs occupying cell-edge PRB by spectrum handoff.

Let $P(i,\ j)$ be the state probability. From figure 2, the following set of balance equations can be obtained:

(i) Four extreme points:

$$(\lambda_c+\lambda_e)P(0,0)=\mu P(1,0)+\mu P(0,1)$$
$$(\lambda_c+C_e\lambda_e)P(0,C_e)=\lambda_c P(0,C_e-1)+\mu P(1,C_e)$$
$$(C_c+C_e)\mu P(C_c,C_e)=\lambda_c P(C_c-1,C_e)+\lambda_e P(C_c,C_e-1)+\lambda_e P(C_c+1,C_e-1)$$
$$(\lambda_e+C\mu)P(C,0)=\lambda_e P(C-1,0)$$

(ii) $i=0,\ 0<j<C_e$:

$$(\lambda_c+\lambda_e+j\mu)P(0,j)=\lambda_e P(0,j-1)+(j+1)\mu P(0,j+1)+\mu P(1,j)$$

(iii) $0<i<C,\ j=0$:

$$(\lambda_c+\lambda_e+i\mu)P(i,0)=\lambda_c P(i-1,0)+\lambda_e P(i,1)+(i+1)\mu P(i+1,0)$$

(iv) $0<i<C_c,\ j=C_e$:

$$(\lambda_c+i\mu+C_e\mu)P(i,C_e)=\lambda_c P(i-1,C_e)+\lambda_e P(i,C_e-1)+(i+1)\mu P(i+1,C_e)$$

(v) $0<i<C,\ 0<j<C_e$:

$$(\lambda_c+\lambda_e+i\mu+j\mu)P(i,j)=\lambda_c P(i-1,j)+\lambda_e P(i,j-1)+(i+1)\mu P(i+1,j)+(j+1)\mu P(i,j+1)$$

(vi) $C_c<i<C,\ 0<j<C_e,\ i+j=C$:

$$(\lambda_c+i\mu+j\mu)P(i,j)=\lambda_c P(i-1,j)+\lambda_e P(i,j-1)+\lambda_e P(i+1,j-1) \quad (1)$$

$P(i,\ j)$ can be found by solving the balance equations together with the following normalization condition:

$$\sum_i\sum_j P(i,j)=1. \quad (2)$$

C. Performance Measures

As performance measures, we consider the aggregate call blocking and forced termination probabilities.

From $P(i,\ j)$, the call blocking probabilities of the cell-center UE and the cell-edge UE are, respectively,

$$P_{B_c}=\sum_{i=C_c}^{C}P(i,C-i),\qquad P_{B_e}=\sum_{i=0}^{C_c}P(i,C_e). \quad (3)$$

From (3), we can calculate the aggregate call blocking probability as follows

$$P_B=(1-\omega)\cdot P_{B_c}+\omega\cdot P_{B_e}. \quad (4)$$

For the forced termination probability, P_f, it is the total UE forced termination rate divided by the total UE connection rate. That is,

$$P_f=\frac{\lambda_e\sum_{i=C_c+1}^{C}P(i,C-i)}{\lambda(1-P_B)}. \quad (5)$$

IV. NUMERICAL RESULTS

In this section, we present the simulation results and compare them with our analysis. For all results, it is assumed that $C=48$ and $C_e=C/3$. In figures 3 and 4, the curves are numerically obtained from the equations given in the preceding analysis, whereas the symbols indicated the corresponding simulation results.

Figures 3(a)-(c) show the call blocking probabilities of the center-, edge-, and aggregate-UE with respect to different values of ω. These numerical examples show that the results of our analysis closely approximate those of the simulations. In figures 3(a)-(c), as the value of ω decreases, the call blocking probabilities of the edge- and the aggregate-UE tend to decrease, whereas the blocking probability of the center-UE increases. This is because as ω decreases, the number of UEs arriving in the edge area decreases, and thus the number of

 ISBN: 978-1-61208-330-8

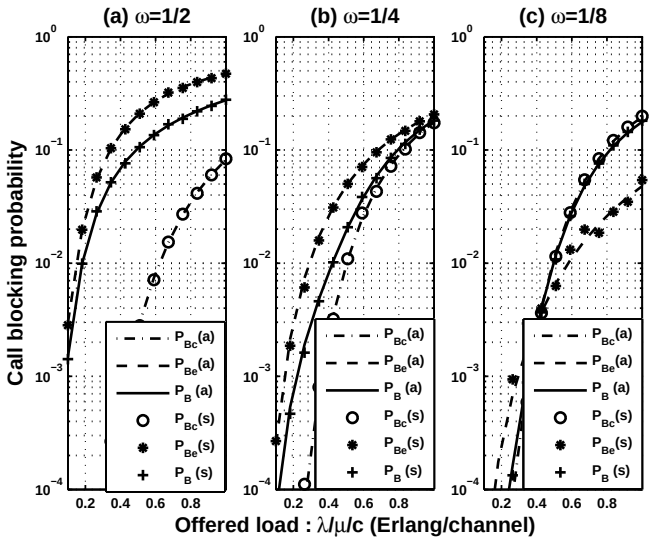


Fig. 3: Call blocking probabilities versus offered load with various ω.

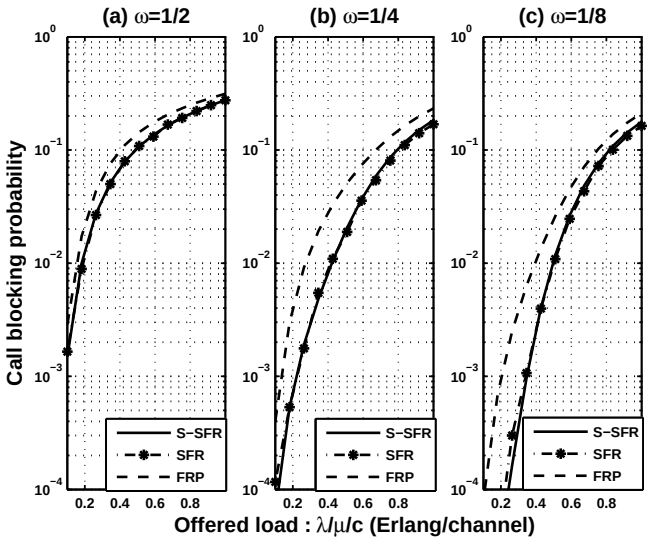


Fig. 5: Call blocking probabilities versus offered load with various schemes.

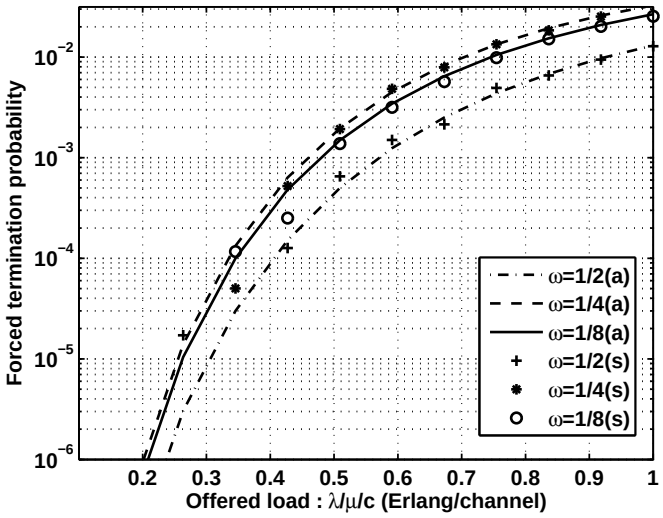


Fig. 4: Forced termination probabilities versus offered load with various ω.

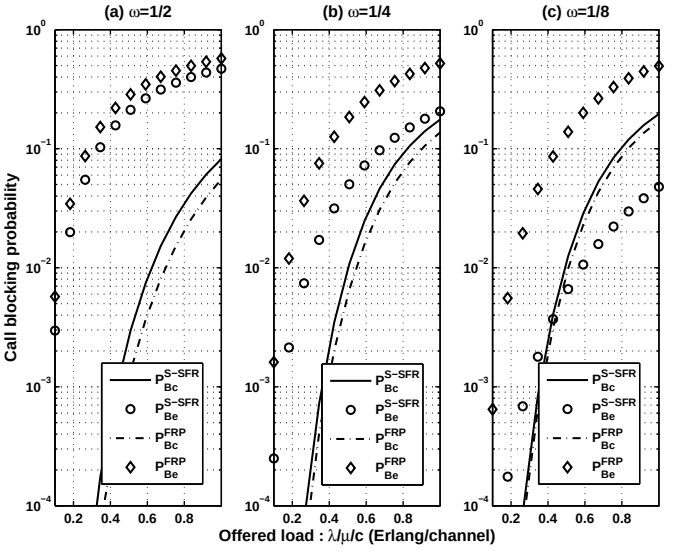


Fig. 6: Cell-center and cell-edge call blocking probabilities versus offered load with various schemes.

blocked cell-edge calls decreases. Figures 3(a)-(c) also show that as the value of ω increases, the difference of the call blocking probabilities between the center- and the edge-UE increases rapidly. We note that the center- and edge-calls are not evenly blocked. The reason is that the cell-edge PRBs are more heavily utilized than the cell-center PRBs.

Figure 4 shows the forced termination probability for S-SFR. It is observed that the results of the mathematical analysis agree reasonably well with those of the simulations. Figure 4 also shows that P_f increases as ω decreases. This is because as ω decreases, the number of cell-center UEs increases, thus leading to the decrement of terminated cell-center calls. When ω is very small, P_f decreases, because the number of cell-edge UEs is reduced.

Additionally, we compare the performance of S-SFR, SFR, and FRP schemes.

Figures 5(a)-(c) show the effect of ω on the aggregate call blocking probabilities for S-SFR, SFR, and FRP. As ω decreases, the aggregate call blocking probabilities of S-SFR, SFR, and FRP tend to decrease. We observed that the call blocking probabilities of both S-SFR and SFR are less than that of FRP. This is because the call blocking probability of the cell-edge UE in SFR-based schemes is decreased by using the forced termination of the cell-center UE using a cell-edge PRB.

Figure 6 shows the cell-center and cell-edge call blocking

 ISBN: 978-1-61208-330-8

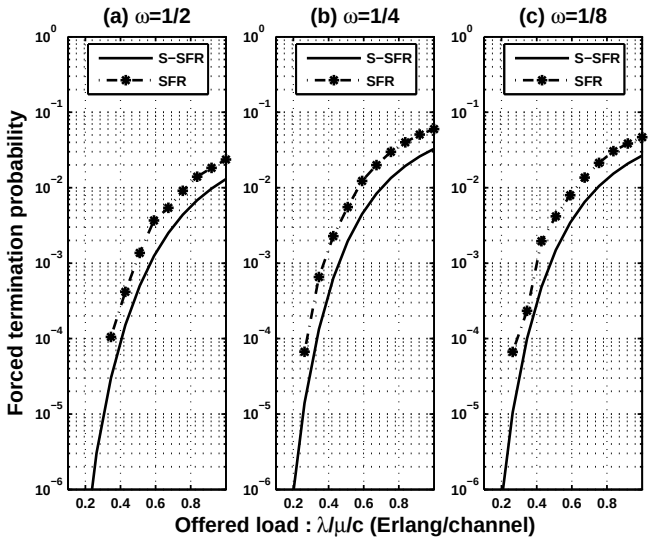


Fig. 7: Forced termination probabilities versus offered load with S-SFR and SFR.

probabilities for S-SFR and FRP schemes. From this figure, we note that the performance of the cell-edge call is improved for S-SFR by using the channel assignment as forced termination technique.

Figure 7 shows the effect of the spectrum handoff technique on the forced termination probabilities for S-SFR and SFR. Because FRP does not consider the forced termination technique, P_f is 0. From this figure, we note that the forced termination probability of S-SFR is less than that of SFR.

V. Conclusion and Future Work

In this paper, we considered a call admission control in SFR-based systems and derived the call blocking and forced termination probabilities using a Markov chain analysis. In addition, SFR with spectrum handoff, called S-SFR, have proposed. The analytical results show good agreement with the simulations. Numerical comparisons among S-SFR, SFR, and FRP schemes have shown that there are differences in the call blocking and forced termination probabilities. By using the forced termination technique, we have shown that S-SFR and SFR have decreased the call blocking probability. For cell-edge calls, these schemes provide an improved call blocking probability. We have also shown that by using spectrum handoff, the forced termination probability of S-SFR is less than that of SFR.

One of the possible research topics is to consider a SFR-based cellular system in the interference scenario. In reality, the system throughput may be calculated by the signal to interference ratio, depending on the level of interference power received from neighboring cell. Therefore, it is worthwhile to study the cases where one UE or BS may have interference signals from neighboring BSs or other UEs.

Acknowledgment

This research was supported by BK21 plus.

References

[1] S-P. Chung and Y-W. Chen, “Performance Analysis of Call Admission Control in SFR-Based LTE Systems,” IEEE commun. Lett., vol. 16, Jul. 2012, pp. 1014-1017.

[2] S-Y. Kim, S. Ryu, C-H. Cho, and H-W. Lee, “Performance analysis of a cellular network using frequency reuse partitioning,” Perform. Eval., vol. 70, no. 2, Feb. 2013, pp. 77-89.

[3] N. Himayat, S. Talwar, A. Rao, and R. Soni, “Interference management for 4G cellular standards [WIMAX/LTE update],” IEEE Commun. Mag., vol. 48, no. 8, Aug. 2010, pp. 86-92.

[4] T. Novlan et al., “Analytical Evaluation of Fractional Frequency Reuse for OFDMA Cellular Networks,” IEEE Trans. Wireless Commun., vol. 10, no. 12, Dec. 2011, pp. 4294-4305.

[5] S. Tekinay and B. Jabbari, “Handover and channel assignment in mobile cellular networks,” IEEE Commun. Mag., Nov. 1991, pp. 42-46.

[6] T. L. Kahwa and N. D. Georganas, “A hybrid channel assignment scheme in large-scale, cellular structured mobile communication systems,” IEEE Trans. on Commun., vol. 26, no. 4, Apr. 1978, pp. 432-438.

[7] D. Niyato and E. Hossain, “Call admission control for QoS provisioning in 4G wireless networks: issues and approaches,” IEEE network, vol 19, no. 5, Sept.-Oct. 2005, pp 5-11.

[8] K.-M. Chan and T.-S. P. Yum, “The maximum mean time to blocking routing in circuit-switched networks,” IEEE J. Sel. Areas Commun., vol. 12, no. 2, Feb. 1994, pp. 313-321.

 ISBN: 978-1-61208-330-8

New IPv6 Identification Paradigm: Spreading of Addresses Over Time

Florent Fourcot, Laurent Toutain
Frédéric Cuppens and Nora Cuppens-Boulahia
Institut Mines-Télécom; Télécom Bretagne
Université européenne de Bretagne
Email: {first}.{last}@telecom-bretagne.eu

Stefan Köpsell
TU Dresden; Faculty of Computer Science
Email: stefan.koepsell@tu-dresden.de

***Abstract*—The identification of packet flows is a very important feature to provide security on the Internet. This flow identification is traditionally done by the well-know five tuple source IP address, destination IP address, transport layer protocol number and the two source/destination identifiers of transport layer protocols (named ports on UDP and TCP). Unfortunately, the IP source address is not reliable at all. However, we can use new security paradigms based on new IPv6 properties. In particular, IPv6 introduces a large address space. Our solution takes the benefit of this space with a high frequency rotation of IP addresses, that we call *spreading*. This spreading improves the security since only the sender and the receiver are able to generate and follow this temporal sequence. An attacker will not be able to successfully insert malicious packets into a flow or to initialize a flow. It protects against session initialization flooding and against attacks on established connections. In this paper, we describe the architecture of our solution and the protocol to initiate a connection and also performance evaluation of our spreading.**

Keywords–IPv6; security; flow identification; spoofing.

I. INTRODUCTION

A. Flow identification in the current Internet

Flow identification is the base of some Internet mechanisms, like security (filtering of packets) and priority policies for Quality of Service. But since the Internet is a datagram network and IP is not a connection oriented protocol, the notion of flow is not explicit at the network layer. Each packet is independent, and two similar packets can follow two different routes. It is why a flow is defined in the RFC 2722 [1] as "an artificial logical equivalent to a call or connection". It is "artificial", and there are no easy ways to discriminate a flow in an IP network.

At the transport layer, the concept of flow is more natural. This is a mandatory function to sort packets, reassemble segments and detect errors, like the popular protocol Transmission Control Protocol (TCP) [2] does. To have this notion of flow, one can use Transport Layer addresses, named ports in case of TCP.

This is why we need a tuple of five elements in order to extract the notion of flow on the network. The first two are the source's and destination's IP addresses, directly available in the IP headers. The next one is the transport layer number, available in the field `next header` for a Internet Protocol version 6 (IPv6) packet without extensions. With the knowledge of the transport protocol, it is possible to parse the transport header and also to read the port numbers. They complete the tuple, already filled with the three identifiers of IP header.

This well known five tuple is the basic identification of a flow. The identification can be more complex. For example, a stateful firewall will follow the TCP states of each connection, and it will discard packets if they do not follow the TCP standard.

B. Address spreading benefits

These five members of the identification tuple $(IP_{src}, IP_{dst}, NextHeader, Port_{src}, Port_{dst})$ are not authentic by nature. The source IP address and the source port especially can be manipulated easily by an attacker. If an attacker is able to send packets with a spoofed source IP address, he will be in good position to try TCP reset attacks, to inject packets on the destination network, to try a targeted attack to the destination, etc.

With IPv6, the large IP address space allows new security opportunities, like Cryptographically Generated Address (CGA) [3]. If all IPv4 solutions had to minimize the number of IP addresses in use, it is now possible to use a lot of addresses. Our solution provides security thanks to the *spreading* of addresses. In our solution, source and destination IP addresses of a flow are renewed frequently, according to a temporal sequence. If this sequence is known only by the sender and the receiver, it adds a new identification feature.

Since we only modify IP addresses, our solution is pretty simple. It does not need some complex encapsulation (like IPsec tunnel does), and can be followed by a firewall with the knowledge of a shared secret.

C. Attacker model

Our attacker can inject some traffic with a spoofed source IP address. He can be on the transmission path and also able to read the legitimate traffic.

We do not try to protect against a rebuilding of a flow. This means that the attacker can use upper layers information like TCP ports and sequence numbers to rebuild the real flow. Our protection is against the spoofing: our spreader will recognize packets from the attacker. However, we provide a protection against correlation of flows, since we obfuscate addresses. An attacker can not guess the real source and destination addresses of a flow, and can not group several flows to one source or destination just with the help of the information available at the network layer.

 ISBN: 978-1-61208-330-8

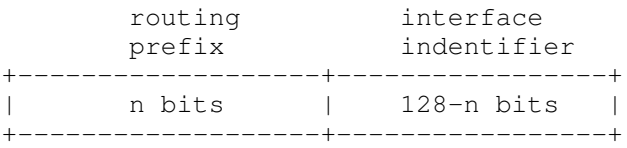


Figure 1: The two parts of IPv6 addresses

D. The two IPv6 address parts

An IPv6 address is divided in two parts, depicted in Figure 1. The routing part (usually called "prefix") is not rewritable without connectivity issues, since such modification on the destination address will prevent upstream routers to send packets to the destination. In the same way, the source address has to be compliant with the anti-spoofing rules of the RFC 2827 [4] and could not be arbitrary rewritten.

There is no recommended size of prefix for a end network [5], but the length can not be more than 64 bits for compatibility with the IPv6 autoconfiguration. Indeed, the IPv6 Stateless Address Autoconfiguration (SLAAC) derives the second part of the address (named interface identifier) from the hardware MAC address to an identifier of 64 bits. A prefix larger than 64 bits breaks this autoconfiguration.

We choose to rewrite only the last 64 bits of each address, giving a total of 128 bits. Since 64 bits is the maximum size for a prefix compatible with autoconfiguration, this value will be compatible with almost all networks.

E. Related work

There is some previous work on dynamical IPv6 addresses. The nearest of our work is the "Moving target IPv6 Defense" publication [6]. This solution uses an User Datagram Protocol (UDP) tunnel to often rotate addresses and to encapsulate the real IP packet. An encryption is optional to protect the payload of the packet. The main drawback of this paper is the choice to make an encapsulation. It means that the solution has to fragment big packets to prevents Maximum Transmission Unit (MTU) problems and to reassemble it at the end of the tunnel. On the other hand, additional headers can have a bandwidth performance impact. This proposition does not use a temporal window for old address to prevent false positive.

An other idea is the publication "An Architecture for Network Layer Privacy" [7]. It uses an Site Multihoming by IPv6 Intermediation (SHIM6) extension to spread addresses over time. Since SHIM6 is an end-to-end solution, the end device has to allocate all possible addresses before using the connection. It is probably not desirable that a computer allocates 1000 addresses on one network interface, because it will send a lot of Neighbor Discovery packet.

An other paper [8] uses SHIM6 for a protection against deny of services. The idea is to switch from an address (under attack) to another one (not under attacks) for all connections already established.

F. Organization of this paper

In this paper, we first presents some problems of address spreading and our solution of architecture to overcome them

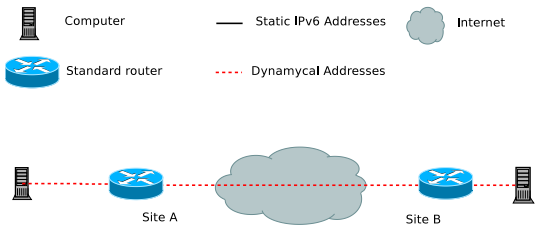


Figure 2: Spreading of addresses without extra device

in Section II. Second, we introduce principles of the spreading and some notations in Section III. We describe step by step the initialization of a connection in Section IV, that we complete with a description of packet processing on spreaders in Section V. We introduce the theoretical source of performance issues in Section VI. We close the paper with the description of performance evaluation in Section VII.

II. Localization of the spreader

A. Difficulties to spread addresses on the end devices

The first idea of spreading is to generate and follow the address sequence on the end devices (see Figure 2). This strategy allows a end-to-end security, and the computer does not need to delegate the security to someone.

However, this solution implies an upgrade of the local router. The main issue arises if the end device does not get a delegated prefix, but share the local prefix with several end devices. The use of a lot of addresses will:

- flood the network with Neighbor Discovery packets. The router is not aware of the spreading, and can not know the link between the temporal sequence of IP addresses and the static MAC address;
- saturate the Neighbor table of the router. With frequently switching of addresses, the router will not be able to store all mappings between IP addresses and MAC addresses;
- introduce a latency at each IP address switching, due to the Neighbor Discovery.

To solve these issues, a solution is to patch the router to follow the sequence of IP destination addresses in the Neighbor table. Thus, the router does not need to know the sequence of source addresses received from the Internet, and is not able to insert a packet in a flow.

B. The prefix delegation solution

With IPv6, we could have enough addresses to delegate one address prefix to each end device. It solves the problem of the mapping between MAC addresses and IP addresses, since the intermediate only send all IP packets matching a prefix to a MAC address.

In this case, the end device is in charge of Interface Identifiers management and can actually be seen as a "router", and it is the same architecture at the network layer shown in Figure 3. The delegation of address prefixes is the best

 ISBN: 978-1-61208-330-8

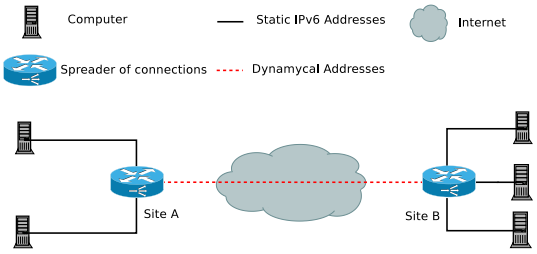


Figure 3: Architecture of the solution: spreader on the border of the local network

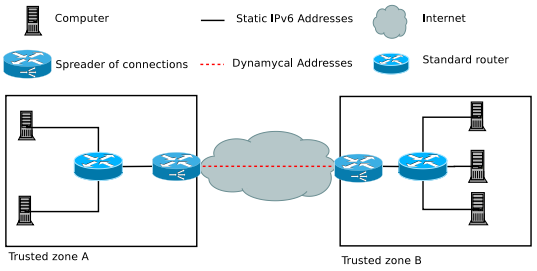


Figure 4: Architecture of the solution: spreader at the border of the trusted zone

architecture for the simplicity of the solution and from a security point of view. However, we need to provide a solution for a standard network without delegation of prefixes.

C. Simplification with a spreader on the path

For a first approach, we propose to simplify the problem by adding some new "spreaders", devices on the path of the communication. These devices are able to rewrite a flow of packets with stable addresses to a flow of packet with dynamical addresses. The spreader can be directly on the border of the network (see Figure 3) or at the border of the trusted zone (see Figure 4).

The first positive argument for this architecture is the simplicity to deploy the solution. An administrator does not need to upgrade and configure each end device, but can simply insert the spreader in the network. It has the same benefit than a modified router following the relation between IP and MAC address, and less drawback.

The second point is the ability of bad packets filtering. Since malicious packets consume resources, and can be send to simply saturate the bandwidth of the network, we have to discriminate malicious packets as soon as possible. With an introduction of the spreader at the border of the trust zone, we achieve efficiently this goal.

We choose this architecture for this paper to simplify concepts and experimentations.

III. General Principles

A. Prerequisite of the solution

To enable the spreading, we need at least to configure two networks, for the mapping between the dynamical addresses and stable identifiers.

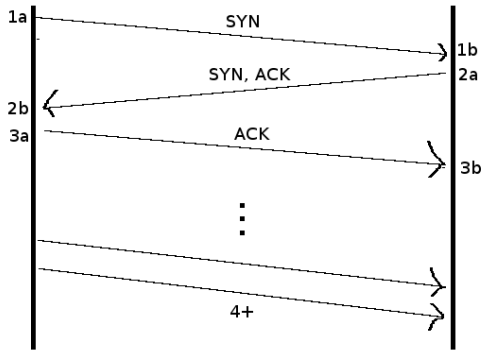


Figure 5: Digital timing diagram of a connection initialization

This configuration is done by adding one spreader at the border of each network. The two spreaders have to share a secret, that an attacker can not guess. The communication of this secret is out-of-scope of this paper.

B. Initialization of the spreader

The initialization of the spreader has to create a configuration for each compatible peer with a shared secret. This configuration contains the prefix list of the destination (to catch packets to be rewritten) and a function to derive cryptographic keys from the shared secret.

C. Exchange of session data

One of our goals is to spread each flow of data with a unique sequence of addresses, make it more difficult for an attacker to group all flows of one end device. To provide it, both spreaders have to exchange session data at each flow initialization. There are several ways to accomplish it. The first one is to add several extras packets to initiate a context for each flow. It increases latency of connection initialization, and costs some bandwidth.

The second one is to add extra information on a real packet, like by adding one extra IPv6 extension header. Since this extension will be added by the spreader and not by the end devices, it could result in some maximal transport unit problem. Indeed, we can not add this header on a big packet, and have two choices. The first one is to fragment the packet on the spreader, IPv6 RFCs do not allow this solution. The second one is to send a "too big" error to the end devices, which reduce the performance of all packets for the session.

These solutions are not satisfying. We choose another one, our spreaders encode all information in IPv6 source and destination addresses, and do not add any extra data to packets payload. It limits the amount of exchangeable data, but does not have any cost of bandwidth or latency due to extra packets.

D. Notations

The description of the protocol follows the same steps than a TCP connection initialization, depicted in Figure 5.

 ISBN: 978-1-61208-330-8

We introduce the following notation for the packet rewriting (summarized in Table I): P_A and P_B are the prefixes of networks for hosts A and B. IP_A and IP_B are the real IP addresses of host A and B, concatenation of prefixes and interface identifiers IID_A and IID_B. IP^n_{src} is the rewritten source IP address of the packet in step n, and IP^n_{dst} the destination IP address in the same step. Since we can not rewrite the prefixes, IP^n_{src} and IP^n_{dst} are concatenation of one stable prefix (P_A or P_B) and one rewritten value.

TABLE I: IP ADDRESSES NOTATION

	Local network		On the Internet	
Steps	Real IP Source	Real IP Destination	Rewritten IP source	Rewritten IP Destination
1a $\rightarrow$ *1b* SYN	IP_A	IP_B	$IP^1_{src} = P_A \vert IID^1_{src}$	$IP^1_{dst} = P_B \vert IID^1_{dst}$
2a $\rightarrow$ *2b* SYN, ACK	IP_B	IP_A	$IP^2_{src} = P_B \vert IID^2_{src}$	$IP^2_{dst} = P_A \vert IID^2_{dst}$
3a $\rightarrow$ *3b* ACK	IP_A	IP_B	$IP^3_{src} = P_A \vert IID^3_{src}$	$IP^3_{dst} = P_B \vert IID^3_{dst}$

IV. STEP BY STEP INITIALIZATION

A. Initialization of a connection - first packet

1) Symmetrical rewrite on spreaders: The rewriting begins with step *1a*, when the spreader receives a packet with an destination IP address matching one of the prefix in the spreader configuration.

At the first packet of a connection, the local spreader computes new source and destination IP addresses with the help of a cryptographic function. We choose the Advanced Encryption Standard (AES) encryption, but it could be another one allowing encryption of blocks length of 128 bits.

The AES function takes as input a block of both interface identifiers of hosts A and B, and the actual key $K(t)$ derived from the shared secret for the encryption. This AES function has an output of 128 bits, that the spreader divides in two blocks of 64 bits to replace last 64 bits of IP_A and IP_B.

$$IID^1_{src} = AES(IID_A|IID_B, K(t))[0-63] \quad (1)$$

$$IID^1_{dst} = AES(IID_A|IID_B, K(t))[64-127] \quad (2)$$

After the rewriting of addresses, the packet follows the standard routing and filtering process. This ends step *1a*.

On the destination spreader, stable addresses are recomputed in the same way (AES is a symmetric function) in the step *1b*. After the computing, the destination spreader checks the validity of the transport layer checksum (this checksum is mandatory for UDP and TCP with IPv6). If the checksum is valid, the packet follows the standard policy of routing and filtering.

If the checksum value is not valid, it can of course be a sign of transmission problem. One other possible cause of this invalid checksum is a try of an attacker to inject one packet in the network, with a spoofing of the source address. Indeed, IPv6 addresses are part of the checksum computation and if addresses after the second spreader are not the same than addresses sent by the source device, it invalids the checksum. Since the attacker does not know the shared secret, he can not compute the AES encryption and the generated packet will be detected by the spreader.

In more details, the Table II depicts the status of the checksum around the Internet. It looks invalid between the two spreaders, but this is not a problem since nobody needs to have a look on the checksum in the path of the communication. On the contrary, we see in Table III that the checksum of a spoofed packet looks valid on the Internet, but invalid after the rewriting of the second spreader.

TABLE II: CHECKSUM VALIDITY FOR A REAL PACKET

Position	Checksum validity	Remarks
Local Network	Valid	Computed by the end device
Internet	Invalid	Addresses have been rewritten, it invalids the checksum
Remote Network	Valid	Addresses have been rewritten by the second spreader

TABLE III: CHECKSUM VALIDITY - SPOOFED PACKET

Position	Checksum validity	Remarks
Attacker's Network	Valid	Computed by the attacker
Internet	Valid	No rewriting if the attacker is not aware of protection
Remote Network	Invalid	Addresses have been rewritten by the second spreader

2) Security analyze of the rewriting: If the attacker is aware of this spoofing protection, he can try to guess the checksum modification added by the AES encryption of spreaders. The length of the checksum field is 16 bits, which it gives one chance on 65 536 to find the good one. This value is only valid for a short time and for a given address couple, the next value of $K(t)$ at $t+1$ will give another checksum modification implied by the AES encryption.

This security mechanism is not good enough to filter all packets of an attacker, and some packets can bypass this protection. But it is important to notice that if the checksum is valid, the attacker can not guess the rewritten addresses and can not know what is the rewritten destination address. The chance to successfully contact a real computer with a valid address is very low. Indeed, if we assume that the rewriting is fully random, an attacker has first to bypass the checksum (one chance on 65 536). If the checksum is valid, a targeted attack on a computer with an address on the remote network has one chance on 2^{64} to reach the targeted computer, since an attacker can not guess the rewritten value after the AES decryption.

B. Initialization of the connection - reply of the remote

The goal of the addresses rewriting on the first packet is to protect an initialization of a connection by an attacker. For the

 ISBN: 978-1-61208-330-8

TABLE IV: REWRITING IN INITIALIZATION STEPS

Steps	Rewritten IID Source	Rewritten IID Destination
1a → *1b* SYN	$IID^1_{src} = AES(IID_A \vert IID_B, K(t))[0-63]$	$IID^1_{dst} = AES(IID_A \vert IID_B, K(t))[64-127]$
2a → *2b* SYN, ACK	$IID^2_{src} = random()$	$IID^2_{dst} = g(t, secret, IID^1_{src})$
3a → *3b* ACK	$IID^3_{src} = g(t, secret, IID^1_{src})$	$IID^3_{dst} = g(t, secret, IID^2_{src})$

next packets, we create a pseudo-random sequence for each flow of data, generated by a function g. This function g is a generator of a random temporal sequence, one example is a hash function like SHA1.

This creation begins in step *2a*. To accomplish it, the second spreader rewrites the first reply packet of a client with a random value as IP source, and a value computed from the source IP address value in the first packet for the destination.

$$IID^2_{src} = random() \quad (3)$$

$$IID^2_{dst} = g(t, secret, IID^1_{src}) \quad (4)$$

g takes as input the current time, the shared secret between networks and another value with the same size than an IP address. This rewriting introduces a random value for the sequence, but the flow is still easy to identify for both spreaders with the IP address destination sets to a value that the first spreader can recognize.

In step *2b*, the first spreader recognizes the IP destination address IP^2_{dst} with the help of a context previously stored. This packet is an acknowledgment of the initialization, the spreading can now really begin. The spreader saves the value of the IP source (randomized in step *2a*) and rewrites the source and destination IP addresses to the real stable value stored in the context.

It ends the second step. The first spreader is now sure of the initialization of the connection, and can use the random value to bootstrap a new random sequence.

C. Initialization of the connection - Acknowledgment of the second spreader

The step *3a* begins at the next packet sent by the device from the first network. Both source address and destination address are now spread with:

$$IID^3_{src} = g(t, secret, IID^1_{src}) \quad (5)$$

$$IID^3_{dst} = g(t, secret, IID^2_{src}) \quad (6)$$

In step *3b*, the second spreader recognizes the couple with the help of the stored context. This packet is an acknowledgment of the random value send in step *2a*, and the second spreader is now aware of the success of the initialization. Steps of initialization are summarized in Table IV.

D. Rewriting during the life of the connection

After the step *3b*, both spreaders follow the same sequence of rewriting with $g(t, secret, IID^1_{src})$ and $g(t, secret, IID^2_{src})$. The rewriting is symmetrical and both end devices receive stable addresses. An attacker can not inject some traffic since he does not have the knowledge of the next addresses in use.

E. Definition of a flow, division of connections in several temporal sequences

Our goal is to create one sequence of address for each flow of data. By flow, we mean a sequence of packets where informations of upper layers are enough for an attacker to correlate and to rebuild the sequence.

The first trivial idea is to make one flow for each couple of IP_{src}, IP_{dst}. It does not take a lot of resources, but do not prevent correlation if more than one flow are send between devices. Nevertheless, it can be desirable to obfuscate this information.

IPv6 introduces a new header field to give information about a flow of packets: the flow label field [9]. It is a 20 bits header, that can be used for Quality of Service or other usages (the RFC 6294 [10] tries to give of survey of usages). This flow label can be rewritten on the path of the communication, and is not part of the checksum. Other usages than Quality of Service are allowed, which means that we respect standard specifications [9] with our proposition.

This is why we define a flow by a tuple of $(IP_{src}, IP_{dst}, flowlabel)$. If the end device set a different value for two flows, it will be spread into two different sequences. The flow label is set by the end device itself, which has enough information to know if a sequence of packets should not be grouped with another connection.

V. Detailed processing on spreaders

Our description of the protocol in Section IV is the ideal situation. We do not have any loss of packets, and the end devices use the TCP protocol. It helps to understand how our protocol works, since it follows the same handshake mechanism than TCP.

But we have to be robust against loss of packets and retransmission of data. In the same way, we have to be compatible with transport protocols, like UDP, where a connection does not follow a rigorous initialization procedure, or ICMP with short session like a simple Echo request. We detail in this section the processing of packets in spreaders, which actually support loss of packets and is transport layer protocol independent.

A. Detailed steps of packets processing (outgoing packets)

The processing of packets from the local network to the Internet is depicted in Figure 6. For each outgoing packet, the spreader extracts the tuple $(IP_{src}, IP_{dst}, flowlabel)$ (step 1). It checks if one context already exists (step 2) and if we already received an acknowledgment (step 3). If both conditions are true, we are in the case of an established connection and we can rewrite IP_{src} and IP_{dst} with the help of the function g

 ISBN: 978-1-61208-330-8

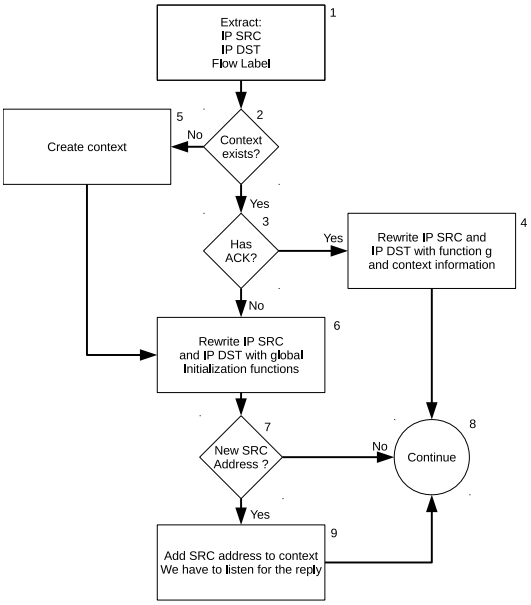


Figure 6: Processing of outgoing packet (to the Internet).

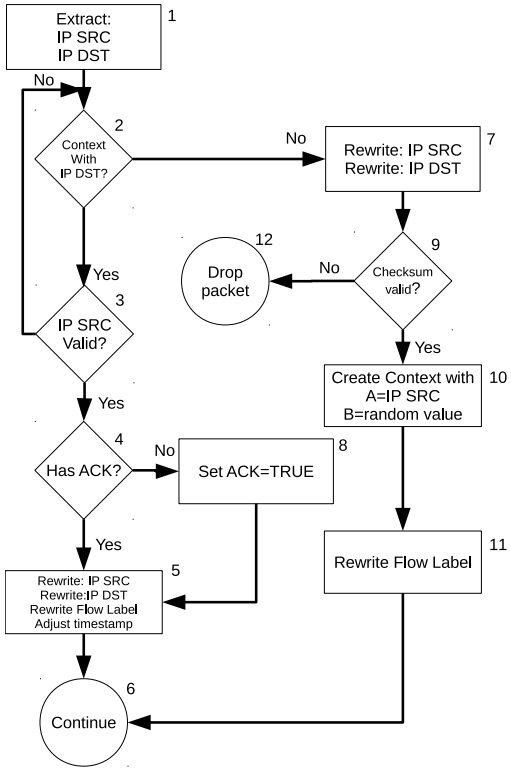


Figure 7: Processing of incoming packet (from the Internet).

and information stored in the context. After that, the packet continues the standard processing.

If the context does not exist, we have to create one. It contains the real IP source address and real IP destination address, as well as the flow label value (step 5). We move to step 6, the case of the context exists but we do not have yet any acknowledgment. We have to use AES encryption with $K(t)$ to rewrite IP addresses.

Since the rewritten IP_{src} is used as parameter for a reply from the remote network, we have to store it (step 7 and step 9). We can store more than one address if we send severals packets before we receive any acknowledgment. The packet follows afterwards the standard packet processing.

B. Detailed steps of packets processing (incoming packets)

The Figure 7 depicts the packet processing for all packets from the Internet. The processing of incoming packets begins with the extraction of the couple (IP_{src}, IP_{dst}). We do not extract the flow label, this value can not be trusted outside of the local network. This flow label will be rewritten to an internal value to make the future flow identification of local packets going to the Internet. The goal is to know if a context already exists for this connection (step 2 and 3). If this is the first packet for this context, it is an acknowledgment of the initialization and we have to change the status of the context (step 8). We end the processing with the rewriting of dynamical addresses to stable addresses and we return the packet for standard processing (step 6).

If we do not have a context, we have to try a decryption of IP addresses with the AES function and $K(t)$ in step 7. This decryption is followed by the computation and the verification of the transport layer protocol checksum in step 9. A bad checksum implies to drop the packet, since it is probably a try of an attacker to send a packet with a spoofed address. If it is valid, we initiate a context (step 10) and we return the rewritten packet for standard processing.

VI. Theoretical loss of packets

Our spreading solution drops packets if they are not following the sequence of addresses. This spreading protects against an attacker, but valid packets sent by the real device can be dropped. Indeed, the latency in the network is the main problem: if a packet is too long in transmission across the network, it will be drop by the receiver, since he had already switched to the next addresses.

The second source of problems is the time desynchronization between two spreaders: if the clocks are not synchronized, a valid packet will be detected as spoofed even if the sender is not an attacker.

In this section, we describe the theory of this packet loss. We first estimate the loss of packets in case of latency in the network with perfectly synchronized spreaders. Second, we add the problem of time desynchronization between spreaders. Next, we explore the consequences for a simple ICMP echo request/echo reply communication. We conclude with solutions to mitigate both problems.

A. Latency effect

The latency is the time needed for a packet to go from the source to the destination. The latency can be less than one millisecond on a local network (LAN), and several seconds between both points on the Internet. If we assume that the latency is stable for all packets and is the same in both

directions, it is easy to estimate the proportion of the packet loss. All packets sent at the end of the lifetime of one address will be drop by the receiver spreader. The duration of this black hole in the communication is exactly the value of the latency. If we consider that packets are uniformly send over time, we have them a packet proportion loss of:

$$loss = \frac{latency}{lifetime} \tag{7}$$

For example, with a configuration of 1 second for the lifetime of addresses, with 100 ms of latency on the network, 10% of packets will be dropped in both direction of the transmission between the spreader A and the spreader B.

B. Desynchronization effect

It is not easy to perfectly synchronize two computers on the Internet. Even with the Network Time Protocol, clocks of computers are not perfect and there are always a little desynchronization. In one way of the transmission, this desynchronization is good, since it reduces the observed latency between both computers. In the other direction, the latency is added to the latency and it implies a longer duration of black hole for the communication.

If we assume that the spreader A is desynchronized in the future with spreader B, the loss in the direction A to B is:

$$loss_{AB} = \frac{latency + desync}{lifetime} \tag{8}$$

In the other direction from B to A, the loss has been reduce to:

$$loss_{BA} = \frac{|latency - desync|}{lifetime} \tag{9}$$

The perfect case for this direction is when the latency is equal to the desynchronization, there are no more packet loss in this direction. If the desynchronization is bigger than the latency, some packets come too early to the spreader A (A did not yet switch to the "new" address) and packets are dropped.

With the same configuration of 1 second for the lifetime of addresses, with 100 ms of latency on the network, and a desynchronization of 50 ms, 15% of packets are dropped between A and B (150ms of black hole) and 5% between B and A.

C. ICMP echo request/echo reply communication

The evaluation of packet loss in both directions is not enough to evaluate the impact on a communication. On the Internet, a unidirectional payload transmission is very unusual. The most popular protocol TCP sends a lot of acknowledgment packets even for a unidirectional transmission, and a simple ICMP echo request is replied with an echo reply packet.

Our loss of packets is not distribute like a standard network error, each packet does not have a probability of failure, but the connection seems to be broken for a short duration, in one or two ways.

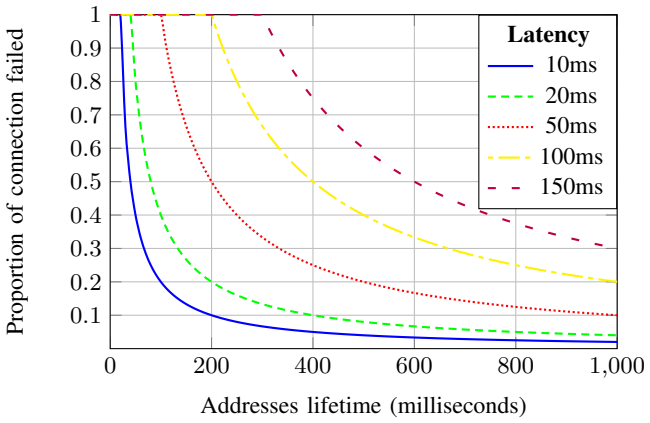


Figure 8: ICMP echo failure rate in function of latency

For example, in case of ICMP echo transmission between spreaders A and B, the total duration of one of packet echo reply/request will dropped is

$$latency * 2 + desync \tag{10}$$

We plot in Figure 8 the theoretical ICMP echo request/reply transmission failure for a given latency in function of address lifetime, without any desynchronization of spreaders.

D. Temporal windows one old and next addresses

To reduce the loss of packets, it is possible to accept the old address of time $t-1$ in a temporal windows where both current address t and $t-1$ addresses are accepted. With a temporal windows larger than the latency, no packets are dropped by synchronized spreaders.

In case of desynchronization smaller than the latency, we have to add this desynchronization duration to the temporal windows to accept all packets send in the communication. A spreader can not know if a packet is delayed due to the latency or due to a desynchronization problem.

If the desynchronization is larger than the latency, a spreader will receive packets to early. To solve it, we can add in the same way a temporal windows where both addresses of t and $t+1$ are valid. If a spreader receives a lot of packets on the $t+1$ address, it is a sign of desynchronization and it could help to resynchronize both spreaders. We present our experimental results on temporal windows to accept old and future addresses in Section VII-D.

VII. Performance tests

A. Test beds

1) Implementation: Our implementation is based on a Netfilter module for Linux. It follows steps explained in Section V. The implementation supports configuration of the validity lifetime of one address as well than temporal windows for packets out of the current time sequence.

 ISBN: 978-1-61208-330-8

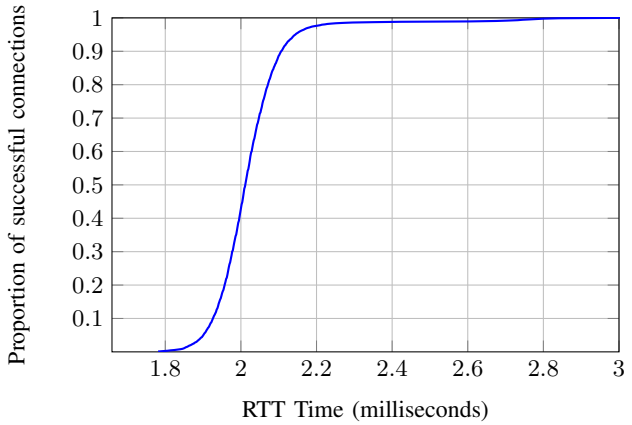


Figure 9: RTT of UDP echo requests on the LAN

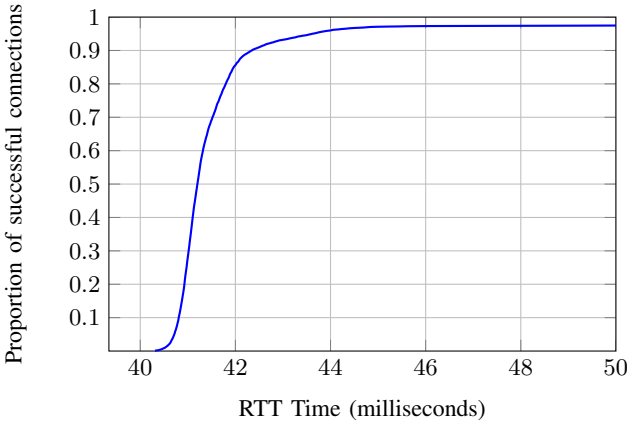


Figure 10: RTT of UDP echo requests on the 6to4 network

2) Networks tests: We have done several experimentations on several test beds, and this paper presents results from two of them. Our first test bed is the ideal one, in a LAN. The typical round-trip delay time (RTT) is around 2 milliseconds (ms), there are no loss or desequencing of packets. The Figure 9 depicts the cumulative distribution function of the RTT.

The second one is between a server in Germany using a 6to4 tunnel and a server with native IPv6 connectivity in France. The network has a poor quality: there are some natural packet loss (around 1%) and desequencing (around 0.5% on high network load). In our preliminary tests, the 6to4 tunnel uses to be less congested in the night, and we ran our tests in the night to avoid some random congestion issue. The Figure 10 depicts the cumulative distribution function of the RTT.

All devices of our networks are time synchronized on the same Network Time Protocol server. It does not provide a perfect synchronization.

B. Spreading consequences

1) Tests description: To evaluate the consequences of our spreading, we ran for each network three tests. The first one send a standard UDP echo packet, it provides a good evaluation of the network quality for the loss of packets and the latency. The second one is a simple TCP handshake initialization, without data transfer. The last one is a TCP connection with data transfer (65535 bytes). Since TCP is the most popular protocol on the Internet and that the test involves transfer of data, it is the best test to evaluate the user experience on a network with spreading. We set a timeout of 4 seconds on both TCP tests, and we consider it loss after this time.

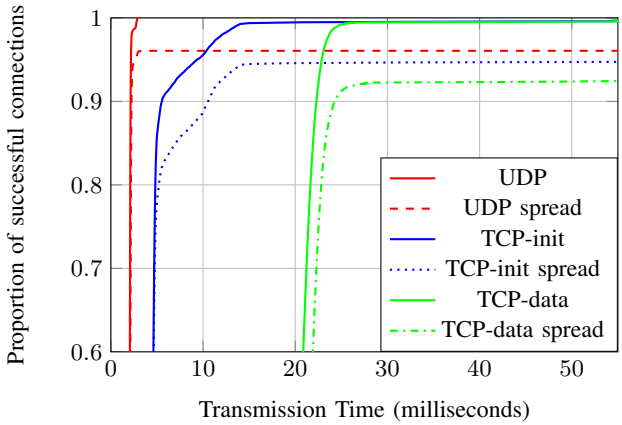


Figure 11: Consequences of a simple spreading on the LAN

2) Simple spreading: Since computers are not perfectly synchronized and due to the network latency, our spreading implies some loss of packets and it has consequences on the network quality. We set first a lifetime of one second for each address, without any temporal windows for old and future addresses. The Figure 11 depicts the results on the LAN and Figure 12 on the 6to4 network.

On the LAN, there are no loss of UDP packet without spreading. With spreading enabled, the percentage of failure is around 4%. UDP does not provide retransmission of data and the proportion of success does not increase with the time.

The TCP handshake needs three packets to be completed, and the opening time of the majority of TCP connection is consistent with the UDP test. Some openings are delayed, and will be successfully completed with retransmission. They are not displayed on Figure 11, but we observe the same kind of results that are shown in the 6to4 network.

On the 6to4 network, loss of packets has big effect on TCP performance. For the opening of the connection, we see some steps corresponding to standard time of Linux retransmission strategy for TCP.

3) High frequency addresses switching, consequences on loss of packets: Of course, the lifetime of addresses has a big impact on the quality of the connection. We tried lifetime values between 50 ms and 2 seconds, and we summarize the percentage of packet loss in Table V.

We compare it to the theoretical result of Section VI-C in Figure 13. For the 6to4 network, the typical latency is around 21 ms, and we measure a desynchronization of 8 ms at the end of the experiment, it gives us around 50 ms of black hole

 ISBN: 978-1-61208-330-8

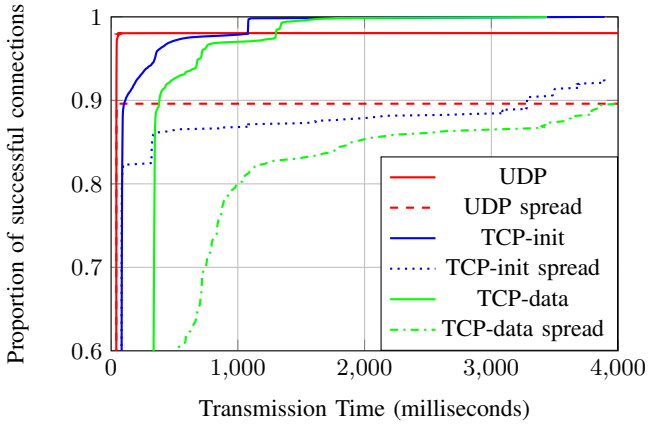


Figure 12: Consequences of a simple spreading on 6to4

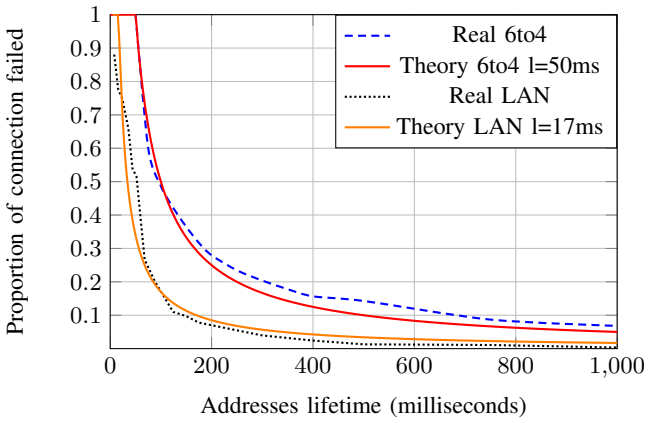


Figure 13: Proportion of failed ICMP echo transmission

TABLE V: PROPORTION OF PACKET LOSS

Lifetime	LAN	6to4
50 ms	51%	100%
75 ms	13.2%	60%
100 ms	10%	48.8%
150 ms	9.7%	36.5%
200 ms	5.7%	28%
300 ms	3.9%	20.4%
500 ms	1.3%	14.3%
750 ms	1.1%	8.7%
1000 ms	0.3%	6.8%
2000 ms	0.1%	5.1%

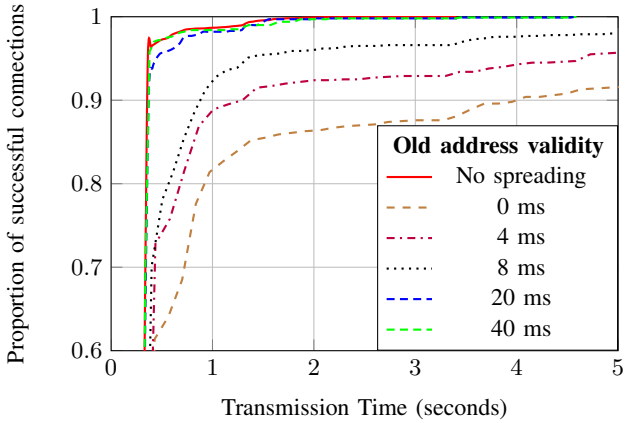


Figure 14: TCP test with transmission of data on the 6to4 network in function of temporal windows for the old address

for the transmission. Our experimental result is very close to this theoretical result. But since the 6to4 network has some natural packet loss, there are more failure than excepted when the spreading effect decrease.

On the LAN, the latency is small, and the desynchronization is the main issue. We measure a desynchronization between 0 and 30 ms, it is not stable in time of experiences. We took the middle value to plot the theoretical value with a black hole of 17 ms at each address switching.

C. Delayed packets: temporal window for the last old address

To prevent loss of delayed packets, we add a temporal window for the old address. In this temporal windows, the old and the current addresses of a sequence are accepted by the spreader. It is very efficient to decrease the loss of packets on the 6to4 network, like we can see in Figure 14.

With a temporal windows larger than the sum of the latency and the desynchronization of the network, we get the same performance than without spreading.

As depicted in Figure 15, it is not enough to prevent loss of packets on the LAN. The desynchronization is bigger than the latency and some packets come to early for a spreader. In this case, we need to accept the previous address on the spreader desynchronized in the future as well than the next address on the other spreader.

D. Desynchronization: temporal window for next address

To solve the desynchronization issue on the LAN, we add a temporal windows for the next address. During this temporal windows, both actual address and next address of the sequence are accepted by the spreader.

We plotted the results of UDP test in Figure 16, with a lifetime of 200 ms and a temporal windows for the old address of 60 ms.

Adding a temporal windows for the next address is not enough to avoid any loose of packet. We need to accept the old address on the spreader with a clock ahead of the real time. We wrote in Table VI the loss of packets with respect of both temporal windows for the old and the future address on the LAN. Since the desynchronization is not stable in the time of the experience, some values can be confusing. The loss can decrease if we increase the temporal windows. However, if

 ISBN: 978-1-61208-330-8

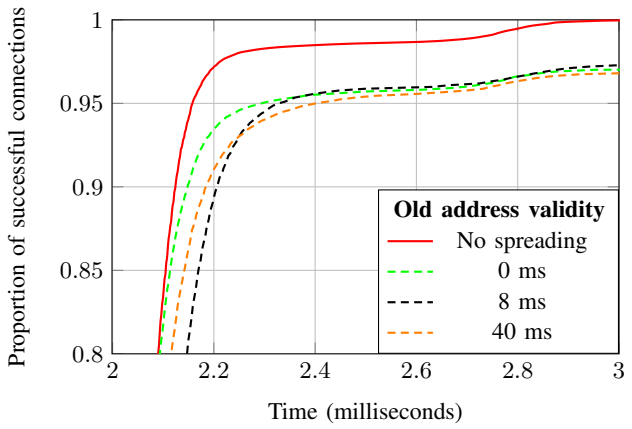


Figure 15: UDP test in function of temporal windows for the old address on the LAN

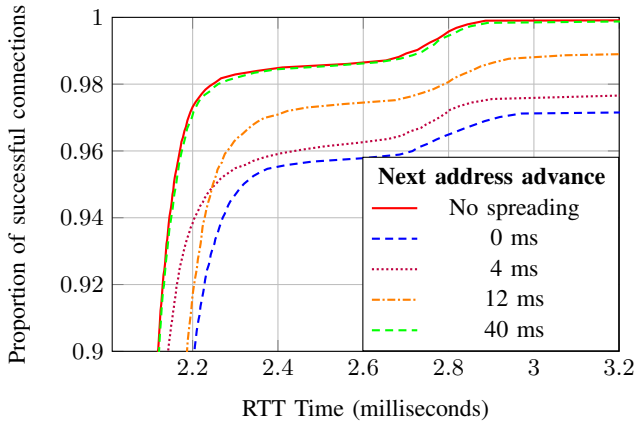


Figure 16: Next address temporal windows on the LAN

both temporal windows are larger than the desynchronization, there are no more loss of packets. A value of 64 milliseconds is enough here.

VIII. Conclusion

The spreading of addresses is an innovative and new solution to identify a connection. This is a new mechanism to protect against spoofing attack. Our spreading protects against initialization of a connection from an attacker, as well than injection of packet inside a established connection.

We described a complete protocol to securely initiate connection between spreaders, with one initialization of temporal sequences of addresses per flow. We did a step by step description of the spreader internal functionality, and we explain the theoretical loss of packets without any temporal windows.

With the use of temporal windows for the old address, we can protect against false positive detection of packets due to the network latency. With the use of a temporal window for the next address, we protect our solution against desynchronization of devices. We can use this information to resynchronize spreaders without external source of time.

Thanks to these temporal windows, we can achieve a very high frequency of address switching. An address is valid only for several duration of the latency in the network.

In the future works, we will evaluate algorithms to generate sequence of addresses. To simplify the work of the network administrator, the auto-configuration and the exchange of the secret between the spreaders have to be considered.

TABLE VI: INFLUENCE OF NEXT AND OLD ADDRESSES ON PACKET LOSS

% loss		Next address temporal windows (ms)							
		0	**4**	**8**	**12**	**20**	**40**	**64**	**80**
Old address validity	**0**	3.5	5	3.9	3.3	3.3	4.8	3	4.9
	4	6	3.3	0.8	0.9	0.5	1.3	1.1	1.5
	8	2.7	2.3	0.6	0.7	0.7	0.1	0	0
	12	6.4	3	2	0	0	0	0	0
	20	3	2.3	0.5	0	0	0	0	0

References

[1] N. Brownlee, C. Mills, and G. Ruth, "Traffic Flow Measurement: Architecture," RFC 2722 (Informational), Internet Engineering Task Force, Oct. 1999. [Online]. Available: http://www.ietf.org/rfc/rfc2722.txt

[2] J. Postel, "Transmission Control Protocol," RFC 793 (INTERNET STANDARD), Internet Engineering Task Force, Sep. 1981, updated by RFCs 1122, 3168, 6093, 6528. [Online]. Available: http://www.ietf.org/rfc/rfc793.txt

[3] T. Aura, "Cryptographically Generated Addresses (CGA)," RFC 3972 (Proposed Standard), Internet Engineering Task Force, Mar. 2005, updated by RFCs 4581, 4982. [Online]. Available: http://www.ietf.org/rfc/rfc3972.txt

[4] P. Ferguson and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing," RFC 2827 (Best Current Practice), Internet Engineering Task Force, May 2000, updated by RFC 3704. [Online]. Available: http://www.ietf.org/rfc/rfc2827.txt

[5] T. Narten, G. Huston, and L. Roberts, "IPv6 Address Assignment to End Sites," RFC 6177 (Best Current Practice), Internet Engineering Task Force, Mar. 2011. [Online]. Available: http://www.ietf.org/rfc/rfc6177.txt

[6] M. Dunlop, S. Groat, W. Urbanski, R. Marchany, and J. Tront, "Mt6d: A moving target ipv6 defense," in MILITARY COMMUNICATIONS CONFERENCE, 2011 - MILCOM 2011, Nov 2011, pp. 1321–1326.

[7] M. Bagnulo, A. Garcia-Martinez, and A. Azcorra, "An architecture for network layer privacy," in Communications, 2007. ICC '07. IEEE International Conference on, June 2007, pp. 1509–1514.

[8] X. Cheng, J. Bi, and X. Li, "Swing - a novel mechanism inspired by shim6 address-switch conception to limit the effectiveness of dos attacks," in Networking, 2008. ICN 2008. Seventh International Conference on, April 2008, pp. 267–272.

[9] S. Amante, B. Carpenter, S. Jiang, and J. Rajahalme, "IPv6 Flow Label Specification," RFC 6437 (Proposed Standard), Internet Engineering Task Force, Nov. 2011. [Online]. Available: http://www.ietf.org/rfc/rfc6437.txt

[10] Q. Hu and B. Carpenter, "Survey of Proposed Use Cases for the IPv6 Flow Label," RFC 6294 (Informational), Internet Engineering Task Force, Jun. 2011. [Online]. Available: http://www.ietf.org/rfc/rfc6294.txt

 ISBN: 978-1-61208-330-8

An Improved Hybrid Scheduler for WiMAX and its Performance Evaluation

Anju Lata Yadav, Prakash D. Vyavahare, Prashant P. Bansod
Electronics and Telecomm. Engg., Electronics and Telecomm. Engg., Biomedical Engg.
Shri G.S. Institute of Technology and Science
Indore (M.P), India
yadawanjulata@rediffmail.com, prakash.vyavahare@gmail.com, ppbansod43@gmail.com

***Abstract* - IEEE 802.16 based Worldwide interoperability for Microwave Access (WiMAX) provides broadband wireless access with integration of variety of applications such as voice, video and data with different Quality of Service (QoS) requirements. In WiMAX, MAC layer scheduling is an integral part of providing QoS to variety of services. In this paper, a hybrid WiMAX scheduler has been proposed and analytically modeled using Markov chain. The proposed scheduler consists of homogeneous scheduler, Weighted Fair Queuing (WFQ), strict priority and a Round Robin (RR) scheduler. Analytical modeling is carried out to investigate whether the proposed scheduler achieves the design goals namely low complexity, satisfying QoS requirements of a variety of applications, maintaining fairness among all applications. Markovian model balance equations are obtained and solved using matrix multiplication method to derive the performance metrics for comparison, such as throughput, mean queuing delay, and packet loss probability of each traffic and interclass fairness. The analytical results show that the proposed hybrid scheduler satisfies not only the QoS requirements of various applications but improves fairness among services.**

Keywords-WiMAX, Quality of Service (QoS), Scheduling algorithms, Analytical Model.

I. INTRODUCTION

Increased demand for support of triple play services (voice, video and data) and high speed internet access led to the development of Wireless Metropolitan Area Networks (WMAN). WiMAX provides advantages of long range communication and support for QoS at the MAC layer as compared to other access technologies. Various QoS constraints are achieved at MAC layer by differentiation of traffic types through five service classes as defined by the standard: Unsolicited Grant Services (UGS), extended real time Polling Service (ertPS), real time Polling Service (rtPS), non real time polling service (nrtPS) and Best Effort (BE). In WiMAX, Base Station (BS) implements the scheduling for both uplink and downlink connections to allocate slots for channel utilization [1]. Although standard has defined the service classes, scheduling mechanism has not been defined. Scheduler at BS transforms the QoS requirements of Subscriber Stations (SSs) into appropriate number of slots. BS then informs each SS about the scheduling decision through UpLink MAP (ULMAP) and DownLink (DLMAP) messages at the beginning of each frame.

WiMAX schedulers can be broadly classified into homogeneous, hybrid and opportunistic. Homogeneous schedulers are based on legacy scheduling algorithms. Homogeneous schedulers are also known as channel unaware schedulers as they do not take care of channel error, packet loss rates and power level into consideration. They serve as intraclass schedulers. Hybrid scheduler is a combination of legacy schedulers while opportunistic scheduler considers the variability in channel condition [2].

In this paper, an improved hybrid scheduler is proposed and analyzed, for four different applications such as voice, video, data and web traffic. The proposed improved scheduler not only satisfies QoS requirements of real time and non real time applications, but also provides high fairness among all applications as compared to the existing three queue scheduler proposed in the literature [3]. The hybrid scheduler is analyzed using Markovian model and the balance equations obtained from Markovian model are solved using matrix multiplication method to derive the performance metrics for comparison such as throughput, queuing delay and packet loss probability of each traffic and fairness among various applications. The average rate at which packets pass through a scheduler at steady state is termed as throughput. Packet loss probability of a traffic class is the probability that no space is available in the corresponding buffer.

Various studies have focused on analytical modeling of hybrid schedulers. In the field of networking, Markov models have been widely used to model the behavior of communication networks under variable traffic load conditions. In [4], queue decomposition method was proposed which divided the Priority Queuing (PQ) system into a group of Single-Server Single-Queue (SSSQ) to obtain their service capacities. Queue length distributions of individual traffic are investigated through analytical modeling. In this paper, the hybrid scheduler is modeled as a Markov chain and stationary probabilities are obtained using closed loop expressions and matrix geometric method.

In [5], priority queuing traffic is divided into heterogeneous Long Range Dependent (LRD) self similar and Short Range Dependent (SRD) Poisson traffic and is analytically modeled by Priority Qeueing- Generalized Processor Sharing (PQ-GPS) scheduling mechanism. In order to deal with heterogeneous traffic, they extended the Schilder's theorem and developed the analytical upper and lower bounds of queue length distributions of individual traffic flows. A model based on G/M/1 queuing system is

 ISBN: 978-1-61208-330-8

proposed in [6] to take into account multiple classes of traffic that exhibit long range dependencies and self similarity among traffic. The authors in [6] developed a Markov chain for non preemptive priority scheduling which is solved to extract a numerical solution for proposed analytical framework. The accuracy of the model is demonstrated by comparing the numerical solution of the analytical modeling to simulation experiments and compared with the actual test bed results.

A discrete time priority queue with two layered general arrival process is analyzed in [7] in which higher priority is given to small fraction of the requests applications that generate large revenues and to applications with small size request. By using probability generating functions, performance measures of the queue such as the moments of the packet delays of both classes are calculated. Through analysis, [8] developed several upper bounds on the queue length distribution of GPS with Long Range Dependent (LRD) traffic, extending the same to packet based GPS. These bounds show that long range dependency and queue length distribution of LRD traffic is not affected by the presence of other sources. A notion of LRD isolation has been introduced in [8] that broadens the range of services offered by GPS system by admitting the traffic of low priority.

A hybrid scheduler based on the integration of PQ and WFQ schedulers is proposed in [3]. Traffic arrival processes are represented by non bursty Poisson process and bursty Markov Modulated Poisson Process (MMPP). The model consists of three separate buffers representing three traffic flows. The strict priority scheduler assigns highest priority to ertPS traffic and the scheduler WFQ schedules the rtPS traffic and nrtPS traffic.

A scheduler for UL and DL is proposed in [9]. Based on the QoS requirements and priority of services classes needed, resources are calculated and granted in terms of number of slots. The calculation of resources depend on bandwidth requirements of each connection and ensuring that it does not exceed maximum requirement of each connection. UL scheduler calculates number of slots needed by taking into account the polling interval while DL scheduler considers packet size for calculation of number of slots needed. Simulation results show that scheduler fulfills delay and throughput requirements of all service classes of WiMAX. An analytical model for performance evaluation of WiMAX networks is developed in [10]. A one dimensional Markov chain is developed which takes into account frame structure, precise slot sharing based scheduling and channel conditions. Closed form expressions are derived for the scheduling policies to satisfy slot fairness and throughput fairness. Opportunistic scheduling is considered to derive the performance metrics.

A mathematical analysis is performed for an uplink scheduling algorithm, named “courteous algorithm” which gives advantage to lower priority traffic class without affecting traffic of higher priority [11]. In this scheme scheduling of nrtPS traffic is improved while satisfying the delay constraints of rtPS traffic. In case maximum packet loss rate has not reached, lower priority traffic is served over higher priority as long as the maximum waiting time of higher priority traffic is not violated. Various performance parameters such as mean waiting time, waiting time for courteous packets, maximum burst size and packet priority are calculated using mathematical model and validated by a simulation. An analytical modeling of MAC protocol of WiMAX network is applied through queuing theory and Markov chain in [12] to analyze average message delays for real time and non real time applications.

The remainder of the paper is organized as follows: Section 2 provides a brief description of proposed hybrid scheduler. The analytical modeling of proposed hybrid scheduler is outlined in Section 3. The results are analyzed and discussed in Section 4. Finally the paper is concluded in Section 5.

II. DESCRIPTION OF PROPOSED HYBRID SCHEDULER

In this section, proposed hybrid scheduling algorithm for the BS and SS in WiMAX system is presented. The main objective of the proposed scheme is to provide QoS to both real time and non real time applications. The scheduler is implemented at the BS for uplink scheduling. Hybrid scheduler, consisting of WFQ, RR and strict priority, combines the advantages of well known scheduler strategies to provide QoS in a communication network [13].

Figure 1 illustrates the proposed hybrid scheduler that provides scheduling among four service classes of WiMAX namely ertPS, rtPS, nrtPS and BE. The incoming traffic is assigned to four separate buffers (Q_1, Q_2, Q_3 and Q_4) and each queue is served in First Come First Serve (FCFS) manner. As the packets of real time traffic such as voice and video, cannot tolerate higher delay they are given higher priority than non real time traffic. ErtPS and rtPS packets are served before nrtPS and BE packets of non real time traffic such as data traffic.

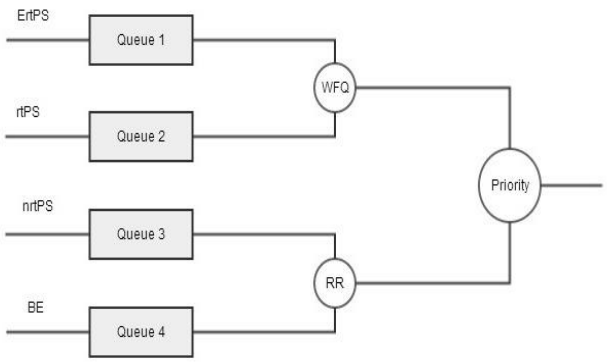


Figure 1: Architecture of proposed hybrid scheduler for providing QoS at MAC layer of WiMAX

The packets of ertPS and rtPS are buffered in queues Q_1 and Q_2 respectively and are served by the scheduler WFQ. Packets of nrtPS and BE traffic are placed in queues Q_3 and Q_4 respectively and are served by the scheduler RR. Queues Q_1 and Q_2 are each assigned a weight representing the maximum number of packets to be served in each round. The weight for queue Q_1 is varied from lowest value of 10% to highest value of 90% for ertPS traffic and simultaneously weight of queue Q_2 for rtPS traffic is varied from 90% to 10%.

III. ANALYTICAL MODELLING OF PROPOSED HYBRID SCHEDULER

The analytical model of the proposed scheme consists of single server multiple queues one for each service class subjected to hybrid scheduling. The system is represented by M/M/1/K queuing with queues (Q_1, Q_2, Q_3 and Q_4) of size K_1, K_2, K_3 and K_4 respectively. The input traffic is a Poisson process and the arrival rate for class c is λ_c. The service time has exponential distribution with service rates μ_1, μ_2, μ_3 and μ_4 and the service capacity of the system is s. The hybrid scheduler is represented by a Markov chain with the state space diagram for the proposed hybrid scheduler diagram shown as in Figure 2. A state P_{mnrs} represents the number of packets m, n, r, s for classes ertPS, rtPS, nrtPS and BE in their respective buffers. A packet added to queue Q_1 changes the state from P_{mnrs} to P_{m+1nrs}. Servicing a packet from queue Q_1 will change the state from P_{mnrs} to P_{m-1nrs}. Similarly the state of each queue changes depending on whether a packet is added to it or serviced from it.

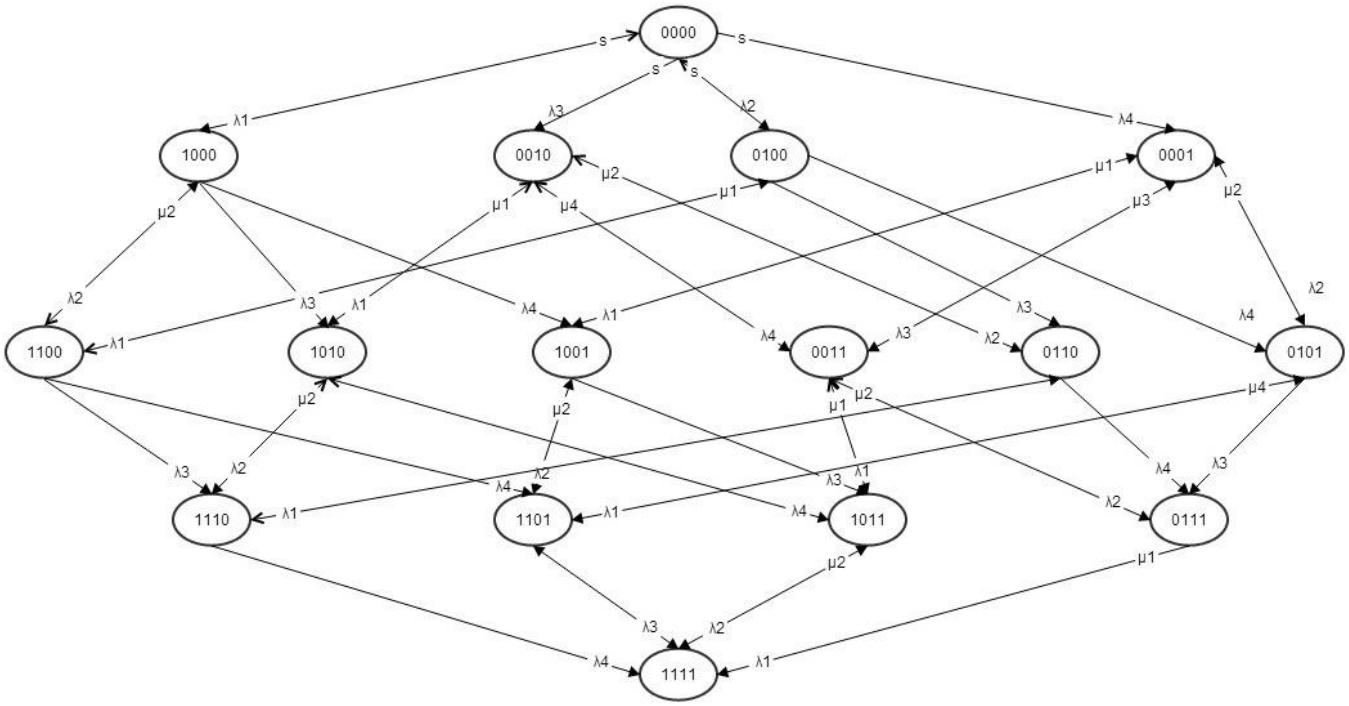


Figure 2: State space diagram for the proposed hybrid scheduler with single buffer queues

Matrix multiplication method approach is used to determine the steady state or limiting state probability of the system. The π matrix represents the steady state probability vector and satisfies equations (1) and (2) [3].

$$\pi Q = 0 \quad (1)$$

$$\pi e = 1 \quad (2)$$

Where $e = (1, 1, 1\ldots\ldots1)^T$ is a unit column vector of length $(K_1+1)x(K_2+1)x(K_3+1)x(K_4+1)x1$ and Q is a generator matrix and π is a steady sate probability column vector. The multiplication of these matrices provide balance equations which are further solved to obtain the steady state probabilities.

The probability p_m^c of having m packets of class c (c = 1, 2, 3, 4) in the corresponding buffer can be calculated from joint state probability p_{ijkl} using equations (3) to (6) [3]:

$$p_m^c = \sum_{l=0}^{K_4}\sum_{k=0}^{K_3}\sum_{j=0}^{K_2} p_{mjkl} \qquad for\ class\ 1 \qquad (3)$$

$$p_m^c = \sum_{l=0}^{K_4}\sum_{k=0}^{K_3}\sum_{i=0}^{K_1} p_{imkl} \quad for\ class\ 2 \qquad (4)$$

$$p_m^c = \sum_{l=0}^{K_4}\sum_{j=0}^{K_2}\sum_{i=0}^{K_1} p_{ijml} \quad for\ class\ 3 \qquad (5)$$

$$p_m^c = \sum_{k=0}^{K_3}\sum_{j=0}^{K_2}\sum_{i=0}^{K_1} p_{ijkm} \quad for\ class\ 4 \qquad (6)$$

With the help of the probabilities p_m^c for each class, performance metrics can be calculated from equations (7) to (10). The mean number of packets in the buffer of class c, L^C, is given by equation (7).

$$L^C = \sum_{m=0}^{L_c}[m * p_m^c] \qquad (7)$$

Similarly the throughput T^C is given by equation (8).

$$T^C = \lambda_c * \sum_{m=0}^{L_c-1}[p_m^c] \qquad (8)$$

The mean queuing delay D^C is given by equation (9).

$$D^C = \frac{L^c}{T^c} \qquad (9)$$

The fairness index F is given by equation (10).

$$F = \frac{(\sum_{i=1}^{c} T^i)^2}{C * \sum_{i=1}^{c}(T^i)^2} \qquad (10)$$

The steady state balance equations (equations 11-27) are derived from equations (1) and (2).

$$\sum_{i=0}^{n-1}\pi_i = 1 \qquad (11)$$

$$-(\lambda_1+\lambda_2+\lambda_3+\lambda_4)\pi_0 + s\pi_1 + s\pi_2 + s\pi_3 + s\pi_4 = 0 \qquad (12)$$

$$\lambda_4\pi_0 - (\lambda_1+\lambda_3+\lambda_2+s)\pi_1 + \mu_1\pi_7 + \mu_3\pi_8 + \mu_2\pi_{10} = 0 \qquad (13)$$

$$\lambda_3\pi_0 - (\lambda_2+\lambda_1+\lambda_4+s)\pi_2 + \mu_1\pi_6 + \mu_4\pi_8 + \mu_2\pi_9 = 0 \qquad (14)$$

$$\lambda_2\pi_0 - (\lambda_4+\lambda_3+\lambda_1+s)\pi_3 + \mu_4\pi_5 = 0 \qquad (15)$$

$$\lambda_1\pi_0 - (\lambda_4+\lambda_2+\lambda_3+s)\pi_4 + \mu_2\pi_5 = 0 \qquad (16)$$

$$\lambda_1\pi_3 + \lambda_2\pi_4 - (\lambda_3+\lambda_4+\mu_4+\mu_2)\pi_5 = 0 \qquad (17)$$

$$\lambda_1\pi_2 + \lambda_3\pi_4 - (\lambda_2+\lambda_4+\mu_1)\pi_6 + \mu_2\pi_{11} = 0 \qquad (18)$$

$$\lambda_1\pi_1 + \lambda_4\pi_4 - (\mu_1+\lambda_2+\lambda_3)\pi_7 + \mu_2\pi_{12} = 0 \qquad (19)$$

$$\lambda_3\pi_1 + \lambda_4\pi_2 - (\mu_4+\mu_3+\lambda_2+\lambda_1)\pi_8 + \mu_1\pi_{13} + \mu_2\pi_{14} = 0 \qquad (20)$$

$$\lambda_2\pi_2 + \lambda_3\pi_3 - (\mu_2+\lambda_1+\lambda_4)\pi_9 + \mu_1\pi_{11} = 0 \qquad (21)$$

$$\lambda_2\pi_1 + \lambda_4\pi_3 - (\mu_2+\lambda_3+\lambda_1)\pi_{10} + \mu_1\pi_{12} = 0 \qquad (22)$$

$$\lambda_3\pi_5 + \lambda_2\pi_6 + \lambda_1\pi_9 - (\mu_2+\mu_1+\lambda_4)\pi_{11} = 0 \qquad (23)$$

$$\lambda_4\pi_5 + \lambda_2\pi_7 + \lambda_1\pi_{10} - (\mu_2+\mu_1+\lambda_3)\pi_{12} = 0 \qquad (24)$$

$$\lambda_4\pi_6 + \lambda_3\pi_7 + \lambda_1\pi_8 - (\mu_1+\lambda_2)\pi_{13} = 0 \qquad (25)$$

$$\lambda_2\pi_8 + \lambda_4\pi_9 + \lambda_3\pi_{10} + \mu_1\pi_{15} - (\mu_2+\lambda_1)\pi_{14} = 0 \qquad (26)$$

$$\lambda_4\pi_{11} + \lambda_3\pi_{12} + \lambda_2\pi_{13} + \lambda_1\pi_{14} - \mu_1\pi_{15} = 0 \qquad (27)$$

IV. ANALYSIS OF RESULTS

This section deals with investigation on the impact of weights of traffic flows scheduled by WFQ on the performance metrics: throughput, queuing delay, packet loss probability of each traffic class and fairness. The size of the buffer for queues (Q_1, Q_2, Q_3 and Q_4) is set to be 1 for each queue. The mean service rate s is 10. The service rate for queue 1 increases from 1 to 9 corresponding to increase in weight ratio from 10% to 90%. Simultaneously, the service rate for queue 2 varies from 9 to 1 corresponding to reduction in weight ratio from 90% to 10%. The idea behind varying weights is to find the optimum weight at which the design goals are achieved for proposed hybrid scheduler model Queues Q_3 and Q_4 are scheduled by RR; therefore, both are serviced with same weight.

Class 1 traffic, carried by ertPS service class, represents audio traffic which requires a lower delay, while higher or lower throughput does not affect much the quality of audio traffic. Thus requirement of low delay is mandatory for ertPS. RtPS service class, representing streaming video traffic, also requires a low delay but the requirement is not so stringent. Class 3 traffic, carried by nrtPS, represents the web traffic for which the requirement of high throughput is mandatory and low delay and low throughput is optional. Class 4 traffic, carried by BE, is the data traffic requiring high throughput.

Figure 3 shows the variation of mean queuing delay of each traffic class with respect to variation in weight ratio of WFQ scheduler. The graph shows that the queuing delay of class 1 traffic is reduced with the increase in weight of class 1 traffic since with the increase in weight more packets are serviced from queue and the queue delay is reduced. When the weight of class 1 traffic increases, the weight of class 2 traffic is reduced, causing an increased delay as a smaller number of packets is being serviced. Class 3 and class 4 traffic show a similar behavior as both are scheduled by RR. Figure 4 shows the variation of throughput for each traffic class with respect to variations in the weight ratio of scheduler WFQ. Figure 4 shows that as the weight for class 1 or class 2 traffic is increased, throughput also increases as more number of packets depart from the buffer. Class 4 has the highest arrival rate (λ_4) of 4 while class 3 has arrival rate of 3. Therefore, it is observed that the throughput for class 4 is the highest.

Figure 5 shows the variation of packet loss probability of each traffic class with respect to variation in weight ratio of WFQ scheduler. The figure shows that as the weight of

class 1 traffic increases, the packet loss probability reduces as mean number of packets in the queue reduces, due to higher service rate. Similar behavior is displayed for class 2 traffic. Class 3 and class 4 traffic have highest packet loss probability as both are given lower priority than class 1 and class 2 traffic. As long as there are packets in class 1 or class 2 traffic buffer queues, lower priority class 3 or class 4 traffic is not served.

Figure 6 shows the variation of fairness with respect to increase in weight ratio of WFQ scheduler. The interclass fairness is the highest for a weight ratio of 9:1. For satisfying low delay requirement of class 1 and class 2 traffic, weight ratio 8:2 can be considered to be optimum as at this weight class 1 traffic has lower delay while class 2 traffic also has lower delay. At a weight ratio of 9:1 class 1 has lowest delay but class 2 has maximum delay verifying that it is not the optimum choice. At a weight ratio of 8:2, class 3 and class 4 have higher throughput than class 1 and class 2. Though they have the highest throughput at a weight ratio of 5: 5, at this ratio, class 1 traffic does not have a lower delay. Therefore, weight ratio of 8:2 satisfies the delay requirement of class 1 and class 2 traffic. It also satisfies the throughput requirement of class 3 and class 4. With respect to interclass fairness, weight ratio 9:1 is optimum as it provides highest fairness but at this weight ratio QoS requirements of all traffic classes are not satisfied. The proposed model not only satisfies QoS requirements of all service classes but also provides high fairness among all the traffic classes with a complexity of O(N) [13].

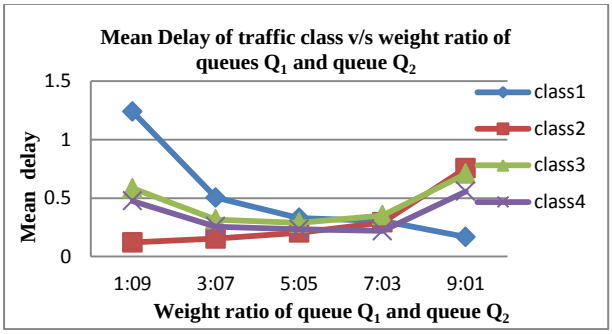


Figure 3: variation of mean Queuing delay for four classes of traffic v/s weight ratio for improved hybrid scheduler

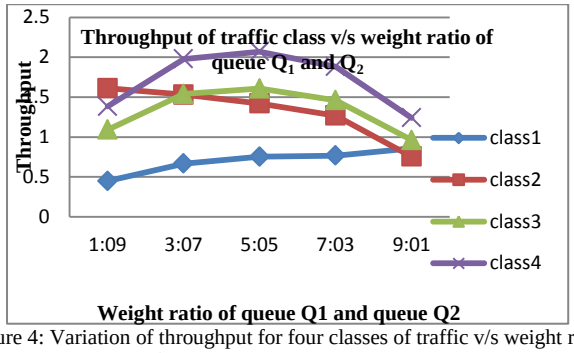


Figure 4: Variation of throughput for four classes of traffic v/s weight ratio for improved hybrid scheduler.

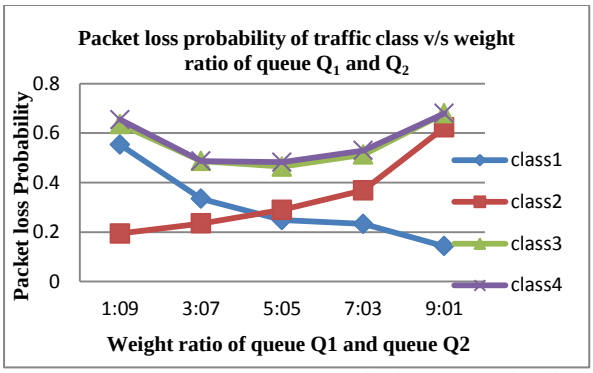


Figure 5: Variation of packet loss probability for four classes of traffic v/s weight ratio for improved hybrid scheduler.

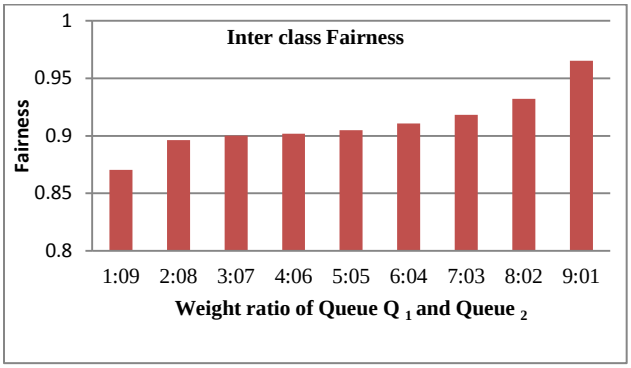


Figure 6: Variation of Interclass fairness v/s weight ratio for improved hybrid scheduler.

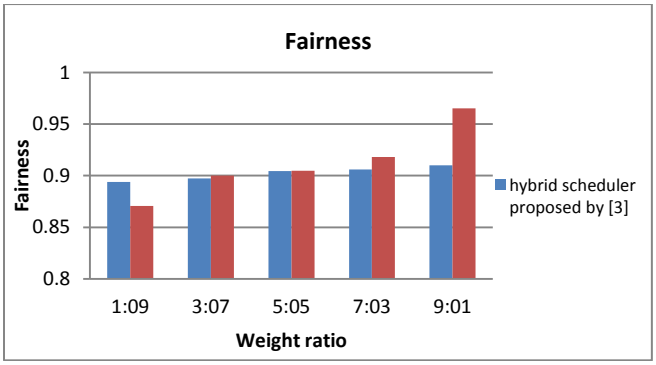


Figure 7: Variation of fairness of proposed improved hybrid scheduler and hybrid scheduler proposed in [3] v/s weight ratio.

Figure 7 shows the variation of fairness of proposed improved hybrid scheduler and the hybrid scheduler proposed in [3] with respect to variation in weight ratio of WFQ scheduler. The figure shows that the proposed improved scheduler provides better fairness at high weight ratio and almost the same fairness at lower weight ratio. Hence the proposed improved scheduler is more fair to all service classes than the scheduler proposed in [3].

Table 1 compares the performance of a three queue hybrid scheduler proposed in [3] (case I) and the four queue improved hybrid scheduler (case II) proposed in this paper. Performance metrics considered for comparison are Packet Loss Probability (PLP) and queuing delay of real time

 ISBN: 978-1-61208-330-8

applications served by class 1 and throughput of non real time application served by class 3. The class 1 traffic in case I is provided highest priority while class 1 traffic in case II is scheduled by WFQ. Therefore, the PLP and queuing delay of class 1 for case I are lower as compared to that of class 1 of case II. Class 2 traffic in both cases is served by WFQ. Case II provides lower PLP and queuing delay for lower weight.

The comparison shows that the proposed improved hybrid scheduler satisfies the major objective of fairness among all service classes along with satisfying the QoS requirements of voice, video and data traffic.

TABLE 1: COMPARISON OF PERFORMANCE METRICS OF A HYBRID SCHEDULER PROPOSED IN [3] (CASE I) AND THE IMPROVED HYBRID SCHEDULER PROPOSED IN THIS PAPER (CASE II).

Weight ratio / Performance metric of a class	1:9	3:7	5:5	7:3	9:1
Case- I ,class 1 PLP	0.09	0.08	0.09	0.08	0.08
Case- II, class 1 PLP	0.55	0.33	0.24	0.23	0.14
Case- I, class 1 Delay	0.1	0.09	0.1	0.09	0.11
Case- II,class 1 Delay	1.23	0.50	0.33	0.30	0.16
Case- I, class 2PLP	0.25	0.23	0.22	0.20	0.19
Case- II, class 2 PLP	0.19	0.23	0.29	0.36	0.62
Case- I, class 2 Delay	0.17	0.14	0.14	0.12	0.11
Case- II, class 2 Delay	0.12	0.15	0.20	0.29	0.75
Case- I, class 3 throughput	2.20	2.20	2.12	2.15	2.00
Case- II, class 3 throughput	1.09	1.53	1.60	1.46	0.96

V. CONCLUSION

WiMAX supports deployment of triple play services by providing QoS through scheduling at MAC layer. This paper proposes a WiMAX hybrid scheduler for triple play services. WFQ, RR and strict priority schedulers are integrated in proposed scheduler to perform scheduling. Analysis of the proposed scheduler is carried out using queuing theory and Markov chain for deriving expressions for performance metrics, such as throughput, mean queuing delay, and packet loss probability to investigate the effectiveness of proposed hybrid scheduler in satisfying design goals. Weight ratio is changed to investigate its effect on performance metrics and to discover the optimum weight which satisfies design goals of a scheduler. Proposed scheduler satisfies QoS requirements of a variety of applications and it maintains fairness among all applications at a complexity of O (N). It can be used in non bursty applications where fairness is the main issue along with satisfaction of QoS requirements. The proposed scheduler can provide the flexibility in varying the weights of various traffic classes for achieving better design objective as per the application and requirement.

REFERENCES

[1] L. Nuaymi, Wi-MAX Technology for Broadband Wireless Access, John Wiley and Sons Ltd, 2007.

[2] C. So-In, R. Jain and A. K. Tamimi, "Scheduling in IEEE 802.16e Mobile WiMAX Networks: Key Issues and a Survey," IEEE Journal On selected Areas In Communications, 27(2), Feb 2009, pp. 156-171.

[3] L. Wang, G. Min, D. D. Kouvatsos and X. Jin, "Analytical Modelling of an Integrated Priority and WFQ Scheduling scheme in Multi Service Networks" , Journal of Computer communications, 33(11), Nov 2010, pp. 93-101.

[4] X. Jin and G. Min, "Modeling and analysis of priority queuing systems with multi-class self-similar network traffic: a novel and efficient queue-decomposition approach", IEEE Transactions on Communications, 57 (5), 2009, pp. 1444–1452.

[5] X. Jin and G. Min, "Performance Modeling of hybrid PQ–GPS systems under long-range dependent network traffic", IEEE Communications Letters, 11(5), 2007, pp. 446–448.

[6] M. Iftikhar, T. Singh, B. Landfeldt and M. Çaglar, "Multiclass G/M/1 queuing system with self-similar input and non-preemptive priority", Journal of Computer Communications, 31(5), Mar 2008, pp. 1012-1027.

[7] J. Walraevens, S. Wittevrongel and H. Bruneel, "Analysis of Priority Queues with Session-based Arrival Streams", Proc. 7th International Conference on Networking (ICN 08), Cancun, Mexico, April 2008, pp. 503–510.

[8] X. Yu, I.L.-J. Thng, Y. Jiang and C. Qiao, "Queueing processes in GPS and PGPS with LRD traffic inputs", IEEE/ACM Transactions on Networking, 13(3) 2005, pp. 676–689.

[9] K. A. Noordin and G. Markarian, "Providing QoS Support through Scheduling in WiMAX Systems", Proc. International Journal of the Physical Sciences, 6 (16), 18 August 2011, pp. 4070-4081.

[10] B. Baynat, G. Nogueira, M. Maqbool and M. Coupechoux, " An Efficient Analytical Model for the Dimensioning of WiMAX Networks supporting multi-profile Best Effort Traffic", Journal of Computer Communications, 33(10), June 2010, pp. 1162-1179.

[11] K. Michel and C. Tata, "Courteous algorithm: Performance optimization in WiMAX networks", Proc. 4th international conference on Communications and information technology, World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, 2010, pp. 23-32.

[12] L. F. M. de Moraes and D. L. F. G. Vieira, "Analytical Modeling and Message Delay Performance Evaluation of the IEEE 802.16 MAC Protocol", Proc. 18th Annual IEEE/ACM International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, Florida, Aug 2010, pp. 180-192.

[13] Pratik, Dhrona, Najah Abu Ali and Hossam Hassanein, "A Performance study of Uplink scheduling Algorithms in point to Multipoint WiMAX Networks", Journal of Computer Communication, 32(3), 2009, pp. 511-521.

 ISBN: 978-1-61208-330-8

Meshed Tree Protocol for Faster Convergence in Switched Networks

Kuhu Sharma, Bruce Hartpence, Bill Stackpole, Daryl Johnson, Nirmala Shenoy
College of Computing and Information Sciences
Rochester Institute of Technology,
Rochester, NY, USA
kxs3104@rit.edu, bhhics@rit.edu, Bill.Stackpole@rit.edu, daryl.johnson@rit.edu, nxsvks@rit.edu

***Abstract*—Loop free forwarding is a continuing challenge in switched networks that require link and path redundancy. Solutions to overcome looping frames are addressed by special protocols at layer 2, which block ports in the bridges to build a logical spanning tree for frame forwarding. However, due to the continuing convergence issues in the Spanning Tree algorithm, IETF RFC 5556 *Transparent Interconnection of Lots of Links* on RBridges (router bridges) and IEEE 802.1aq Shortest Path Bridging both use link state routing techniques to build Dijkstra trees from every switch. Both techniques have the expense of higher processing complexity. In this paper, a novel meshed tree algorithm (MTA) is investigated to address convergence issues faced by STA while also avoiding the complexity of Link State routing. The MTA based protocol is compared with Rapid Spanning Tree Protocol using OPNET simulations. The significant reduction in convergence time combined with the simplicity in implementation indicates that the Meshed Tree Protocol would be superior candidate to resolve looping issues in switched networks.**

Keywords- Loop Avoidance; Switched Networks; Meshed Trees.

I. INTRODUCTION

Loop free forwarding is a continuing challenge in switched networks. The mandate for link and path redundancy to provide a continued communications path between pairs of end switches in the event of switch or link failure often results in a physical network topology that has loops. The physical loops in turn cause broadcast storms when forwarding broadcast packets. Implementing a loop free logical topology over the physical topology is one way to avoid broadcast storms. The first logical loop-free forwarding technique based on Spanning Tree Algorithm (STA) was proposed by Radia Perlman [1]. Spanning tree in switched networks was constructed by logically blocking some of the bridge's ports. The Rapid Spanning Tree protocol (RSTP) was subsequently developed to reduce the convergence times on topology changes in the basic STP. Transparent interconnection of lots of links (TRILL) on Rbridges (router bridges) was proposed by the same researcher to overcome the disadvantages of STA-based loop avoidance. This came at the cost of some overhead and implementation complexity through the adoption of the Intermediate system to Intermediate system (IS-IS) routing protocol. IS-IS related messages are encapsulated in special frames by Rbridges. This is currently an Internet Engineering Task Force (IETF) draft [4]. Shortest Path Bridging (SPB) was developed along similar lines, adopting the IS-IS protocol. Its specifications can be found in the IEEE 802.1aq standard.

The premise for these solutions is that a single logical tree from a root switch that operationally eliminates physical loops is necessary to resolve the conflicting requirements of physical link redundancy and loop free forwarding. In the event of link failure, the tree has to be recomputed. While spanning tree is a single tree constructed from a single elected root switch, the Dijkstra algorithm adopted in IS-IS allows building a tree from every switch. IS-IS requires link state information in the whole network to be made available to every switch so that each can build its own tree. The Dijkstra algorithm uses the connectivity information to compute the tree.

In this paper, a novel meshed tree algorithm (MTA) is proposed to address the convergence issues faced with STA based protocols and at the same time avoid the complexity in adopting *Link State* routing at layer 2. MTA allows creation and maintenance of *multiple* overlapping tree branches from *one* root switch. The multiple branches mesh at switches, and of failure of a link (or branch) the switch can immediately fall back on another branch. Packet forwarding can continue while the broken branch is pruned. This eliminates temporary inconsistent topologies and latencies resulting from tree reconstruction. It is important to have a tree (logical or physical) for forwarding broadcast packets. But, that should not preclude the construction of multiple tree branches simultaneously or the overlapping of the tree branches if this can be achieved without loops. Redundant tree branches will thus take over packet forwarding seamlessly in the event of a link or failure.

Meshed trees (MT) can be implemented through a simple numbering scheme called MT_VIDs (virtual IDs) that will be assigned to a switch in the bridged network. The MT_VID defines a tree branch or logical packet-forwarding path from the root switch to the switch with the MT_VID. A switch can acquire several MT_VIDs as it is allowed to join multiple tree branches. In this way, meshed trees leverage the redundancy in meshed topologies to set up several loop-free logical forwarding paths without blocking switch ports. Meshed trees can also be built from multiple root switches although this aspect of the Meshed Tree Protocol (MTP) is not covered here.

In this paper, the implementation details of MTP are presented. The performance of MTP is evaluated and compared with RSTP. The comparison was conducted using OPNET simulation tool [7]. RSTP models are available

 ISBN: 978-1-61208-330-8

with OPNET and hence the comparison studies were limited to RSTP. However, the significant improvement in the convergence times and the hops taken by frames to reach destinations indicate the superior capabilities of MTP. The operational simplicity of MTP also provides advantages over complex Link State solutions. MT loop free forwarding at layer 2 is currently the IEEE 1910.1 working group [8] and the authors lead the effort. The rest of the paper is organized as follows. Section II discusses related work in the context of STP and *Link State* based solutions highlighting the comparable features of MT based solutions. In Section III, operational details of the MT algorithm and protocol are presented. Section IV describes the optimized unicast frame forwarding schema adopted in MTP. Section V provides the simulation details and performance results. Section VI follows with conclusions.

II. RELATED WORK

In this section, we focus on the two primary techniques adopted for loop resolution in bridged networks. The first of these is based on the (Rapid) Spanning Tree Protocol (STP and RSTP) and the second is based Link State (LS) Routing. STP and RSTP both use the spanning tree approach. TRILL on RBridges and SPB are two efforts based on LS routing. The presentation in this section focuses on some distinct features of these techniques without describing operational details as such information is publicly available.

A. Protocols on Spanning Tree Algorithm

The STP is based on the STA. To avoid loops in the network while maintaining access to all the network segments, the bridges collectively elect a root bridge and then compute a spanning tree from the root bridge. In STP, each bridge first assumes that it is the root and announces its bridgeID. This information is used by the neighboring bridges to elect the root bridge. The unique bridgeID is a combination of a bridge priority and the bridge medium access control (MAC) address. A bridge may supplant the current root if its bridgeID is lower. Once a root bridge is elected, other bridges then resolve their connection to the root bridge by listening to messages from their neighbors. These messages also include path cost information. This continues until the topology converges on a single tree.

STP has high convergence times subsequent to topology changes. To reduce the convergence times the *Rapid Spanning Tree* protocol (RSTP) was proposed [2]. RSTP is a refinement of the STP and therefore shares most of its basic operation characteristics, with some notable differences. The differences are: 1) The detection of root bridge failure is 3 ‘hello’ times. 2) Response to Bridge Protocol Data Units (BPDUs) are sent only from the direction of the root bridge, allowing RSTP bridges to ‘propose’ their spanning tree information on their designated ports. This allows the receiving RSTP bridge to determine if the root information is superior, and set all other ports to ‘discarding’ and send an ‘agreement’ to the first bridge. The first bridge, can rapidly transition that port to forwarding bypassing the traditional listening/learning states. 3) Backup details regarding the discarding status of ports are maintained to avoid failure timeouts of forwarding ports.

Advantages: STA based implementation is simple as the spanning tree is executed with the exchange of BPDUs that carry *tree formation* information.

Disadvantages: Several disadvantages of STA based protocols are noted in [2]. These include: 1) Traffic is concentrated on the spanning tree path, and all traffic follows that path even when other more direct paths are available. This causes traffic to take potentially sub-optimal paths, resulting in inefficient use of the links and reduction in aggregate bandwidth. 2) Spanning tree is dependent on the way the bridges are interconnected. Small changes due to link failure can cause large changes in the spanning tree. Changes in the spanning tree take time to propagate and converge, especially for non-RSTP protocols. 3) Though 802.1Q supports multiple spanning trees, it requires additional configuration, the number of trees is limited, and the defects apply within each tree [3].

B. TRILL Protocol on RBridges and Shortest Path Bridging

These two techniques overcome the shortcomings of RSTP as they combine the routing functionality of layer 3 by using the IS-IS protocol [4] at layer 2 to compute pair-wise optimal paths between two bridges. The computed pair-wise optimal paths will be used for forwarding frames at layer 2. Inconsistencies and loop formations during topology change are overcome by a *hop count* used in inter-bridge forwarding. TRILL encapsulates link state routing messages in special headers and uses protocols to learn end station addresses. SPB has two versions; one which creates shortest path trees that are identified by the base VLAN ID called SPBV, and the other which uses the source MAC address to identify the trees and uses MAC in MAC encapsulation. The second technique requires MAC address information dissemination

Advantages over 802-style bridging [4]: 1) Frames travel via an optimal path. 2) Transit frames are routed with a hop count; temporary loops will result in frames being discarded when the hop count reaches zero. 3) Route changes can be made quickly and safely based on local information.

C. Meshed Tree Protocol

Tree like structures imposed on topologies may reduce or eliminate loops but also create an environment in which there are failover delays to alternate links. These topologies also lack redundancy or the ability to load balance. Protocols such as SPB and TRILL work to alleviate these problems but are complex, incorporating routing at layer 2 and requiring additional encapsulation. Link State routing requires that link state database be stable for a certain interval of time before running the Dijkstra algorithm to

 ISBN: 978-1-61208-330-8

compute forwarding paths; routing and forwarding can be unstable during this time.

MTP seeks to address these same issues with less complexity and even shorter failover times upon discovery of link failure. The core of the protocol is the ability of each Meshed Tree Switch (MTS) to be a member of more than one tree. This provides redundancy and optimized traffic forwarding to hosts, while supporting redundant paths that takeover upon link or switch failures.

III. THE MESHED TREE ALGORITHM

The *meshed tree* algorithm allows construction of logically *meshed trees* from a single root switch in distributed fashion and with local information [5]. The discussion presented in this article does not include the election of a root bridge as the focus is on the loop resolution / avoidance capability of MT algorithms. A process similar to that adopted by STA can be used to elect a root bridge. In this article we assume a designated root bridge, which is an option advocated in IEEE1910.1.

Bridge ID: For the operation of the MT algorithm bridgeIDs are necessary. These have to be unique only within the switched network (a simple MAC address derivative can be used). The MT_VIDs would be thus simple, and the first value in the MT_VID will be the root bridgeID. In this article without loss of generality we used a single digit ID for the root switch. Resolution upon root failure is not included in this work.

An MT_VID describes a path that connects the root to a particular switch. The elements of the MT_VID are derived from the root bridgeID and the outbound port numbers of the switches in the path to that particular switch. In a single physical topology, a switch can be associated with more than one MT_VID and thus:

- A *Meshed Tree* could contain ***all*** of the possible paths from the root to each switch.
- More than one path to each switch is supported

Consider a three-switch single loop topology shown in Figure 1. In the upper left is the physical loop topology. In order to prevent traffic from looping, we might impose any one of several logical tree topologies like those shown. In the upper right, the topology is optimized for transmissions associated with switches connected to the root. But in the lower left and lower right, the topology is optimized for nodes connected to switches A and B, respectively. These tree topologies do not provide for redundancy. Meshed trees utilize all of the pathways and because the pathways are pre-established, *failover times to redundant links are near zero.*

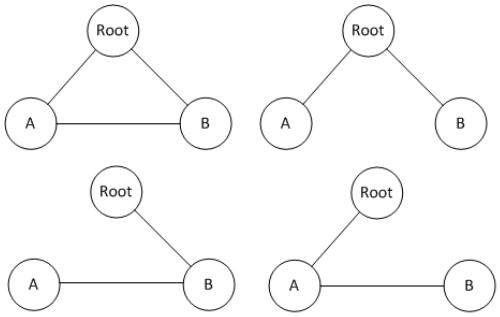


Figure 1: One physical topology - three logical tree topologies

A. Protocol Description

The topology resolved under MTA will have multiple paths between the root and other switches. These overlapping trees are created and maintained through the MT_VIDs. A Meshed Tree Switch (MTS) that has membership on a tree will be assigned an MT_VID that is associated with that tree and a particular path back to the root. Critically, switches having more than one pathway back to the root will have primary, secondary, tertiary, etc., memberships in multiple trees, each having a separate and unrelated MT_VID. MT_VIDs are stored in a table and have an association with ports through which they were established. Examples of trees from a single root and associated MT_VIDs are shown in Figure 2.

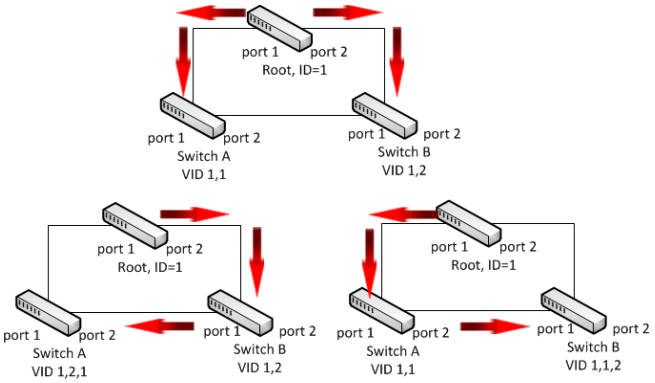


Figure 2 MT topologies and MT_VID Creation

On the top we can see that the topology is optimized to the root. The MT_VIDs (identified in the figure as VIDs) and the tree are derived based on this perspective. However, in a looped topology, the downstream or child switches have alternate paths. In the bottom left and bottom right we see the MT_VIDs that would be derived in these alternate logical topologies.

Another way to look at this is to consider the traffic that might flow between switches A and B. Clearly the topology that would be derived per spanning tree would be suboptimal. It is noteworthy that these alternate paths might be used to optimize transmissions between the hosts connected to the switches. So, another important aspect of MTP is that meshed tree switches do not possess source address tables or SATs. Instead they use a virtual SAT or

 ISBN: 978-1-61208-330-8

VSAT. MAC addresses of nodes connected locally will be learned in much the same way as described in 802.1D. Neighboring switches can exchange VSAT information in order to obtain more efficient pathways to the end hosts. This is possible as the MTP does not block ports. Within the VSAT, nodes are associated with an MT_VID for forwarding. Ports connecting the switch to a host are the *Host* ports. A port connecting a switch to another switch participating in the MTP is called an MT port because it is active in the MT topology. Port roles are shown in Fig. 3.

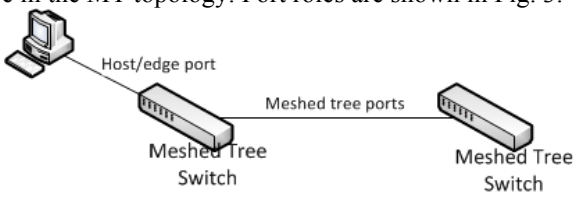


Figure 3 – Meshed Tree Switch Port Roles

B. Basic Protocol Operation

Switches join a meshed tree topology by either advertising themselves or hearing an advertisement from another MTS. Switches exchange *Hello* messages and establish an MT_VID. The MT_VID is derived from the parent MTS and the port transmitting the *Hello* message. This is explained with two switches in Figure 4.

Once all switches have at least one MT_VID, the forwarding topology can be viewed as an MT_VID tree. One of these MT_VID trees will be identified as the primary VID (PVID) tree. Unknown MAC addresses, broadcast and multicast traffic will be forwarded via this tree.

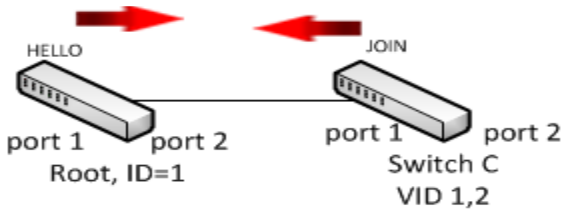


Figure 4 – Meshed Tree Hello and Join Process

Once switches have joined the MT topology and understand their parent and child relationships via the MT_VID, they exchange information contained in their VSATs via *VSAT Update* messages (VUM). Upon receipt, the VSAT in the receiving switch is modified in order to provide optimized forwarding to destination host MAC addresses. In more complex topologies, there will be superior pathways between some hosts and these can easily be identified through the VID structure. For example, parent and child switches are direct neighbors.

On discovery of a link failure or other problem, the meshed tree topology responds by deleting MT_VIDs from a Switch's MT_VID table and any VSAT entry associated with the lost MT_VID. Because redundant paths are permitted, the topology may have an alternative pathway immediately available. This path may now be elevated to the PVID. Generally speaking, shorter MT_VIDs are preferred as they represent a shorter path, though allowance for cost can be implemented.

Broadcast Packets: For forwarding broadcast packets or packets to unknown destinations, the switches should associate the MT_VIDs to the ports through which they were acquired. Thus, when forwarding to an MT_VID, the switch is correctly and efficiently forwarding the frame. Non-root switches forward broadcast frames using the following guidelines; If the broadcast frame is received from the port of PVID, it is sent out on all ports that have an MT_VID derived from the PVID and all host ports. However, if the broadcast frame is received from any other port, it is sent out on ports associated with the PVID and all host ports.

IV. Optimized Forwarding

All switches that have MT_VIDs populate a VSAT that is indexed by Host MAC address. Locally connected hosts are added to the VSAT and in this case the port field is populated with the local switch port. Hosts connected to other switches will be represented in the VSAT with a field listing all of the MT_VIDS of switches handling traffic for the hosts. This indicates that at VSAT entry for a host may have more than one possible pathway back to the host. For non-local hosts the port field will also contain the egress port for packets destined for that host MAC address. Every time a VSAT entry is changed the forwarding port field is updated to reflect this change.

A. VSAT Update Message

When a Host leaves, its timer expires, or when a new host connects on a port, the switch creates a VSAT Update Message (VUM) and sends the VUM as shown in Fig. 5. A VUM;

- Includes only the changes to the VSAT
- Is sent out on all MT ports using an MT multicast destination address
- Includes Host MAC addresses and list of MT_VIDs of the associated switch
- Includes a flag to indicate addition or removal
- Contains a sequence number to avoid duplication of activity and ordering

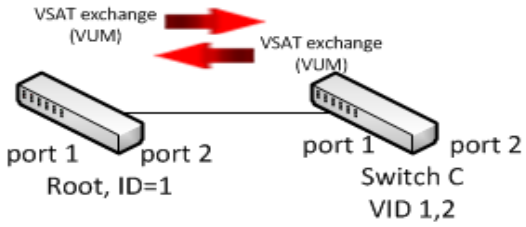


Figure 5 – VSAT updates

For each Host MAC address in the received VUM, MTS processes the message as follows;

- If the information is different than an existing VSAT entry – replace if the VUM sequence number is higher

 ISBN: 978-1-61208-330-8

- If not already in the VSAT – add an entry
- If a matching entry exists in the VSAT – do nothing

If changes were made to the VSAT, the switch creates a new VUM to reflect the changes and multicasts the VUM on all MT ports except the port that received the change. In this way, all of the switches in the topology learn of the VSAT changes.

C. Frame Delivery Process

Following cases can occur when forwarding frames:

1. Destination is this switch
2. Destination is in this MT branch away from root
3. Destination is in this MT branch towards root
4. Destination is in a different MT branch off of a switch towards the root
5. Destination is in another MT branch off of the root
6. Destination is on a switch that no longer exists
7. Destination has moved

The switch receiving a frame to forward will look up the destination MAC in its VSAT to obtain the switch's MT_VIDs associated with the Host. The switch must then follow a standard decision tree.

Case1: Is there an exact match in the local MT_VID table to the destination host switch's MT_VID?

- YES – the host is on the local switch and the frame has to be delivered through the local port.
- NO – frame forwarding will be handled by one the following cases

Case 2: Find shortest entry in the forwarding switch's MT_VIDs that is a parent (or grandparent, etc.) to the destination MT_VID. Select the next digit from the MT_VID after the matching pattern – this will be the port to forward the frame.

Case 3: Find shortest entry in the forwarding switch's MT_VID for which the destination switch's MT_VID is a parent (or grandparent, etc.). If there is a tie, pick one. Retrieve the port from the VID table – this will be the port to forward the frame.

Case 4: Find an entry in the forwarding switch's VID list that has a common parent (or grandparent, etc.) with the destination switch's MT_VID. This will resolve to the forking switch that leads to the destination. When that switch receives the frame it will use case 3 to direct the frame down the correct branch.

Case 5: This is a special instance of Case 4 where the common parent (or grandparent, etc.) is the root switch. When the root switch gets the frame it will follow case 2 to determine correct branch to send the frame on.

The above process can be executed on receiving a VUM and the ports associated with the host MAC address can be populated in the VSAT. A typical VSAT entry would be as shown in Fig. 6.

MAC	port	VID
00:01:02:03:04:05	23	1,1 1,2,3

Figure 6. VSAT entry

V. SIMULATIONS AND PEFORMANCE

The models for MTP were developed in OPNET using two scenarios; one with four switches and 1 loop, the other with six switches and 2 loops. For comparison the OPNET model for RSTP was utilized. The following performance parameters were targeted; ;.

MTP Single Tree Creation (MSTC) Time: The interval required for all switches to receive at least one MT_VID and can start forwarding frames.

MTP Meshed Tree Creation (MMTC) Time: Each Switch was allowed a maximum of three MT_VIDs. The time taken by all switches to record a maximum of the three different best paths was recorded. In MTP this would be the time when on link failures the backup paths can be used without new tree resolution.

MTP VSAT Update (MVSAT) time: The time taken for all switches to record a path to all hosts subsequent to receiving VUMs. At this time unicast frames can be optimally forwarded via other VIDs other than the primary.

RSTP initial convergence (IC) time was recorded when the spanning tree was formed. RSTP broadcasts unicast frames to unknown destinations at this time, as learning time is removed to improve convergence time.

Maximum hops taken by frames.

The converged topologies for MTP and RSTP in the case of the 4-switch scenario are shown in Fig. 7.

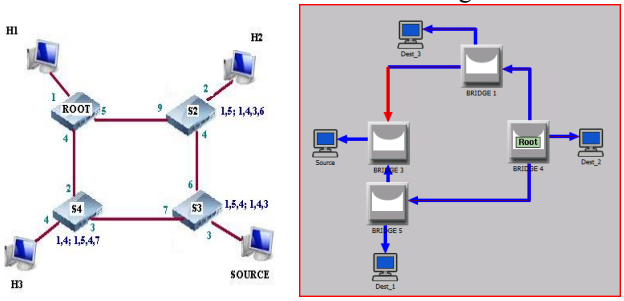


Figure 7. Meshed trees (left), spanning tree (right)

The MT_VIDs in [Figure 7] identify the three trees on which switches S2, S3 and S4 reside. The red line indicates the blocked port in the spanning tree. A host was connected to every switch. One host was identified as the source, which sent packets continuously, while the other hosts sent only for 3 seconds from the start of the simulation.

Packet exponential inter-arrival time at the hosts was set to 0.01 sec. At the switches the control traffic service rate was set to 100,000 packets per sec, while the data traffic service rate was 500,000 packets per sec. Duplex Link speed were maintained at 100 Mbps. Packet sizes were1500 bytes. The duration of simulation was set to 20 secs.

 ISBN: 978-1-61208-330-8

A. 4-Switch Single Loop Scenario

In this scenario, MSTC was recorded as 0.000037 sec, MMTC = 0.000047 sec, while MSAT was 0.0209882 sec. In the case of RSTP, IC was recorded to be 0.55 seconds. In MTP even if we avoided the flooding during the time that switches learn the host addresses through VUMs, the improvement in convergence is 26 times faster than RSTP. If we allow for frame flooding then the convergence time improvement is several thousand times. The hops taken by packets in MTP were recorded to be a maximum of 3 hops. In the case of RSTP the maximum hops would be 4.

Table 1: CONVERGENCE TIMES IN MTP

SEED	MSTC	MMTC	MSAT
127	0.000037	0.000047	0.028708
317	0.000037	0.000047	0.007826
509	0.000037	0.000047	0.024935
1009	0.000037	0.000047	0.019308
1721	0.000037	0.000047	0.024164

Note in Table 1, for seed 317, the MSAT was as low as 0.007826. The reason for the variance; when the switch gets the first data packet, it may not have had an MT_VID and hence that packet would have been discarded. The arrival of the second data packet would depend on the seed since the inter-arrival time for data packets is an exponential distribution. So if the second data packet were to trigger VSAT updates from some of the switches, the convergence time would be different for different seeds. Hence this convergence time depended on the packet inter-arrival at the host. If the inter-arrival were low then the MSAT would be also very low.

B. 6-Switch – Two Loop Scenario

In this scenario, the MSTC, MMTC and MVSAT were recorded to be 0.000047 sec, 0.000070 sec and 0.0225622 seconds. The RSTP IC time was 0.56 seconds. MTP records several thousand times improvement if packets could be forwarded before learning end host addresses (i.e. without a VSAT update) and 24 times better after all host addresses were recorded in all switches. The hop counts for packets were recorded to be 6 hops as compared to a maximum of 4 hops with MTP.

The convergence times noted and the hop counts depend on the topology. With more complex and meshed topologies the convergence times and hop counts can vary significantly. For example, in a full meshed topology the maximum hop count for frames in MTP would be 2, whereas for RSTP the frames will have to travel through the root switch. The control message overhead and excess traffic due to frame flooding also would significantly differ.

C. Comparison with Link State Protocols

In the case of TRILL on RBridges and SPB, optimal pairwise paths are computed and used for frame forwarding. However, the processing complexity has increased by several magnitudes. In the case of single meshed tree MTP, optimal paths can be computed based on the MT_VIDs acquired by the switches. Since switches may not record all MT_VIDs offered, some paths may not be the shortest.

In terms of convergence, link state routing requires that all link state information to be flooded to all switches. Subsequently the Dijkstra algorithm will be run to compute the forwarding paths. During this time the SAT may not be updated and could result in unstable operation. Comparatively in MTP, the tree is built using information received from neighbor switches and flooding of information is avoided for tree resolution. In the event that tree pruning is required, the switches can still use the backup paths to forward frames.

VI. CONCLUSIONS AND FUTURE WORK

Loop free forwarding in networks with redundant paths has been hitherto addressed on the premise that a single logical tree topology originating from a root switch is essential. This resulted in the spanning tree algorithm, which had high convergence delays. This was addressed by RSTP, which continued to face several disadvantages as stated by their inventors. More complex IS-IS based routing solutions are being adopted at layer 2. This article describes a simple solution that can replace STA algorithm at layer 2, without its disadvantages, while at the same time avoid the complexity from adopting layer 3 routing solutions at layer 2. Specification of the MTP is currently being developed under a new IEEE standard [8].

MTP performance has been compared with RSTP in terms of convergence times and path hop counts taken by framed. The superior performance achieved with MTP can be noted from these results. These results can also be used as benchmark when TRILL and SPB are evaluated.

REFERENCES

[1] LAN/MAN Standards Committee of the IEEE Computer Society, ed. (1998). *ANSI/IEEE Std 802.1D, 1998 Edition, Part 3: Media Access Control (MAC) Bridges*. IEEE

[2] W. Wodjek, "Rapid Spanning Tree Protocol: A new solution from old technology", http://www.redes.upv.es/ralir/MforS/RSTPtutorial.pdf Reprinted from CompactPCI Systems / March 2003 [Retreived Feb, 2014]

[3] R. Perlman, D. Eastlake, G. D. Dutt and A. G. Gai, "Rbridges: Base Protocol Specification", RFC 6325, July 2011.

[4] J. Touch, R. Perlman, "Transparent Interconnection of Lots of Links (TRILL): Problem and Applicability Statement", RFC 5556.

[5] N. Shenoy, Y. Pan, D. Narayan, D. Ross and C. Lutzer, "Route robustness of a multi-meshed tree routing scheme for internet MANETs", Proceeding of IEEE Globecom 2005. 28th Nov – 2nd Dec. 2005 St Louis, pp. 3346-3351.

[6] N. Shenoy and S. Mishra, "Multi Hop routing and load balancing in Mobile Ad hoc networks", Book Chapter in Encyclopedia on Ad Hoc and Ubiquitous Computing, Published by World Scientific Book Company, 2008.

[7] Network Simulation (OPNET Modeler Suite), Riverbed Technologies

[8] 1910 Working Group for Loop-Free Switching and Routing, Project 1910.1 Standard for Meshed Tree Bridging with Loop Free Forwarding.

 ISBN: 978-1-61208-330-8

Measuring Quality and Penetration of IPv6 Services

Matěj Grégr, Miroslav Švéda
Brno University of Technology
Faculty of Information Technology
Božetěchova 2, 612 66 Brno, Czech republic
Email: igregr,sveda@fit.vutbr.cz

Tomáš Podermański
Brno University of Technology
Center of Computer and Information Services
Antonínská 1, 601 90 Brno, Czech republic
Email: tpoder@cis.vutbr.cz

***Abstract*—Much has been written about the depletion of IANA's pool of IPv4 addresses and the needs of urgent transition to IPv6. Several of the biggest content and service providers have enabled IPv6, however, there is still the lack of information about worldwide IPv6 adoption and quality of services that are measured in large scale. The aim of this paper is to present our methodology to measure penetration and quality of IPv6 adoption amongst web, mail and DNS service providers. The system is built to provide an open online access to IPv6 adoption overview for the whole community. The paper discusses our methodology and measurement system and compares them with other known solutions. The analysis of collected data is presented to help to understand the IPv6 penetration.**

Keywords–ipv6,dns,mail,speed,measurement

I. Introduction

Two crucial events concerning IPv6 happened in last two years. The IPv6 Day on June 8th 2011 and IPv6 Launch on June 6th 2012. These events should have motivated the activity of service and content providers to enable IPv6. The IPv6 Day was considered as a testing day on which mainly the content providers enabled IPv6 for 24 hours. The event should have tested how many clients would connect to their dual stack web servers, how big would be the bandwidth shift from IPv4 to IPv6 and how many clients would have problems with their IPv6 connectivity. Observation confirmed an increase of IPv6 traffic and changes in application mix. Google kept IPv6 enabled for several YouTube servers, thus the main contributor for IPv6 traffic since then is YouTube as Sarrar et al. [1] showed in their report and our observation confirms their results. Thanks to the results from IPv6 day, big content providers decided to turn on IPv6 permanently one year later on IPv6 Launch day.

However, big content providers such as Google, YouTube, Akamai etc. do not represent the whole Internet. Millions of other websites still remain without IPv6 addresses. What is the ratio of enabled IPv6 websites and services (e.g., mail servers, name servers) and is the ratio increasing or stagnant? Even more, if IPv6 is enabled, is the quality of service (e.g., response time) better, worse or the same as in IPv4? These are the questions the paper tries to answer.

The paper is organized as follows. Section II describes related work. Methodology, system's architecture and implementation are described in section III. Data analysis and comparison of projects measuring IPv6 penetration are discussed in section IV. Section V evaluates validity of results. Comparison of results with other projects is presented as well. Conclusion and future work are discussed in section VI.

II. Related Work

There are several approaches to measure IPv6 adoption. Measurement can be performed on the content provider's side. The methodology is typically based on inserting a small invisible image [2] or JavaScript fragment [3] into the content of content provider's web page. Client's browser executes the code or tries to download the image using IPv4, IPv6 or other transition technology (6to4, Teredo). The analysis of requests can show the number of clients that can or cannot connect to dual-stack Web servers and their latency. This methodology measures clients. It shows, if IPv6 is supported by application (web browser), operating system and client's ISP (Internet Service Provider). The number of "consumers" is essential for content providers, because without IPv6 active customers they will not invest their time and money to IPv6 transition. Currently, the numbers are around 2.5 %, as it is shown in Google's IPv6 statistics, with some exceptions like Romania and France as shown in Geoff Huston's statistics [4].

Another method is based on measuring the number of autonomous systems announcing an IPv6 prefix. The statistics inform how ISPs and transition networks are prepared to provide IPv6 connectivity for their customers. The good analysis was presented by Karpilovsky et al. [5]. Their study has shown, that almost half of prefixes are not used at all and the rest of them is announced long after the allocation. The drawback of the methodology is that the presence of an ISP's IPv6 prefix in the global BGP (Border Gateway Protocol) table does not mean availability of IPv6 for ISP's customers. Despite of that, the number of IPv6 prefixes in global routing table is increasing and can be seen as IPv6 transition progress.

One way of measuring the quality of IPv6 service is to measure the one-way delay. Zhou et al. [6], [7] published a study comparing IPv4 and IPv6 one-way delay between several measurement points. They analyzed the RIPE IPv6 data, which include traceroute, delay, and loss measurements among a list of IPv6 sites since 2003. Their conclusion is that native IPv6 paths have small 2.5 percentile and median end-to-end delay, and comparable delay to their IPv4 counterparts. The study presented by Arthur Berger [8] found that the latency is less over v4 than v6. For example, for destinations in the North America, the mean latency is 55 ms over v4 but substantially higher, 101 ms, over v6. The difference between the IPv4 - IPv6 performance is more likely correlated with a different forward AS-level path as was reported by Amogh Dhamdhere et al. [9]. The measurement by Mehdi Nikkhah et al. [10] compares performance of IPv6 and IPv4 protocol by measuring the web page download time. They found that control plane

 ISBN: 978-1-61208-330-8

(routing) is responsible for differences between IPv4 and IPv6, because the data plane performs comparably.

The methods described above provide information about IPv6 prefix allocation, measures of clients, or test the path delay. The paper presented by Jakub Czyz et al. [11] tries to analysis the IPv6 adoption from several perspectives - allocation of IPv6 prefixes, clients readiness etc. However, the ISPs are more interested in the number of content providers that enabled IPv6 for their services. The number can indicate how much IPv6 traffic service providers will expect in case they decide to deploy IPv6. The next important information is the quantity of sites reachable from IPv6 only networks. Several measurements have been published to describe these information [12]–[17]. These papers will be analyzed in more detail in the Section V.

This paper describes the measuring platform for gathering long-term statistics about IPv6 penetration amongst content providers. The measuring is focused on the penetration of web services, mail services and name services available over IPv6. The paper compares also IPv4 and IPv6 one-way delay for web services to measure the quality of connection from the users perspective. The results obtained during the information gathering are discussed together with comparison to other services using similar methodology.

III. Methodology, Architecture and Implementation

IPv6 penetration amongst content providers can be measured by checking the appropriate resource records (RR) in DNS. Table I briefly describes the RRs tested in our methodology. If any record contains `CNAME` record it is followed until a valid record for IPv4 (`A`) or IPv6 (`AAAA`) is found. The web services are also checked if an alternative record for IPv6 is available such as `www6, ipv6, www.v6`. The alternative records are sometimes used by network administrators for testing purposes. Availability of records in DNS is tested periodically and is described in more detail in the section III-A.

TABLE I
Resource records checked

Service	Record type	Test
Web	A	`www.<domain>`
	AAAA	`www.<domain>`
	AAAA	`www6|ipv6|www.v6.<domain>`
Mail	A for MX	`<domain>`
	AAAA for MX	`<domain>`
DNS	A for NS	`<domain>`
	AAAA for NS	`<domain>`

The response time of web services is measured for sites announcing web services over IPv4 and IPv6 protocol as well. Using IPv4 and IPv6, the system tries to connect to the remote web server. The time is measured between the first packet initiating relation (SYN) and the answer from the server (SYN, ACK). Comparing the results obtained via IPv4 and IPv6, it can be observed, which protocol has a better response. We are aware of the fact, that the IPv6/IPv4 response time can vary from location to location (due to asymmetric routing, missing IPv4/IPv6 peering, different number of hops, link quality, etc.), however, the goal is to measure the quality from the perspective of our users.

A. Architecture and Implementation

The architecture of the system is divided into several blocks as is depicted in the Figure 1. The core of the system is SQL database containing list of domains and statistical data. There are two subsystems connected to the database. The first one is responsible for querying DNS system. It takes the list of domains from the database and periodically updates data with the information gathered from DNS. The history of changes for each record is stored as well, allowing us to provide current and historic data for each domain in the database. The next subsystem performs the check of IPv6 quality by measuring the one path delay as was described in the previous section. It also updates information into domains database and stores historical information of each measurement.

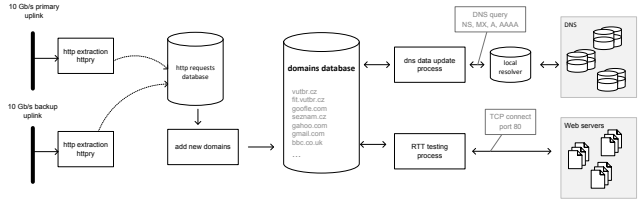


Figure 1. Architecture of the system.

B. Frequency of Updates

The DNS update process is executed every day, running in parallel using 100 threads to get the maximum performance. To update records and gather all data related to one domain, approximately 10 queries are needed. The system is able to obtain data for more than 100 domains per second. That means, the system must perform approximately 1000 DNS queries per second. Using database containing almost 9 million records, the whole update process takes approximately 2 days. Domains are queried in a random order to avoid the overloading of DNS servers. As was stated in the previous section, web domains supporting both protocols are checked to measure the quality of service. The update process is executed every day for domains that enabled IPv6 for a web service in past 24 hours. All dual stacked domains in the database are checked periodically once a week. The whole update process takes close to 6 hours for the database containing currently approximately 415 000 of dual stacked domains. However, the DNS system contains errors causing data inaccuracies. To get rid of invalid data, the system uses following rules:

- Queries returning addresses that belong to the reserved address space (private, link local) are ignored e.g., `192.168.0.0/16, 10.0.0.0/8, fe80::/10`.
- Address of loopback is ignored (`127.0.0.1/8, ::1/128`)
- Records with endless loop are ignored - a `CNAME` record points to another `CNAME` which points back to the original `CNAME`.
- The `CNAME` chain is followed up to 10 levels
- Manual patterns for domain names, e.g., domains containing random names like

 ISBN: 978-1-61208-330-8

www.335297538.acaoradical.com, www.335325653.acaoradical.com etc. are ignored. The list of patterns is managed manually.

C. Data Sources

The total number of tested domains will determine the precision of the system as is shown in the section IV. The results will be more accurate with the growing number of tested domains. A popular source of domains is the list of domains provided by Alexa The Web Information Company. It is possible to download daily updated list of the top 1 million sites for free in csv format from the Alexa's webpage. However, we believe that only the top list is not enough. For example, there are about 4 500 domains within Czech TLD in the Alexa's top list. The total number of domains in Czech TLD is approximately 1000000 thus the list contains less than 0.5 % registered CZ TLD domains. We believe that this extensively small number of domains is not enough to provide accurate results of IPv6 penetration. Another drawback of using Alexa data set only is that all sub domains are aggregated to appropriate TLD. It is quite common, that subdomains are used for different services, e.g., scholar.google.com and maps.google.com, however, subdomains are included neither in top 1 million sites nor in the list of top 500 sites per country. Using only top 1 million sites, it is impossible to check if subdomains are accessible via IPv6 or not. In the previous example, scholar.google.com is accessible only over IPv4 and maps.google.com is dual stacked website.

Other sources populating the database can be logs from mail server, DNS cache and reverse lookup of addresses accessing a network. The amount of domains that can be collected using these methods depends mainly on the network design and the number of users. In our case, we were able to collect initially approximately 100 000 of valid domains. Unfortunately, the growth of domains in the database was slow. The main reason for the slow growth is that we are a campus network without control of DNS and mail servers used by faculties of the university. To extend the number of domains and overcome the drawbacks of using only Alexa top sites, and DNS cache, the following solution was developed. Several probes are used in the Brno University of Technology's campus network listening to all HTTP requests performed by users. The requests are inspected using the httpry HTTP analyzer [18]. The advantage of this solution is that the probes are able to intercept all HTTP traffic from the whole university. The output from the analyzer is sent to a collector where the requests are stored in the *HTTP requests database* as depicted in the Figure 1. Once per day, the update process adds the new unique domains into the main database. Another source of new domains is a web interface that we provide on our site. Anybody can use the interface to check data of any domain and if the requested domain is not found, it will be added into the database. Lastly, the system is open to add any other source - e.g., import of whole zone file.

It is also important to remove domains that do not exist or have disappeared from DNS. If the domain does not contain any valid record (A, AAAA, CNAME, MX, NS), the domain is removed from the database including all related historical records. If the domain is added later using one of the approaches described above, it is treated as a new record in the database.

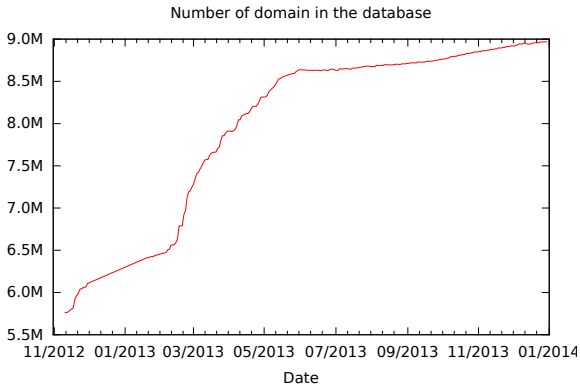


Figure 2. The number of domains in the database.

As of December 2013, the database contains approximately 9 million of domains [19]. The number is depicted in the Figure 2 and it is still growing as connected users are actively building the database with their HTTP requests. To summarize, the benefits of this approach is following:

- Independence on third party data sets (Alexa list).
- Visibility of IPv6 enabled sites, which are interesting for our users.
- Visibility of sub domains in a TLDdomain. This is useful because Alexa list does not provide information, about visited subdomains, only aggregated data.
- Long term solution with minimum maintenance.

IV. Data Analysis

Data collection is still ongoing. This paper presents data collected up to August 2013. The total number of web domains is defined as NWT, the number of web domains supporting web services over IPv6 as NWV6, the number of web domains supporting dualstack as NWDS and the number of web domains supporting IPv6 web through alternative name as NWA6.

- NWT - domains having at least one IPv4 or IPv6 record announced for www.<domain>.
- NWV6 - domains having at least one IPv6 (AAAA) record and do not have IPv4 record (A) announced for www.<domain>.
- NWV4 - domains having at least one IPv4 (A) record and do not have IPv6 record (AAAA) announced for www.<domain> and do not have alternative IPv6 record www6|ipv6|www.v6.<domain>
- NWDS - domains having both IPv6 and IPv4 records announced for www.<domain>.
- NWA6 - domains announcing any of www6|ipv6|www.v6.<domain> via IPv6 (AAAA) and do not announce IPv6 for a record www.<domain>.

 ISBN: 978-1-61208-330-8

TABLE II
IPV6 PENETRATION AMONGST WEB, MAIL AND DNS SERVICES - RATIO ON 15TH OF AUGUST 2013

	IPv4 only	IPv6 only	Dual stack	IPv6 alt. name
Web service	94.77 %	0.00150 %	4.78 %	0.45 %
Mail service	87.29 %	0.00031 %	12.71 %	
DNS service	74.58 %	0.00032 %	25.42 %	

The penetration ratio of IPv4 only sites is computed as shown in equation 1. Other ratios are computed using the same formula but the numerator is changed accordingly to NWV6 for IPv6 only sites ratio, NWDS for dualstack ratio etc.

$$NWV4ratio = \frac{NWV4}{NWT}100 \quad (1)$$

Based on these rules, we can analyze the data in our database to obtain the IPv6 penetration amongst web service as shown in Table II. As we can see, the majority of web pages is accessible using the IPv4 protocols or more precisely 99.9985 %. Despite the two big IPv6 events in last two years and the fact, that IPv4 addresses are depleted in APNIC and RIPE regions, the number of web sites accessible over IPv6 still stays on a very low level. The IPv6 only domains are usually without any meaningful content for end users. These are test websites used for testing user's IPv4/ IPv6 connectivity [4], sites where a www.<domain> has only AAAA record, but there is also record for <domain> which is accessible via IPv4 etc. DNS and mail services accessibility using IPv4 protocol is 99.99969 % and 99.99968 % but the availability using IPv6 protocol is much higher in comparison to web sites. The higher penetration of these services corresponds to deployment strategy for a new service, where an administrator goes from the key services to the less important ones.

V. VALIDITY OF THE RESULTS

Every measuring system must prove that data provided by the system are trustworthy. To provide the most accurate penetration of IPv6 services, all DNS data would have to be collected. This approach is however impossible thus we compared our data with similar projects. All projects with no exception use list of domains provided by Alexa and usually only a subset of the top 1 million is used.

The projects are compared in the Table III. The **Tests** column describes which tests projects run. The web test obtains the evidence of an AAAA record for selected domain. alt test checks an existence of alternative names for the domain (e.g., v6.<domain>). MX and NS tests are testing presence of AAAA record for mail and DNS services, DNSSEC and SPF verify the support for these services and avail test measures the quality of connection using both IPv4 and IPv6.

Results for global penetration and selected TLDs that in most statistics indicate the largest IPv6 penetration are compared in the Table IV. The comparison is based on the data from 17th August 2013 or latest available. As we can see in the Table IV the obtained results are very different. The Figure 3 shows one of the reasons for such distinction. The chart can be interpreted as follows: IPv6 penetration (y axis) is calculated for every number of domains (x axis). For example, IPv6 penetration for first 3 domains according to Alexa order

TABLE III
COMPARISON OF PROJECTS MEASURING IPV6 PENETRATION

Project	Data	Records	Tests	Freq.
IPv6matrix [12]	Alexa	top 1 million	web, alt, MX, NS	month
IPv6observatory [13]	Alexa	top 500/TLD	web, alt, MX, NS	daily
Eric Vyncke [14]	Alexa	top 50/TLD	web, alt, MX, NS	daily
Hurican Electric [15]	Alexa, zones	165 million	web, MX, NS	daily
Ari Keranen [16]	Alexa	10000	web, alt, avail	twice
6lab.cisco.com [17]	Alexa	top 500/TLD	web, alt	daily
cz.nic [20]	.cz TLD	1 million	web, MX, NS, DNSSEC	month
6lab.cz [19]	Alexa, Users	8.7 million	web, alt, MX, NS, avail, DNSSEC, SPF	daily

TABLE IV
RESULTS OF IPV6 PENETRATION PROVIDED BY DIFFERENT PROJECTS

Project	Global	.cz	.de	.fr	.ch
ipv6matrix	18.26 %	58.55 %	49.44 %	46.28 %	34.32 %
ipv6observatory	–	13.6 %	7.1 %	5.7 %	–
Eric Vyncke	7 %	30 %	10 %	8 %	46 %
Hurican Eletric	3.06 %	–	16.9 %	–	–
Ari Keranen	3 %	–	–	–	–
6lab.cisco.com	7.96 %	60.63 %	46.83 %	47.75 %	50.63 %
cz.nic	–	18.4 %	–	–	–
6lab.cz	4.79 %	14.45 %	12.09 %	3.08 %	2.34 %

is 100 % in .com TLD. If the ratio is evaluated for top 5 domains, it drops to 60 %. If we use the same data source and top 500 domains, the penetration decreases to 4.86 %. The results show that it strongly depends on the number of involved records used for calculation of IPv6 ratio. This also means, that IPv6 penetration calculated only on several top websites will tend to over-estimate the overall IPv6 penetration in web domains.

It can be seen that the IPv6 ratio becomes more stable for more than 1000 records. The increasing IPv6 penetration for some TLDs after 5 000 domains is caused by the fact, that the rest of the data set does not have any ranking, thus is sorted randomly. The number of requests must be increased at least to 10 000 records in the case of global IPv6 penetration where all domains and countries are involved. Using a smaller number of domains will again tent to over-estimate the IPv6 penetration. Similar observation was published by IPv6 Observatory [13]. A minimum of 10 000 domains is necessary to estimate the ratio of domains having an AAAA record as calculated from the top 1 million domains with an error in the range [-0.5;0.5]. The IPv6 Observatory, however, runs analysis using top 500 sites per TLD as was described in the Table IV.

Another reason for such a big difference between the projects' IPv6 penetration is that the projects use different methodologies. For example, the IPv6matrix project shows 58.55 % of IPv6 penetration for .cz TLD. This is due to the fact that IPv6matrix counts a domain as IPv6 enabled if the domain has an AAAA record for NS or MX or web or NTP server. As shown in the Table II, the penetration of AAAA records for NS and MX records is high, thus IPv6matrix shows much higher penetration than others. The 6lab.cisco.com project uses methodology which counts a weight of a domain. The weight of a domain is an approximation based on pageviews from Alexa statistics. The consequence of the weight is that more visited domains with IPv6 increase the IPv6 ratio for the country. This is based on an idea, that users are more likely to connect, spend time and access content on a very small number of sites as stated in their report [17]. To analyze the assumption, we used all HTTP requests made by all users in the

 ISBN: 978-1-61208-330-8

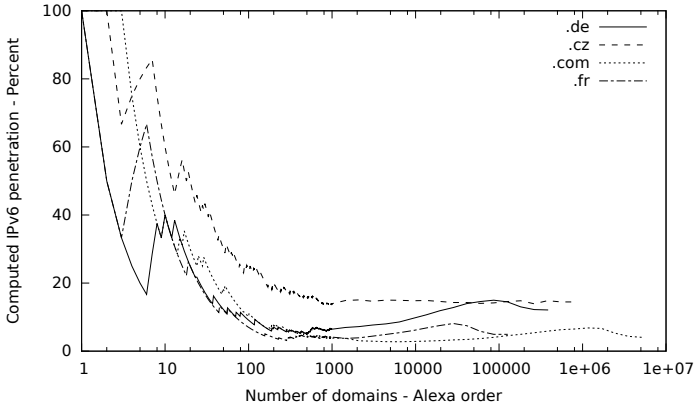


Figure 3. Dependence of the IPv6 ratio on the number of domains.

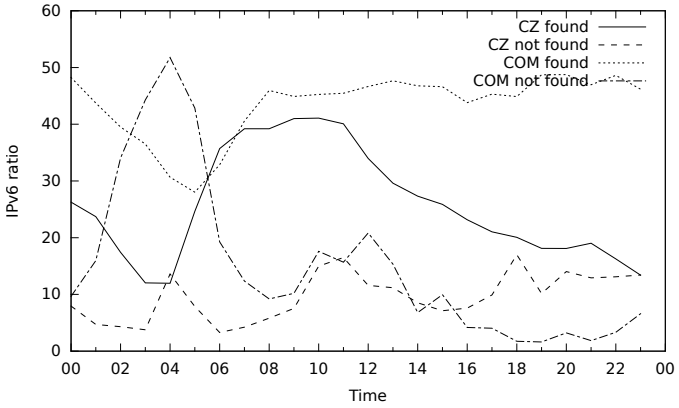


Figure 4. The IPv6 ratio of requests (not) matching Alexa list in a day.

TABLE V
HTTP REQUESTS FOUND/NOT FOUND IN THE ALEXA LIST

	Found	Not found	Ratio
All	79 216 045	15 555 245	83.58 %
.com	37 273 163	2 946 005	92.1 %
.cz	24 973 190	7 082 944	71.6 %

BUT network in a day. The requested domains were aggregated to the first subdomain after TLD - e.g., `maps.google.com` to `google.com`. This is due the fact, that Alexa list does not contain subdomains. The Table V describes the requested domains which were found or were not found in the Alexa list. The column **Ratio** stand for the chance, that a domain is found in the list. The results confirm that a majority of requests is found in the Alexa top list, but there is still a significant number of requests that is not found - almost 30 % of requests for .cz TLD are not found in the Alexa list. Relying only on Alexa list, the IPv6 penetration could be overestimated, because the domains which are not found in the Alexa list show much lower IPv6 penetration.

The `HTTP` requests for TLDs `.com` and `.cz` were further analyzed and the Figure 4 shows the difference between IPv6 penetration within the `HTTP` requests which were/were not found in the Alexa list. For example, the line `COM found` shows that IPv6 ratio is almost `50 %` amongst the `HTTP` requests which were found in the Alexa list; however, it is bellow `10 %` for the `HTTP` requests which were not found as shows the line `COM not found`. The higher IPv6 penetration during the night is mainly because of automatic updates, tests of IPv6 connection etc. - e.g., connections to `ds.download.windowsupdate.com`. These requests are not triggered by users. Although, it seems significant that approximately half of `HTTP` requests in `.com` TLD made by our users is IPv6 enabled, it is necessary to understand, that the reason lies in extensive usage of services like Google ads, Google analytics and social plugins (Facebook's 'like' button, etc.). These services are IPv6 enabled therefore, regarding a user visiting a domain which uses these services, there would be one request to obtain the domain and several other subrequests connecting to Facebook, Google and Twitter services.

The Figures 5 and 6 depict the performance between IPv4 and IPv6 connectivity using data from last two years. The round trip time is measured as described in the section III. The difference between these two protocols is counted as IPv4 - IPv6, thus negative values represents measurements where IPv6 is faster then IPv4. The graph in Figurer 5 plots the distribution of the difference using cumulative distribution function. The Figure 5 shows the frequency of measurements using histogram to provide better insight into the numbers in the <-5, 5>interval. Further analysis of the data shows, that IPv6 is more than 1 ms faster in 16% of measurements in 2012 and 22 % in 2013. IPv4 is faster more than 1 ms in 40 % of measurements in 2012 and 18 % in 2013. It can be seen, that the performance is usually similar - majority of measurements fits in the <-20, 20>interval. During the last year, the performance of IPv6 improved and usually is almost the same as IPv4. This can be a sign, that IPv6 is starting to be used for production traffic, however there is a substantial number of sites (3.51 %) we were not able to connect to. These non-working domains or slow IPv6 connectivity are nowadays usually handled by implementation of Happy Eyeballs in modern web browsers, however there is still a large fraction of clients without Happy Eyeballs implementation.

VI. CONCLUSION AND FUTURE WORK

The paper describes a system and a methodology used for measuring IPv6 penetration amongst content providers. The contribution of the paper is following. The system gathers data from various sources and the latest data were analyzed. The system does not rely on the Alexa list only but the database is built actively from users' HTTP requests in a fully automatic way. The methodology for computing the IPv6 penetration has been compared with other projects measuring the IPv6 adoption with finding that IPv6 penetration amongst domains can vary from 13 to 60 percent. These differences are caused by distinct methodologies and strongly depend on the number of measured records. Unfortunately, only a subset of available data set is used, thus the IPv6 adoption presented by different projects is usually overestimated.

Based on the analyzed data, we are recommending to use at least 1000 domains to compute the penetration for a TLD. Global IPv6 ratio should be computed using as many domains as possible. The all domains should be treat equally without using any weight or only top x sites. The current IPv6

 ISBN: 978-1-61208-330-8

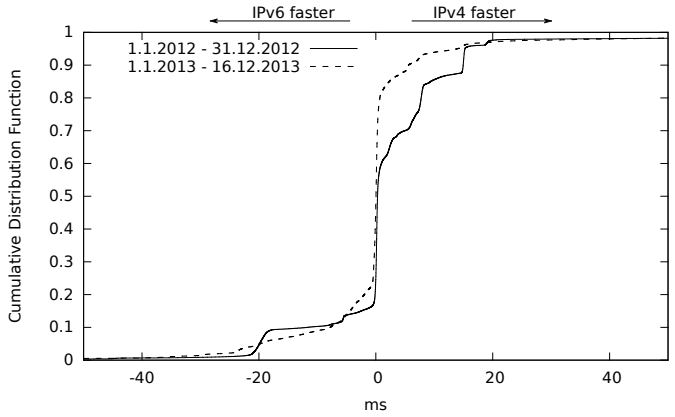


Figure 5. IPv4,IPv6 performance - CDF function

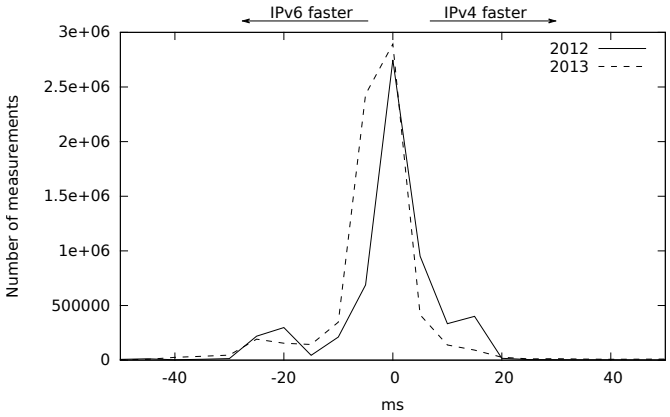


Figure 6. IPv4,IPv6 performance - number of measurements

methodologies measuring IPv6 penetration per TLD use top 50 or top 500 which overestimates the IPv6 penetration.

The presented system measures also quality of IPv6 connectivity and the results are analyzed with finding that overall IPv6 performance has been improved, although there are still number of sites (3.51 %) that are not accessible over IPv6 even though the website announce `AAAA` record.

The access to measured data together with all historical data is available and open to everybody on the project webpage [19]. As far as we know, our framework and methodology is the only one, which save all historical data of every page, thus we are able to see, when the `AAAA` record is added, removed etc. Other methodologies measuring the IPv6 adoption update the global IPv6 penetration only and does not keep the historical records. The future work could check sites' unavailability from different measuring points. We are also collecting the information, if a domain is signed with DNSSEC [21] or if a mail server provides SPF record [22]. This could also be a part of a future analysis. We plan to keep collecting and processing data, together with checking IPv4 and IPv6 performance.

ACKNOWLEDGMENT

This work is a part of the project VG20102015022 supported by Ministry of the Interior of the Czech Republic. This work was also supported by the project FIT-S-14-2299 Research and application of advanced methods in ICT.

REFERENCES

[1] N. Sarrar, G. Maier, B. Ager, R. Sommer, and S. Uhlig, "Investigating IPv6 Traffic," in Passive and Active Measurement. Springer Berlin / Heidelberg, 2012.

[2] G. Huston, "Testing IPv6 for World IPv6 Day," May 2011, [cit. 27.2.2014]. [Online]. Available: http://www.potaroo.net/ispcol/2011-05/ip6test.html

[3] L. Colitti, S. Gunderson, E. Kline, and T. Refice, "Evaluating IPv6 Adoption in the Internet," in Passive and Active Measurement. Springer Berlin / Heidelberg, 2010.

[4] G. Huston, "Measuring IPv6 - Country by Country," [online], July 2012. [Online]. Available: http://www.potaroo.net/ispcol/2012-07/v6report.html

[5] E. Karpilovski, A. Gerber, D. Pei, J. Rexford, and A. Shaikh, "Quantifying the Extent of IPv6 Deployment," in Network Measurement. Springer Berlin, 2009.

[6] X. Zhou and P. Van Mieghem, "Hopcount and E2E Delay: IPv6 Versus IPv4," in Passive and Active Measurement. Springer Berlin / Heidelberg, 2005.

[7] X. Zhou, M. Jacobsson, H. Uijterwaal, and P. Van Mieghem, "IPv6 delay and loss performance evolution," International Journal of Communication Systems, vol. 21, no. 6, 2008, pp. 643–663.

[8] A. Berger, "Comparison of performance over ipv6 versus ipv4," Akamai Technologies, 2011.

[9] A. Dhamdhere, M. Luckie, B. Huffaker, k. claffy, A. Elmokashfi, and E. Aben, "Measuring the deployment of IPv6: topology, routing and performance," in Proceedings of the 2012 ACM conference on Internet measurement conference. ACM, 2012, pp. 537–550.

[10] M. Nikkhah, R. Guérin, Y. Lee, and R. Woundy, "Assessing ipv6 through web access a measurement study and its findings," in Proceedings of the Seventh COnference on emerging Networking EXperiments and Technologies. ACM, 2011, p. 26.

[11] J. Czyz, M. Allman, J. Zhang, S. Iekel-Johnson, E. Osterweil, and M. Bailey, "Measuring IPv6 Adoption," Technical Report TR-13-004, ICSI, Tech. Rep., 2013.

[12] O. M. Crpin-Leblond, "IPv6 Matrix Project," [cit. 27.2.2014]. [Online]. Available: http://www.ipv6matrix.org

[13] IPv6 Observatory, "Top-500 websites with AAAA records," [cit. 27.2.2014]. [Online]. Available: http://www.ipv6observatory.eu/indicator/

[14] E. Vyncke, "IPv6 Deployment Aggregated Status," [cit. 5.3.2014]. [Online]. Available: http://www.vyncke.org/ipv6status

[15] M. Leber, "Global IPv6 Deployment Progress Report," [cit. 16.12.2013]. [Online]. Available: http://bgp.he.net/ipv6-progress-report.cgi

[16] A. Keranen and J. Arkko, "Some Measurements on World IPv6 Day from an End-User Perspective," RFC 6948, 2013. [Online]. Available: http://www.ietf.org/rfc/rfc6948.txt

[17] H. Kaczmarek, "Internet IPv6 Adoption: Methodology, Measurement and Tools," [online], July 2012. [Online]. Available: http://6lab.cisco.com/stats/information

[18] J. Bitte, "HTTP logging and information retrieval tool," [cit. 27.2.2014]. [Online]. Available: http://dumpsterventures.com/jason/httpry/

[19] T. Podermaski and M. Grgr, "Worldwide online IPv6 penetration," August 2013, [cit. 27.2.2014]. [Online]. Available: http://6lab.cz/live-statistics/data-source

[20] CZ.NIC, ".cz statistics," [cit. 12.2.2014]. [Online]. Available: http://stats.nic.cz

[21] R. Arends, R. Austein, M. Larson, D. Massey, and S. Rose, "Resource Records for the DNS Security Extensions," RFC 4034, Mar. 2005. [Online]. Available: http://www.ietf.org/rfc/rfc4034.txt

[22] M. Wong and W. Schlitt, "Sender Policy Framework (SPF) for Authorizing Use of Domains in E-Mail, Version 1," RFC 4408, Apr. 2006. [Online]. Available: http://www.ietf.org/rfc/rfc4408.txt

 ISBN: 978-1-61208-330-8

On Service-Oriented Architectures for Mobile and Internet Applications

Sathiamoorthy Manoharan
Department of Computer Science
University of Auckland
New Zealand

***Abstract*–Service-oriented architectures have been around for long now, but the surge in the Smartphone and tablet market and the wide availability of fast mobile networks now cast new light on service-oriented architectures. The diversity of mobile platforms demand application abstraction. Such abstraction can be made possible by adapting a service-oriented architecture where the bulk of the business logic is hosted as a service. A service-oriented architecture may appear to be unsuitable when mobile networks are slow or unreliable. However, most modern mobile networks are reliable and reasonably fast, and thus applications employing a service-oriented architecture do not necessarily reduce user experience when compared to native, device-hosted applications. This paper reviews some of the technological advantages and challenges arising from the use of a service-oriented architecture for mobile and Internet applications.**

***Keywords*-Service-oriented architecture (SOA), Software architecture, Application layer, Application security.**

I. INTRODUCTION

The surge in the Smartphone and tablet market and the wide availability of fast mobile networks now cast new light on service-oriented architectures.

The diversity of mobile platforms poses interesting challenges to application developers. To reach all potential users of the application, the developers need to make the application available for a number of platforms where the operating systems [1], [2], development environments and languages [3], [4], [5] may substantially differ. This may therefore require the developers to 'replicate' code for these different platforms. However, code replication, in whatever form, does not follow good software engineering principles. Depending on the type of the application, a possible alternative that mitigates code 'replication' is to employ a service-oriented architecture [6]. In this case, the bulk of the business logic is centralized in a service and is published on a centrally located server (typically in the Cloud). Thin clients are then developed for the chosen number of platforms; these clients consume what the service offers and typically share little code.

See Figure 1 that shows a broad overview of service-oriented architecture. The clients can either be native or browser-based, depending on the application requirements.

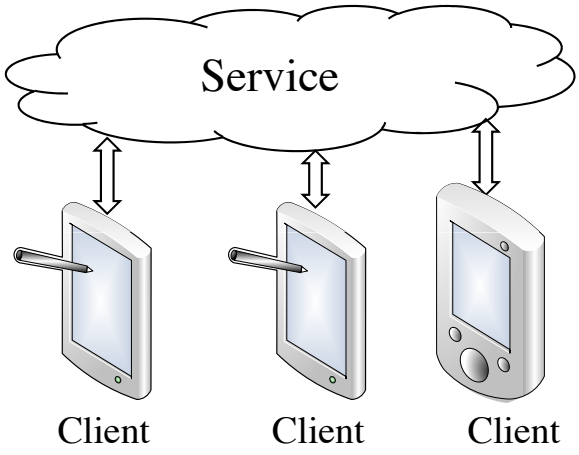


Figure 1. An overview of service-oriented architecture

Two immediately recognizable challenges with such a service-oriented approach is availability and performance. The thin client may not be available for use if there is no or limited connectivity to the service. The amount and frequency of data transfer between the client and the service may lead to performance bottlenecks. Not only that the application logic needs to take into account its algorithmic efficiency, but also it needs to take into account its data transfer efficiency.

This paper reviews some of the technological advantages and challenges arising from the use of service-oriented architectures for mobile and Internet applications. Note that this is a review paper drawing from the practice and experience of constructing services and clients. It does not propose new research results.

The rest of the paper is organized as follows. Section II reviews the fundamental aspects of services and service-oriented architectures. Section III discusses performance of various segments of service-oriented architectures. Section IV explores how service-oriented architectures can be secured using standard Internet authentication mechanisms. Section V presents a case study illustrating some of the concepts discussed in this paper. The final section concludes the review with a summary.

 ISBN: 978-1-61208-330-8

II. Services and Service-oriented Architectures

The World Wide Web Consortium (W3C) defines service-oriented architecture as a "set of components which can be invoked, and whose interface descriptions can be published and discovered" [7]. Sprott and Wilkes extend this to say that "services can be invoked, published and discovered, and are abstracted away from the implementation using a single, standards-based form of interface" [8]. The service interface defines the operations provided by the service and the message types and formats to be used with these operations.

SOAP [9] has traditionally been the message format used by service operations. The SOAP envelope has a header and body and the body wraps the object requested of or sent to the service. SOAP is based on XML, and is verbose. The verbosity can result in slower transfer speeds as well as slower parsing of messages. In addition, being a text format, it requires binary objects to be encoded in text, typically using base-64 [10] which expands the binary object by a factor of $\frac{4}{3}$.

SOAP-based Web Services consume SOAP messages and produce SOAP message. Since SOAP messages are verbose (and thus large), the messages are generally sent to the service using HTTP POST. A downside of using HTTP POST is that POST responses are not easy to cache.

SOAP-based Web Services are therefore not quite efficient. SOAP messages are heavy-weight and the responses from SOAP-based Web Services are difficult to cache.

Representational state transfer (REST) is an HTTP-based light-weight protocol that provides loose-coupling between the service and the client consuming the service [11], [12]. REST does not have a fixed message format, and thus a REST-based service (commonly known as a *RESTful service*) can deliver any message (i.e., any MIME type). For instance, binary objects such as a PNG image need not be encoded into a textual format in a RESTful service, thus there is no message bloat required by REST. Besides, objects can be delivered to the service using HTTP GET (i.e., as part of the URL). For example, the following URL describes a service operation that searches for the term "soa" and returns the top 10 results:

http://search.site.org/?term=soa&results=10

This URL passes to the service at *serach.site.org* two objects: a string named *term* with value *soa*, and an integer named *results* with value *10*.

HTTP GET has the advantage of letting its response cacheable. For example, the response from the above URL may be cached in the client (e.g., browser) or any other intermediaries (e.g., proxy) for a period of time.

Modern Web Services are based on REST [13]. The service operations of a RESTful Web Service are defined by URLs and are therefore a lot simpler than the equivalents in a SOAP-based service. The messages in a RESTful service are more efficient than SOAP messages. Besides, service responses can often be cached for re-use, thus saving on server resources, network bandwidth, and latency.

Overall, a RESTful Web Service is more efficient than a SOAP-based Web Service. Service-oriented architectures are therefore best-implemented using RESTful services [14].

III. Performance Considerations in Service-oriented Architectures

As we have already seen in the previous section, RESTful services outperform SOAP-based services. Thus, it is desirable to use RESTful services unless there is a strong reason not to (e.g., compatibility with legacy systems).

Given the inefficiencies of SOAP-based services, we will discuss performance considerations only for RESTful services.

One of the most important performance consideration in RESTful services is to reduce latencies involved in data transfer. Principles of latency reduction can be summarized by the three "R"s:

- Reuse
- Repetition avoidance
- Redundancy removal

In the context of service-oriented architecture, reuse is achieved by exploiting temporal locality. Items that have been used in the past are likely to be used again, and thus it pays to keep those items cached for later reuse. Effective caching can result in large savings in latencies, for there is little data transfer involved when the item is found locally in the cache.

The transport protocol HTTP supports caching extensively [15], [16]. All service responses should be considered carefully to permit caching at the service so as to optimize repeat-requests. Such caching reduces load on the service as well as reducing latency for the service consumer. Personalized responses (e.g., bank balances) will need to be cached on a per-user basis at the service; if this is not possible, then such responses must not be cached. Caching at the client end (i.e., service consumer end) should always be allowed.

When a response is cacheable, a suitable expiry time stating how long a cache could keep the response fresh should be indicated by the service. The service is in the best position to determine such expiry times. An item that is past its expiry date may not always be stale. In this case, it can still be used from the cache so long as it is re-validated to be fresh by the service. To help such re-validation, the service should consider setting a validator (e.g., HTTP *ETag* [16]).

Repetition avoidance and redundancy removal are typically achieved through compression. HTTP supports compressing the payload in responses. RESTful services thus can benefit from this. However, payload in a request is not typically compressed (by a client): payloads are allowed in HTTP POST but not in HTTP GET. If POST has to be used

 ISBN: 978-1-61208-330-8

to transfer data, then it is important to consider compressing the data prior to the transfer; and this has to be done by the service consumer manually and the service should be designed to accept compressed payloads.

HTTP headers are never compressed. The service and the consumer have to be mindful of this and ensure the headers and header values are optimally used. For instance, unnecessary headers and header values should not be exchanged.

Similarly, it is beneficial to keep service URLs terse. Given that the URLs are seldom manually processed, terseness will not be an issue. Smaller URLs are quicker to send. Besides, there are limits on URL length imposed by HTTP entities (such as browsers). For example, the sample service URL we saw in the previous section

http://search.site.org/?term=soa&results=10

can be shortened to

http://search.site.org/?t=soa&r=10

Such shortening will allow room for larger data input to the service still using HTTP GET.

Output of a service operation is typically a custom object. Both XML and JSON (JavaScript Object Notation) can be used to serialize a custom object so that the object can be transferred to the service consumer. In a RESTful service, the XML response does not include a SOAP envelope and therefore the response is more compact. However, XML is still verbose. The other widely used alternative is JSON and JSON-serialized objects are generally more compact than XML-serialized objects. BSON is binary-encoded JSON and thus is even more compact than JSON.

With compression, the size difference between JSON-serialized objects and XML-serialized objects may not be apparent. However, generating and parsing large serialized forms require additional resources, so usually it is beneficial to use a compact serialized form.

When choosing the serialization format for custom objects in a service-oriented architecture, we must consider both

1) data transfer time, and
2) serialization and de-serialization times.

IV. Security Considerations in RESTful Service-oriented Architectures

In many systems, the service operations should only be made available only to those who have access rights.

One of the simplest form of limiting service consumers is based on their IP addresses. The service can be configured to either allow or deny accesses from service consumers originating at a given range of IP addresses.

The other simple form of limiting access is based on username and password. A consumer in this case will first need to authenticate herself by supplying the pre-registered username and password combination. Once authenticated, the service may decide either to authorize the consumer or not based on the access rights setup for the user.

HTTP servers support two standard authentication mechanisms: basic and digest access authentication [17].

With basic authentication, the username and password are supplied in clear in the HTTP headers. Basic authentication is thus prone to man-in-the-middle attacks and is not secure. However, used with HTTPS which encrypts all the transactions including the authentication headers, basic authentication is secure enough.

Digest authentication never transfers the password across the transport channels [17]. It only sends the digest of the password and a number of other entities (such the user name and realm) using a challenge posed by the server. Therefore, digest authentication is stronger than basic authentication, and is a candidate to consider if using HTTPS is not possible (e.g., because of setup costs or speed).

Another aspect to consider is the security of data during transmission. Sensitive data should not be available to eavesdrop. Use of HTTPS is the widely-interoperable and simplest means of protecting data from eavesdroppers.

V. Case Study

The case study involves revamping a University website to using a service-oriented architecture so that both native mobile and web applications can be created providing a richer user experience than what is currently available with the site through a standard browser.

Most of our websites are not easily viewable on smart devices with a small screen (e.g., Smartphones). As an example, accessing the Computer Science website at the University of Auckland from a Smartphone shows the difficulties of reading and navigation [18]. This case study involves re-architecting Computer Science's information services so that the information can be rendered to suit the target device.

To this end, we (1) separate information content from the user-interface and make available the information as a service, and (2) construct applications for smart devices that consume the information service and render them to fit within the interface paradigm of the device. This is an approach taken by a number of news media providers to provide a richer experience to the readers.

A number of data sources that supply key information content from the current site have been identified. These enable separation of information content from the presentation.

A. Service Operations

The descriptions of the operation contracts supported by the service are as follows.

1) Get a list of offered courses. The URL corresponding to this service operation resembles *http://www.site.org/css/courses*.

 ISBN: 978-1-61208-330-8

2) Get a list of staff IDs (or keys). The URL corresponding to this service operation resembles *http://www.site.org/css/people*.
3) Given the ID (or key), further details, such as email, phone number and picture, of staff can be obtained in the form of a vCard [19]. The URL corresponding to this service operation resembles *http://www.site.org/css/vcard?id=jbon077* where *jbon007* is an ID within the list of staff IDs.
4) Get a home screen image. The home screen image is one that changes from invocation to invocation. The image is randomly picked from a repository of images. The URL corresponding to this service operation resembles *http://www.site.org/css/himg*. This service operation returns an image (of type PNG, JPG or GIF, which can be inferred from HTTP Content-Type header).
5) Get an RSS feed of active seminars. The URL corresponding to this service operation resembles *http://www.site.org/rss?c=seminars*. The RSS feeds have the MIME type `application/rss+xml`.
6) Get an RSS feed of active events. The URL corresponding to this service operation resembles *http://www.site.org/rss?c=events*.
7) Get an RSS feed of current news items. The URL corresponding to this service operation resembles *http://www.site.org/rss?c=news*.

The service employs some of the performance and security considerations outlined in sections III and IV.

- All of the service operations except operation 4 (home image) return compressed (gzipped) data. Note that since PNG, GIF and JPG images are already compressed, further compression is unlikely.
- The custom objects (course list and people list) are serialized to JSON.
- All operations permit client-side caching
- Given the simplicity of the operations, there is no server side caching. However, there is provision to turn on server-side caching should this become necessary.
- IP-based access restriction is supported. However, given the public nature of the data, the restriction is not turned on.
- Both basic and digest access authentication are supported. Again, given the public nature of the data, no authentication is turned on.

B. Clients

Rich mobile apps that render data provided by the service operations are then constructed. These apps fit within the UI paradigm of a smart device with a small screen (e.g., Smartphone).

The client app includes logical spaces displaying course information, staff details (including photo, email, and phone number), current seminars, current events, and published news stories.

Where possible, the client app provides simple smarts to render a rich user experience. These smarts include the following.

1) If an email is selected, the Mail app is fired to compose an email to the selected address.
2) If a phone number is selected and if the device is capable of making an outgoing call, an outgoing call is initiated.
3) Being able to add the contact details from the vCard to the Address Book.
4) Being able to add events and seminars to the Calendar.

VI. Summary and Conclusion

The "write-once run-anywhere" application development paradigm does not always provide the best user experience: the application may be slow (for instance, because of layering or poor virtualization); and the user-interface may be poor (for instance, due to being unable to use native device capabilities).

Using a service-oriented architecture gets close to the principle of the "write-once run-anywhere" paradigm, but still permits providing the best possible user experience. This is through striking a balance between shared code on the server and specialist code on the client. The client code could be native and could exploit native device capabilities.

This paper reviewed service-oriented architectures in the context of mobile and Internet applications with a particular emphasis on performance and security. Following is the summary of the key points.

- RESTful service-oriented architecture is seen to outperform SOAP-based service-oriented architecture in terms of simplicity and efficiency.
- When choosing the serialization format for custom objects, we must consider the times taken to (1) serialize the data, (2) transfer the serialized data, and (3) deserialize the received serialized data.
- HTTP caching can increase the performance of services, and caching naturally lends itself to RESTful services. Both server side and client side caching are possible and are highly recommended to reduce (1) network latency, (2) use of bandwidth, and (3) server load.
- Setting a cache validator (e.g., HTTP *ETag*) can increase the effectiveness of caching.
- All data transfers, including response and request, will benefit from compression. While responses can be compressed out-of-the-box, request compression will need custom handling. However, a well-designed service will predominantly use HTTP GET and thus may not always require request compression.
- User access control can be achieved using HTTP basic access authentication (with HTTPS) or using HTTP

 ISBN: 978-1-61208-330-8

digest access authentication. Both of these are light-weight protocols. In addition, IP-based access control can also be employed.

- Data can be secured from eavesdroppers using HTTPS.

The paper presented a case study of an application conforming to service-oriented architecture, and illustrated some of these key points using the case study.

REFERENCES

[1] Z. Mednieks, L. Dornin, G. B. Meike, and M. Nakamura, *Programming Android*. O'Reilly Media, 2011.

[2] M. Neuburg, *Programming iOS 7*, 4th ed. O'Reilly Media, 2013.

[3] J. Gosling, B. Joy, G. Steele, G. Bracha, and A. Buckley, *The Java Language Specification*. Oracle, 2013.

[4] A. Hejlsberg, M. Torgersen, S. Wiltamuth, and P. Golde, *The C# Programming Language*, 4th ed. Addison-Wesley Professional, 2010.

[5] S. G. Kochan, *Programming in Objective-C*, 6th ed. Pearson Education, 2013.

[6] N. M. Josuttis, *SOA in Practice: The Art of Distributed System Design*. O'Reilly, 2007.

[7] H. Haas and A. Brown, *Web Services Glossary*, February 2004, W3C recommendation.

[8] D. Sprott and L. Wilkes, "Understanding service-oriented architecture," *The Architecture Journal*, pp. 10–17, January 2004.

[9] M. Gudgin *et al.*, *SOAP Version 1.2 Part 1: Messaging Framework*, 2nd ed., April 2007, W3C recommendation.

[10] S. Josefsson, "The base16, base32, and base64 data encodings," *The Internet Eng. Task Force RFC 4648*, October 2006.

[11] R. T. Fielding, "Architectural styles and the design of network-based software architectures," Ph.D. dissertation, University of California, Irvine, 2000.

[12] J. Webber, *REST in Practice: Hypermedia and Systems Architecture*. O'Reilly, 2010.

[13] L. Richardson and S. Ruby, *RESTful web services*. O'Reilly, 2007.

[14] T. Erl, B. Carlyle, C. Pautasso, and R. Balasubramanian, *SOA with REST: Principles, Patterns &Constraints for Building Enterprise Solutions with REST*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2012.

[15] D. Wessels, *Web Caching*. O'Reilly & Associates, Inc., 2001.

[16] R. Fielding *et al.*, "Hypertext transfer protocol – HTTP/1.1," *The Internet Eng. Task Force RFC 2616*, June 1999.

[17] R. Franks *et al.*, "HTTP authentication: Basic and digest access authentication," *The Internet Eng. Task Force RFC 2617*, June 1999.

[18] U. of Auckland, "Department of computer science," http://www.cs.auckland.ac.nz/. Last Visited February 2014.

[19] S. Perreault, "vCard format specification," *The Internet Eng. Task Force RFC 6350*, October 2006.

 ISBN: 978-1-61208-330-8

Collaborative Wireless Access to Mitigate Roaming Costs

Carlos Ballester Lafuente, Jean-Marc Seigneur, Thibaud Lyon
Institute of Services Science
University of Geneva
Geneva, Switzerland
{carlos.ballester, jean-marc.seigneur, thibaud.lyon}@unige.ch

Abstract— **Environments such as ski slopes are highly dynamic, as users are constantly moving at high speeds and in different directions, and also many users are foreign tourists, not locals, thus having to roam if they want to access the Web. Unfortunately, this introduces costs that discourage the roaming users to connect. In order to solve this issue, a collaborative wireless access service has been designed and implemented on Android. Simply put, locals to the environment become hotspots on-the-fly thanks to our application, which works on all recent Android smartphones, without requiring to root the smartphone, which shares their mobile data access with the foreigners for the period of time that they are in range whilst legally protecting the sharer from any potential illegal use of the foreigner, e.g., illegal download of copyrighted music through peer-to-peer. We have validated our service with agent-based simulation results, users feedback through an online survey supported by an EU Future Internet testbed and real performance tests on the ski slopes regarding client-to-hotspot connection times, distance and energy consumption.**

Keywords - Crowd augmented; Wi-Fi; Smart Ski; mobility; wireless access; simulation.

I. INTRODUCTION

According to the International Telecommunication Union (ITU) and Juniper Research [1], the number of subscribers using mobile Internet services is going to rise from the current 577 million to an impressive 1.7 billion users by 2013, accounting for almost the 50% of the world's Internet usage. This previous fact and the emergence and fast growth of applications such as social networking, user generated content, location services, collaborative tools, augmented human and augmented reality applications etc., has fueled the user's need for permanent connectivity wherever she/he is, and under any circumstances.

While in regular day-to-day environments this need can be fulfilled with regular wireless access provided via hotspots (wireless access points) or mobile data transmission technologies, highly dynamic and changing environments have different requirements that might not be fulfilled with regular hotspots [10]. Also, situations on which the user is on roaming (does not have access to his mobile operator because of being in a different country or out of the area of network coverage), might deter the user to connect through such previous mentioned mobile technologies, as the cost can be very high.

TEstbed for Future Internet Services (TEFIS) [2] is a large-scale integrating project, which will support Future Internet of Services Research by offering a single access point to different testing and experimental facilities for communities of software and business developers to test, experiment, and collaboratively elaborate knowledge. As a part of TEFIS, the Smart Ski Resort experiment run in Megève during 2011/2012 and 2012/2013 winters aims to launch the next generation of intelligent ski resorts providing them with mobile applications and with the resources to create a sustainable development. For our testing purposes, we have used the Smart Ski Resort experiment to obtain the performance results shown later in Section V.

Environments such as ski slopes are highly dynamic, as users are constantly moving at high speeds and in different directions, and also many users are not locals, thus having to roam in order to be able to connect through mobile data. These two previous reasons make connectivity through regular means to be difficult to attain, thus impeding the use of such smart mobile applications, augmented reality applications, or the mere upload of data and statistics for user tracking or measuring purposes.

In order to solve such a challenge, we have envisioned a crowd augmented wireless access [9]. Simply put, locals to the environment become hotspots on the fly, sharing their mobile data access with a foreigner for the (rather short or not) period of time that they might be in range. In this way, all the foreign skiers in the slopes, and more broadly speaking foreigners in general, are still able to upload fundamental data and statistics and even use applications on places where normally they would not be able to get connectivity through their own means or would be too expensive to do so. All of this, without having to deploy real fixed wireless access points and signal amplifiers, and not limiting the area of coverage, as the access points are carried by the local people, which might be static or on the move. In this paper, we present the technical feasibility and performance results of our collaborative wireless access service, which also legally protects the sharer from any potential illegal use of the connection done by the foreigner, e.g., illegal download of copyrighted music through peer-to-peer. However, the description of how we legally enforce the protection of the sharer is beyond the scope of this paper.

The rest of the document is organized as follows. First, Section II presents the related work and then Section III describes the simulation experiment and presents the results obtained from it. After, Section IV presents the user survey feedback regarding several aspects on roaming, costs, and risk awareness. Next, Section V shows quantitative assessment in terms of performance of the Android application tested in real scenarios, both a dynamic one, in the ski slopes, and static one, in a cafeteria. It also presents performance results regarding battery consumption and CPU usage. Finally, Section VI concludes the paper.

 ISBN: 978-1-61208-330-8

II. RELATED WORK

There are several applications and projects that aim as well to enable and to make easier the task of sharing a mobile data access through Wi-Fi in order to solve similar challenges as the ones we want to tackle.

Air Mobs [3] is an application that enables users to share their excess data with users who might be running up against their monthly limits. Essentially, one user agrees to let their mobile device act as a tethering hub that will send data from their LTE smartphone over Wi-Fi to any users nearby. In exchange, the central hub user gets a "data credit" that gives them access to other users' data in the future. Put another way, the new app creates a sort of "cap-and-trade" market for mobile data that helps users exceed the hard limits set on their consumption by rationing data with one another based on their needs at given times. Compared to our solution, it does not provide any sort of protection for the sharer against any risk that might arise from the sharing and it does not provide a WPA2 secured connection.

The Open Garden application [4] application enables users to access the most appropriate connection without configuring their devices or jumping through hoops. It also enables users to access Internet as cheaply as possible. Users can find the fastest connection and most powerful signal without checking every available network, and can move between networks seamlessly. Open Garden provides a way to access more data at faster speeds in more locations. Consumers actually become part of the network, sharing connections when and where they provide the best possible access. Compared to our solution, it does not provide any sort of protection for the sharer against any risk that might arise from the sharing and it does not provide a WPA2 secured connection.

The User-Centric Local Loop (ULOOP) [5] FP7 European project brings in a fresh approach to user-centricity by exploring user-provided networking aspects in a way that expands the reach of a multi-access backbone. ULOOP addresses the user as a key component of networking services in future Internet architectures. Building upon current (commercial) examples ULOOP explores not only the adequate technical sustainability of user-centric models, but also legislation implications and the potential of community-driven services and how these new aspects may give rise to novel business models both from a user and from an access perspective. The aim of ULOOP if to seamlessly expand the backbone of the network through the end users' devices, extending the area of coverage while offloading the often saturated provider networks [11]. From a preliminary assessment, we believe that the project functionality is not yet mature enough and fully developed, and that they do not provide legal protection for the sharer as well.

III. SIMULATION

This section presents the simulation experiment in order to preliminarily evaluate the feasibility of the real application and the results obtained from running the simulation.

A. Simulation Experiment

The experiment shows the simulation of a ski slope, with local skiers which have connectivity through 2G, 3G, etc. and foreign skiers which in principle do not have connectivity of any kind. The simulation has been carried out using Any Logic's [6] agent based simulation capabilities, assigning real values and proportions to the scenario.

When the simulation is started, an introduction screen that depicts the scenario and asks for the skier concentration rate is shown. The concentration rate defines how busy the ski slopes are, and thus influences the availability of connectivity for foreign skiers. It ranges from 1 (not busy), which will deploy zero to one skier, either foreign or local, for each ski lift arrival, to 4 (very busy), which will always deploy four skiers (maximum ski lift capacity) per ski lift arrival. The foreign-to-local skier's ratio is 30%-to-70%, in order to reflect the real statistics of Megève ski resort.

The ski lifts are modeled with a discrete-event approach, having the appropriate inter-arrival time and speed. Every time one ski lift arrives to its "sink element", 3 skiers are deployed according to the concentration rate, as explained previously. We have chosen 3 as the concentration rate as it reflects a moderately full ski resort, which is the case we want to analyze. When the skiers are deployed, they are assigned an initial random speed, which ranges from the slowest to the highest average speed of a regular downhill skier, which ranges from 25 to 40 km/h, and a final destination at the end of the slope. In order to make the simulation more realistic, the trajectory between the deployment point to the end point is not set as a straight line but as a sinusoidal function which imitates the real movement of a skier. The ski slope's length is set to 800m and the width to 80m, as it is a good estimate of the length of a typical ski slope according to Megève Tourism board and Megève ski lifts' company.

Foreign skiers' terminals scan for access points, and when they are in range of one (or many) of them, they try establishing a connection as depicted in the process of Figure 1. The range of the portable access points has been set to 40 meters in the simulation, as it is the typical average value of that of a mobile device such as an Android or an iPhone terminal, and besides it has been confirmed through real experimentation cases as explained afterwards in Section V.A.

The aim of the simulation is to measure how effective a solution of this kind can prove to be in a highly dynamic environment such a ski slope, as if it is working in such an environment, it will also work in a more static environment such as a bar or a local shop. In order to study the feasibility of our approach, we measure for each foreign skier his or her connectivity duration time and their connectivity status, be it "connectivity setup" or "connected".

 ISBN: 978-1-61208-330-8

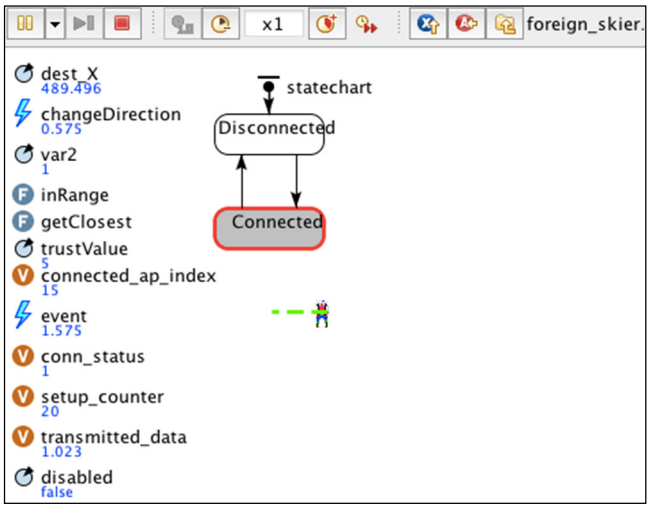


Figure 1. Foreign skier state chart and events.

Once connected, the foreign skier proceeds to transmit for the duration of the connection (assuming regular 3G/HSDPA rates), and when she or he is disconnected the process starts over again until the skier reaches the end of the slope and goes into a ski lift. This full simulation process is depicted in Figure 2, where local skiers are represented by blue avatars surrounded by a green circle which displays their portable hotspot range (as previously said set to 40m), foreign skiers as red avatars, and the connectivity status between them is represented by dashed lines, yellow in the case of a connection setup ongoing and green in the case of an already established connection.

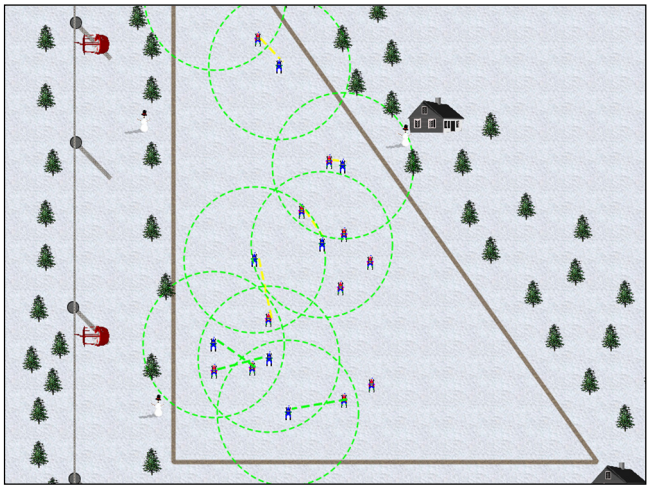


Figure 2. Main simulation screen.

Connectivity setup time has been set up to 50 seconds, as it is what we expect our application to perform like. The following section presents the results derived from the simulation.

B. *Simulation Results*

In order to study the feasibility of a real application on the ski slopes, we have measured during simulation experiments 3 sets of different data. For all the graphs, the X axis represents the simulation time:

- Global amount of foreign and local skiers in the ski slope at any given time of the simulation.
- Amount of foreign skiers in "connection setup" state vs. the amount of foreign skiers in "connected" state at any given time of the simulation.
- The maximum amount of time a skier has been connected at any given point of the simulation time and the global average time a skier is in "connected" status.

The first graph, depicted in Figure 3, shows the statistics for the global amount of local and foreign skiers present in the simulation slope at any given point of the simulation time, being the X-axis the time and the Y-axis the amount of skiers. As can be seen in the graph, the amount of local skiers ranges in average from 30 to 40 while the amount of foreign skiers ranges from roughly 10 to 20. This gives a ratio of approximately 67% of local skiers and 33% of foreign skiers, which adequately reflects the real ratio between French local skiers and foreign skiers in Megève ski resort according to Megève Tourism board. As can be inferred from this numbers, the total amount of skiers in the slope at any given point of time ranges from 50 to 60, which is a realistic measure for what a moderately busy ski slope would look.

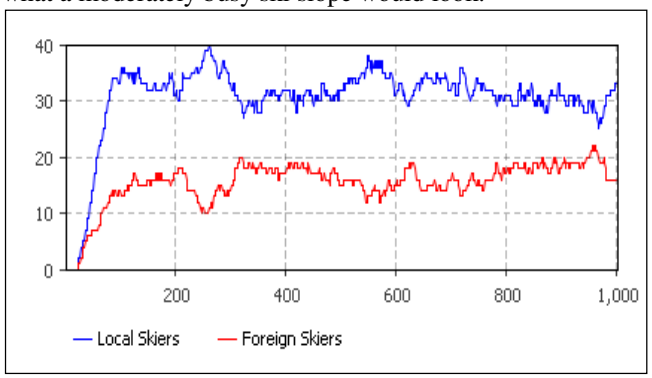


Figure 3. Amount of local and foreign skiers in the slope at any given point of the simulation time.

The second graph in Figure 4 shows the amount of foreign skiers that are setting up a connection with a local skier versus the amount of foreign skiers that are already connected and thus, transmitting data, being the X-axis the simulation time and the Y-axis the amount of skiers in each of the two states. As can be seen in the graph, from the 10 to 20 foreign skiers present in the slope at any time, around a 20%-25% have an established connection, while 70%-75% are into "connection setup" state. The remaining percentage accounts for those who have no local skier to try to connect to. This data correlates properly both with the total amount of foreign skiers present in the slope at the given simulation time and with the expected results of the simulation, as the connection setup phase takes 50 seconds, thus allowing only skiers that have been for at

 ISBN: 978-1-61208-330-8

least those 50 seconds in the slope to have an established connection.

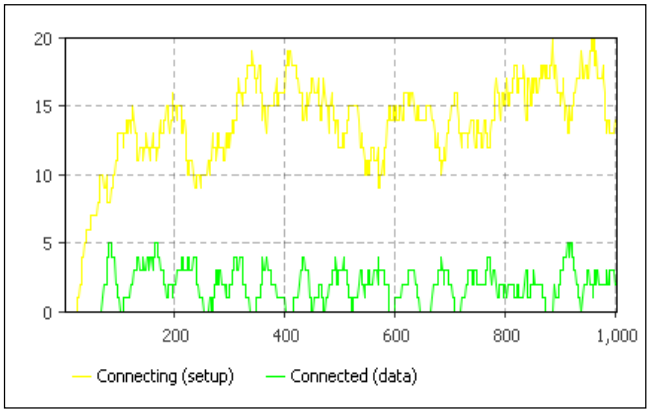


Figure 4. Amount of foreign skiers in connection setup and connected states at any given point of the simulation time.

Finally, the third graph in Figure 5 shows the maximum amount of time (in seconds) a skier has been already connected in a given point of time, and the global average of the time (in seconds) skiers in general are in a connected state, being the X-axis the simulation time and the Y-axis the amount of seconds of achieved connectivity. We can see in the graph that in average, skiers are connected during 10 seconds to a local skier, reaching maximum connection times of over 20 seconds. This fits into the expected results, as the speed of a foreign skier is in the 25 to 40 km/h range as previously said before, which makes an average speed of 32.5 km/h (9 m/s). Taking into account that the length of the slope is 800m, we can establish that it takes nearly 88 seconds to complete the full length of the slope.

A perfect descent would imply that the foreign skier is in "connection setup" state for 50 seconds, having 38 extra seconds to be in a connected status, but taking into account that the trajectories and speeds of the skiers in the slopes are not uniform, that connectivity setup can start later in time than when the skier first reaches the slope and that the connectivity setup phase can be broken if the foreign skier goes out of the area of coverage of the local skier's hotspot, we assume that the results obtained in the simulation adjust to the reality of the situation well, deeming them as valid.

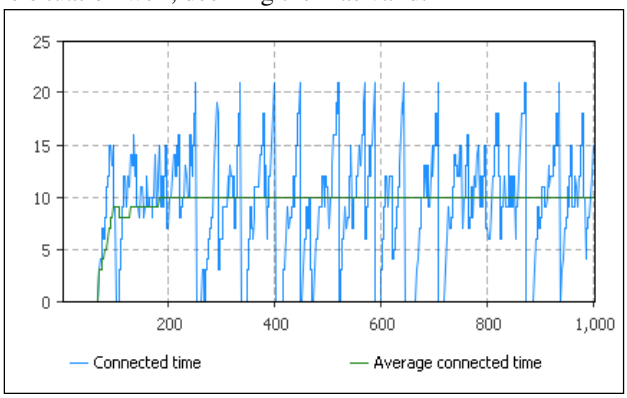


Figure 5. Average connected time for any foreign skier during the simulation time.

All in all, the results obtained in the simulation phase imply that such an application is feasible, as foreign skiers have enough time to at least do some light data exchange, as the average 3G/4G connection averages real data rates from 200Kbps to 1Mbps [7], in order to for example update slope maps and status in real time, which can take approximately the order of 0.98 to 2.86Mb with a decent zoom level [8] and upload meaningful statistics of an order of a couple of hundred Kbs about speed, distance and such to a server when using a Smart Ski Resort application which offers this capabilities, all without having to take care to connect manually or use their smartphone actively.

Also, these results imply that in a static situation, such a local cafeteria, restaurant or shop, connectivity would be obtained without any significant problem, as the situation is much less demanding than that of a ski slope.

IV. User Surveys Feedback

In order to estimate the need and awareness of users regarding Wi-Fi sharing and to determine which mobile platform would be the best to deploy our application, we carried out three sets of surveys with the real users in Megève ski resort thanks to Megève Tourism board. The first one aimed at determining the best mobile platform, both taking into account the amount of users and technological constraints inherent to the platform itself, the second one to better understand the needs, requirements and risk awareness of the users regarding Wi-Fi sharing and the third survey, that was carried out on a marketing database of users who are interested in computer programming in order to cover a broader audience than Megève users, gave more insight regarding that risk awareness of Wi-Fi sharing.

A. *First User Survey*

In this user survey, we wanted to determine which mobile platform in regards of popularity in Megève ski resort would better suit our needs, while providing the least possible technological constraints, such as API extension, access to system functions and the like.

The survey was carried out was open from the 1st of February to the 5th of March 2012 in the ski resort of Mègeve, and involved the participation of 3458 users, from which 58.7% were female and 41.3% were male, filling out an online questionnaire as can be seen in Figure 6. We worked in collaboration with Megève Tourism board to distribute the online form to as many tourists and locals as possible through its different communication channels adapting the form to each channel: a standard Web form on Megève main Web site, a Facebook form app on Megève Facebook page and a mobile Web version of the form for Megève mobile users. Following TEFIS living lab methodology [2], we motivated the users to fill the form by adding a prize draws. They could win a week-end for 2 persons in a 5-star hotel in Megève including a ski pass as well as ski gears.

 ISBN: 978-1-61208-330-8

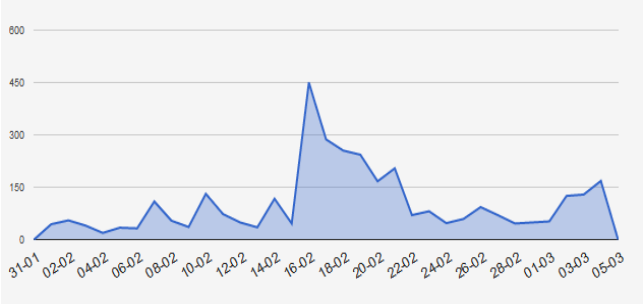


Figure 6. Amount of users filling the survey per date.

From those 3458 users, the age distribution was as depicted in Figure 7 and their preferred mobile OS can be found in Figure 8.

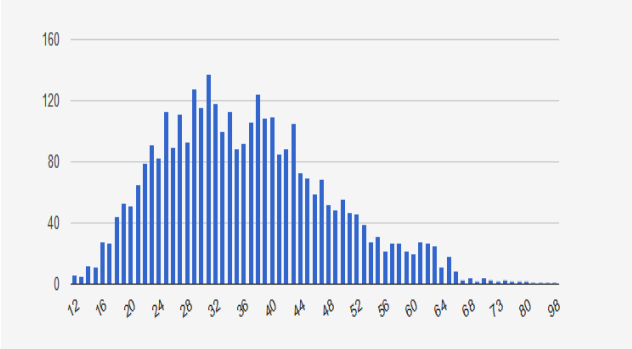


Figure 7. Distribution of users based on their age.

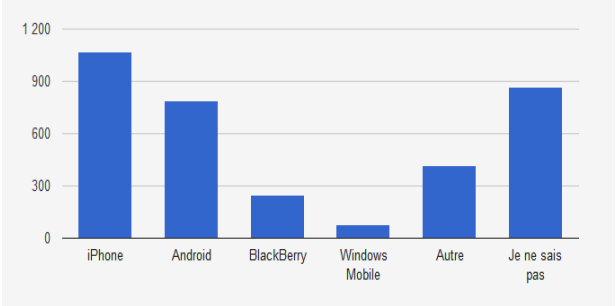


Figure 8. Number of users per mobile operating system.

Where "Autre" means other mobile operating system and "Je ne sais pas" means that the user did not know which mobile operating system she or he was using.

After obtaining these results, and looking carefully to the two most used mobile OSes, namely iPhone (iOS) and Android, we decided that due to the technological constraints which iOS imposes, such as the impossibility to access low-level system functionality and stricter rules in order to place applications in the Apple Store, we would target Android as the mobile operating system of choice.

B. Second User Survey

The second survey, carried out during January 2013 and still online, is aimed at determining the awareness of the users regarding Wi-Fi sharing in general, roaming costs and the risks derived from sharing wireless connections.

Even though the survey contains to the date only the answers of 22 users, it already provides a valuable insight in these matters. The survey is divided into different sets of questions, each set related to one topic. Following, we present the results gotten so far from the survey.

1) General questions

This set of questions is of general purpose, to determine several main characteristics of the users taking the survey. The questions contained in this section are as follow:

1. What country do you come from?
2. How old are you?
3. Which mobile phone operating system do you use?
4. Would you like that we inform you by email if we have a proposal to lower your costs to access the Internet when you are abroad?
5. What is your email address?

We have plotted in a two charts the answers of questions 3 and 4 as they are the only relevant ones for this paper's purpose. Those can be seen in Figures 9 and 10.

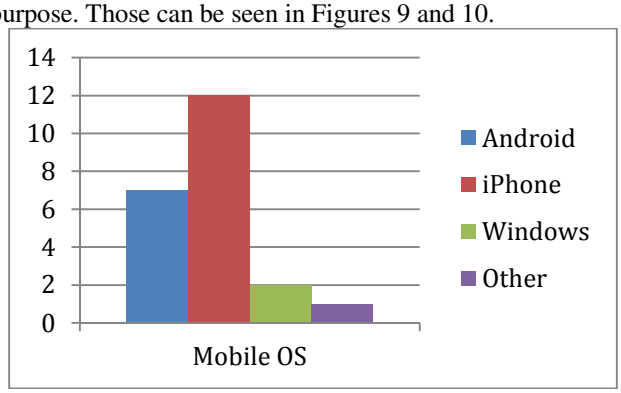


Figure 9. Number of users having filled the form per mobile operating system.

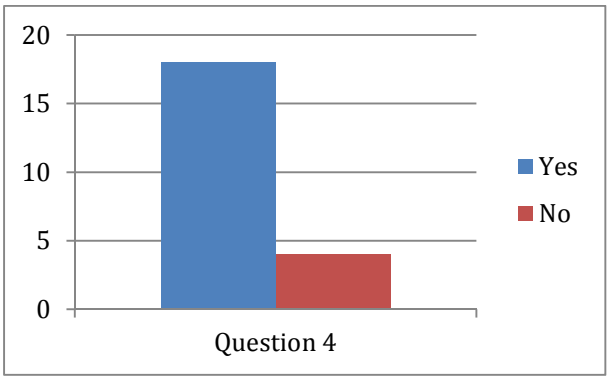


Figure 10. Amount of users interested on updates on our proposal.

2) Roaming cost questions

The goal of these questions is to determine whether the users know the cost associated with roaming and their willingness to still access the Internet even at a cost. The questions contained in this section are as follow:

6. How much do you pay on average for accessing the Web, email and social networks from your phone when abroad during your holidays or a trip?

 ISBN: 978-1-61208-330-8

7. Would you like to access the Web, email and social networks from your mobile phone when abroad?
8. How much is the maximum that you would like to pay per day to access the Web, email and social networks when abroad?
9. Do you have a special option in your mobile monthly subscription that allows you to access the Web, email and social networks when abroad?
10. How much data can you access with your mobile phone subscription with that extra access to the Web, email and social networks when abroad?
11. How much do you pay in your mobile phone subscription for that extra access to the Web, email and social networks when abroad?

For question number 6, 5 out of 22 users estimated that they pay less than 5 euros to use mobile data when abroad, four said that they pay between 6 and 15 euros, two stated between 15 and 25 euros, two from 25 to 50 euros, and one from 50 to 100 euros. Six users answered that they were too afraid of the cost so they were not using data when roaming at all, and two stated that they did not know how much they were paying.

For question 7, 21 out of 22 users said they would like to be able to use mobile data while abroad, and regarding question 8, the answers ranged in between 0 to 15 euros in order to access the Internet while abroad being the average 5 euros per day.

Regarding question number 9, seven users stated that they had an especial option in order to be able to use mobile data when abroad, twelve did not have any special option and three did not know. From those people with the special option, three of them could use from 16 to 50 MB of data per month while abroad, two from 5 to 16 MB, one from 1 to 5 MB and one did not know how much data while roaming the special option allowed for. Finally from those seven users with the special option, four had to pay less than 10 euros for it and three in between 10 to 20 euros.

3) Wi-Fi sharing questions

The goal of these questions is to determine whether the user knows how much they pay for their data access and if they incur into any extra cost when sharing their mobile data access through a portable hotspot in their own country and also to know the willingness of the user regarding sharing their data access. The questions contained in this section are as follow:

12. Do you know how much you pay to be able to access the Internet in your home country with your current telecom operator subscription?
13. Do you know much data per month you can access in your home country with your current telecom operator subscription?
14. According to your telecom operator mobile phone contract, are you allowed to share your mobile phone data access with your other devices?
15. Do you know what happens if you go over your monthly data access quota with your current telecom operator subscription?
16. Have you already shared your Wi-Fi access in order to let someone else access the Internet?
17. Would you mind sharing your mobile phone access to the Web through its Wi-Fi connection with someone else?

For question number 12, the most common answer amongst users was from 31 to 50 euros per month with 8 out of 22 users, followed by five users paying in between 16 to 30 euros, four users paying below 15 euros per month, one paying from 76 to 100 euros and four not knowing how much they were paying per month.

Answer to question 13 ranged from below 20 MB per month up to 3 GB per month, being the average around 1 GB of data per month. For question 14, eight users were allowed to share their mobile data access with other devices, six were not allowed by their operator and eight did not know whether they could share it or not.

Question 15 most common answer was that the user had to pay if going over the monthly data quota with 8 out of 22 users answering that, six users had decreased speed after passing the monthly quota, three had their data closed until the beginning of the next month and five did not know what happened.

For question number 16, fourteen users had already shared their Wi-Fi access and eight had not, and finally for question 17 eighteen out of the twenty-two users taking the survey stated that they would not mind to share their mobile phone data access through Wi-Fi with either family, friends or colleagues, and four would not share it with anybody.

4) Wi-Fi sharing risks

The goal of these last set of questions is to determine whether the user is aware of the risks associated with sharing her or his own data connection through a portable hotspot, whether she or he would be willing to share it if there were no associated risks and up to which percentage of their data allowance they would be willing to share free of risk. The questions contained in this section are as follow:

18. Did you know that you take risks when you share your Wi-Fi access with someone else because she/he could carry out illegal actions such as illegal download of copyrighted music?
19. Due to those risks to share your Wi-Fi connection, would you decide not sharing your Wi-Fi connection?
20. If a new way of sharing your Wi-Fi access without the risk for you to be responsible for the illegal actions that the person might do through the connection, would you share your mobile phone Wi-Fi connection?
21. How much percentage of your monthly mobile data access would you be prepared to share if there is no legal risk for you and it contributes to the tourism quality of service of your region?
22. How much percentage of your mobile monthly data access would you be prepared to share if there is no legal risk for you and you are in return paid more than what your data access costs?

For question number 18, nineteen out of the twenty-two users stated that they know they take risks when sharing their Wi-Fi access, while three answered no. Regarding question 19, twelve users said that due to those risks they would

consider not sharing their Wi-Fi connection while ten would still share it.

In question 20, fourteen users would share their Wi-Fi connection if they would not be legally responsible for any illegal action performed by the person they would share with, while eight still would not share it.

Finally, for question 21 users would be willing to share from a 10% to a 90% of their monthly data access and in average a 35.75% if it would contribute to the tourism quality of their region, and in question 22 users would be willing to share from a 10% to a 100% of their monthly data access and in average a 45% if they would be paid more than what their data access costs.

C. Third User Survey

During summer 2012, we created a short survey and sent it to a list of users who are subscribed to a marketing database and who are interested in computer programming and speak English or French. 1767 users answered, which is quite a large number of answers. We asked them the following question "Do you know that a Wi-Fi hotspot public access point name can be easily impersonated and that it can be a security risk for you?" They could reply one of the following answers "Yes; No; I don't care" and optionally add a textual comment. 5 of them used that comment option and answered: yes with the following comment "but it is possible to secure the link"; yes with the following comment "VERY COMMON AND IT CAN CAUSE HAVOC!!!!!" yes with the following comment "Obvious: P"; yes with the following comment "Honeypot :-)"; no with the following comment "Yes, now I know :P". Among the English speaking people, 540 replied "yes", 185 replied "no" and 1017 replied "I don't care". The image below on Figure 11 indicates the percentages for each answer type.

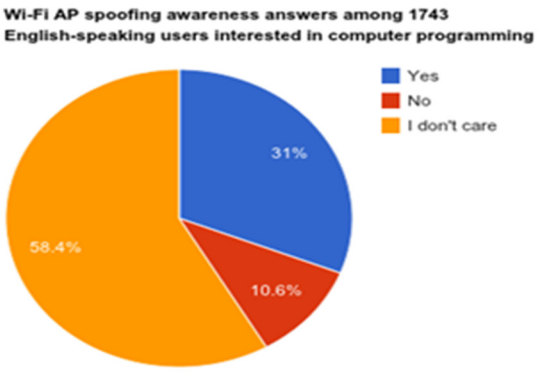


Figure 11. Answers for question 4.

Although these users are interested in computer programming, it is surprising to see that 58.4% of 1743 English-speaking users did not care really care about this issue and that 10.6% did not know it. Concerning the comment on securing the connection, few users would know how to really secure their connection. Furthermore, the fact that many of them answered that they do not care, leaves us to think that they would not take the time to secure it if it is not automated, which is not the case today with current Wi-Fi connections. It is the reason that we decided that our Android application would automatically share the Wi-Fi connection with encryption (such as the one provided by WPA2) enabled by default.

V. Feedback from Application Real Testing

In this section we present the performance results obtained when testing the real Android application both in the ski slopes and in a cafeteria, plus some preliminary tests about Android portable hotspot connectivity range.

As said above at the end of Section IV.C, we have chosen to protect the connection between the foreign skier Android phone and the local skier Android phone sharing its connection through Wi-Fi with WPA2 security. Unfortunately, it is longer than a normal non-encrypted client-to-hotspot connection but we needed to know how much longer it would take and if it was still compatible to share while moving and skiing on the ski slopes.

Once a stable prototype of the application was achieved, we proceeded to test it under real conditions. In order to do this, we performed tests in Megève ski resort located in France, both at the slopes while skiing and in a cafeteria, while not moving at all.

A. Client and Hotspot Measurement Results

In order to determine the required parameters for both the simulation and the real application, we have carried out a few simple tests to determine the range of a portable hotspot such as the one present in Android phones. Table I shows the results from these first tests.

TABLE I. Preliminary client-to-hotspot distances and signal strength.

Distance to AP (m)	RSSI
5	-29
10	-57
15	-57
20	-57
25	-57
30	-57
35	-57
40	-57
45	-65
50	-62
55	-57
75	-70
100	-74

As shown in the table, we have tested how the signal strength of the portable hotspot evolves for a set of distances, ranging from 5 to 100 meters, being lower values a better signal strength. As can be seen, the coverage area of such a device goes well up to 100 meters, even though the signal strength is already too low in order to achieve a meaningful and reliable data transmission. Thus, according to this results, we have considered for both the simulation experiment and the real application tests a range of 40 meters maximum, as

 ISBN: 978-1-61208-330-8

we think it is a realistic approach on what the capabilities of a portable hotspot are.

B. Ski Slope Test Results

In order to perform the tests in the ski slopes, we first divided them into two sets of tests, carried out in two consecutive days. The results of the first day of testing can be seen in Table II, while the results from Table III correspond to the second day of testing.

In the first day of testing, we focused more actively in testing the overall performance and response of the application than in performing intensive data consuming operations (heavy download, HD video streaming, etc.), which was left for the second testing day. As can be seen in Table II we underestimated slightly the connection setup time in the simulation, being sometimes higher than the previously 50 seconds used. Nevertheless, we obtained good results while skiing inside the appropriate distance which the hotspot covers, achieving connections lasting up to 5 minutes and a good amount of data download and upload.

The short lived connections present in the table account for the cases where either the connection setup phase broke due to surpassing the adequate distance in between the skiers or due to the local skier not having mobile data connectivity at the moment, or due to automated sharing protection mechanisms not being established properly during the setup phase, or due to HSDPA to 3G failover or vice versa.

TABLE II. UPLOAD, DOWNLOAD, CONNECTION SETUP TIME AND CONNECTED TIME IN THE SKI SLOPES, 1ST DAY.

Data upload (Bytes)	Data download (Bytes)	Connection Setup	Connected Time
634744	3719567	1 min, 6 sec	5 min, 18 sec
0	10171	1 min, 24 sec	0 min, 6 sec
641559	11364516	0 min, 40 sec	5 min, 59 sec
0	0	0 min, 55 sec	0 min, 0 sec
203325	2287543	1 min, 40 sec	0 min, 59 sec
6338	24506	0 min, 50 sec	0 min, 55 sec
144730	1151956	0 min, 50 sec	2 min, 49 sec
889	13057	0 min, 54 sec	0 min, 37 sec
1131749	39404562	0 min, 50 sec	4 min, 57 sec
0	3099	0 min, 40 sec	0 min, 19 sec
0	0	1 min, 5 sec	0 min, 6 sec
10580	28864	0 min, 42 sec	1 min, 9 sec
0	72	1 min, 14 sec	0 min, 6 sec
80058	1470702	0 min, 49 sec	0 min, 47 sec
210412	6311026	0 min, 44 sec	0 min, 55 sec

We dedicated the second day of testing to perform more controlled and more data intensive experiments, trying to stay at all times inside the appropriate range to the local skier while performing data consuming tasks such as HD streaming and the like. We aimed to maintain long lived connections (as long lived as can be while in a ski slope and a ski lift) to see how the application would perform, even though we still got some short or non-existent connections due to the facts previously mentioned previously in the first testing day. The results can be seen in Table 3.

TABLE III. UPLOAD, DOWNLOAD, CONNECTION SETUP TIME AND CONNECTED TIME IN THE SKI SLOPES, 2ND DAY.

Data upload (Bytes)	Data download (Bytes)	Connection Setup	Connected Time
0	11215	0 min, 41	0 min, 6 sec
713780	24391110	0 min, 47 sec	7 min, 56 sec
3161848	102125792	1 min, 18 sec	13 min, 18 sec
1234	9215	0 min, 45	0 min, 8 sec
644697	22112428	0 min, 44 sec	6 min, 39 sec
0	0	0 min, 56 sec	0 min, 0 sec
148083	1547139	1 min, 32 sec	29 min, 18 sec

As shown in the table, we successfully achieved to maintain quite long connections (up to almost 30 minutes), while skiing and in the ski lifts. Also, it was possible to upload and download a good amount of data without further problem while being connected, streaming HD video and browsing internet.

C. Cafeteria Test Results

To finalize the set of tests, we carried out a session in a less dynamic place than the ski slopes. We tested the application in less demanding and more static conditions by performing some tests inside a cafeteria of the ski resort while not moving any of the terminals at all. The results of these last set of tests can be seen in Table IV.

TABLE IV. UPLOAD, DOWNLOAD, CONNECTION SETUP TIME AND CONNECTED TIME IN A CAFETERIA.

Data upload (Bytes)	Data download (Bytes)	Connection Setup	Connected Time
551012	2611660	0 min, 46 sec	11 min, 40 sec
590794	2528282	0 min, 46 sec	8 min, 27 sec
0	5460	0 min, 53 sec	0 min, 2 sec
468920	1916792	0 min, 47 sec	4 min, 25 sec
1227882	4608589	0 min, 52 sec	6 min, 16 sec
0	3465	0 min, 59 sec	0 min, 0 sec
1321702	3829081	1 min, 30 sec	21 min, 27 sec
1118089	3206335	0 min, 57 sec	55 min, 30 sec
265781	1844596	0 min, 48 sec	19 min, 7 sec

As displayed in the table, we were able to achieve long lived connections, up to 55 minutes without it breaking, and

 ISBN: 978-1-61208-330-8

successfully doing light browsing and casual social network and app use. Again, some of the shorter or broken connections account for the cases where either the local phone lost connectivity or switched from one type of network to another, deeming impossible to perform the appropriate steps in order to protect the user sharing her or his mobile connection, or to establish a connection at all.

D. Battery and CPU Usage Results

In order to assess the performance of the application regarding battery consumption and CPU usage, we measured those values using the built-in functionality to monitor per application battery and CPU usage that can be found in any Android phone. High battery consumption values could deter users from adopting our application in the future, and thus were of a high concern for us.

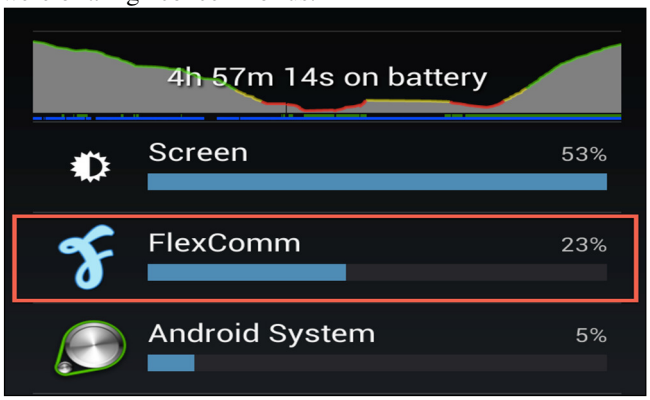


Figure 12. Percentage of battery consumed by the Android application service.

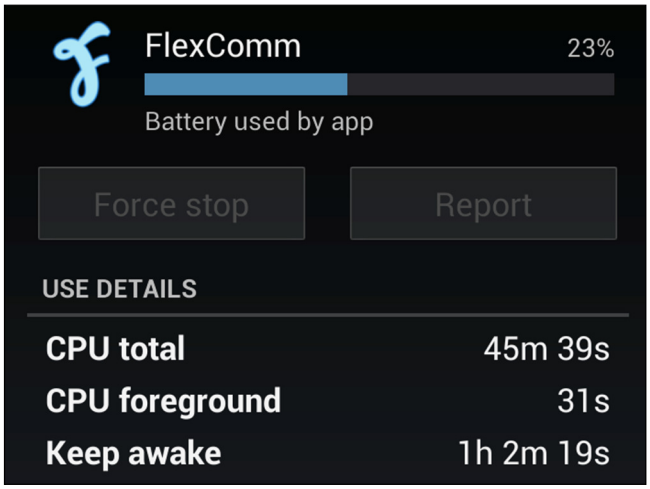


Figure 13. CPU statistics for the Android application.

As can be seen in Figure 12 and Figure 13, the battery usage during the testing sessions falls inside an acceptable range, accounting for the 23% of the total battery use, while the CPU usage both while performing active operations and just in a keep awake status are also inside acceptable values. This said, the battery usage level could and should be improved not to impact the overall user experience and this is one of the points we will work in following versions of the application.

VI. CONCLUSION

In this paper we have presented a complete simulation and real testing results for an Android application to achieve collaborative wireless access to mitigate roaming costs while protecting the sharer.

Both the simulation and the real testing results plus the data acquired from the surveys are encouraging and prove the feasibility and need of such an application. In terms of connection ranges, hotspots perform well up to 50m range, and regarding connection establishment times and data upload, our results show that the times and amount of data are enough to make the system reasonably useful. Finally, regarding battery and CPU consumption results are inside acceptable ranges for the app to be usable.

Future work will involve improving the application in order to solve some of the drawbacks found from the testing sessions, such as better detection of actual connectivity failure, ensuring that the local phone switching from and to different networks does not introduce false positives, better battery use performance and a desktop client.

REFERENCES

[1] I. Chard, “Mobile Web 2.0 - Leveraging 'Location, IM, Social Web & Search' 2008-2013”, Juniper Research, Mobile Content and Applications, May 2008.

[2] The TEFIS Project, http://www.tefisproject.eu/, 2013.

[3] Air Mobs, http://eeiiaa.com/blog/?p=785, 2013.

[4] Open Garden, http://opengarden.com/, 2013.

[5] ULOOP Project, http://www.uloop.eu, 2013.

[6] AnyLogic Simulation Framework, http://www.anylogic.com/overview/, 2013.

[7] http://www.pcpro.co.uk/features/379402/mobile-data-how-much-do-you-need, 2013.

[8] http://www.microimages.com/documentation/TechGuides/76googleMapsStruc.pdf, 2013.

[9] Carlos Ballester, Jean-Marc Seigneur, "Crowd Augmented Wireless Access", Augmented Human 2012, 08-09 March 2012, Megeve, France.

[10] Sánchez, Ricardo, Joseph Evans, and Gary Minden. "Networking on the battlefield: Challenges in highly dynamic multi-hop wireless networks." Military Communications Conference Proceedings, 1999. MILCOM 1999. IEEE. Vol. 2. IEEE, 1999.

[11] Mendes, Paulo. "Cooperative Networking in ULOOP." EU FP7 IST ULOOP project (grant number 257418) white paper (2012).

 ISBN: 978-1-61208-330-8

Traffic Offloading Improvements in Mobile Networks

Tao Zheng, Daqing Gu
Orange Labs International Center
Beijing, China
e-mail: {tao.zheng; daqing.gu}@orange.com

***Abstract* – The exponential increase in mobile IP data usage causes a shortage in the mobile bandwidth. Traffic offloading is regarded as a solution to the exploding growth of mobile broadband data traffic in the mobile networks. In this paper, a content aware traffic offload scheme is proposed to implement using multiple access paths simultaneously. Moreover, the process flow based-on the scheme in Long Term Evolution (LTE) and other traffic offloading improvements are also presented. This scheme utilizes Content Centric Networking (CCN) concept and Digital Fountain Codes to handle the multi-path control and reduce the complexity of traffic offload implementation.**

Keywords- Traffic offloading, CCN, Fountain Codes, LTE

I. INTRODUCTION

Mobile broadband devices such as smart phones, tablets, wireless dongles and some data-intensive apps have resulted in an exponential increase in mobile IP data usage, which is expected to cause a shortage in the mobile bandwidth. According to the Cisco Visual Networking Index Global Mobile Data Forecast[1], the global mobile data usage tripled in 2010 as compared to 2009. Further, the Cisco forecast predicts that by 2015, there is going to be a 26 fold increase as compared to 2010 levels.

To anticipate this problem, operators have deployed or are deploying "mobile data offloading" solutions to alleviate network congestion quickly. Traffic offloading refers to the ability to move mobile data traffic from cellular to alternative network such as WiFi. So, the data traffic management will be an important issue in traffic offloading.

The 3rd Generation Partnership Project (3GPP) has defined and is defining some traffic offloading mechanisms in various releases. For example, Local IP Access[2] in Release 9, Selected IP Traffic Offload[2], IP Flow Mobility[3], Access Network Discovery and Selection enhancements[4][5][6] and Multiple Access PDN Connectivity[6][7] in Release 10, Broadband Access Interworking using WLAN/H(e)NB and Broadband Access Interworking using H(e)NB[8] in Release 11, WLAN Network Selection for 3GPP Terminals[9], LIPA Mobility and Selected IP Traffic Offload at the Local Network[10], S2a mobility based on General packet radio service (GPRS) Tunneling Protocol (GTP) and WLAN Access[11] and IP Flow Mobility support for S2a and S2b Interfaces[12] in Release 12. Traffic management in these mechanisms focuses on networks and terminals.

The objective of the paper is to study and compare the IP traffic offloading and management solutions in 3GPP. A proposal based on content aware traffic offloading is finally presented.

This paper is organized as follows. In Section 2, we study and compare the IP traffic offloading and management solutions in 3GPP. In Section 3, a content aware traffic offload scheme and its verification are presented. In Section 4, potential extensions in the proposed scheme and other improved points in multi-attachment network are studied. Finally, Section 5 summarizes the conclusions.

II. TRAFFIC OFFLOADING IN 3GPP

Traffic offloading is regarded as a solution to the exploding growth of mobile broadband data traffic in the deployed 3GPP mobile networks. The reason why traffic offloading by WiFi is considered to be a viable solution for mobile data traffic explosion is that there is a lot of available WiFi spectrum with a very large number of compatible devices. And it simplifies the complexity as well as cost of managing and deploying a Cellular network. In 3GPP, from Release 9 to 12, some traffic offloading mechanisms are defined.

A. Release 9

- Local IP Access (LIPA)

LIPA was introduced in 3GPP Release 9, with the discussion on architecture options and impacts to procedures being continued in 3GPP Release-10 and 3GPP Release-11 actively.

LIPA is a mechanism by which a User Equipment (UE) connected to a Home NodeB or Home eNodeB (H(e)NB), is able to transfer data to a local data network connected to the same H(e)NB system directly, without the data traversing the cellular network, and accordingly reduce the load on the mobile core network. LIPA also allows the User Equipment (UE) to access any external network that is connected to the local network.

Considering an IP based corporate wireless network with multiple devices such as laptops, tablets, printers, servers, video conferencing units and IP based telephones which all need to connect to each other and also connect to the internet. The network is implemented using a femto cell gateway and a private gateway to which all these devices connect. If a user needs to print from a laptop, LIPA helps in routing the print request internally, without routing it through the femto cell gateway. Also, email could be sent directly through the private gateway.

 ISBN: 978-1-61208-330-8

LIPA is a simple architecture well suited to local networks. LIPA is applicable only to H(e)NB access, not for macro cell access. It needs the additional Local Gateway function in H(e)NB.

B. Release 10

- Selected IP Traffic Offload (SIPTO)

SIPTO is a mechanism where portions of the IP traffic on a H(e)NB access or cellular network are offloaded to a local network, in order to reduce the load on the core network. SIPTO is applicable to H(e)NB access or another gateway in the cellular network that is closer to the UE. SIPTO can be triggered by events like UE mobility, special occasions that lead to concentration of traffic or other network rules.

Compared to LIPA, SIPTO is applicable in both femto and macro networks use cases. However, SIPTO doesn't help radio congestion.

- IP Flow Mobility (IFOM)

IFOM is a mechanism where the terminal has data sessions with the same Packet Data Network (PDN) connection simultaneously over a 3GPP and a WLAN access network. Under this situation, the UE could add or delete data sessions over either of the access methods, effectively offloading data. Unlike LIPA and SIPTO, where the data offload is largely transparent to the UE, the logic of data offloading in IFOM is more UE centric and largely transparent to the Radio Access Network (RAN).

IFOM helps in both radio and core network congestion. However, compared to LIPA and SIPTO, IFOM needs support of Dual Stack Mobile IPv6 (DSMIPv6) and WiFi or other non 3GPP access network and is more complicated to be implemented.

- Access Network Discovery and Selection (ANDSF) enhancements in Release10

ANDSF is a module within an Evolved Packet Core (EPC) of the System Architecture Evolution for 3GPP mobile networks. ANDSF enables consumer-side devices such as notebooks, modems and mobile phones to discover and communicate with non-3GPP networks such as WiFi or WiMAX and enforce network policy controls. Standards Related to ANDSF in 3GPP are TS 24.312[4], TS 22.278[5] and TS 23.402[6].

- Multiple Access PDN Connectivity (MAPCON)

MAPCON provides the capability for terminals to establish multiple connections to different PDNs via different access methods and a selective transfer of PDN connections between accesses. MAPCON feature is characterized by multiple packet core IP addresses at the UE, any of which may be moved (but unchanged) between 3GPP cellular and WiFi access without impacting the 3GPP access connectivity of the other IP addresses. This allows IP traffic to multiple PDNs through the use of separate PDN-GateWays (PDN-GWs) or a single PDN-GW. The usage of multiple PDNs is typically controlled by network policies and defined in the user subscription of TS 23.401[7].

Multiple PDN connections would need to be supported when the UE is using LTE for part of data connection and WiFi for other part. In fact these two (or multiple) connections should be under the control of the same EPC core that can help support seamless mobility once the terminal moves out of the WiFi hotspot.

C. Release 11

The main problems of the interworking between a 3GPP system and a fixed broadband access arose from the different methods of policy control. 3GPP TS 23.139[8] specifies the interworking between a 3GPP system and a fixed broadband access network defined by Broadband Forum (BBF) to provide the IP connectivity to a 3GPP UE using a WLAN and a H(e)NB connected to a fixed broadband access network. It covers the system description including mobility, policy, Quality of Service/Quality of Experience (QoS/QoE) aspects between a 3GPP system and a fixed broadband access network as well as the respective interactions with the Policy and Charging Control (PCC) frameworks.

3GPP identified the initial use cases for Fixed-Mobile Convergence (FMC) in Release 9 and finalized the work on phases 1 and 2 within Release 11, and phase 3 will be addressed in Release 12 and later releases.

- Broadband Access Interworking using WLAN/H(e)NB

When WLAN is being used for interworking with a fixed broadband access network, 3GPP Evolved Packet System (EPS) considers both EPC routed traffic and non-seamless WLAN offloaded traffic; both traffic types can coexist during network operation. That is, UE can simultaneously have a connection to both the EPC and the non-seamless WLAN offloaded traffic.

For the purpose of interworking with a fixed broadband access network, a 3GPP system has to recognize the local IP address of the UE connected to the fixed broadband access network. The S9a interface session can be established and perform the policy interworking when the local policy of the BBF network indicates that the policy control for non-seamless WLAN offload is allowed for the UE, as well as the 3GPP home operator's policy.

- Broadband Access Interworking using H(e)NB

In contrast to the interworking scenario using WLAN access, the interworking architecture using H(e)NB supports only EPC-routed traffic. For the purpose of interworking with a fixed broadband access network, this architecture basically uses S9a similar to the architecture using WLAN described above. Then the S9a interface also carries the H(e)NB's local IP address and User Datagram Protocol (UDP) port number(s), and/or the Fully Qualified Domain Name (FQDN) of the fixed broadband access network to the Broadband Policy Control Framework (BPCF) from the Policy and Charging Rules Function (PCRF).

D. Release 12

- WLAN Network Selection for 3GPP Terminals

3GPP TR 23.865[9] studies WLAN network selection for 3GPP terminals in Release 12. The solutions are based on

 ISBN: 978-1-61208-330-8

architectures as specified in TS 23.402[6] and will take into account Hotspot 2.0 specifications developed by the Wi-Fi Alliance (WFA) [13]. 3GPP operator's policies for WLAN network selection will be provisioned on 3GPP terminals via pre-configuration or using the ANDSF server for their delivery.

- LIPA Mobility and SIPTO at the Local Network

3GPP TR 23.859[10] studies on the support of mobility for LIPA between the H(e)NBs located in the local IP network and functionality to support SIPTO requirements at the local network, including mobility. The report is intended to document the analysis of the architectural aspects to achieve these objectives in order to include the solutions in the relevant technical specifications.

- Study on S2a mobility based on GTP and WLAN Access

In 3GPP TR 23.852[11], S2a mobility based on GTP and WLAN Access is studied:

The addition of an S2a based on GTP option. In particular this Study Item will develop the necessary stage 2 message flows to support S2a based on GTP and mobility between GTP-S5/S8 and GTP-S2a; Supporting WLAN access to EPC through S2a via mechanisms; The study item gives some solutions separately for GTP based S2a and WLAN access to EPC through S2a.

- Study of IP Flow Mobility support for S2a and S2b Interfaces

3GPP TR 23.861[12] studies the scenarios, requirements and solutions for UEs with multiple interfaces which will simultaneously connect to 3GPP access and one, and only one, non-3GPP access. Solutions to be studied include the possibility of dynamically routing to specific accesses individual flows generated by the same or different applications belonging to the same PDN connection. The study of solutions to support routing of different PDN connections through different access systems is also in the scope of this item. This study item also investigates the mechanisms for provisioning the UE with operator's policies for multiple access PDN connectivity and flow mobility.

III. CONTENT AWARE TRAFFIC OFFLOAD SCHEME

This section proposes a content aware traffic offloading scheme in 3GPP. In the scheme, fountain codes [14] are used to transport the information from servers to terminals through different access paths. Fountain codes technique having the slight overhead and impressive codec, is exploited to generate the segments for delivery. The data streams can be transported smoothly in all different access paths simultaneously or part of them when breakdown happens and terminals do not need to handle the breakdown and recovery.

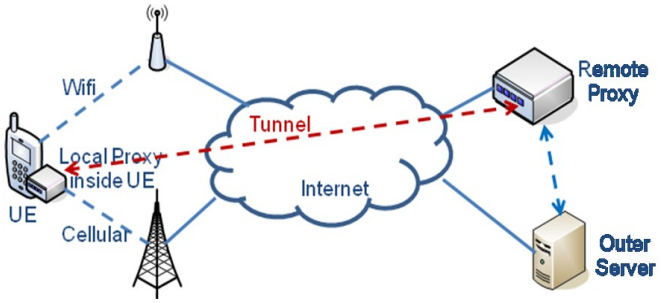


Figure 1. System Overview of the scheme

A. The scheme overview

The system overview is illustrated in Figure 1. It is composed of terminal and remote general Hypertext Transfer Protocol (HTTP) proxies. A local proxy built in the terminal is responsible for the conversion between the HTTP request/response and the Interest/Data packets. The remote proxy exchanges Interest/Data packets with the local HTTP proxy through the Transmission Control Protocol (TCP) or UDP based tunnel established over multiple heterogeneous links in the content centric way. The remote proxy fetches the content from the outer server in the Internet to satisfy the terminal's request. The main reason for using two proxies is that the tunnel between the two proxies can help Content Centric data stream penetrate the network.

The working environment can be IPv4 or IPv6. In IPv6 network, the remote proxy can be assigned with different IPv6 addresses including unicast address, multicast address and anycast address. Given the unicast address, the terminal establishes the tunnel with the single specific remote proxy. The multicast addressing refers to the configuration where the terminal can set up the tunnels concurrently connecting the collection of remote proxies. The anycast addressing makes the terminal build the tunnel to the nearest remote proxy among several candidates.

Since the service session in the scheme is identified with the Uniform Resource Identifier (URI) that is independent of the IP addresses associated with the different connections to the network, the terminal can keep the ongoing session alive as long as the content identifier is invariant. The dynamics due to the connection switching in the mobile scenario is only visible in the tunnel running over the TCP or UDP sessions managed by the local HTTP proxy and shielded from the perspective of the HTTP session in the normal web-based client of the terminal.

CCN is an alternative approach to the architecture of networks which was proposed by Xerox PARC within the CCNx [15] project. From the network perspective, in CCN, network entities in ordinary network are replayed by data entities.

Unlike state-of-the-art multi-path approaches such as Multi-Path Transmission Control Protocol (MP-TCP), CCN can help us to hide some network control implementing details with the help of the 'connection-less' nature of CCN. In the proposed solution, we utilize the CCN concept rather than CCN protocol,

 ISBN: 978-1-61208-330-8

thus it's not necessary to consider the change of network access paths through fountain codes to encapsulate data packets.

The service session coupling with the remote proxy avoids the problem of Domain Name System (DNS) resolution potentially confronted with the multi-homed terminal. Since the remote proxy acts as the agent of the terminal for content acquisition, the terminal has no need for DNS resolution except to forward the request message to the remote proxy. The update on the access router is additionally not mandatory any more because the conversion between IP address and URI is executed in the sense of application layer.

B. The scheme in LTE

When the scheme is applied in LTE network, the remote proxy can be deployed within PDN-GW and the local proxy is still in terminals. The architecture is showed in Figure 2. Terminals can retrieve information from outer server through different access system simultaneously and can switch smoothly among them without any breakdown. Multiple data streams can be transported though multiple access systems, such as LTE RAN, trusted non-3GPP access and non-trusted non-3GPP access.

In this scheme, fountain codes are used to transport data packets between two proxies. The reasons of choosing fountain codes as the encapsulation technology are showed as follows. Fountain codes are rateless in the sense that the number of encoded packets that can be generated from the source message is potentially limitless. Regardless of the statistics of the erasure events on the channel, the source data can be decoded from any ser of K' encoded packets, for K' slightly larger than K. Fountain codes can also have very small encoding and decoding complexities and automatically adapt the change of multiple access paths to avoid the implementation of path control details [14].

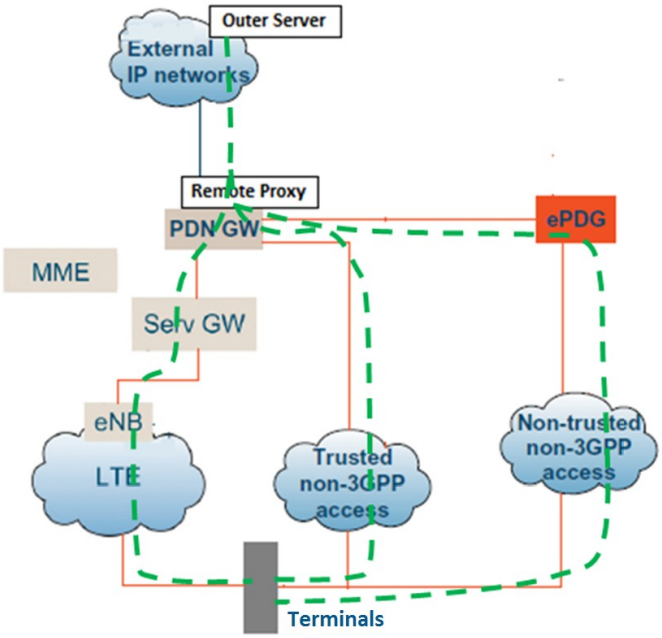


Figure 2. The content aware traffic offload in EPC

The source hosts simply transfer the packets with the different coding schemes as many as possible to the destination hosts without concerns over the reordering induced in various paths. It increases the data throughput and avoids the complicated reconciliation between the multiple paths.

Figure 3 gives the process of the content aware traffic offload.

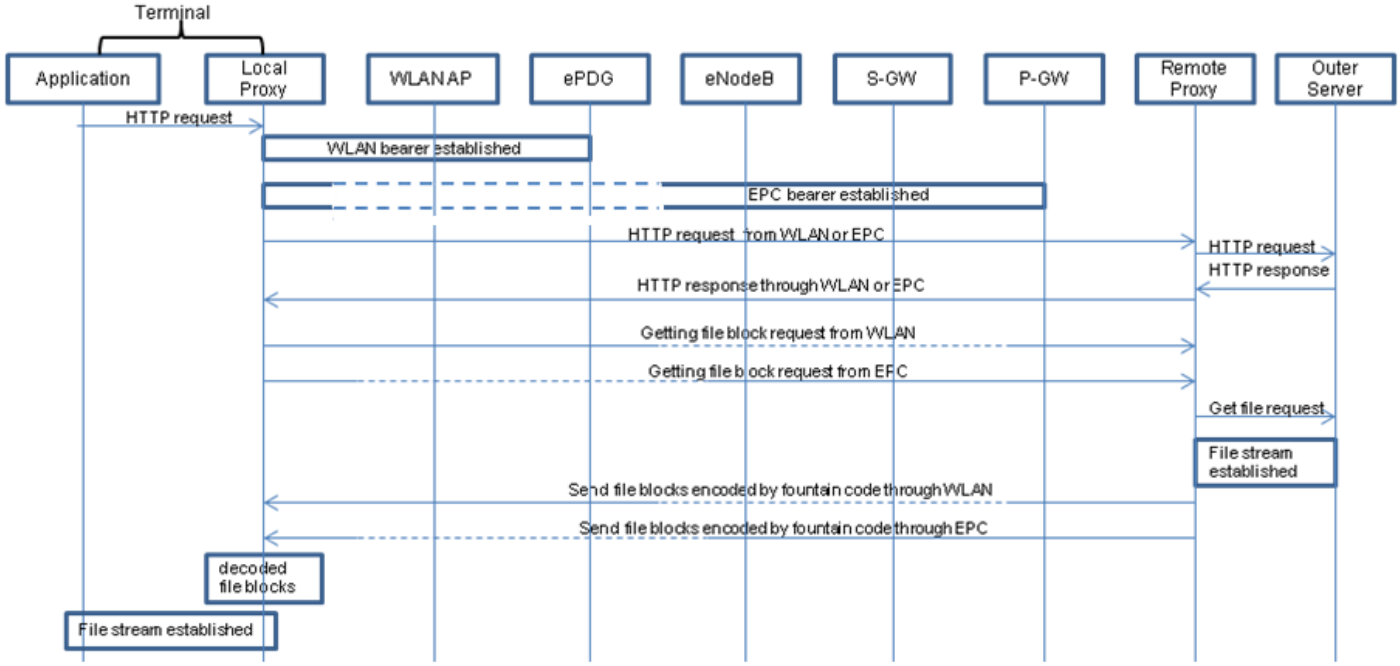


Figure 3. The process flow generating IPv6 flow label

 ISBN: 978-1-61208-330-8

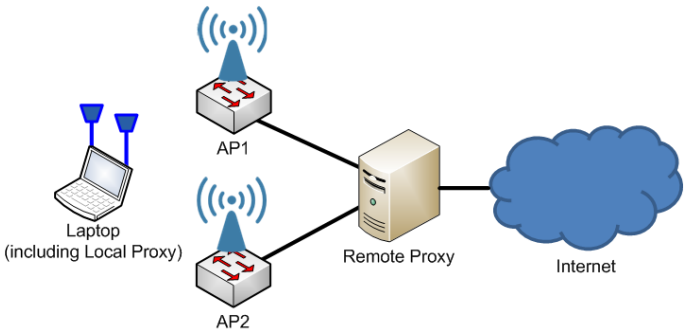


Figure 4. The test-bed topology in lab

In the terminal, the local proxy handles the application's request towards outer server through HTTP. The bearer of WLAN or EPC is established in advance or when the local proxy receives the request. After the local proxy received the response from outer server, it sends the requests getting file blocks to the remote proxy, which acquires the file from outer server, encodes the file blocks by fountain codes and sends them to the local proxy though different access paths. The local proxy decodes the received blocks and sends them to application. During the transportation, the change of paths between the remote proxy and the local proxy will not impact the blocks' decoding due to the fault-tolerance property of fountain codes.

C. *The verification of this scheme*

According to the traffic offloading scheme mentioned above, a verification test was implemented in our lab. Because there is no mobile data access in the lab, we chose two WiFi interfaces and Access Points (AP) to simulate two access paths. Figure 4 shows the test topology, which is composed of a laptop (including a local HTTP proxy) with two WiFi interfaces, two WiFi APs and a remote proxy connecting to Internet.

The local proxy and remote proxy, fountain encoding and decoding were implemented through programming. On this test-bed, we tested various combinations of two access paths and handover between them. The data delivery between the laptop and Internet cannot be interrupted in these scenarios. The test validated the feasibility and validity of this scheme.

IV. POTENTIAL EXTENSIONS AND IMPROVED POINTS FOR MULTI-ATTACHMENT NETWORKS

The proposed scheme in Section 3 can be easily extended in other use cases and there are other improved points for multi-attachment networks.

A. *Potential Extensions*

The potential extensions of this scheme include mobile content distribution and mobile data offloading in addition to the mobility management.

- Mobile content distribution

It is simple and practicable to implement the content distribution in the IPv6 mobile network with our solution. The remote proxy introduced in our method may take the role of content cache in the Content Distribution Network (CDN). The content centric networking relying on the URI identifier and many-to-many transport enables the tight coupling between request routing and load balancing in the simplified way where the remote proxy can realize the distributed load sharing by simply tuning its transmission rate and forwarding the content request to other proxies.

- Mobile data offloading

Mobile data offloading, also called traffic offloading can be smoothly supported by our solution without interrupting the ongoing session. The session maintenance with URI is able to shield the negative side-effect yielded by the IP address variation due to the link switching triggered by the traffic offloading. Given the single service session, the many-to-many transport implemented with the Digital Fountain coding makes the data delivery through the complementary network independent of the original one in the cellular networks without fearing the packet reordering that may deteriorate the QoS/QoE performance.

B. *Improved points for multi-attachment network*

- IP aware traffic management

In EPC network, the IP packets are transported in GTP Tunnel. Figure 5 shows the IP packet structure through GTP Tunnel. EPC entities just handle GTP Tunnel instead of IP packet.

The UE and the PDN-GW (for GTP-based S5/S8) or Serving-GW (for Proxy Mobile IP (PMIP)-based S5/S8) use packet filters to map IP traffic onto the different bearers. Some EPC's mechanisms, such as QoS, PCC, are based-on bearers, the packets contained in the same bearer share same QoS profile and be treated in the same way. It's hard to handle IP packets in such mechanisms.

When UE attaches multiple networks, except for EPC, other access methods handle IP packets directly. Therefore in EPC, the transport layer is not able to know and use the information, e.g., QoS related information, hidden in (1) the tunnel header and (2) original IP packet. In the context of this paper, by applying IP aware traffic management, the transport layer will be able to read and use this hidden type of information. So, IP aware traffic management in multi-attachment network will be beneficial to reduce the complex and promote the efficiency of the traffic offloading mechanism.

- IPv6 application at multi-attachment radio network

As a network evolution goal, IPv6 is deployed in mobile network, including access network, core network and mobile carrier IP network. IPv6 introduction in mobile network will impact on QoS of mobile services.

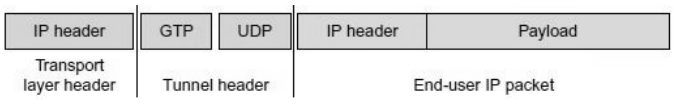


Figure 5. IP packet structure in GTP Tunnel[16]

 ISBN: 978-1-61208-330-8

In EPC, there are two IP headers, which correspond to the transport layer and the end-user IP packet respectively. That is, two IPv6 flow label fields can be handled by mobile network entities and IP carrier network entities respectively.

A sequence of packets sent from a particular source to a particular unicast, anycast, or multicast destination constitute a flow. In IPv4 network, the 5-tuple of the source and destination addresses, ports, and the transport protocol type is able to identify a flow. IPv6 has introduced a field named flow label. The 20-bit flow label in the IPv6 header is used by a node to label packets of a flow. General rules for the flow label field have been documented in RFC 3697 [17].

In multi-attachment radio network, the IPv6 Flow Label can be employed to identify the different access methods and provide traffic management based-on flow for further processing in EPC.

V. CONCLUSION AND FUTURE WORK

This paper presents the study on the IP traffic offloading and management in the multi-attachment network in 3GPP, where some mechanisms are investigated and compared. Then a content aware traffic offloading scheme based on CCN and fountain codes is proposed to handle different access systems simultaneously and automatically adapt the change of multiple paths to avoid the implementation of path control details. Tunnel and proxy can help deploy the scheme when CCN is not supported in current networks. Some potential extensions of this scheme and improved points for multi-attachment network are proposed too.

We tested this scheme in our lab. Traffic was able to be transported smoothly among multiple WiFi access paths. The test results verified the feasibility and showed the features of the scheme. In the future work, we consider testing the extension use cases of this scheme and applying other improved points listed in Section 4. In addition, we will focus on the performance of this scheme.

REFERENCES

[1] http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html, retrieved: February 2014.

[2] 3GPP TS 23.829 Local IP Access and Selected IP Traffic Offload (LIPA-SIPTO) (Release 10), October 2011.

[3] 3GPP TS 23.261 IP flow mobility and seamless Wireless Local Area Network (WLAN) offload (Release 10), March 2012.

[4] 3GPP TS 24.312 Access Network Discovery and Selection Function (ANDSF) Management Object (MO) (Release 10), June 2012.

[5] 3GPP TS 22.278 Service requirements for the Evolved Packet System (EPS) (Release 10), October 2010.

[6] 3GPP TS 23.402 Architecture enhancements for non-3GPP accesses (Release 10), September 2012.

[7] 3GPP TS 23.401 General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access (Release 10), March 2013.

[8] 3GPP TS 23.139 3GPP system - fixed broadband access network interworking; Stage 2 (Release 11), March 2013.

[9] 3GPP TR 23.865 WLAN network selection for 3GPP terminals (Release 12), December 2013.

[10] 3GPP TR 23.859 Local IP access (LIPA) mobility and Selected IP Traffic Offload (SIPTO) at the local network (Release 12), April 2013.

[11] 3GPP TR 23.852 Study on S2a Mobility based on GTP & WLAN access to EPC (Release 12), September 2013.

[12] 3GPP TR 23.861 Network based IP flow mobility (Release 12), November 2012.

[13] https://www.wi-fi.org/hotspot-20-technical-specification-v100, retrieved: February 2014.

[14] J. Byers, M. Luby, M. Mitzenmacher, and A. Rege, A digital fountain approach to reliable distribution of bulk data, Proc. ACM SIGCOMM '98, September 1998, pp. 56-67.

[15] http://www.ccnx.org, retrieved: February 2014.

[16] 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.

[17] J. Rajahalme, A. Conta, B. Carpenter, and S. Deering, IPv6 flow label specification, IETF RFC 3697, March 2004.

 ISBN: 978-1-61208-330-8

Security Issues in Cooperative MAC Protocols

Ki Hong Kim
The Attached Institute of ETRI
Daejeon, Korea
e-mail: hong0612@ensec.re.kr

***Abstract*—A lot of cooperative media access control (MAC) protocols have been proposed to support cooperative communications in wireless networks in the last few years. In this paper, the security vulnerabilities in some cooperative MAC protocols (e.g., COSMIC, VC-MAC, BTAC, and cooperative MAC for IEEE 802.11g) are analyzed. Channel-assisted authentication approach is also discussed to verify entities in cooperative MAC protocols. These analytical results should be significantly useful in the design of efficient authentication solutions for secure, cooperative MAC protocols.**

Keywords–Cooperative MAC protocols; vulnerability; authentication; physical layer security.

I. INTRODUCTION

A cooperative wireless network (CWN) is an emerging communication mechanism that takes advantages of spatial diversity among neighboring relay nodes, to achieve gains in performance and improved reliability. CWNs have attracted much attention within the last decade. In CWNs, when the source sends data to the destination, some nodes serve as relays by forwarding replicas of the source data to the destination. The destination receives multiple sets of data from the source and the relays, and then combines them. There are three major methods for forwarding from relay. First, for the amplify-and-forward (AF) method, after receiving a noisy version of the original data, the relay amplifies and retransmits noisy data to the destination. Second, for the decode-and-forward (DF) method, the relay decodes data from the source and then retransmits the decoded data to the destination. Finally, in the compress-and-forward (CF) method, the relay forwards incremental redundancy of the original data to the destination. The destination receives multiple data sets from the source and multiple relays; then it combines them to achieve gains in performance and quality [1][2].

Due to the rapid growth and evolution of CWNs, much research has been done to propose a cooperative MAC protocol that supports cooperative communication in wireless networks such as wireless sensor networks (WSNs) and vehicular networks. In other work, a new carrier-sense, multiple-access, collision-avoidance (CSMA/CA)-based MAC protocol, called the cooperative MAC protocol for WSN with minimal control message (COSMIC), was proposed to support cooperative relaying with minimum overhead [3]. The vehicular cooperative MAC (VC-MAC) was designed for gateway downloading in vehicular networks [4]. It leverages the advantages of both cooperative communication and spatial re-usability, maximizing system throughput. A busy-tone-based cooperative MAC protocol (BTAC) for wireless local area networks (WLANs) has also been proposed [5]. An efficient cooperative MAC protocol based on IEEE 802.11g was proposed [6], which can be extended to 802.11n. Easy comparison has been made possible by an analytical model of the power consumption of the various MAC protocols [7].

CWNs are vulnerable to security attacks due to the open broadcast nature of the wireless channel and the use of cooperative transmission involving multiple transmitters. There have been a number of studies regarding security issues, including attacks and vulnerabilities, in the cooperative MAC protocols. One study introduced the case study of security attacks based on control-packet vulnerabilities in Synergy MAC [8], while another addressed the potential security issues and vulnerabilities that arise in CoopMAC [9]. The security vulnerabilities found in traffic adaptive-cooperative, wireless sensor-MAC (CWS-MAC) have been identified and analyzed [10]. Coordinated denial-of-service (DoS) attacks against data packets on IEEE 802.22 have been studied from the perspective of malicious nodes [11]. A detection scheme to mitigate malicious relay behavior in a cooperative environment has been proposed [12][13][14]. Similarly, the selfish-behavior attack/detection model and the attack strategies of smart selfish nodes have been analyzed [15]. A secure, cooperative-data-downloading framework for paid services in vehicular ad hoc networks (VANETs) has also been proposed [16].

In spite of all the work mentioned above, security vulnerabilities in many cooperative MAC protocols have not yet been analyzed (i.e., COSMIC, VC-MAC, BTAC, and cooperative MAC for IEEE 802.11g). In this paper, some security attacks against COSMIC, VC-MAC, BTAC, and cooperative MAC for IEEE 802.11g are disclosed, and security vulnerabilities that arise in them due to attacks, are then analyzed. The emerging, channel-assisted authentication mechanism using physical layer characteristics is also discussed to verify entities in cooperative MAC protocols. To my knowledge, this is the first comprehensive case study of security issues caused by possible security attacks on COSMIC, VC-MAC, BTAC, and cooperative MAC for IEEE

 ISBN: 978-1-61208-330-8

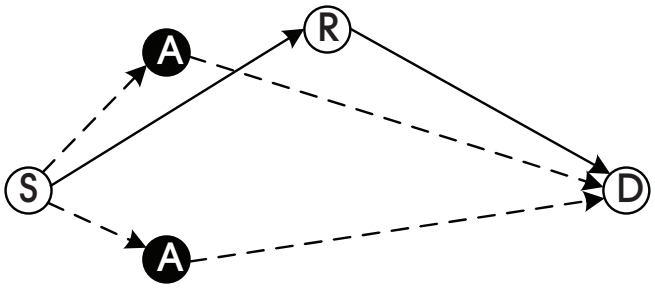


Figure 1. Example of Security Attack in Cooperative Wireless Networks.

802.11g.

The remainder of this paper is organized as follows. In Section II, a brief description of some cooperative MAC protocols is provided. In Section III, some possible security attacks are disclosed and then the security vulnerabilities caused by these attacks are analyzed. In Section IV, channel-assisted authentication mechanism is discussed to authenticate entities in cooperative MAC protocols. Finally, in Section V, conclusions are presented along with plans for future work.

II. Cooperative MAC Protocols

Cooperative wireless communication is an innovative communication scheme that takes advantage of the open broadcast nature of the wireless medium, and its spatial diversity, to improve channel capacity, reliability, robustness, delay, and coverage. It is known to be essential for making ubiquitous communication connectivity a reality. Multiple protocols in the MAC layer have been suggested to utilize the concept of cooperative transmission.

COSMIC is a cooperative MAC protocol for WSN with minimal control packets. It uses only one control packet, request-for-relay (RFR), for relay selection. In COSMIC, the relay selection is decided using both the channel-state information (CSI) and the remaining energy. COSMIC is able to increase network lifetime by about 25% and the delivery ratio by 5 times [3].

VC-MAC is a cooperative MAC protocol for vehicular networks. It is composed of four stages, namely, the gateway broadcast period, information exchange period, relay set selection period, and data forwarding period. VC-MAC exploits the concept of cooperative communication and takes advantages of the broadcast nature of the wireless medium to maximize throughput. This protocol also leverages spatial diversity and user diversity to overcome the unreliability under many broadcast scenarios. VC-MAC significantly increases system throughput compared with existing strategies [4].

BTAC is a cooperative MAC that increases throughput in multi-rate WLANs. A busy tone signal of only one-time-slot length is used to improve the throughput performance and reduce relay. It is known to improve throughput performance by at least 35% and reduce system delay, compared to the

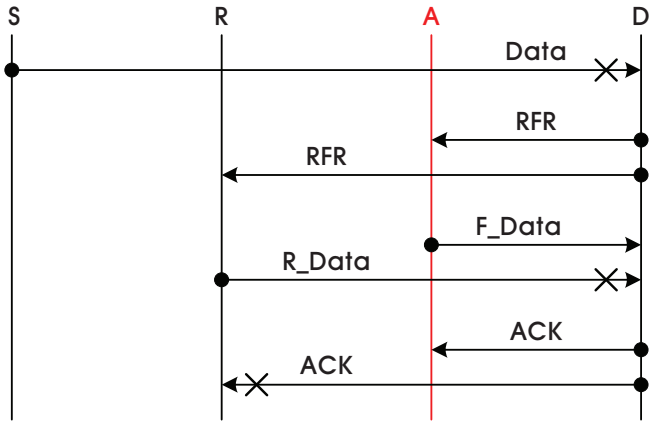


Figure 2. Security Vulnerability in COSMIC.

IEEE 802.11b MAC protocol. BTAC is compatible with IEEE WLAN [5].

To increase performance and reduce energy consumption in previous versions of cooperative MAC for IEEE 802.11b, a new cooperative MAC protocol for IEEE 802.11g (being extended to 802.11n) was proposed. It can support ten different transmission rates (1, 2, 6, 9, 12, 18, 24, 36, 48, and 54 Mbps) and can efficiently reduce the time for selecting better relays, by partitioning the relays with similar transmission rates, into the same groups [6].

III. Security Vulnerability in Cooperative MAC Protocols

Cooperative MAC protocols suffer from vulnerability to various security attacks due to the open broadcast nature of the wireless channel and the use of cooperative communication with multiple relays [8][9].

For example, in Fig. 1, let us assume that the attacker is closer to source than to the relay, or that it is between source and relay. In this environment, attacker can disguise itself as relay to allow its illegal packet to get to source and destination. There is no suitable countermeasure to prevent this attack, nor any way to authenticate legitimate relay. Therefore, the result is disruption of the normal cooperative transmission between source and destination.

Attackers are focused on network performance, which means they want to disturb the communication between source and destination. They would exploit the weakness in the cooperative procedure, especially in the control packet exchange, and disguise themselves as legitimate relays to disturb the network operation, and to degrade the wireless channel quality. Security attacks based on control packets can be classified into two categories: faked request-to-send (RTS) attacks and faked clear-to-send (CTS) attacks. The former generates a false RTS packet in order to achieve virtual jamming of source, while the latter generates a false

 ISBN: 978-1-61208-330-8

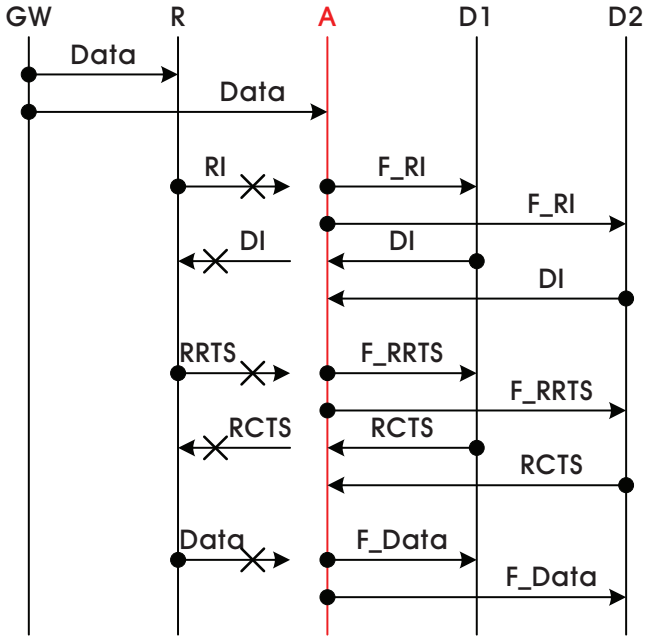


Figure 3. Security Vulnerability in VC-MAC.

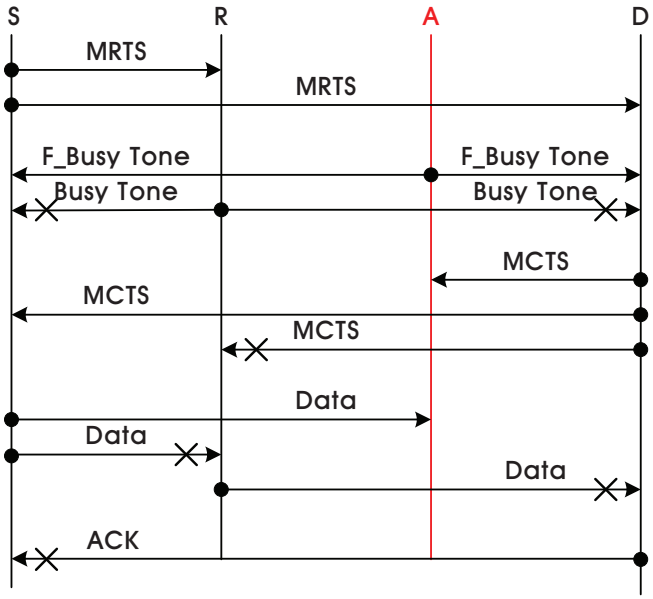


Figure 4. Security Vulnerability in BTAC.

CTS packet in order to disguise attacker as legitimate relay or destination.

A. Security Vulnerability in COSMIC

Fig. 2 shows a destination attack caused by faked data (F_Data) from attacker. In this case, attacker deliberately transmits its F_Data to destination, informing it that attacker is a legitimate relay.

As shown in Fig. 2, in COSMIC, source sends data to destination, which receives the data. Neighbors, relay and attacker overhear it. If destination is able to decode the data, it sends an acknowledgment (ACK) to source. In this case, no relaying is needed. However, if destination is not able to decode the data, a cooperative relaying is engaged.

When destination doesn't successfully receive data from source, it sends a request-for-reply (RFR) to relay to express its need for a relaying. Destination then waits for the data (R_Data) from relay. The R_Data is the relayed copy of the data. Since attacker is close to relay, it is able to receive the RFR. After receiving the RFR from destination, attacker sends its faked data (F_Data) to destination. Finally, destination sends an ACK to attacker to notify that it successfully received the data. This blocks the transmission of R_Data from relay. Consequently, cooperative communication between source and destination is not established.

B. Security Vulnerability in VC-MAC

Fig. 3 illustrates a security attack using faked relay information (F_RI) from an attacker in the VC-MAC.

In the VC-MAC, after the gateway, which is deployed along the roadside, senses the channel is idle, it sends data directly with no handshaking procedure. After the broadcast of the gateway, relay and attacker, which both received the data, become potential relays. Attacker sends faked relay information (F_RI) to two destinations (Destinations 1 and 2) before the transmission of relay information (RI) from relay. Attacker then waits for destination information (DI) from destinations 1 and 2. Since the authentication (or integrity) mechanism is not applied to the control packets exchanged between relay, and destinations 1 and 2, the legal RI from relay may be rejected by destinations 1 and 2 due to illegal previous F_RI received from attacker. This means that because destinations 1 and 2 have already received RI from attacker, they reject additional RI from relay. Once attacker receives two sets of DI, it transmits a faked relay request-to-send (F_RRTS). After receiving the relay clear-to-send (RCTS), attacker makes a faked data (F_Data) transmission to destinations 1 and 2. As a result, normal cooperation between relay and destination 1 or 2 cannot be guaranteed.

C. Security Vulnerability in BTAC

The potential security attack in BTAC is also shown in Fig. 4. An attacker sends a faked busy tone (F_BusyTone) to inform the source and destination that it is an intended helper to forward the data received from source.

Source sends modified RTS (MRTS) to relay and destination. Attacker near relay or destination comes to know of this. The F_BusyTone is sent from attacker to source and destination. This means that attacker is an intended legitimate relay for forwarding data. Accordingly, since the authentication (or integrity) mechanism is not applied to F_BusyTone, the legal busy tone (BusyTone) from relay is denied by source and destination due to the previous illegal

 ISBN: 978-1-61208-330-8

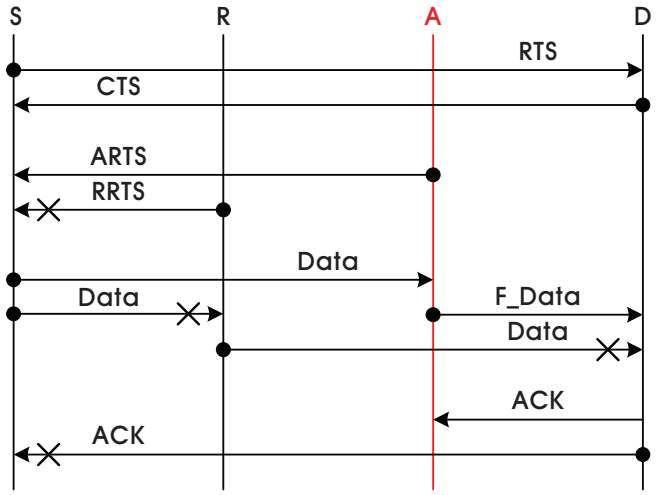


Figure 5. Security Vulnerability in Cooperative MAC for IEEE 802.11g or 802.11n.

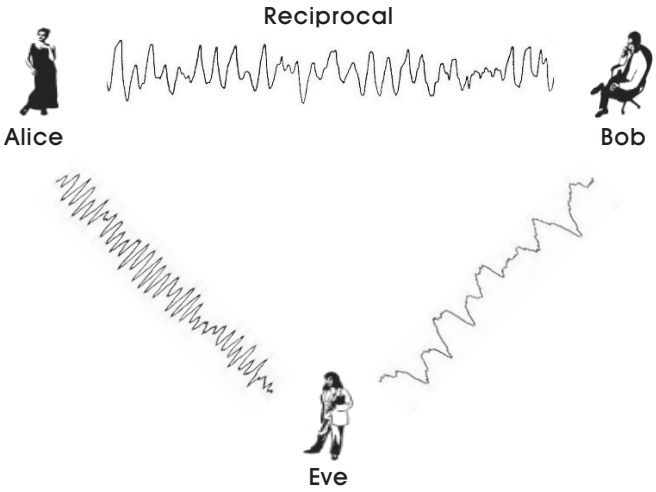


Figure 6. Example of Security Community with Alice (Legitimate), Bob (Legitimate), and Eve (Illegitimate).

F_BusyTone received from attacker. Then, destination sends its modified CTS (MCTS) to attacker and source. Source sends data to attacker, not to relay. Finally, attacker denies cooperative communication service to source by simply dropping the data it receives from source, or forwarding faked data to destination. Due to this false data transmission from source to attacker, cooperative communication between source and destination via relay is not established.

D. Security Vulnerability in Cooperative MAC for IEEE 802.11g or 802.11n

Fig. 5 shows a security vulnerability caused by the illegal RTS packet (ARTS) from attacker in cooperative MAC for IEEE 802.11g.

When source finds a free channel and it can send data to destination, it will send an RTS to destination and wait for CTS from destination. Since attacker, as well as relay, can overhear both RTS and CTS, attacker can communicate with both source and destination so that it can serve as a legitimate helper candidate between source and destination. Just after overhearing both RTS and CTS, attacker calculates the data rates from itself to source and destination. Attacker then replies ARTS to tell source that it can help with transmission. This means that attacker is an intended legitimate relay forwarding data. Since source receives illegal RTS (ARTS) from attacker, it rejects the legal RTS (RRTS) from relay. Then, source sends data to attacker, not relay. If attacker receives data from source, it simply drops the data received or forwards faked data (F_Data) to destination. It may also spoof an acknowledgment (ACK), causing destination to wrongly conclude a successful cooperative transmission via relay.

IV. Channel-Assisted Authentication Mechanism

In order to prevent the security attacks inherent in cooperative MAC protocols and verify entities more efficiently, a channel-assisted authentication mechanism using physical layer properties of wireless channel is discussed. The following four main characteristics of wireless channel can allow the wireless channel to be used as a means to authenticate the legitimate entity [17][18].

- The impulse response for time-variant wireless channel decorrelates quite rapidly in space.
- Wireless channel also changes in time, which results in a natural refresh for a channel-assisted security mechanism.
- The wireless channel is reciprocal in space.
- The time variation is slow enough so that the channel response can be accurately estimated within the channel coherent time.

In the typical environment shown in Fig. 6, three entities (Alice, Bob, and Eve) are potentially located in spatially separated positions. Alice and Bob are the two legitimate entities, and Eve is the illegitimate entity. Alice is the transmitter that initiates communication and sends data, while Bob is the intended receiver. Eve is an attacker that injects false signals into the channel in the hope of spoofing Alice. The main security goal is to provide authentication service between Alice and Bob. The legitimate receiver (Bob) should have to distinguish between legitimate signals from legitimate transmitter (Alice) and illegitimate signals from attacker (Eve).

As depicted in Fig. 6, let us suppose that Alice transmits data to Bob at a sufficient rate to ensure temporal coherence between successive data sets. In addition, while trying to impersonate Alice, Eve wishes to convince Bob that she is

 ISBN: 978-1-61208-330-8

Alice. To provide authentication between Alice and Bob, Bob first uses the received signal from Alice to estimate the channel response. He then compares this signal with a previous signal version of the Alice-Bob channel. If the two channel responses are close to each other, Bob concludes that the source of the data is the same as that of the previously transmitted data. Otherwise, Bob concludes that the transmitter is not Alice [18][19]. Using this uniqueness of the Alice-Bob wireless channel, it is possible to distinguish between legitimate transmitter (Alice) and illegitimate one (Eve). It is caused by the fact that the wireless channel decorrelates in space, so the Alice-Bob channel is totally uncorrelated with the Alice-Eve and Bob-Eve channels if Eve is more than an order of a wavelength away from Alice and Bob.

V. CONCLUSIONS

Security is the principal issue that must be resolved in order for the potential of CWNs to be fully exploited. This work provides a comprehensive analysis of the security issues caused by attackers for cooperative MAC protocols such as COSMIC, VC-MAC, BTAC, and cooperative MAC for IEEE 802.11g. Security vulnerabilities are analyzed at each handshaking stage, while attacking control packets are being exchanged among nodes (source, destination, and relay). It also discusses that a channel-assisted authentication mechanism is applicable to enhance and supplement conventional cryptographic authentication mechanism for cooperative MAC protocols. These results should be significantly useful in the design of efficient authentication mechanisms for secure, cooperative MAC protocols.

In the future, the author plans to design and implement a lightweight, low-power authentication (or integrity) mechanism using physical layer properties suitable for CWNs. The plan is then to examine the effects of the proposed mechanism on security cost, power consumption, and transmission performance.

REFERENCES

[1] A. Nosratinia, T. E. Hunter, and A. Hedayat, "Cooperative Communication in Wireless Networks," IEEE Communication Magazine, vol. 42, October 2004, pp. 74-80.

[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, March 2006, pp. 659-672.

[3] A. B. Nacef, S.-M. Senouci, G.-D. Yacine, and A.-L. Beylot, "COSMIC: A Cooperative MAC Protocol for WSN with Minimal Control Messages," IFIP International Conference on New Technologies, Mobility and Security (NTMS 2011), February 2011, pp. 1-5.

[4] J. Zhang, Q. Zhang, and W. Jia, "VC-MAC: A Cooperative MAC Protocol in Vehicular Networks," IEEE Trans. on Vehicular Technologies, vol. 58, March 2009, pp. 1561-1571.

[5] S. Sayed and Y. Yang, "BTAC: A Busy Tone Based Cooperative MAC Protocol for Wireless Local Area Networks," International Conference on Communication and Networking in China (ChinaCom 2008), August 2008, pp. 403-409.

[6] J.-P. Sheu, J.-T. Chang, C. Ma, and C.-P. Leong, "A Cooperative MAC Protocol Based on 802.11 in Wireless Ad hoc Networks," IEEE Wireless Communications and Networking Conference (WCNC 2013), April 2013, pp. 416-421.

[7] J. Rousselot, A. El-Hoiydi, and J.-D. Decotignie, "Low Power Medium Access Control Protocols for Wireless Sensor Networks," European Wireless Conference (EW 2008), June 2008, pp. 1-5.

[8] K. H. Kim, "Security Attack based on Control Packet Vulnerability in Cooperative Wireless Networks," IARIA International Conference on Networking and Services (ICNS 2013), March 2013, pp. 123-128.

[9] K. H. Kim, "Analysis of Security Vulnerability in Cooperative Communication Networks," IARIA International Conference on Networking and Services (ICNS 2011), April 2011, pp. 80-84.

[10] T. O. Walker III, M. Tummala, and J. McEachen, "Security Vulnerabilities in Hybrid Flow-specific Traffic-adaptive Medium Access Control," Hawaii International Conference on System Sciences (HICSS 2012), January 2012, pp. 5649-5658.

[11] 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, April 2011, pp. 890-902.

[12] Y. Mao and M. Wu, "Tracing Malicious Relays in Cooperative Wireless Communications," IEEE Trans. on Information Forensics and Security, vol. 2, June 2007, pp. 198-207.

[13] S. Dehnie, H. T. Sencar, and N. Memon, "Detecting Malicious Behavior in Cooperative Diversity," Conference on Information Sciences and Systems (CISS 2007), March 2007, pp. 895-899.

[14] S. Dehnie and S. Tomasin, "Detection of Selfish Nodes in Networks Using CoopMAC Protocol with ARQ," IEEE Trans. on Wireless Communications, vol. 9, July 2010, pp. 2328-2337.

[15] H. Li, M. Xu, and Y. Li, "The Research of Frame and Key Technologies for Intrusion Detection System in IEEE 802-11-based Wireless Mesh Networks," International Conference on Complex, Intelligent and Software Intensive Systems (CISIS 2008), March 2008, pp. 455-460.

[16] Y. Hao, J. Tang, and Y. Cheng, "Secure Cooperative Data Downloading in Vehicular Ad Hoc Networks," IEEE Journal on Selected Areas in Communications/Supplement, vol. 31, September 2013, pp. 523-537.

[17] S. Mathur, A. Reznik, Y. Chunxuan, and R. Mukherjee, "Exploiting the Physical Layer for Enhanced Security," IEEE Wireless Communications, vol. 17, October 2010, pp. 63-70.

[18] L. Xiao, L. J. Greenstein, N. B. Mandayam, and W. Trappe, "Using the Physical Layer for Wireless Authentication in Time-Variant Channels," IEEE Trans. on Wireless Communications, vol. 7, July 2008, pp. 2571-2579.

[19] K. Zeng, K. Govindan, and P. Mohapatra, "Non-Cryptographic Authentication and Identification in Wireless Networks," IEEE Wireless Communications, vol. 17, October 2010, pp. 56-62.

 ISBN: 978-1-61208-330-8

UBI-CA : A Clustering Algorithm for Ubiquitous Environments

Rim Helali
Higher Institute of Management of Tunis
Research Laboratory SOIE
Tunis, Tunisia
rim.helali@gmail.com

Nadia Ben Azzouna
National School of Computer Science
Research Laboratory SOIE
Manouba, Tunisia
nadia.benazzouna@ensi.rnu.tn

Khaled Ghedira
Higher Institute of Management of Tunis
Research Laboratory SOIE
Tunis, Tunisia
khaled.ghedira@isg.rnu.tn

***Abstract*—This paper describes a new Clustering Algorithm for UBIquitous environments (UBI-CA). These types of environments are dynamic in nature due to the mobility of nodes. The constant motion of devices and their unexpected appearance or disappearance could perturb the stability of the network topology. Clustered architectures seem to be adequate to address such a challenge. In the proposed algorithm, inspired by the Weighted Clustering Algorithm (WCA), we aim for the reduction of the computation and communication costs by electing the most suited cluster heads on the base of a calculated weight. We propose to emphasize on the mobility and the number of neighboring nodes while calculating its weight in order to cope with ubiquitous environments specificities. The cluster structure is maintained dynamically as nodes move by defining a nodes dynamicity management method. Simulation results prove that the proposed algorithm ensure a good scalability for our ubiquitous system as the number of nodes increases. Furthermore, when compared with the original algorithm WCA, UBI-CA shows a better stability with increasing transmission range of nodes.**

Keywords-ubiquitous environments; clustered architecture; mobility; scalability.

I. INTRODUCTION

Ubiquitous computing aims to exchange and share services anywhere, anytime. In addition to the abundant number of users, ubiquitous environments are characterized by the heterogeneity of devices (computers, mobile phones, personal digital assistants, etc.) and their dynamicity (the mobility of devices in the environment and their unexpected appearance or disappearance [1]). A distributed architecture is well adapted for this type of networks in order to prevent the problems of bottleneck and Single Point Of Failure (SPOF) caused by the big number of connected devices. Furthermore, the mobility of nodes should be considered in the deployed architecture. Clustered architectures are proven to be adequate for ubiquitous environments thanks to their distributed deployments and to their ability to be robust in the face of topological changes caused by nodes motion [2]. In that context, a clustered architecture for ubiquitous environments should be able to dynamically adapt itself with the changing network configurations [3] and to reduce the communication costs. Therefore, a clustering algorithm UBI-CA is proposed in this paper allowing the division of the geographical region into small zones [3] in order to reduce the cost of searching services. Hence, we aim in this work to define a clustering algorithm that manages the high dynamicity of the environment while guaranteeing the scalability and the stability requirements [4].

The rest of this paper is organized as follows. Section 2 addresses related work and identifies some limitations. Section 3 describes the proposed algorithm. Section 4 presents simulation results. Finally, Section 5 concludes the paper and discusses some future directions.

II. RELATED WORK

Several clustering algorithms have been proposed in the literature especially for ad hoc networks. According to Agarwal et al. in their review of clustering algorithms [5], clustering algorithms for mobile ad hoc networks (MANET) could be classified in two categories, single metric based clustering and multiple metrics based clustering. In single metric based clustering, only one performance factor is considered to select cluster heads. A number of clustering algorithms for this category has been proposed in the literature. In Lowest ID cluster algorithm (LIC) [6], the node with the lowest id is chosen as a cluster head. In Highest connectivity clustering algorithm (HCC) [7], each node broadcasts its id to the nodes that are within its transmission range. The node with the highest number of neighbors is elected as cluster head. In K-CONID [8], two methods are considered in order to choose cluster heads. Connectivity is used as a first criterion and lower ID as a secondary criterion. The above mentioned algorithms suffer from the problem of cluster head overloading since each algorithm deals with only one parameter [3]. To solve this problem, multiple metrics based clustering has been proposed. In this category, a number of metrics, such as node degree, residual energy capacity and, moving speed are taken into account [5]. In this context, Basagni proposed two clustering algorithms [9], distributed clustering algorithm (DCA) and distributed mobility adaptive clustering algorithm (DMAC). In DCA, a node is chosen to be a cluster head if its weight is higher than any of its neighbors weight; otherwise, it joins a neighboring cluster head. The weight is generic and is calculated according to nodes mobility related parameters [5]. It is assumed in DCA that the network topology does not change during the execution of the algorithm. The DMAC algorithm was proposed as an extension to DCA allowing the adaptation of the algorithm to the network topology. In Weighted Clustering Algorithm (WCA) [3], the authors focused mainly on the stability of

 ISBN: 978-1-61208-330-8

the network topology. They take into consideration the ideal degree which refers to the ideal number of nodes a cluster head can handle, transmission power, mobility, and battery power of mobile nodes. The cluster head election procedure is invoked on-demand and aims to reduce the computation and communication costs. Nevertheless, WCA suffers from several drawbacks especially if applied in ubiquitous environments. For example, using Global Positioning System (GPS) in order to calculate nodes mobility appears to be not adequate due to devices high heterogeneity. Furthermore, considering the cumulative time during which a node acts as a cluster head to calculate the battery power is of a low relevance in our case because knowing the consumed battery power of a node does not reflect the time during which a node can play the role of a cluster head after being elected.

Thus, in our proposed algorithm, we adapt the Weighted Clustering Algorithm (WCA) originally proposed for ad hoc mobile networks to meet ubiquitous environments challenges.

III. PROPOSED ALGORITHM

In this section, we present our defined clustering algorithm UBI-CA which is inspired from the WCA algorithm. We justify the choice of WCA by the fact that this algorithm considers several parameters like transmission power, mobility and battery power of mobile nodes [3]. These parameters are very significant for the election of the cluster heads, especially in ubiquitous environments where nodes are characterized by their high dynamicity. In our proposed architecture for a semantic and dynamic cluster based web service discovery system presented in [10], each cluster is organized in a two level hierarchical architecture containing a Directory Server (DS) and a number of web service servers (WSS) [11]. Thus, we are dealing, in this algorithm, with three types of nodes :

- A Directory Server (DS): responsible of a certain region in the ubiquitous environment. It maintains descriptions of all existing web services distributed within this region [11].
- A backup Directory Server (backup DS): used to replace the DS in case of failure in order to guarantee service availability.
- A Web Service Server (WSS): containing a number of web services and their descriptions. A WSS is assigned to a DS in the ubiquitous environment. This DS becomes its gate to the entire ubiquitous environment [11].

The following subsection describes in details the UBI-CA algorithm in order to form the proposed architecture. A preliminary version of this algorithm appeared in [10].

A. DSs election and clusters formation

Initially, there are no elected DSs in the environment. The DSs list is generated for the first time by invoking the DSs election procedure at system activation time. The procedure begins by nodes broadcasting to their immediate neighbors (i.e., one-hop neighbors) their ID. Then, only the nodes with a transmission delay lower than a predefined threshold are chosen. The motivation behind this restriction is essentially

Algorithm 1 UBI-CA

Input : listID, n, TD, P_v, δ
Initialization : dv=0
for $i = 1$ to n **do**
  for $j = 1$ to $n-1$ **do**
    if Transmission Delay $\langle$ TD **then**
      $d_v = d_v + 1$
      $add(listID(j), N)$
    end if
  end for
  $\Delta_v = |d_v - \delta|$
  $M_v(t-\Delta t) = \frac{1}{|U|}\sum_{i,j\in U}\frac{M_{ij}(t-\Delta t)}{\Delta t}$
  $W_v = \frac{(1+b\Delta_V + cM_v)^{\alpha}}{P_v}$
end for

listId = list of the identifiers of one-hop neighbors, TD = a predefined transmission delay threshold that a node cannot exceed to be considered as a neighbor, n = the total number of one-hop neighbors, N = the set of one-hop neighbors of the node v with transmission delay $\langle$ TD, P_V = Battery power at the time t.

Figure 1: UBI-CA

the reduction of the distance between a node and its neighbors. The procedure, as illustrated in Fig. 1, consists of electing the DSs on the base of a calculated value W_v. In the original WCA algorithm, the value W_v calculated for every node is a sum of components with certain weighting factors chosen according to the system needs [3]. Due to the dynamicity of the ubiquitous environment and in order to guarantee the system stability, we judge that the nodes mobility M_v and the degree of nodes (i.e., the number of neighboring nodes D_v) are more relevant for DS stability than battery power in the DSs election procedure. Therefore, the value W_v is calculated as (1) :

$$W_v = \frac{(1+b\Delta v + cMv)^{\alpha}}{Pv} \quad (1)$$

After calculating its W_v, each node v sends its value to its neighbors. This exchange of messages permits the nodes to be aware of the node with the lowest W_v in their neighbors set which is elected as a DS. Non elected nodes (i.e., WSSs) send a membership query to their local DS. Thus, each new elected DS forms a list of its cluster members. If a node receives at least one membership query despite the fact that it has not the lowest W_v in its neighbors set (it has the lowest W_v in another node neighbors set), it has to disjoin the cluster to which it has been assigned as a WSS by sending a disjunction query to the local DS and should be elected as a DS. Each elected DS chooses from the cluster the node that has the second lowest W_v as a backup DS. If a node has an empty neighbors set due to transmission delay restriction, it is then elected as a DS in its region [10].

In the next subsections, we describe in details the three components forming the W_v value.

B. The degree difference Δ_v

The first component Δ_v computes the degree-difference for each node v by calculating the difference between the ideal number of nodes a DS can support δ and the number of

 ISBN: 978-1-61208-330-8

Algorithm 2 Nodes Mobility

Input : d_v, array x, array y
for $i - 1$ to d_v **do**
  for $j = i + 1$ to $d_v - 1$ **do**
    if $i \in N_j$ and $j \in N_i$ **then**
      $add(i, j, U)$
      $\phi_{vj} = cos^{-1} \frac{d_{vi}^2 + d_{ij}^2 + d_{vj}^2}{2 d_{vi} d_{ij}}$
      $l = \sqrt{d_{vi}^2 + (\frac{d_{ij}}{2})^2 - d_{vi} d_{ij} cos\phi_{vj}}$
      $\theta = cos^{-1} \frac{l^2 + (\frac{d_{ij}}{2})^2 - d_{vj}^2}{l d_{ij}}$
      $y'(i) = l cos\theta$
      $x'(i) = l sin\theta$
    end if
  end for
end for
for $i = 1$ to $|U|$ **do**
  $M_i = \sqrt{(y'^2 - y^2) + (x'^2 - x^2)}$
  $M_v = M_v + \frac{M_i}{\Delta t}$
end for
$Mv(t - \Delta t) = \frac{1}{|U|} M_v$
return M_v

Figure 2: Nodes mobility

neighbors d_v. It helps, thus, choosing the node with the nearest degree to the ideal degree δ as a DS. This is to ensure that the DSs are not overloaded and the efficiency of the system is maintained at the expected level [3].

C. Nodes mobility

The component M_v represents the mobility of nodes (i.e., nodes changing their location over time). To calculate nodes mobility, we choose, rather than using the mobility metric proposed in the original algorithm that is based on a localization system, to use the mobility metric proposed in [12] that does not depend on any location system (e.g., GPS) and can fully capture the relative motions between a node and its neighborhood, in real-time, using simple triangulation [12].

As illustrated in Fig. 2, the set U represents for each node v the pair of nodes that are both neighbors to the node v as well as one-hop neighbors to each other. Each node periodically measures the distances to its one-hop neighbors using the "transmission delay" between nodes (i.e., d_{vi}, d_{ij}, d_{vj}) . The measurements are sent periodically with a frequency of sending of $1/\Delta t$ [12]. l calculates the distance from node v to the midpoint of the line between nodes i and j as well as the included angle θ to interpret the relative position between node v and $pair(i, j)$ [12].

D. Battery power

The last component P_v, represents the battery power of a node. We choose in our algorithm rather than computing the consumed battery power used in WCA, to focus on the remaining battery power and to elect the node with the highest value. We assume that the remaining battery power is more relevant in our case due to the fact that a node may initially have limited battery power. So, knowing the consumed battery power of a node does not reflect the time during which a node can play the role of a DS after being elected.

E. Nodes dynamicity management

Due to nodes mobility caused by the dynamic nature of the ubiquitous environment, a node may be detached from its local DS and looks for a new DS; a reaffiliation is thus needed. New DSs could be added to the set if a node cannot find a DS to which it can be affiliated due to the overload of neighboring DSs. In addition to nodes mobility, unavailability of nodes (i.e., breakdown, battery power extinction) has to be managed especially for DSs and backup DSs. We discuss each of these cases in details in what follows [10].

- A node detecting a new DS with higher signal strength than its local DS: The node sends an affiliation request. The DS updates its list and informs the old DS after accepting the request.
- The arrival of a new node which sends an affiliation request to all its neighboring DSs but no response is received (due to the overloading of neighboring DSs): The DS election procedure is invoked, but only neighboring clusters are involved. After reforming the clusters, each old DS sends to its local DS its registered service descriptions, and each new DS sends a DS replacement query to all other DSs in the environment.
- The failure of a DS (If no messages are received by the backup DS during a specific period of time): The backup DS initiates the DS election procedure by sending a DS election query to the WSSs members of the cluster. After electing the DS, the backup DS sends its registered service descriptions to the new DS which chooses the node with the second lowest W_v as its new backup DS and sends to the other DSs in the environment a message informing them of its new status.
- The failure of a backup DS (if no acknowledgements are received during a specific period of time) : The DS selects the node that has the smallest W_v from its WSSs to be the new backup DS (Nodes have to recalculate their W_v)
- The failure of both DS and backup DS (the unavailability of the DS is detected by the WSSs): The WSSs have to elect a new DS among them by recalculating their W_v and choosing the WSS with the smallest W_v. The second WSS to have the smallest W_v is chosen as a backup DS.

IV. Simulation Study

In order to study the performance of our algorithm UBI-CA, a simulator was developed using omnet++, an extensible, modular, component-based C++ simulation library and framework. OMNeT++ provides a component architecture for models. Components (modules) are programmed in C++, and then assembled into larger components and models using a high-level language (NED) [13]. For the simulation study, the following aspects are observed: (i) the scalability of the system, using UBI-CA, in terms of maintaining a good performance as the number of nodes increases, (ii) the stability of the proposed algorithm in terms of DSs number while the number of nodes and transmission range increase.

 ISBN: 978-1-61208-330-8

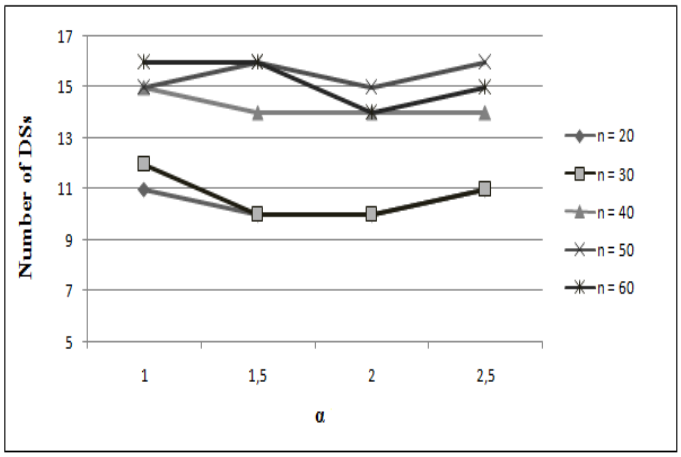


Figure 3: Number of DSs variation with α (transmission range =100)

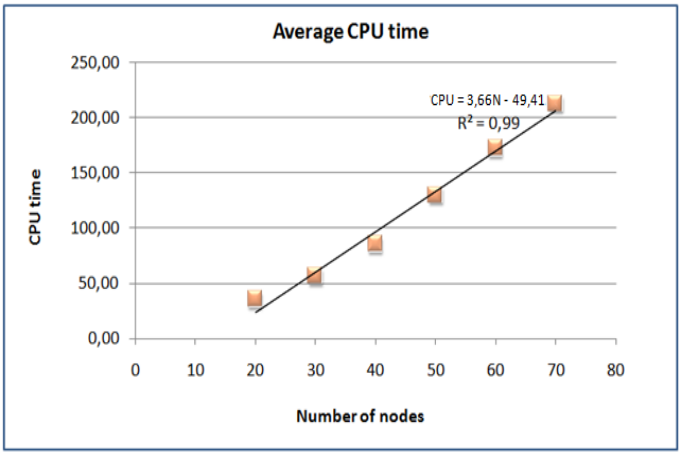


Figure 4: Average CPU time variation

A. Simulation Environment

The experiments were all conducted using an Intel (R) Pentium (R) 2.00 GHz computer with 3 Go RAM, running Windows 7. We simulate a system of N nodes on a 900 x 900 grid. In our simulation experiments, N was varied between 20 and 70, and the transmission range was varied between 10 and 100. The values used for the simulation are c = 0.5 and b = 0.5, and for the ideal degree δ =10 . Note that these values are arbitrary and should be adjusted according to the system requirements. The value of α (used in (1)) is set to be equal to 2. Note that this value is obtained by varying the parameter between 1 and 2.5 and, as shown in Fig. 3, α = 2 provides the lowest number of DSs.

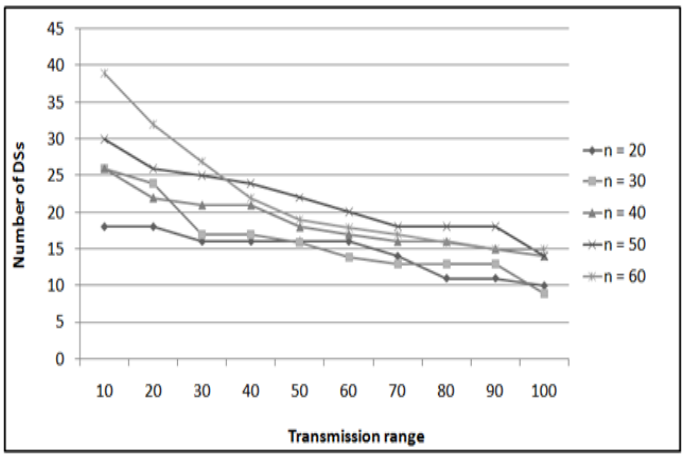


Figure 5: Number of DSs variation

B. Experimental results

In order to analyze how our system scales with increasing number of nodes, we varied the transmission range between 10 and 80 meters and measured the variation of the average CPU time with respect to the number of nodes N. We observe, as shown in Fig. 4, that the increase of the average CPU time is linear. The linearity is proven by the correlation coefficient R^2 value (calculated according to PEARSON function [14]) which is close to 1 indicating an excellent linear fit. This shows that our system has a good performance as the number of nodes increases.

The stability of our system is tested by measuring the variation of the number of DSs with respect to the transmission range. The results are shown for varying N. We observe, as noticed in Fig. 5, that the number of DSs decreases with the increase of the transmission range. This is due to the fact that a DS with a large transmission range will cover a larger area. The results show that the number of DSs is not influenced by the variation of the number of nodes which emphasizes the stability of the system.

Comparing our results with those found using WCA algorithm (as shown in Fig. 6), we notice that the UBI-CA results are proven to be better in terms of number of DSs. This shows that the emphasis put on both nodes mobility and degree difference while calculating the W_v value is confirmed to be adequate for ubiquitous environments.

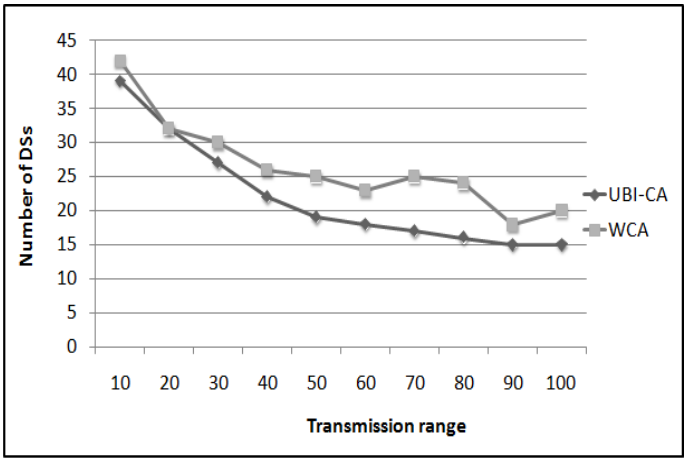


Figure 6: Number of DSs variation for UBI-CA and WCA (N = 60)

 ISBN: 978-1-61208-330-8

V. Conclusion and future work

In this paper, we introduced a new clustering algorithm for ubiquitous environments named UBI-CA which is inspired from the original Weighted Clustering Algorithm (WCA). The choice of WCA is based on the fact that this algorithm takes into consideration several parameters like transmission range, mobility and battery power of mobile nodes. However, this algorithm suffers from several problems that we tried to resolve. Therefore, to calculate nodes mobility, we used the mobility metric proposed in [12] which is independent from any location system (e.g., GPS). Furthermore, battery power in our algorithm represents the remaining battery power of a node rather than the consumed power used in the original algorithm. In order to calculate the W_v value, we focused mainly on the nodes mobility M_v and the degree of nodes Δ_v. In order to evaluate our approach, we observed the scalability and the stability of the proposed algorithm. Our future work includes the extension of the proposed system to support quality of service during the web service selection in order to meet the challenges of ubiquitous environments such as invisibility.

References

[1] I. Abdennadher, "Une approche pour l'assurance des qualites de services des systemes publier/souscrire deployes sur un reseau mobile ad-hoc," Master's thesis, National Engineering School of Sfax, Tunisia, 2008.

[2] C. R. Lin and M. Gerla, "Adaptive clustering for mobile wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 15, pp. 1265–1275, 1997.

[3] M. Chatterjee, S. K. Das, and D. Turgut, "Wca: A weighted clustering algorithm for mobile ad hoc networks," *Journal of Cluster Computing (Special Issue on Mobile Ad hoc Networks)*, vol. 5, pp. 193–204, 2001.

[4] M. Satyanarayanan, "Pervasive computing: Vision and challenges," *IEEE Personal Communications*, vol. 8, pp. 10–17, 2001.

[5] R. Agarwal1, R. Gupta, and M. Motwani3, "Review of weighted clustering algorithms for mobile ad hoc networks," *GESJ: Computer Science and Telecommunications*, vol. 33, pp. 71–78, 2012.

[6] D. J. Baker and A. Ephremides, "A distributed algorithm for organizing mobile radio telecommunication networks." in *In proceeding of the 2nd International Conference on Distributed Computing Systems*, 1981, pp. 476–483.

[7] M. Gerla and J. T. chieh Tsai, "Multicluster, mobile, multimedia radio network," *Journal of Wireless Networks*, vol. 1, pp. 255–265, 1995.

[8] G. Chen, F. Nocetti, J. Gonzalez, and I. Stojmenovic, "Connectivity-based k-hop clustering in wireless networks," in *Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02)-Volume 7 - Volume 7*, ser. HICSS '02, 2002, pp. 2450–2459.

[9] S. Basagni, "Distributed clustering for ad hoc networks," in *International Symposium on Parallel Architectures, Algorithms, and Networks*, 1999, pp. 310–315.

[10] R. Helali, N. B. Azzouna, and K. Ghdira, "Towards a semantic and dynamic cluster based web service discovery system for ubiquitous environments." in *ICEIS (2)*. SciTePress, 2012, pp. 295–300.

[11] T. Xu, B. Ye, M. Kubo, A. Shinozaki, and S. Lu, "A gnutella inspired ubiquitous service discovery framework for pervasive computing environment," *8th IEEE International Conference on Computer and Information Technology, CIT*, pp. 712 – 717, 2008.

[12] Z. Li, L. Sun, and E. C. Ifeachor, "Gps-free mobility metrics for mobile ad hoc networks," *IET Communications*, vol. 5, pp. 1–18, 2007.

[13] "Omnet++ documentation," http://omnetpp.org/documentation, accessed: 16.12.2013.

[14] K. Pearson, "Notes on regression and inheritance in the case of two parents," 1985, pp. 240–242.

 ISBN: 978-1-61208-330-8

Protocol Independent Multicast in OMNeT++

Vladimír Veselý, Ondřej Ryšavý, Miroslav Švéda
Department of Information Systems
Faculty of Information Technology, Brno University of Technology (FIT BUT)
Brno, Czech Republic
e-mail: {ivesely, rysavy, sveda}@fit.vutbr.cz

***Abstract*—Multicast transmission for one-to-many data delivery and for online media streaming are becoming more and more popular. Interest in proper simulation and modeling has increased together with those two trends. This paper introduces simulation modules for dynamic multicast routing, namely Protocol Independent Multicast and its variants Dense Mode and Sparse Mode. Both of them are now parts of our ANSA extension developed within the INET framework for OMNeT++ discrete event simulator.**

Keywords–Multicast Routing; PIM-DM; PIM-SM.

I. INTRODUCTION

The multicast transfers prove to be more efficient for one-to-many data delivery if there is one (or more) known source(s) and a number of unknown destinations ahead [1]. Multicast spares network resources, namely bandwidth. Sender and receivers communicate indirectly instead of many separate connections between them. Because of that, multicast traffic is carried across each link only once and the same data are replicated as close to receivers as possible. However, this effectiveness goes concurrently with increased signalization and additional routing information exchange which is done by the following protocols:

- IGMP (Internet Group Management Protocol) [2] /MLD (Multicast Listener Discovery) [3] – End-hosts and first hop multicast-enable routers are using IGMP and MLD protocols for querying, reporting and leaving multicast groups on local LAN segments – they announce their willingness to send or receive multicast data. IPv6 MLD is descendent of IPv4 IGMP, but both protocols are identical in structure and message semantic.
- DVMRP (Distance Vector Multicast Routing Protocol) [4], MOSPF (Multicast Open Shortest-Path First) [5], PIM (Protocol Independent Multicast) – All of them are examples of multicast routing protocols that build multicast topology in router control plane to distribute multicast data among networks. DVMRP and MOSPF are closely tight to the particular unicast routing protocol (RIP, OSPF), whereas variants of Protocol Independent Multicast (PIM) are independent by design and they are using information inside unicast routing table more generally.

End-hosts and routers maintain multicast connectivity with the help of previously mentioned protocols. Nowadays, computer network design suggests deployment of IGMP/MLD on the access layer and PIM on the distribution layer of hierarchical internetworking model.

The project ANSA (Automated Network Simulation and Analysis) running at the Faculty of Information Technology is dedicated to develop the variety of software tools that can create simulation models based on real networks and subsequently allow for formal analysis and verification of target network configurations. One of our future goals is to model multicast flows in the Brno University of Technology network and thus implementing PIM models is one of our milestones. This report outlines two simulation modules (first one revisited and second one new), which are part of the ANSA project and which are extending functionality of the INET framework in OMNeT++.

This paper has the following structure. The next section covers a quick overview of existing OMNeT++ simulation modules relevant to the topic of this paper. Section 3 treats about design of the relevannt PIM models. Section 4 presents validation scenarios for our implementations. The paper is summarized in Section 5 together with unveiling our future plans.

II. STATE OF THE ART

The current status of multicast support in OMNeT++ 4.3.1 and INET 2.2 framework is according to our knowledge as follows. We have merged functionality of generic IPv4 `Router` and IPv6 `Router6` nodes, so that we created the dual-stack capable router – **ANSARouter**.

The module `RoutingTable` has been updated and since INET 2.0 it supports multicast routes and appropriate functions enabling to find the best matching record for the target multicast group.

The basic goal behind our research is to support native multicast transfers together with dynamic multicast routing using PIM. Hence, we have decided to start with IPv4 multicast and to add missing functionality in form of simulation modules directly connected to `networkLayer` module, as depicted on Figure 1.

We have searched in scientific community around simulation and modeling for other PIM implementations prior to our work. Limited versions exist for NS-2 [6] or OPNET [7]. However, none of them provide robust implementation. Also, existing OMNeT++ multicast attempts proved to be depreciated [8].

 ISBN: 978-1-61208-330-8

OMNeT++ state of the art prior to this paper is the result of our ongoing research covered in our other articles about IGMPv2 and initial PIM Dense Mode multicast support [9].

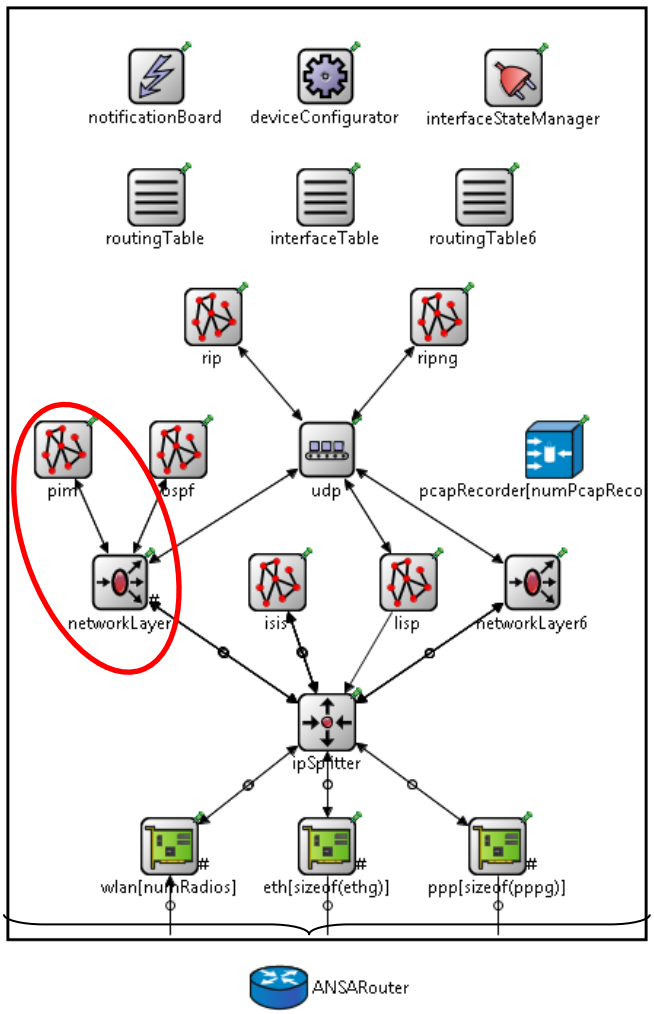


Figure 1. ANSARouter structure with highlighted contribution

III. IMPLEMENTATION

A. Theory of Operation

All multicast routing protocols provide a function to answer the question, "How to create routing path between sender(s) and receivers?" Baselines for this functionality are distribution trees of the following two types:

Source trees – The separate shortest path tree is built for each source of multicast data. A sender is the root and receivers are the leaves. However, memory and computation overhead causes this type is not scalable in the case of a network with many sources of multicast. In these situations usually the Shared tree is used.

Shared trees – A router called Rendezvous Point (RP) exist in a topology that serves as a meeting point for the traffic from multiple sources to reach destinations. The shared tree interconnects RP with all related receivers.

There are four PIM operational modes: PIM Dense Mode (PIM-DM), PIM Sparse Mode (PIM-SM), Bidirectional PIM (BiDir-PIM) and PIM Source-Specific Multicast (PIM-SSM). All of them differ in signalization, employed distribution trees and suitable applications.

Multicast routing support is performed by one dedicated router on each LAN segment elected based on *PIM Hello* messages. This router is called **designated router (DR)** and it is the one with highest priority or highest IP address.

PIM-DM is recommended for topologies with only one multicast source and lots of receivers. PIM-DM can be easily deployed without burdening configuration on active devices. However, PIM-DM does not scale well when number of sources increases. For this situation or for topologies with sparsely connected receivers, PIM-SM is suggested to be employed. Sparse mode scales much better in large topologies comparing to Dense mode, but configuration and administration is more complicated. PIM-SSM suits for multicast groups containing multiple sources providing the same content where client using IGMPv3 or MLDv2 may specify from which particular source it wants to receive data. BiDir-PIM is intended for topologies where many-to-many communication occurs. Currently, PIM-DM and PIM-SM are widely deployed PIM variants. Hence, we decided to implement them as the first.

PIM-DM idea consists of initial data delivery to all multicast-enable destinations (to flood multicast traffic everywhere), where routers prune themselves explicitly from the distribution tree if they are not a part of the multicast group. PIM-DM is not taking advantage of RP; thus, it is using source trees only.

PIM-DM routers exchange following messages during operation:

- *PIM Hello* – Used for neighbor detection and forming adjacencies. It contains all settings of shared parameters used for DR election;
- *PIM Prune/Join* – Sent towards upstream router by downstream device to either explicitly prune a source tree, or to announce willingness to receive multicast data by another downstream device in case of previously solicitated *PIM Prune*;
- *PIM Graft* – Sent from a downstream to an upstream router to join previously pruned distribution tree;
- *PIM Graft-Ack* – Sent from an upstream to a downstream router to acknowledge *PIM Graft*;
- *PIM State Refresh* – Pruned router refreshes prune state upon receiving this message;
- *PIM Assert* – In case of multi-access segment with multiple multicast-enabled routers one of them must be elected as an authoritative spokesman. Mutual exchange of *PIM Asserts* accomplishes this operation.

On the contrary to PIM-DM, **PIM-SM** works with different principle where initially no device wants to receive multicast. Thus, all receivers must explicitly ask for multicast delivery and then routers forward multicast data towards end-hosts. PIM-SM employs both types of multicast

 ISBN: 978-1-61208-330-8

distribution trees. Sources of multicast are connected with RP by source trees – source of multicast is the root of a source tree. RP is connected with multicast receivers by shared trees – RP is the root of shared tree. Multicast data are traversing from sources down by source tree to RP and from here down by shared tree to receivers. PIM-SM cannot work properly as long as all PIM routers in a network do not know exactly which router is RP for a given multicast group.

PIM-SM exchanges subsequent message types:

- *PIM Hello* – same as PIM-DM;
- *PIM Register* – Sent by source's DR towards RP whenever new source of multicast is detected;
- *PIM Register-Stop* – Solicited confirmation of *PIM Register*. It is sent by RP in reverse direction that source's DR can stop registering process of a new source. RP is aware of multicast data and may send them to receivers via shared tree;
- *PIM Prune/Join* – This message forms the shape of source and shared distribution trees. Multiple sources could provide data to the same multicast group – each one of them sends data via own source tree towards RP, from here data are reflected to receivers via shared tree;
- *PIM Assert* – same as PIM-DM.

The thorough survey on PIM-DM and PIM-SM message exchange scenarios are out of scope of this paper. More can be found in RFC 3973 [10] and RFC 4601 [11]; let us state that our implementations (i.e., finite-state machines, message structure, etc.) fully comply with IETF's standards.

B. Design

We have synthetized multiple finite-state machines that describe behavior of PIM-DM and PIM-SM with reference to used timers and exchanged PIM messages [12].

Figure 2 shows implemented architecture of the `pim` module:

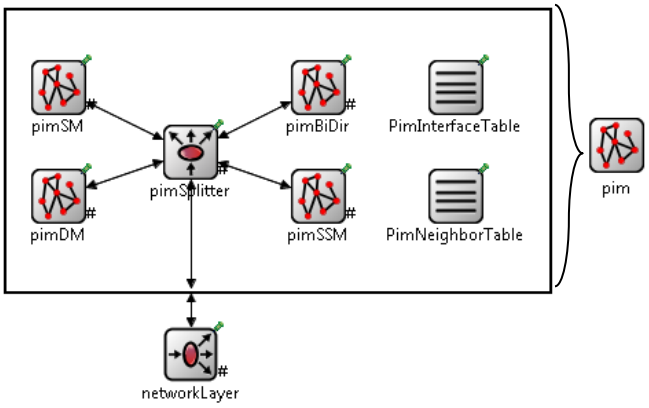


Figure 2. Proposed PIM module design

Besides previous modules, there were also some minor alternations to IPv4 `networkLayer` as well as to IPv4 `routingTable` module.

Implementation is done in NED (model design) and C++ (model behavior) languages. Brief description of implemented components is summarized in following table:

TABLE 1. DESCRIPTION OF PIM SUBMODULES

Name	Description
`pimSplitter`	This submodule is connected with INET `networkLayer`. It inspects all PIM messages and passes them to appropriate PIM submodules.
`pimDM`	The main implementation and logic of PIM-DM protocol is over here.
`pimSM`	The main implementation and logic of PIM-SM protocol is over here.
`pim InterfaceTable`	Stores all PIM relevant information for each router's interface.
`pim NeighborTable`	Keeps state of formed PIM adjacencies and information about neighbors (PIM version they are using, priorities, neighbors IPs).
`pimSSM`, `pimBiDir`	Prepared as a placeholder for upcoming implementations of BiDir-PIM and PIM-SSM variants.

IV. TESTING

In this section, we provide information on testing and validation of our implementations using several test scenarios. We compared the results with the behavior of referential implementation running at Cisco routers. We have built exactly the same topology and observed (using transparent switchport analyzers and Wireshark) relevant messages exchange between real devices (Cisco 2811 routers with IOS c2800nm-advipservicesk9-mz.124-25f.bin and host stations with FreeBSD 8.2 OS).

A. PIM-DM

In this testing network (topology is shown on Figure 3), we have three routers (R1, R2 and R3), two sources of multicast (Source1 and Source2) and three receivers (Host1, Host2 and Host3).

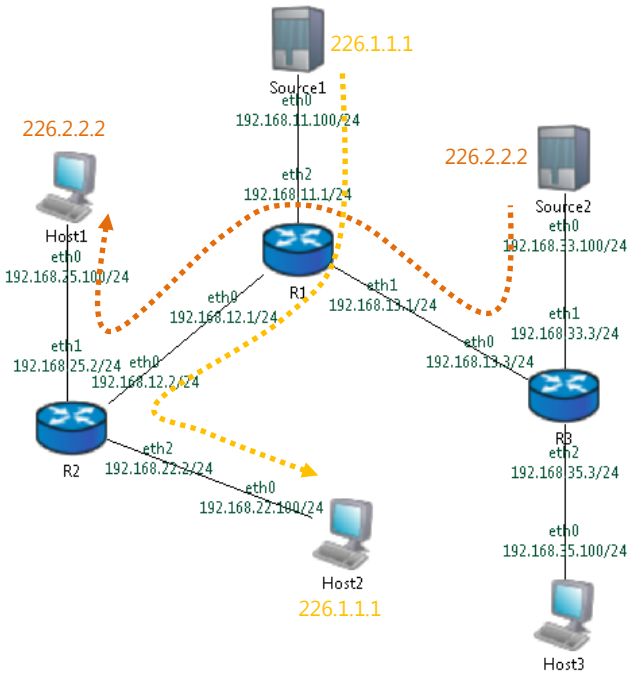


Figure 3. PIM-DM testing topology

We scheduled actions like source's start and stop sending and host's start and stop receiving of multicast data. Scheduled scenario is summarized in Table 2.

TABLE 2. PIM-DM EVENTS SCENARIO

Phase	Time [s]	Device	Multicast action	Group
#1	0	Host1	Starts receiving	226.2.2.2
#2	87	Source1	Starts sending	226.1.1.1
#3	144	Host2	Starts receiving	226.1.1.1
#4	215	Source2	Starts sending	226.2.2.2
#5	364	Host2	Stops receiving	226.1.1.1
#6	399	Source2	Stops sending	226.2.2.2

Hosts sign themselves to receive data from particular multicast group via *IGMP Membership Report* message during phases #1 and #3. Similarly, the host uses *IGMP Leave Group* message to stop receiving data during phases #5 and #6.

#1) There are no multicast data transferred. Only *PIM Hellos* are sent between neighbors.

#2) First multicast data appear but, because of no receivers, routers prune themselves from source distribution tree after initial flooding.

#3) Host2 starts to receive data from group 226.1.1.1 at the beginning of #3. This means that R2 reconnects to source tree with help of *PIM Graft* which is subsequently acknowledged by *PIM Graft-Ack*.

#4) The new source starts to send multicast data. All routers are part of the source distribution tree with R3 as the root. R3 acts as RP that is illustrated on Figure 4.

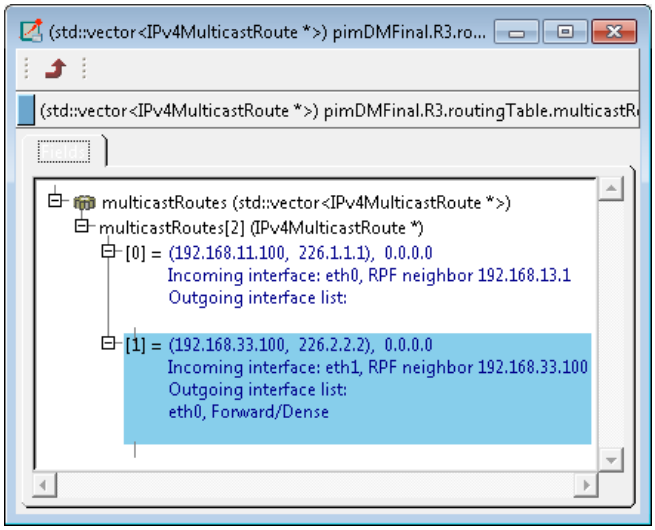


Figure 4. R3 multicast routing table after phase #4

#5) Host2 is no longer willing to receive multicast from 226.1.1.1 and, because Host2 is also the only listener to this group, R2 disconnects itself from distribution tree with *PIM Prune/Join*.

#6) Finally Source2 stops sending data to the group 226.2.2.2 at the beginning of #6. Subsequent to this, no PIM message is generated. Routers just wait for 180 seconds and then wipe out an affected source tree from the multicast routing table.

The confluence of messages proved correctness of our PIM-DM implementation by simulation as well as by real network monitoring, which can be observed in Table 3.

TABLE 3. TIMESTAMP COMPARISON OF PIM-DM MESSAGES

Phase	Message	Sender	Simul. [s]	Real [s]
#1	*PIM Hello*	R1	30.435	25.461
#2	*PIM Prune/Join*	R3	87.000	87.664
#3	*PIM Graft*	R2	144.000	144.406
	PIM Graft-Ack	R1	144.000	144.440
#5	*PIM Prune/Join*	R2	366.000	364.496

A. PIM-SM

For testing purposes of PIM-SM, topology is more complex. We have two designated routers (DR_R1, DR_R2) for receivers (Receiver1, Receiver2), two DRs (DR_S1, DR_S2) for sources (Source1, Source2) and one rendezvous point (RP). The scenario is depicted on Figure 5.

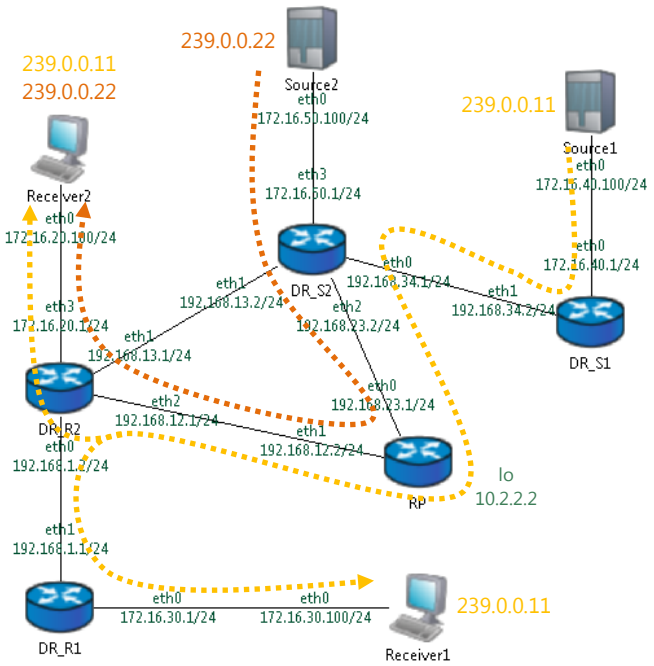


Figure 5. PIM-SM testing topology

A scenario for PIM-SM is summarized in Table 4 and additional description of actions follows bellow:

TABLE 4. PIM-SM EVENTS SCENARIO

Phase	Time [s]	Device	Multicast action	Group
#1	10	Source1	Starts sending	239.0.0.11
#2	20	Receiver1	Starts receiving	239.0.0.11
#3	25	Receiver2	Starts receiving	239.0.0.11
#4	40	Receiver2	Starts receiving	239.0.0.22
#5	60	Source2	Starts sending	239.0.0.22
#6	90	Receiver1	Stops receiving	239.0.0.11
#7	120	Receiver2	Stops receiving	239.0.0.11
#8	220	Receiver2	Stops receiving	239.0.0.22
#9	310	Source1	Stops sending	239.0.0.11
#10	360	Source2	Stops sending	239.0.0.22

 ISBN: 978-1-61208-330-8

Just as in PIM-DM scenario, receivers send *IGMP Membership Report* and *IGMP Leave Group* messages to sign on and off the multicast groups during phases #2, #3 and #6-#8.

#1) Source1 starts to send multicast data. Those data are encapsulated into *PIM Register* message sent by DR_S1 via DR_S2 towards RP. Following next RP responds with *PIM Register-Stop* back to DR_S1, thus registration of new source is finished.

#2) *IGMP Membership Report* for multicast group 239.0.0.11 by Receiver1 turns on joining process of DR_R1 and DR_R2 to shared tree and joining of RP and DR_S2 to source tree by sending *PIM Join/Prune*.

#3) DR_R2 is already connected to a shared tree, thus *IGMP Membership Report* only adds another outgoing interface to shared tree as could be seen on Figure 6.

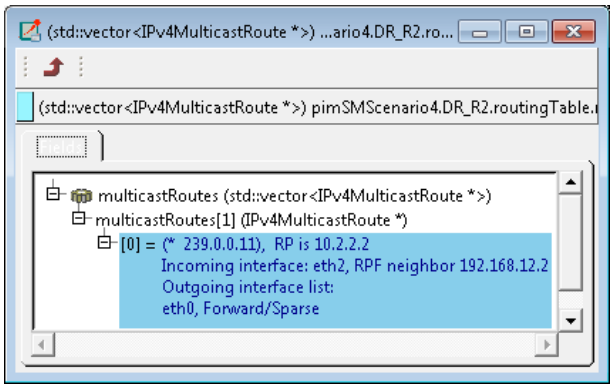


Figure 6. DR_R2 multicast routing table after phase #3

#4) Whenever Receiver2 starts receiving multicast group 239.0.0.22, new multicast route is added on DR_R2 (see Figure 7). Subsequently DR_S2 joins to shared tree via *PIM Join/Prune* sent towards RP.

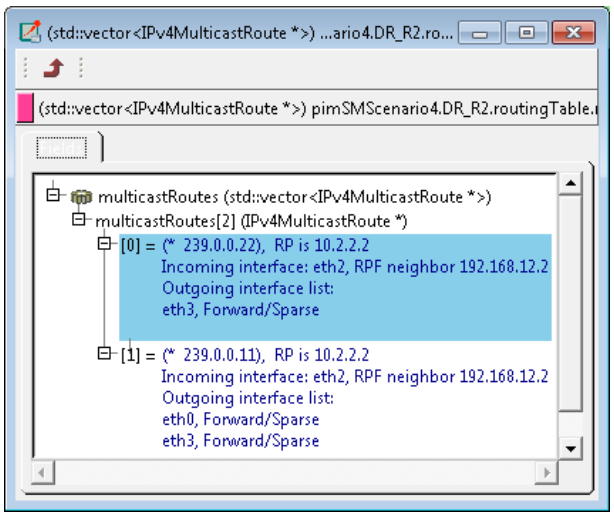


Figure 7. DR_R2 multicast routing table after phase #4

#5) Source2 starts sending multicast data to 239.0.0.22 after Receiver1 already joined the shared tree. DR_S2 registers source with *PIM Register* that contains also multicast data. Those data are decapsulated and sent down via shared tree to receivers. As a next step RP joins the source tree via *PIM Prune/Join* message and a moment later it confirms registration via *PIM Register-Stop* sent towards DR_S2. Multicast routes on RP converged and they could be observed on Figure 8.

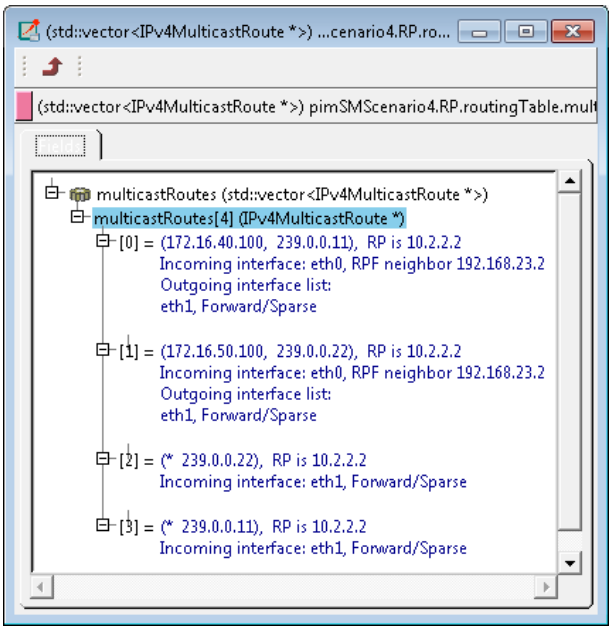


Figure 8. RP multicast routing table after phase #5

#X) Every 60 second after successful source registration, the given DR and RP exchange empty *PIM Register* and *PIM Register-Stop* messages to confirm presence of multicast source. Also every 60 seconds after last receiver joined multicast group, PIM router refreshes upstream connectivity to any tree via *PIM Prune/Join* message. This phase cannot be planned or scheduled; it is default behavior of PIM-SM protocol finite-state machine. It is illustrated only once for Source1 distribution trees but the same message exchange happens also for Source2.

#6) Upon receiving *IGMP Leave Group*, DR_R1 prunes itself from shared tree via *PIM Prune/Join* message sent upstream to DR_R1. DR_R1 then removes interface eth0 as outgoing interface for multicast group 239.0.0.11.

#7) Receiver2 decides not to receive multicast from Source1. Its *IGMP Leave Group* starts pruning process that goes from DR_R2 up to DR_S1. On each interim PIM router, multicast route for 239.0.0.11 is removed via *PIM Prune/Join* message.

#8) Later Receiver2 signs off from receiving 239.0.0.22 which causes similar exchange of *PIM Prune/Join* as in case of #7.

 ISBN: 978-1-61208-330-8

#9) Whenever Source1 stops sending multicast, elimination process starts for a given multicast route. As time goes by, ExpireTimer times out on every PIM router and multicast distribution tree for 239.0.0.11 is wiped out from routing table. The same approach applies for #10.

Validation testing against real-life topology shows just reasonable time variations (around ±3 seconds). This variation observable on real Cisco devices is caused by two factors: a) control-plane processing delay; b) stochastic message jitter to avoid potential race conditions in similar processes. Table 5 outlines results:

TABLE 5. TIMESTAMP COMPARISON OF PIM-SM MESSAGES

Phase	Message	Sender	Simul. [s]	Real [s]
#1	*PIM Register*	DR_R1	10.005	10.127
	PIM Register-Stop	RP	10.006	10.380
#2	*PIM Prune/Join*	DR_R1	20.001	20.422
	PIM Prune/Join	DR_R2	20.002	20.813
	PIM Prune/Join	RP	20.003	21.117
	PIM Prune/Join	DR_S2	20.005	21.320
#4	*PIM Prune/Join*	DR_R2	40.001	43.524
#5	*PIM Register*	DR_S2	60.000	61.459
	PIM Prune/Join	RP	60.003	61.970
	PIM Register-Stop	RP	60.004	62.758
#X	*PIM Register*	DR_S1	70.008	74.304
	PIM Register-Stop	RP	70.009	75.671
	PIM Prune/Join	DR_R1	80.000	83.041
	PIM Prune/Join	DR_R2	80.001	83.647
	PIM Prune/Join	RP	80.003	83.950
	PIM Prune/Join	DR_S2	80.003	84.004
#6	*PIM Prune/Join*	DR_R1	90.000	92.909
#7	*PIM Prune/Join*	DR_R2	120.001	122.311
	PIM Prune/Join	RP	120.002	122.704
	PIM Prune/Join	DR_S2	120.003	123.296

V. CONCLUSION AND FUTURE WORK

In this paper, we discussed options for dynamic multicast routing. We presented an overview of currently existing modules relevant to above topics in OMNeT++. The main contributions are simulation models for PIM-DM and PIM-SM that extend functionality of our ANSARouter and overall INET framework. Also, we introduce simulation scenarios and their results, which show that our implementations comply with relevant RFCs and referential behavior on Cisco devices.

We plan to wrap up our native IPv4 multicast implementation and complete it with integrating MLD support. After finishing this, we would like to focus on IPv6 and Source-Specific Multicast simulations.

More information about ANSA project is available on webpage [13]. Source codes of simulation modules could be downloaded via GitHub repository [14].

ACKNOWLEDGMENT

This work was supported by the Brno University of Technology organization and by the research grant IT4Innovation ED1.1.00/02.0070 by Czech Ministry of Education Youth and Sports.

REFERENCES

[1] G. Phillips et al., "Law, Scaling of Multicast Trees: Comments on the Chuang-Sirbu Scaling", in Proceedings of ACM SIGCOMM ’99, Cambridge-Boston, 1999.

[2] B. Cain et al., "Internet Group Management Protocol, Version 3", October 2002. [Online]. Available: https://tools.ietf.org/html/rfc3376. [Accessed: 2013].

[3] R. Vida and K. Costa, "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", June 2004. [Online]. Available: http://tools.ietf.org/html/rfc3810. [Accessed: 2013].

[4] D. Waitzman, C. Partridge and S. Deering, "Distance Vector Multicast Routing Protocol", November 1988. [Online]. [Accessed: http://tools.ietf.org/html/rfc1075 2013].

[5] J. Moy, "Multicast Extensions to OSPF", March 1994. [Online]. Available: http://tools.ietf.org/html/rfc1584. [Accessed: 2013].

[6] T. Henderson, November 2011. [Online]. Available: http://www.isi.edu/nsnam/ns/doc/node338.html. [Accessed: 2013].

[7] C. Adam, Y. Chien, May 1998. [Online]. Available: http://www.cs.columbia.edu/~hgs/teaching/ais/1998/projects/pim/report.html. [Accessed: 2013].

[8] R. Leal, J. Cacinero and E. Martin, "New Approach to Inter-domain Multicast Protocols", ETRI Journal, vol. 33, issue 3, pp. 355-365, June 2011.

[9] V. Veselý, O. Ryšavý and M. Švéda, "IPv6 Unicast and IPv4 Multicast Routing in OMNeT++", in SimuTools '13 Proceedings of the 6th International ICST Conference on Simulation Tools and Techniques, Cannes, France, 2013.

[10] A. Adams, J. Nichols and W. Siadak, "Protocol Independent Multicast - Dense Mode (PIM-DM)", January 2005. [Online]. Available: http://tools.ietf.org/html/rfc3973. [Accessed: 2013]

[11] B. Fenner et al., "Protocol Independent Multicast - Sparse Mode (PIM-SM)", August 2006. [Online]. Available: http://tools.ietf.org/html/rfc4601. [Accessed: 2013]

[12] V. Rybová and T. Procházka, Brno University of Technology, October 2013. [Online]. Available: https://nes.fit.vutbr.cz/ansa/uploads/Main/pim-fsm.pdf. [Accessed: 2014].

[13] Brno University of Technology, January 2014. [Online]. Available: http://nes.fit.vutbr.cz/ansa. [Accessed: 2014].

[14] GitHub, December 2013. [Online]. Available: https://github.com/kvetak/ANSA. [Accessed: 2014].

 ISBN: 978-1-61208-330-8