ICN 2012

The Eleventh International Conference on Networks

ISBN: 978-1-61208-183-0

February 29 - March 5, 2012

Saint Gilles, Reunion Island

ICN 2012 Editors

Pascal Lorenz, University of Haute Alsace, France

Tibor Gyires, Illinois State University, USA

Iwona Pozniak-Koszalka, Wroclaw University of Technology, Poland

ICN 2012

Foreword

The Eleventh International Conference on Networks [ICN 2012], held between February 29th and March 5th, 2012 in Saint Gilles, Reunion Island, continued a series of events focusing on the advances in the field of networks.

ICN 2012 welcomed technical papers presenting research and practical results, position papers addressing the pros and cons of specific proposals, such as those being discussed in the standard fora or in industry consortia, survey papers addressing the key problems and solutions, short papers on work in progress, and panel proposals.

We take here the opportunity to warmly thank all the members of the ICN 2012 Technical Program Committee, as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to ICN 2012. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the ICN 2012 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that ICN 2012 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of networks.

We are convinced that the participants found the event useful and communications very open. We also hope the attendees enjoyed the charm of Saint Gilles, Reunion Island.

ICN Chairs:

Tibor Gyires, Illinois State University, USA
Eva Hladká, Masaryk University - Brno / CESNET, Czech Republic
Pascal Lorenz, University of Haute Alsace, France
Iwona Pozniak-Koszalka, Wroclaw University of Technology, Poland

ICN 2012

Committee

ICN General Chair

Pascal Lorenz, University of Haute Alsace, France

ICN Advisory Chairs

Tibor Gyires, Illinois State University, USA
Eva Hladká, Masaryk University - Brno / CESNET, Czech Republic
Iwona Pozniak-Koszalka, Wroclaw University of Technology, Poland

ICN 2012 Technical Program Committee

Jalel Ben-Othman, Université de Versailles, France
João Afonso, FCCN - Fundação para a Computação Científica Nacional - Lisboa, Portugal
Max Agueh, LACSC - ECE Paris, France
Kari Aho, University of Jyväskylä, Finland
Pascal Anelli, Université de la Réunion, France
Cristian Anghel, Politehnica University of Bucharest, Romania
Tarun Bansal, The Ohio State University - Columbus, USA
Alvaro Barradas, University of Algarve, Portugal
Carlos Becker Westphall, Federal University of Santa Catarina, Brazil
Zdenek Becvar, Czech Technical University in Prague, Czech Republic
Djamel Benferhat, University of South Brittany, France
Ilham Benyahia, Université du Québec en Outaouais - Gatineau, Canada
Robert Bestak, Czech Technical University in Prague, Czech Republic
Jun Bi, Tsinghua University, China
Fernando Boronat Seguí, Universidad Politécnica de Valencia, Spain
Detlef Bosau, Scientist-Stuttgart, Germany
Matthias R. Brust, Technological Institute of Aeronautics, Brazil
Jorge Luis Castro e Silva, UECE - Universidade Estadual do Ceará, Brazil
Joaquim Celestino Júnior, Universidade Estadual do Ceará (UECE), Brazil
Eduardo Cerqueira, Federal University of Para, Brazil
Marc Cheboldaeff, Alcatel-Lucent AG, Germany
Kwangjong Cho (조광종), Korea Institute of Science and Technology Information (KISTI)- Daejeon, Republic of Korea
Andrzej Chydzinski, Silesian University of Technology - Gliwice, Poland
Javier Del Ser Lorente, TECNALIA-TELECOM, Spain
Daniela Dragomirescu, LAAS/CNRS, Toulouse, France
Matthew Dunlop, Virginia Tech, USA
Inès El Korbi, High Institute of Computer Science and Management of Kairouan, Tunisia
Emad Abd Elrahman, TELECOM & Management SudParis - Evry, France

Jose Oscar Fajardo, University of the Basque Country, Spain
Weiwei Fang, Beijing Jiaotong University, China
Khalid Farhan, Al-Zaytoonah University of Jordan, Jordan
Mário F. S. Ferreira, University of Aveiro, Portugal
Mário Freire, University of Beira Interior, Portugal
Wolfgang Fritz, Leibniz Supercomputing Centre - Garching b. München, Germany
Holger Fröning, University of Heidelberg, Germany
Laurent George, University of Paris-Est Creteil Val de Marne, France
Eva Gescheidtova, Brno University of Technology, Czech Republic
S.P. Ghrera, Jaypee University of Information Technology - Waknaghat, India
Markus Goldstein, German Research Center for Artificial Intelligence (DFKI), Germany
Anahita Gouya, AFD Technologies, France
Vic Grout, Glyndwr University - Wrexham, UK
Mina S. Guirguis, Texas State University - San Marcos, USA
Huaqun Guo, Institute for Infocomm Research, A*STAR, Singapore
Tibor Gyires, Illinois State University, USA
Keijo Haataja, University of Eastern Finland- Kuopio / Unicta Oy, Finland
Jiri Hajek, FEE-CTU - Prague, Czeck Republic
Mohammad Hammoudeh, Manchester Metropolitan University, UK
Hiroyuki Hatano, Shizuoka University, Japan
Haiwu He, INRIA, France
Eva Hladká, Masaryk University - Brno / CESNET, Czech Republic
Osamu Honda, Onomichi University, Japan
Florian Huc, EPFL - Lausanne, Switzerland
Jin-Ok Hwang. Korea University - Seoul, Korea
Muhammad Ali Imran, University of Surrey - Guildford, UK
Borka Jerman-Blažič, Jozef Stefan Institute, Slovenia
Börje Josefsson, SUNET, Sweden
Aravind Kailas, UNC - Charlotte, USA
Omid Kashefi, Iran University of Science and Technology-Tehran Iran
Andrzej Kasprzak, Wroclaw University of Technology, Poland
Abdelmajid Khelil, TU Darmstadt, Germany
Sun-il Kim, University of Alaska Anchorage, USA
Wojciech Kmiecik Wroclaw University of Technology, Poland
Christian Köbel, Technische Hochschule Mittelhessen - Raum, Germany
Leszek Koszalka, Wroclaw University of Technology, Poland
Tomas Koutny, University of West Bohemi-Pilsen, Czech Republic
Polychronis Koutsakis, Technical University of Crete, Greece
Francine Krief, University of Bordeaux, France
Michał Kucharzak, Wroclaw University of Technology, Poland
Radek Kuchta, Brno University of Technology, Czech Republic
Hadi Larijani, Glasgow Caledonian University, UK
Angelos Lazaris, University of Southern California, USA
Steven S. W. Lee, National Chung Cheng University, Taiwan R.O.C.
Jun Li, Qinghua University, China
Yan Li, Conviva, Inc. - San Mateo, USA
Diogo Lobato Acatauassú Nunes, Federal University of Para - Belem, Brazil
Andreas Löffler, Friedrich-Alexander-University of Erlangen-Nuremberg, Germany

Pascal Lorenz, University of Haute Alsace, France
Richard Lorion Université de la Réunion, France
Pavel Mach, Czech Technical University in Prague, Czech Republic
Damien Magoni, University of Bordeaux, France
Ahmed Mahdy, Texas A&M University - Corpus Christi, USA
Zoubir Mammeri, IRIT - Paul Sabatier University - Toulouse, France
Gustavo Marfia, University of Bologna, Italy
Rui Marinheiro, ISCTE - Lisbon University Institute, Portugal
Boris M. Miller, Monash University, Australia
Pascale Minet, INRIA - Rocquencourt, France
Jogesh Muppala , Hong Kong University of Science and Technology, Hong Kong
Katsuhiro Naito, Mie University - Tsu City, Japan
Frank Oldewurtel, Audi AG - Ingolstadt, Germany
Go-Hasegawa, Osaka University, Japan
Constantin Paleologu, University Politehnica of Bucharest, Romania
Konstantinos Patsakis, University of Piraeus, Greece
João Paulo Pereira, Polytechnic Institute of Bragança, Portugal
Kun Peng, Institute for Infocomm Research, Singapore
Yoann Pigné, University of Luxembourg, Luxembourg
Ionut Pirnog, "Politehnica" University of Bucharest, Romania
Marcial Porto Fernandez, Universidade Estadual do Ceara (UECE), Brazil
Iwona Pozniak-Koszalka, Wroclaw University of Technology, Poland
Jani Puttonen, Magister Solutions Ltd., Finland
Victor Ramos, UAM-Iztapalapa, Mexico
Priyanka Rawat, INRIA Lille - Nord Europe, France
Yenumula B. Reddy, Grambling State University, USA
Krisakorn Rerkrai, RWTH Aachen University, Germany
Karim Mohammed Rezaul, Glyndwr University - Wrexham, UK
Wouter Rogiest, Ghent University, Belgium
Simon Pietro Romano, University of Napoli Federico II, Italy
Teerapat Sanguankotchakorn, Asian Institute of Technology - Klong Luang, Thailand
Susana Sargento, University of Aveiro, Portugal
Panagiotis Sarigiannidis , University of Western Macedonia - Kozani, Greece
Masahiro Sasabe, Osaka University, Japan
Thomas C. Schmidt, HAW Hamburg, Germany
Hans Scholten, University of Twente- Enschede, The Netherlands
Dimitrios Serpanos, ISI/RC Athena & University of Patras, Greece
Pengbo Si, Beijing University of Technology, P.R. China
Rajesh Siddavatam, Jaypee University of Information Technology, India
Frank Siqueira, Federal University of Santa Catarina - Florianopolis, Brazil
Peter Skworcow, MontFort University - Leicester, UK
Karel Slavicek, Masaryk University Brno, Czech Republic
Adam Smutnicki, Wroclaw University of Technology, Poland
Arun Somani, Iowa State University - Ames, USA
Lars Strand, Norwegian Computing Center, Norway
Miroslav Sveda, Brno University of Technology, Czech Republic
Nabil Tabbane, SUPCOM, Tunisia
János Tapolcai, Budapest University of Technology and Economics, Hungary

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

Strict Priority Scheduler for Rate Controlled Transmission 1
Lukasz Chrost and Agnieszka Brachman

Network and Services Monitoring: A Survey in Cloud Computing Environments 7
Guilherme Da Cunha Rodrigues, Vinicius Tavares Guimaraes, Glederson Lessa dos Santos, Lisandro Zambenedetti Granville, and Liane Rockenbach Tarouco

The Space-Time Information in the Access Management 14
Jaroslav Kadlec, Radimir Vrba, David Jaros, and Radek Kuchta

A High-level Network-wide Router Configuration Language 18
Miroslav Sveda, Michal Sekletar, Tomas Fidler, and Ondrej Rysavy

A New Approach to NGN Evaluation Integrating Simulation and Testbed Methodology 22
Marcial Fernandez, Sebastian Wahle, and Thomas Magedanz

Impact of Propagation Factors on Routing Efficiency in Wireless Mesh Networks: A Simulation-based Study 28
Ewa Osekowska, Iwona Pozniak-Koszalka, and Andrzej Kasprzak

Comparison of Heuristic Methods Applied to Optimization of Computer Networks 34
Tomasz Miksa, Leszek Koszalka, and Andrzej Kasprzak

Wireless Home Automation Network Stability Testing 39
Radek Kuchta, Radovan Novotny, Jaroslav Kadlec, Radimir Vrba, and Vladimir Sulc

Band-Pass Filters for Direct Sampling Receivers 44
Pavel Zahradnik, Vlcek Miroslav, Simak Boris, and Kopp Michal

Comparison of the Fully-Differential and Single-Ended Solutions of the Frequency Filter with Current Followers and Adjustable Current Amplifier 50
Jan Jerabek and Kamil Vrba

Precision Full-wave Rectifier Using Current Conveyors and Two Diodes 55
Jaroslav Koton, Norbert Herencsar, and Kamil Vrba

Single GCFDITA and Grounded Passive Elements Based General Topology for Analog Signal Processing Applications 59
Norbert Herencsar, Jaroslav Koton, Kamil Vrba, and Abhirup Lahiri

How Many Cores Does Parallel BGP Need in a High-Speed Router 63
Yaping Liu, Shuo Zhang, Zexin Lu, and Baosheng Wang

Multi-level Machine Learning Traffic Classification System 69
Geza Szabo, Janos Szule, Zoltan Turanyi, and Gergely Pongracz

TCP SYN Protection: An Evaluation 78
Pascal Anelli, Fanilo Harivelo, and Richard Lorion

Extension of the Shared Regional PACS Center MeDiMed to Smaller Healthcare Institutions 83
Karel Slavicek, Michal Javornik, and Otto Dostal

Multi-Tenancy Authorization System with Federated Identity for Cloud-Based Environments Using Shibboleth 88
Marcos A. P. Leandro, Tiago J. Nascimento, Daniel R. dos Santos, Carla M. Westphall, and Carlos B. Westphall

A SPIT Avoidance Workflow for SIP-Provider 94
Nicolas Rueger, Sebastian Huebner, and Bettina Schnor

Using PGP Signatures for Securing SIP Infrastructures 102
Sebastian Huebner, Nicolas Rueger, and Bettina Schnor

Developing Trust and Reputation Taxonomy for a Dynamic Network Environment 109
Tanja Azderska and Borka Jerman Blazic

A Reference Model for Future Computer Networks 115
Hoda Hassan

Application Design over Named Data Networking with Its Features in Mind 121
Sen Wang, Jianping Wu, and Jun Bi

Feedback, Transport Layer Protocols and Buffer Sizing 125
Shankar Raman, Shashank Jain, and Gaurav Raina

A Generic Service Model for QoS Management 132
Tatiana Aubonnet and Noemie Simoni

Towards Opportunistic Data Dissemination in Mobile Phone Sensor Networks 139
Viet-Duc Le, Hans Scholten, and Paul Havinga

Towards A Theoretically Bounded Path Key Establishment Mechanism in Wireless Sensor Networks 147
Aishwarya Mishra, Tibor Gyires, and Yongning Tang

Selective Link Cost Alteration in Reservation-Based Multi-Hop Wireless Mesh Networks 153
Christian Kobel, Walter Baluja Garcia, and Joachim Habermann

Reliable Technology for Wireless Mesh Networks with Low System Requirements 159
Vladimir Sulc, Radek Kuchta, and Radimir Vrba

Forwarding and Routing Stateless Multi-Hop Protocol for Wireless Sensor Networks 165
Rivo S. A. Randriatsiferana, Richard Lorion, Frederic Alicalapa, and Fanilo Harivelo

Resource Management for Advanced, Heterogeneous Sensor-Actor-Networks 169
Matthias Vodel, Mirko Lippmann, and Wolfram Hardt

Improving Fairness in Wireless Mesh Networks 175
Jorge Peixoto, Marcial Fernandez, and Luis Felipe Moraes

An Integrated Location Method using Reference Landmarks for Dead Reckoning System 181
Mingmei Li, Kazuyuki Tasaka, and Kiyohito Yoshihara

MANET with the Q-Routing Protocol 187
Ramzi Haraty and Badieh Traboulsi

Estimation of Collision Multiplicities in IEEE 802.11-based WLANs 193
Benoit Escrig

CTC Turbo Decoding Architecture for LTE Systems Implemented on FPGA 199
Cristian Anghel, Valentin Stanciu, Cristian Stanciu, and Constantin Paleologu

Multi-Relay Cooperative NB-LDPC Coding with Non-Binary Repetition Codes 205
David Declercq, Valantin Savin, and Stephan Pfletschinger

A Redundancy Information Protocol for P2P Networks in Ubiquitous Computing Environments: Design and Implementation 215
Rafael Araujo, Hiran Ferreira, Pedro Rosa, and Renan Cattelan

Kernel Module Implementation of IPv4/IPv6 Translation Mechanisms for IPv4-oriented Applications 221
Katsuhiro Naito, Kazuo Mori, and Hideo Kobayashi

A Complementary Approach for Transparent NAT Connectivity 227
Lucas Vella, Lasaro Camargos, and Pedro Rosa

Minimization of Branching in the Optical Trees with Constraints on the Degree of Nodes 235
Massinissa Merabet, Sylvain Durand, and Miklos Molnar

Mitigating Spoofing Attacks in MPLS-VPNs using Label-hopping 241
Shankar Raman and Gaurav Raina

A Proposal of Dynamic RWA Using Ant Colony in Optical Burst Switched Networks 246
Erick Donato, Joaquim Celestino Junior, Antonio Vieira, and Ahmed Patel

Strict Priority Scheduler for Rate Controlled Transmission

Lukasz Chrost and Agnieszka Brachman
Silesian University of Technology
Gliwice, Poland
lukasz.chrost@polsl.pl, agnieszka.brachman@polsl.pl

***Abstract*—The paper proposes a modification to the strict priority scheduler, to guarantee the Quality of Service (QoS) of a network with variable and undetermined bandwidth capacity. The proposed modifications make it possible to accomplish QoS along the path. The presented solution allows providing the minimum guaranteed transmission rates for all active flows with the respect to their priorities and to provide the fair share of the additional bandwidth. The scheduler also rejects flows, for which the minimum rate requirements exceed the available bandwidth. Moreover, a simple algorithm for mapping the WiMAX traffic classes to the strict n-priority scheduler bands is described. This allows providing WiFi - WiMAX internetworking with the QoS support. The presented test results show that the proposed scheduler preserves the defined, minimum transmission rates and improves the performance of the delay and throughput.**

***Keywords*-packet scheduling; strict priority scheduler; Quality of Service; wireless network.**

I. INTRODUCTION

Nowadays, the real-time traffic occupies a significant percentage of the available bandwidth and Internet must evolve to support the new applications. For the newly developed applications and services such as VoD (Video on Demand), VoIP (Voice over IP), VTC (Video-Teleconferencing), interactive games, distributed virtual collaboration, remote classrooms, grid computing, etc., the best effort delivery is unacceptable, since in case of a congestion the Quality of Service (QoS) and Quality of Experience (QoE) declines to an unsatisfactory level. Therefore, the main and crucial objective of the future Internet is to change the best effort network into the Quality of Service controlled network.

Various applications may have different, sometimes stringent requirements in terms of throughput, packet losses and/or delays. It brings out the necessity to provide different priorities to different applications, users or data flows, or to guarantee a certain level of performance to a data flow; in short, to provide the Quality of Service (QoS). Some real time traffic application will not be commercially viable without the QoS guarantees. Enabling the differentiated resource allocation is also very important from the providers point of view. The predominant form of pricing currently in practice in the Internet is per achievable throughput. A fee is charged for the amount of bandwidth to access the network. Therefore, the ability to provide the exact, required part of the available bandwidth is crucial. Accomplishing this task may seem easy, however the problem arises in case of the unpredictable and variable environment.

Providing the minimum transmission rates is particularly difficult in the wireless networks. The varying conditions of the wireless channel lead to the unpredictable transmission channel parameters, i.e., the available bandwidth. The physical radio transmission is based on the emission of the electromagnetic waves. Radio waves decrease in the amplitude as they propagate and pass through the obstacles. In the urban environments the large throughput variations may arise even in a Line-Of-Sight (LOS) conditions. This happens especially due to the moving vehicles in the radio path as well as due to the multi-path effect.

WiFi [1] and WiMAX [2] are two common, low cost technologies for providing the ubiquitous wireless Internet access. WiFi provides high data rates up to 100 Mbps, within the short ranges, usually used within buildings. WiMAX is designed to offer throughput up to 70 Mbps, in 5 km range, used for covering the large outdoor locations. Integrating these two technologies is considered for the next generation network technology [3]. To provide the WiMAX-WiFi internetworking a new solution for the last WiFi hop is necessary.

WiFi is especially prone to the bandwidth degradation due to the varying conditions in the transmission channel, due to the modulation changes. The knowledge of the currently available bandwidth is crucial for providing the guaranteed rates and/or delays. The bandwidth estimation algorithms try to provide an accurate estimation of the available bandwidth. However, due to the high variability of the wireless channel throughput, most current techniques produce relatively inaccurate results and long convergence times.

The main contribution of this paper is the strict priority scheduler designed to provide the minimum guaranteed transmission rate for all active flows with the respect to their priorities and to provide fair share of the additional bandwidth. The scheduler also rejects flows, for which the minimum rate requirements exceed the available bandwidth. The proposed solution is applicable for the WiFi wireless network, to accomplish QoS along the path. It is simple to implement and does not require the bandwidth estimator. Additionally, we provide a simple algorithm for mapping the WiMAX traffic classes to the strict n-priority scheduler in order to provide the WiFi - WiMAX internetworking with

 ISBN: 978-1-61208-183-0

the QoS support.

The rest of the paper is organized as follows. Section II reviews the similar solutions and the priority schedulers designed for the real-time services and link-sharing service provisioning. In Section III, the proposed strict priority scheduler is presented and the proof-of-concept tests are depicted and explained. Section IV is devoted to the description of mapping of WiMAX classes to the strict n-priority scheduler band. Finally, the paper is concluded and the future work perspectives are presented in Section V.

II. Related work

Scheduling algorithm determines the allocation of the bandwidth among the users, flows or the service classes. Packet scheduling algorithms are widely discussed in the literature.

The QoS and packet scheduling are addressed by the DiffServ (Differentiated Services) architecture [4]. For the DiffServ several queuing and scheduling methods are associated, namely the priority scheduling and the Weighted Fair Queuing (WFQ).

The methods based on the priority scheduling are described in [5], [6]. Priority scheduling can reduce the packet, delay, jitter and loss for the high priority traffic. The Strict Priority (SP) scheduling is a simple and common solution. It provides the preferential treatment for the high priority classes, however at the cost of starving the lower priority traffic. The SP serves the high priority traffic queue, until it is empty and then moves to the lower priority queues. SP discipline itself is not controllable, therefore it cannot handle the starvation problem. Several modifications have been proposed to alleviate this problem. Authors in [7] propose to assign a parameter to each priority queue, which determines the extent, to which the priority queue is served. Their Probabilistic Priority (PP) discipline provides the minimum average throughput and the delay guarantees. However the algorithm does not assume that the resources may be scarce.

WFQ attempts to provide a share of bandwidth for each class or flow in proportion to their specified rates. WFQ and its variants are described in [8]–[11].

The Weighted Fair Queuing (WFQ) [9] is a packetized Generalized Processor Sharing (GPS) algorithm [9], [12], which works as follows. All traffic is classified into the so-called traffic classes i. Classes can be either individual flows or a bunch of flows with similar transmission requirements. Each class is assigned a positive weight ϕ_i, which specifies the minimum share of the available bandwidth C. This weight is also used for the distribution of the excess capacity, when a particular class does not fully use its bandwidth's share. Each backlogged traffic class, i.e., the class that has the packets waiting for the transmission in its queue or class, which packet is in service, receives the guaranteed service rate r_i:

$$r_i = \frac{\phi_i}{\sum \phi} C,$$

where $\sum \phi$ is the sum of the weights for all traffic classes.

GPS offers the protection among traffic classes, along with the full statistical multiplexing. GPS is an idealized scheduler, based on the assumption, that the capacity is infinitely divisible, which means that several packets can be served at the same time. Since in reality, the traffic is composed of the discrete packets sent in sequence, GPS cannot be implemented in a real system.

The packet scheduling is crucial for providing any QoS guarantees for the multiple service classes or priorities. Majority of proposed solutions requires information concerning the current bandwidth. The knowledge of bandwidth capacity is elementary. If it is not possible to guarantee the required, adequate performance, i.e., of a voice conversation, it is more beneficial to block the call rather than accept and experience excessive delays and packet drops. Therefore rejecting unfitting flows is a desired feature. To provide the efficient and stable transmission through the heterogeneous network, rate control, applied to all active flows, is necessary.

Providing Qos in WiMAX-WiFi integrated network is very challenging. WiMAX standard incorporates QoS features into the Media Access Control (MAC) layer. It implements the signalling and bandwidth allocation algorithms, thanks to which, the traffic with the different QoS requirements may be jointly regulated to make the best use of the available bandwidth. On the other hand, WiFi provides only the prioritized QoS introduced by the 802.11e enhancement [13], without any bandwidth-share guarantees.

The problem of the WiMAX and WiFi integration is discussed in [14]. Authors present an integrated Access Point, which combines the WiMAX subscriber station and WiFi Access Point. However the authors do not provide any scheduling strategy. In [15], an integration model based on the traffic mapping and signalling is presented. The authors describe the scheduling algorithm, which provides the QoS in terms of the delay bound for the real time traffic and the buffer bound for the non real time traffic. Nonetheless, they do not consider the bandwidth variations.

III. Strict priority scheduler with the minimum and maximum rates guarantees

The proposed strict priority-based scheduler with the minimum and maximum rates guarantees is designed to provide the following:

- Enforcement of the minimum guaranteed transmission rates for the existing flows, according to their priority.
- Rejection of the non-provisioned flows.
- Equal distribution of the additional bandwidth among active flows, up to their maximum transmission rates.
- Fast detection of the bandwidth capacity degradation.

 ISBN: 978-1-61208-183-0

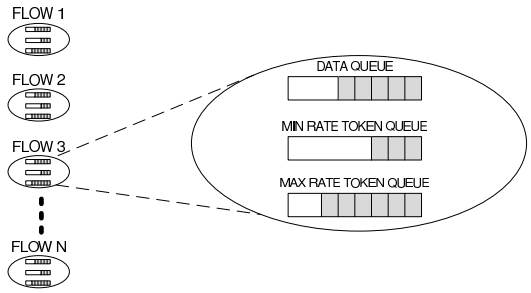


Figure 1. Strict priority based system

The aforementioned functionality is especially important for variable transmission channels, i.e., wireless networks. This method also allows passive estimation of the available bandwidth, however only if the link utilization is high.

The strict priority based system consists of the data queue and two virtual token bucket queues (the minimum rate token queue and the maximum rate token queue) for each transmission flow. The schematic representation of the algorithm is depicted in Figure 1. The virtual token queues are the standard TBF (Token Bucket Filter) queues. Tokens are added to the bucket every $1/rate_{min}$ and $1/rate_{max}$ seconds. If there is no $rate_{max}$ defined, the corresponding queue is always full. The parameters $rate_{min}$ and $rate_{max}$ are determined by the external system. All queues have size limits.

Every cycle requires up to three passes and ends if the packet is dequeued. The first two passes start from the queue with the highest priority. The first pass searches for the packets belonging to the flow with the non-empty minimum rate token bucket queue. The second (optional) pass looks for the packets belonging to the flows, with non-empty minimum rate token bucket queue, which were previously rejected (if the system supports re-enabling rejected transmissions). Third pass picks the packets belonging to the flows with the empty minimum rate token bucket queue and the non-empty maximum rate token bucket queue. This pass may be scheduled according to the round robin algorithm or any other algorithm, which may provide fair share of additional bandwidth regardless of the flows' priorities. In the proposed solution the modified Round Robin (RR) mechanism is used. The modified RR scheduler selects the packets in the third pass, however only from the queues, which were left nonempty in the earlier pass. RR does not provide any fairness, therefore if necessary, another algorithm may be applied, i.e., Deficit Round Robin (DRR).

When the packet of n bytes is dequeued, n tokens are removed from both token buckets - if the packet is selected in the first or in the second pass - or from the maximum rate token bucket - if packet is selected in the third pass. Subsequently, the packet is sent to the network.

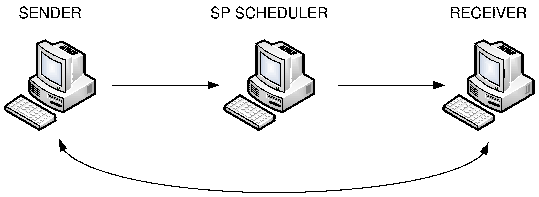


Figure 2. Scheduler testbed

The system is designed for handling flows with the unique priority. If there are two or more identical priorities, several approaches are possible, that is: random priorities, hash priorities or priorities set successively, according to the arrival time. The Type of Service (TOS) field in the IPv4 header may be used to identify and store the packets' priorities.

Implementation is based on the Linux kernel 2.6.39. The priority scheduler and the token bucket filter are implemented as the Traffic Control modules.

A. Test results

Proof-of-concept tests were run to verify the scheduler performance. The test environment is depicted in Figure 2.

SENDER is a station with a single Pentium III 800 MHz under Linux 2.6.39 with the modified traffic control module, described in the previous Section. 150 service flows with the unique priority were governed by SP SCHEDULER. Each queue has capacity for 10ms of traffic with regard to their minimum transmission rates. The overall bandwidth is ~ 90 Mbps. Traffic is generated using D-ITG on VMware to avoid throttles.

Figures 3 and 4 present the results for the test with traffic pattern composed of 150 UDP connections, sent with constant bit rate 1200 kbps each and the packet size set to 1400 B. All flows start at the same time. The minimum rate was set to 600 kbps (fig. 3) and 950 kbps (fig. 4). In the first case all flows fit to the overall bandwidth, in the second approximately the first 90 flows are able to achieve the minimum transmission rate.

When the minimum rates of all flows fit to the overall bandwidth, the minimum transmission rate is provisioned for all flows and the additional bandwidth is shared according to the modified RR algorithm. The average, experienced delay is similar for all provisioned flows.

Under the over-provisioned scenario around 60% of all flows achieves the desired transmission rate. The non-served flows experience large delays till they are blocked.

Figures 5 and 6 present results for a similar test but with the TCP transport protocol.

Since TCP implements its own rate control using the window-based mechanism, it adjusts the sending rate. Flows adjust the transmission rate to the offered bandwidth share.

 ISBN: 978-1-61208-183-0

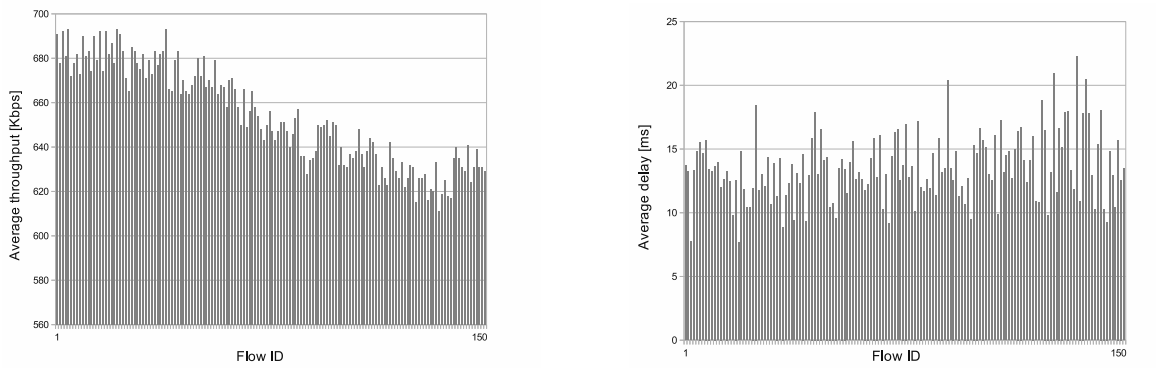


Figure 3. Average throughput and delay for 150 UDP connections, CBR = 1200 Kbps, packet size = 1400B, minimum rate = 600 Kbps

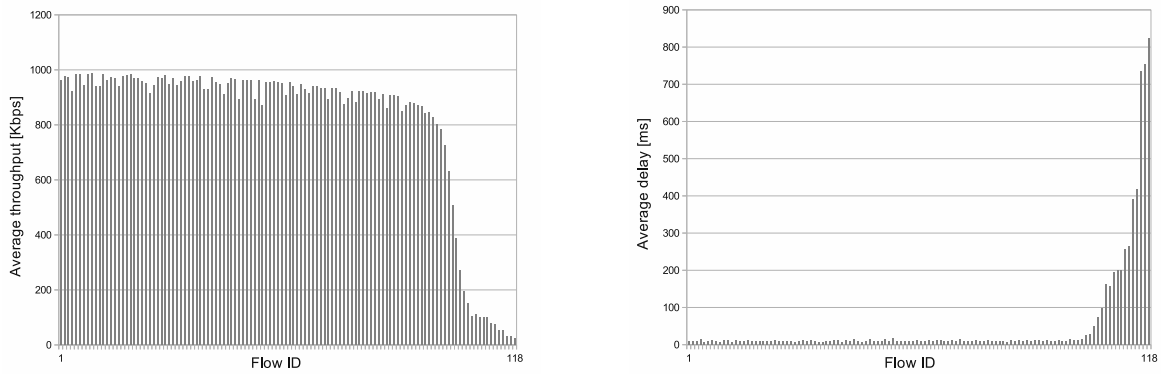


Figure 4. Average throughput and delay for 150 UDP connections, CBR = 1200 Kbps, packet size = 1400B, minimum rate = 950 Kbps

The flows with higher priority experience lower delays in both cases.

IV. Mapping of WiMAX classes to strict n-priority scheduler band

As shown in the previous section, the multi-pass strict priority scheduler can provide good bandwidth for multiple streams with low jitter. However, in case of the heterogeneous networks, a single, consistent QoS configuration system is required. WiMAX already contains an expanded QoS definition set describing different types of traffic classes. However, the classes are not usable neither in the standard WiFi network, nor, directly, in the strict-priority extended one. Fortunately, a simple class-to-priority mapping is possible.

The prosed algorithm provides a simple means of mapping the WiMAX traffic class to the strict n-priority scheduler band. To achieve that, we introduce a Φ vector describing scheduler bands. The $\Phi_i <> 0$ if the particular band has a WiMAX QoS SF mapped, and 0 otherwise. The actual mapping is done by an external, system-wide mechanism with the regard to several rules:

1) The 1's have to form continuous series inside Phi vector, with i denoting the first, and j denoting the last mapped position.
2) For n bands and m classes, $i = n/2 - m/2$, that is the '1' should occupy the central part of the Phi vector.
3) The incoming classes are mapped to the position k, where $(i-1) < k < (j+1)$.
4) If the SF leaves the system, its class should be unmapped.
5) In case of the mapping/unmapping, the vector shift may be required to fulfil 2. The chosen direction shall require the minimum number of bands to be remapped.

The actual mapping is based on WiMAX's QoS class hierarchy described in Table I of ϕ values. The decimal fraction is set according to the maximum delay parameter for a given service flow, while the whole part is defined by its class, i.e., UGS service flow with the maximum delay of 3ms has a ϕ value of 0.003.

The external mechanisms selects new k to form an increasing sequence of $\Phi_{i..j,k}$ SF vector mappings, where $\Phi_k = \phi$.

A schematic diagram depicting mapping process of in-

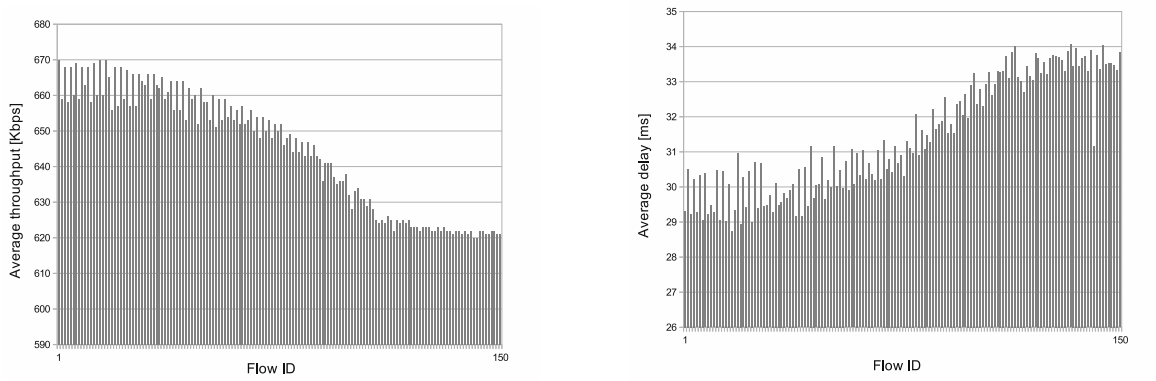


Figure 5. Average throughput and delay for 150 TCP connections, packet size = 1400B, minimum rate = 600 Kbps

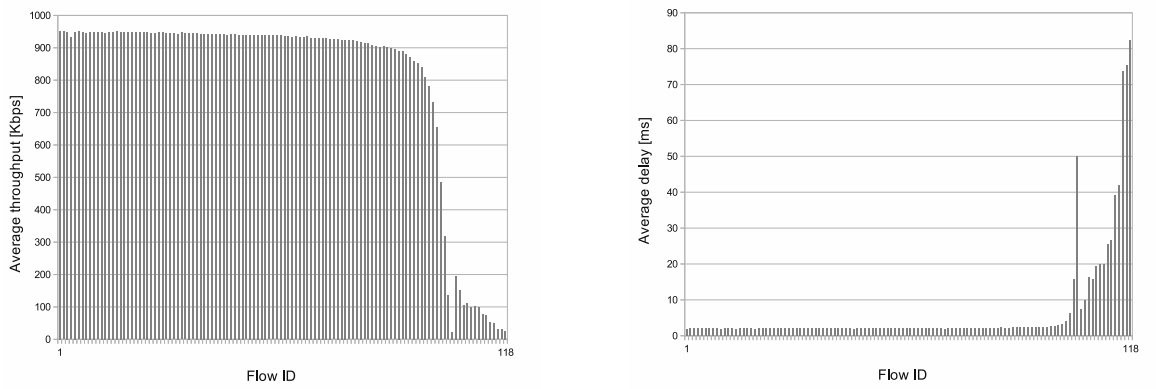


Figure 6. Average throughput and delay for 150 TCP connections, packet size = 1400B, minimum rate = 950 Kbps

Class	ϕ value
Unsolicited Grant Service (UGS)	0.0
Extended Real-Time Polling Service (ertPS)	0.0
Real-Time Polling Service (rtPS)	2.0
non-Real-Time Polling Service (nrtPS)	4.0
Best Effort	4.0

Table I
MAPPING OF THE WIMAX QOS TRANSPORT CLASSES TO ϕ PARAMETER VALUE. THE FRACTION PART IS DEFINED BY THE MAXIMUM DELAY PARAMETER DESCRIBING SPECIFIC SF.

coming SF has been presented in Figure 7. The mapping process requires a k value to be selected for the new SF.

V. CONCLUSION AND FUTURE WORK

The paper proposed a modification to the strict priority scheduler, which provides the minimum and maximum transmission rates for all active flows. Various tests were performed, which include the performance measurements for the UDP and TCP traffic in the provisioned and over-provisioned scenarios. The designed solution has been shown to distribute the available bandwidth according to the predefined requirements.

The main disadvantages of the proposed solutions are: packet-based operation and performance-related concerns - each cycle requires, at worst case, passing each queue three times. Another problem arises if the scheduler rejects the flows and there is no backward communication. In such a case, the rejected flows waste bandwidth up to the hop with the strict priority scheduler.

In further studies, we plan to enhance the scheduler to satisfy the delay requirements using adequate buffer sizing. The proposed scheduler needs also verification under more sophisticated scenarios including: varying bandwidth rate to imitate transmission in the wireless network, varying packet size and diversified minimum rate requirements.

ACKNOWLEDGMENT

The material is based upon work supported by the Polish National Science Centre under Grant No. N N516 479240.

 ISBN: 978-1-61208-183-0

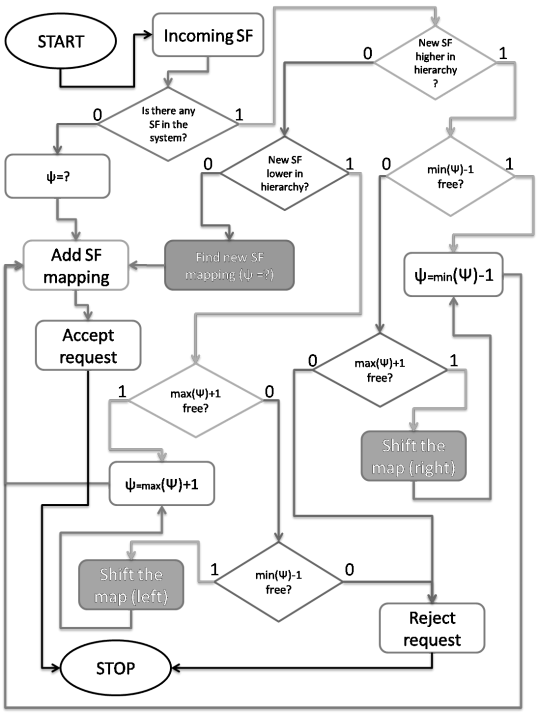


Figure 7. Diagram depicting the process of mapping incoming WiMAX service flow to multi-pass strict priority scheduler band.

REFERENCES

[1] "IEEE standard for information technology-telecommunications and information exchange between systems-local and metropolitan area networks-specific requirements - part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications," *IEEE Std 802.11-2007 (Revision of IEEE Std 802.11-1999)*, pp. C1–1184, December 2007.

[2] "IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems," *IEEE Std 802.16-2004 (Revision of IEEE Std 802.16-2001)*, pp. 0 –857, 2004.

[3] M. Jamil, S. Shaikh, M. Shahzad, and Q. Awais, "4G: The future mobile technology," in *TENCON 2008 - 2008 IEEE Region 10 Conference*, nov. 2008, pp. 1 –6.

[4] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Service," RFC 2475 (Informational), Internet Engineering Task Force, Dec. 1998, updated by RFC 3260.

[5] Z. Rosberg and F. Sabrina, "Rate control of multi class priority flows with end-to-end delay and rate constraints for QoS networks," *Computer Networks*, vol. 53, no. 16, pp. 2810–2824, Nov. 2009.

[6] A. E. Kamal and H. S. Hassanein, "Performance evaluation of prioritized scheduling with buffer management for differentiated services architectures," *Computer Networks*, vol. 46, no. 2, pp. 169–180, Oct. 2004.

[7] Y. Jiang, C. Tham, and C. Ko, "A probabilistic priority scheduling discipline for high speed networks," in *2001 IEEE Workshop on High Performance Switching and Routing*. IEEE, 2001, pp. 1–5.

[8] A. Demers, S. Keshav, and S. Shenker, "Analysis and simulation of a fair queueing algorithm," *SIGCOMM Comput. Commun. Rev.*, vol. 19, pp. 1–12, August 1989.

[9] A. Parekh and R. Gallager, "A generalized processor sharing approach to flow control in integrated services networks-the multiple node case," in *INFOCOM '93. Proceedings.Twelfth Annual Joint Conference of the IEEE Computer and Communications Societies. Networking: Foundation for the Future. IEEE*, 1993, pp. 521–530 vol.2.

[10] J. Bennett and H. Zhang, "WF2Q: worst-case fair weighted fair queueing," in *INFOCOM '96. Fifteenth Annual Joint Conference of the IEEE Computer Societies. Networking the Next Generation. Proceedings IEEE*, vol. 1, Mar. 1996, pp. 120–128 vol.1.

[11] J. F. Lee, M. C. Chen, and Y. Sun, "WF2Q-M: Worst-case fair weighted fair queueing with maximum rate control," *Computer Networks*, vol. 51, no. 6, pp. 1403–1420, 2007.

[12] A. Parekh and R. Gallager, "A generalized processor sharing approach to flow control in integrated services networks-the single node case," in *INFOCOM '92. Eleventh Annual Joint Conference of the IEEE Computer and Communications Societies, IEEE*, May 1992, pp. 915–924 vol.2.

[13] "IEEE Standard for Information Technology - Telecommunications and Information Exchange Between Systems - Local and Metropolitan Area Networks - Specific Requirements Part 11: Wireless LAN Medium Access control (MAC) and Physical Layer (PHY) Specifications Amendment 8: Medium Access Control (MAC) Quality of Service enhancements," *IEEE Std 802.11e-2005 (Amendment to IEEE Std 802.11, 1999 Edition (Reaff 2003)*, pp. 1 –189, 2005.

[14] S. Frattasi, E. Cianca, and R. Prasad, "An Integrated AP for Seamless Interworking of Existing WMAN and WLAN Standards," *Wirel. Pers. Commun.*, vol. 36, pp. 445–459, March 2006.

[15] G. Prasath, K. Raghu, and M. Ma, "Integration of WLAN and WiMAX with base station assisted QoS," in *Wireless and Optical Communications Networks, 2008. WOCN '08. 5th IFIP International Conference on*, may 2008, pp. 1 –5.

 ISBN: 978-1-61208-183-0

Network and Services Monitoring: A Survey in Cloud Computing Environments

Guilherme da Cunha Rodrigues,
Vinicius Tavares Guimaraes,
Glederson Lessa dos Santos
Federal Institute of Education,Science and
Technology Sul Rio-Grandense
Charqueadas, Brazil
Emails: {grodrigues,vicoguim,gledersons}@charqueadas.ifsul.edu.br

Lisandro Zambenedetti Granville,
Liane M. Rockenbach Tarouco
Federal University of
Rio Grande do Sul
Porto Alegre, Brazil
Emails: {granville,liane}@inf.ufrgs.br

***Abstract*—Cloud Computing are currently seen as a trend in the computing area for companies, institutions and research groups that seek provision of computational resources on demand. This solution usually works out by means of virtual systems which are managed through a specific infrastructure, termed as Infrastructure-as-a-Service or IaaS in the case of cloud computing. In an attempt to identify the network monitoring resources in use, this paper aims at presenting features of some Framework for Cloud Computing operating at the IaaS level.**

Keywords-Cloud Computing; IaaS; Network Monitoring.

I. INTRODUCTION

The development of activities in several areas of human knowledge is increasingly dependent on computational resources. These resources must have a minimum of capacity in order to perform these tasks and meet some requirements imposed by the market, such as low cost, availability, flexibility and type of resources.

Hence, concerns with the availability of these resources began in the 80's, with the first Local Area Network (LAN), Wide Area Network (WAN) and high performance computer networks, technologies which have made possible the emergence of connections between computational systems. These systems have been defined as distributed systems, i.e., an independent group of computers that shows itself to their users as a single and coherent system that has as its main characteristics shared resources, transparency, scalability, availability, and flexibility [1].

Thus, computer networks and distributed systems have been developed, and from them other technologies have emerged, which aim to meet these demands, such as grid computing and, more recently, cloud computing.

Nowadays, the concept of Cloud Computing has emerged as a trend in the computer area to professionals, companies and institutions that are concerned both with flexible computational resources with high rates of performance and availability and with the use of hardware devices in rational and sustainable ways [2]. These systems aim to provide services and can be showed at several levels of abstraction, such as software, platform and infrastructure. For example, when it comes to service-oriented systems the term "Software as a Service (SaaS)" has been found in several papers, like Michel Head's 2010, in which he observes that "Cloud computing is a paradigm of computing that offers virtualized resources 'as a service' " [3].

Cloud computing systems usually have a business focus, e.g., resources lease. Some of its characteristics have become important, of which the most prominent appears to be non-functional guarantee provisioning. Applications can thus be executed by considering predefined standards, such as runtime, operations costs, security, privacy, among others. Such guarantees are currently specified in Service Level Agreements (SLA) [4] [5]. However, it will be necessary an effective management in order to maintain the SLAs in a cloud computing environment.

Currently, it is notorious, the main topics in management resources are monitoring and controlling devices; therefore, which activities should be defined and consistent on systems that support cloud computing systems. Thus, this paper aims to demonstrate about the state of art in network monitoring resources for cloud computing systems.

This remaining of this paper is organized as follows. Section 2 presents current technologies in monitoring, such as Simple Network Management Protocol (SNMP) and Management via Web Services. Section 3 presents current systems on Cloud Computing. Section 4 presents the some Frameworks to Cloud Computing. Section 5 presents the network monitoring in Cloud Computing. Finally, conclusions and future works are show in Section 6.

II. MONITORING

A management protocol aims at providing the basis for monitoring and controlling resources. The management process can be divided, according to the OSI management standard, in five functional areas as it follows: fault management, configuration management, accounting management, performance management, and security management [6][7]. Each functional area has the purpose of defining the focus of action of monitoring and controling.

 ISBN: 978-1-61208-183-0

With the growth of computational systems, the demand for better administration methods also increased. The use of Cloud Computing Systems is consequentely affected by such trend and, therefore, the characteristics of each cloud must also be taken into consideration when managing such environments.

In this context, the utilization of management protocols becomes an important tool. Currently, we can cite three relevant management options: Simple Network Management Protocol (SNMP) [6] [8], Network Configuration Protocol (NETCONF) [9] and Management Systems via Web Services [10].

A. SNMP

The SNMP protocol was originally developed for network management. However, due to its flexibility it may be used for other types of management applications. The structure of this protocol is based on managers and agents, where the agents are spread on the resources and the management operations (e.g., read or write of objects) are performed by the manager through direct solicitations to agents. The objects that can be managed are described in a MIB (Manageament Information Base) [11].

The MIB serves as reference for tracking objects that are part of the system with the aim of getting information about resources. It contains a hierarchical structure set according to a numeral sequence [12].

The Simple Network Management Protocol (SNMP) is currently used to refer a collection of specifications for network management, which includes the protocol itself (SNMP), the specification to describe management objects (SMI - Structure of Management Information), and the management objects (MIB). All the operations supported by SNMP are related to reading/writing management objects. The main operations available are the following:

- Get: this operation is used to read an object value.
- Set: this operation is used to write object values.
- Get-Next: this operation is used to read the value of the next available object.
- Trap: this operation is used for communicating special events.

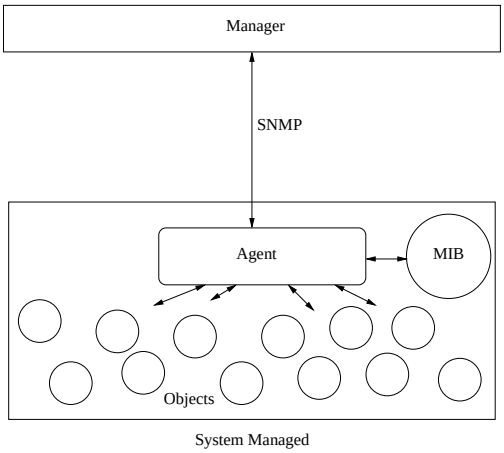


Fig. 1. SNMP Architecture.

The amount of operations is limited, making the protocol simple, flexible and easy to implement. Such operations are used between manager and agents.

The architecture of the SNMP is shown in Figure 1, where it is possible to see the manager and agent interacting. In this figure, the manager sends solicitations using SNMP to the agent, then it observes the reference within the MIB, it accesses the object and it sends a reply to the manager using SNMP.

The development of the agent follows the features of the managed system, being one extension of system. We can notice that the features on the system influence directly the complexity for agent development.

Because of these characteristics, we can define that SNMP is effective to monitoring resources, but is limited to controlling. Thus, there is a need for others technologies. For this, currently, we find the Netconf and Management via Web Services.

B. Web Services

In general, Web Services provide a standard means of interoperating between different software applications, running on a variety of platforms and/or frameworks [13]. Based on this concept, Web Services emerge with different potentials in the network management context. According to Klie et al. [14], in the network management community, many people see Web Services as a possibility to solve some of the most important problems:

- First of all, since Web Services are platform independent and use standard Internet protocols, they help to deal with the heterogeneity.
- Second, they offer a unified communication model for network, application, and management systems.
- Third, with Web Service composition mechanisms, automation can be supported.

Vianna et al. [15] highlights that the motivation for using Web Services in the network management area is that these technologies in fact address problems that SNMP investigations have been trying to solve in years. Basically, a Web service is a software system designed to support interoperable machine-to-machine interaction over a network. Figure 2 shows the general process of engaging a Web service, emphasizing the involved entities.

As shown, the requester and provider entities become known to each other. The requester and provider entities somehow agree on the service description and semantics that will govern the interaction between the requester and provider agents. The service description and semantics are realized by the requester and provider agents. Finally, the requester and provider agents exchange messages, thus performing some task on behalf of the requester and provider entities.

The core technologies of Web Services include Simple Object Access Protocol (SOAP) [16], Web Service Definition Language (WSDL) [17] and Universal Description Discovery and Integration (UDDI) [18], they are expressed in the standard form of XML documents and are built XML-based

 ISBN: 978-1-61208-183-0

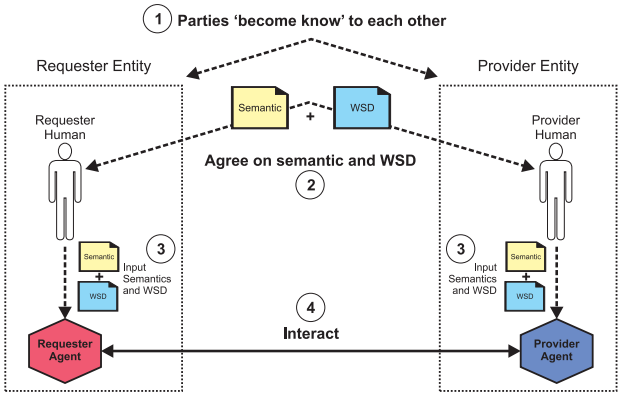


Fig. 2. The general process of engaging a Web service [13].

specification, this ensures the platform-independent, language-independent and human-computer interaction of the architecture [9]. Basically, SOAP is the communication protocol, WSDL describes the service, how to access it and the available operations, while the UDDI allows publish and discover of Web Services directories.

In the network management context, the main standards are the Management Using Web Services (MUWS) [10] by the Organization for the Advancement of Structured Information Standards (OASIS) and the Web Services for Management (WS-Management) [19] by the Desktop Management Task Force (DMTF). Nowadays, specifically in the cloud computing context, the use of Web Services is presented as a potential alternative given the heterogeneity and diversity of environments that composes this new paradigm. This way, the Amazon Web Services (AWS) [20] is a popular example.

III. Cloud Computing

Nowadays, there are many different definitions to Cloud Computing. Rajkumar Buyya defines it as a paralel and distributed system composed of clustered virtual machines interconnected which are allocated dynamically and presented as a unified system of computing resources based on SLAs estabilished through business trading between service provider and clients [21].

On the other hand, Ian Foster understands that the paradigm of cloud computing is not necessarily a new concept, but the result of a symbiosis among different paradigms, such as: grids, clusters, distributed systems and others [22].

This paper does not aim at discussing different concepts as to what cloud computing is. However, in order to present a focus to this work, some definitions need to be taken into account such as types of cloud computing (Deployment models) and service models.

A. Types of Cloud Computing (Deployment Models)

Computing clouds can be classified according to their creation, use and purpose. They may come from the need of short-time computing resources within an organization, which in this context, utilizes resources from another business partner. This classification may also come from underutilization of resources within an enterprise, generating an opportunity to lease its computing availability on demand to a third party through a cloud. The peculiarities described above result in consolidated types of cloud computing described as: Private, Public, Community and Hybrid Cloud [23].

B. Service Models

The concept of cloud computing encompasses different levels of computing services. They may vary from allocated applications within a cloud to making storage resources available and data processing. Such levels are regarded as service models as it follows [24]:

- Software as a Service (SaaS): applications of interest to a large amount of users may be hosted within a cloud as an alternative to local processing. Applications are offered as a service by providers and accessed by clients through applications such as a browser. Managing and controlling of networking infrastructure, operational systems, servers and storage is done by the service provider. Google Apps is a example of SaaS.
- Platform as a Service (PaaS): the structure provided to the user by the provider to develop applications which are going to be executed and made available within clouds. On that regard, another concept emerges known as Utility Computing, used to denominate the whole support platform to the developing and providing of applications in clouds.
- Infrastructure as a Service (IaaS): it is the capacity a provider has to offer a processing infrastructure and storage in a transparent way. The user does not have control of the physical infrastructure, but through virtuliazation mechanisms it is possible to control operational systems, storage, installed applications and possibly a limited control of network resources.

Such virtualized infrastructure must follow some guidelines which enable an effective resource management. Such effectiveness relies on [25]:

- Providing a uniform and homogenous vision of the virtualized resources, regardless of the platform, be it Xen, KVM or Vmware, for example.
- Managing a VM life cycle, including networking communication configurations, in a dynamic way for virtual machine clusters.
- Managing system resource storage.
- Adjustable support to allocation of resources in order to cater for the specific needs of each and every enterprise which comprises or utilizes the cloud, for each one uses a cloud according to its needs, such as: availability, server consolidation, decrease in energy, among others.
- Adapting the organization in order to choose necessary resources, including peaks where the local ones do not cater for. The new choice of resources should include subsidies taking into account the addition of new physical resources, in a dynamic way, or containing a buffer zone tolerant of faults originated by physical resources.

 ISBN: 978-1-61208-183-0

The characteristics above should be presented by frameworks which enable the managing of virtual resources in cloud computing systems. Currently we can use the Amazon EC2 [26], Eucalyptus [27], OpenNebula [28] and Nimbus [29] as examples of available utility computing clouds as an IaaS.

IV. FRAMEWORKS

A. Amazon EC2

The Amazon Elastic Compute Cloud (Amazon EC2) is a private solution for Cloud Computing. This solution was a pioneer on solutions to cloud, thus, become reference for others frameworks.

The Amazon EC2 is mainly characterized by [26]:

- Resource and instances storage capacity in different and distributed allocations.
- Safe cloud computing environment.
- Automatic load scheduling and balancing.
- The ability to import external virtual machines.

B. Eucalyptus

Eucalyptus is a framewrok designed for cloud computing systems which operate on IaaS level. This system enables users and cloud computing system administrators the manipulation of virtual machines through functionalities such as creation, control and finalization [27].

It was developed aiming at the broadening of academic studies of cloud computing, it is one of the first systems of this kind and it has an open code and its structure adapted to a wide range of resources which utilize physical infrastructure (processing, storage, network, among others). It is usually found and available within academic research groups.

In this regard, it is worth pointing out that the functional struture of Eucalyptus is flexible and modular, enabling possible adaptations to the system which aim at best suiting it to testing scenarios, experiments, analysis or studies.

By being modular, Eucalyptus is formed by components which interact among themselves through interfaces. This system modularity enables it to be altered, updated and perfected.

Apart from being modular, Eucalyptus also has a hierarchical structure (as described in the Architecture subsection), which facilitates the usage of resources available in labs, clusters, workstations and servers.

1) Architecture: The architecture of Eucalyptus is simple, flexible and modular. The structure of the system is hierarchical and it is presented with a friendly operational environment. The system enables functionalities such as beginning, access, control and ending of VMs, using a presentation system similar to Amazon's EC2 [30].

By definition, implementing each component of high level in the system is a stand alone web service. Such definition brings some benefits such as:

- Each web service is presented with a defined language from an API (Application Programming Interface) in the shape of a WSDL (Web Service Description Language) containing the operations that the service is able to execute and the structures of an in/out database.
- The system enables the implementation of security policies regarding the communication among its components.

Eucalyptus is formed currently by three components of high level, Instance Manager (IM), Group Manager (GM) and Cloud Manager (CM).

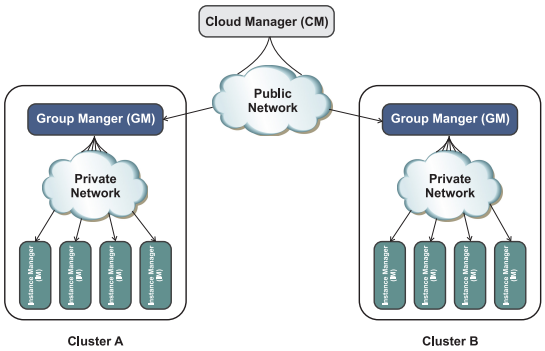


Fig. 3. Eucalyptus Architecture.

Figure 3 shows Eucalyptus Architecture, highlighting the high-level components.

C. OpenNebula

OpenNebula is a virtual structure manager which can be used to deploy and manage virtual machines (VM) or manage VM clusters that can be co-scheduled in local resources or even in external public clouds.

The automatic configuration of virtual machines, the preparation of disc images and the communication network configuration among other functionalities occur independently of the virtualized platform (e.g., Xen [31], KVM [32], Vmware [33]) or the external cloud (e.g.,EC2); therefore, the Opennebula works independently of the virtual system being used [34].

The platform independence is one of the main characteristics of Opennebula, being only possible given the modularity of the system. This is one of the peculiarities of Opennebula.

This framework has functionalities similar to the ones found in similar systems, its focus, however, lies on the attempt to fulfill some gaps still found in other frameworks for cloud computing systems. Among these gaps we can highlight the following [25]:

- The inability to scale external clouds.
- Monolithic and closed architectures that are hard to extend or interface with other applications, not allowing seamless integration with existing storage and network management solutions deployed in data centers.
- A limited choice of preconfigured placement policies, such as First Fit and Round Robin.
- A lack of support for scheduling, deploying and configuring groups of VMs (for example, a group of VMs representing a cluster, where all the nodes either deploy entirely or do not deploy at all, and where some VMs configuration depends on others, such as the headworker relationship in compute clusters).

 ISBN: 978-1-61208-183-0

1) Architecture: The architecture of the Opennebula has modular and specialized components which execute functions to fulfill requirements regarding virtual infrastructure management.

In this regard, the virtual machine life cycles is executed by Opennebula's core, which manipulates three managing modules: Image and storage technology module, Network factory and a thrid module called Underlying Hypervisor's [25].

Among these three modules, we highlight the network factory, which is composed of virtual network equipment like DHCP servers, firewalls and switches, among others equipments. They provide for the VMs an virtual communication network environment.

The Opennebula core communicates with the storage devices, network and virtual systems through the so-called pluggable drivers, so that Opennebula is not bound to a specific environment, providing a uniform managing to the underlying infrastructure layers.

Apart from managing virtual machine life cycles, Opennebula's core has deployment support, which includes interconnected component clusters, such as web services and requested data base in several virtual machines. That way a group of virtual machines can be treated as first class entities in Opennebula. Additional to the managing of virtual machines as a single unit, the core can also be presented with information regarding the context, such as digital certificates and virtual machine software licensing.

A resource scheduler, usually Haizea, determines how the virtual machine allocations are going to be accomplished [25]. More specifically, the scheduler has access to all information about Opennebula requests and based on them it determines future and current allocations, creating and updating resource programming and sending the deployment commands to the Opennebula core.

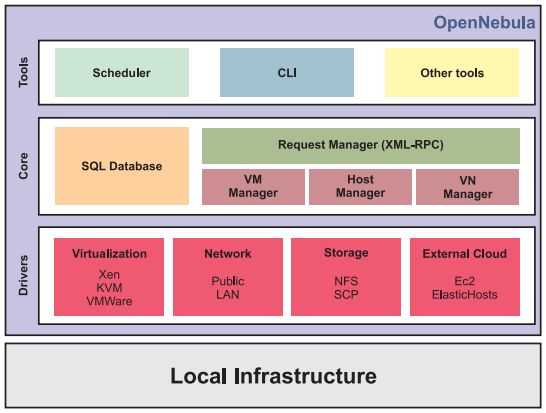


Fig. 4. OpenNebula Architecture.

Figure 4 shows the Opennebula's architecture and its components, highlighting the core, the Haizea scheduler and the drivers which interconnect the system modules and external clouds.

D. Considerations

We will hold a brief comparison between Eucalyptus and Opennebulla, for this, we defined some functions and topics will be analyzed in both systems, such as, Architecture, Virtual Systems, Virtual Infrastructure Management and Allocation and Resources Allocation.

1) Architecture: We understand architecture as organization structure for operation of framework, in this way, both, Eucalyptus and Opennebulla have diferent organization forms. The architecture of Eucalyptus is hierarchical and have three high level components. The components communicate with each other through networks public and private. On the other hand, Opennebula have a modular architecture, because uses the modular resources (e.g., drivers, cores) to interoperate between them.

2) Virtual Systems: Clouds are composed by sets of virtual machines. Thus, frameworks for cloud computing need have the capacity to manage virtual machines. Nowadays, we found several systems to deploy the virtualization, such as, Xen and KVM, for example. In this context, a good framework need provide support for these systems. The Eucalyptus, currently, have support to Xen, KVM, and VMware, among others. Additionally, Eucalyptus, have support to Virtual Machine Manager (VMM). The Virtual Machine Manager is a device that allow operate and manage a virtual infrastructure over virtual machine (e.g., memory, hard disk). On the other hand, Opennebula have support to several virtual systems, such as, Xen, KVM and Vmware, among others, but, we can not found reference to VMM support.

3) Virtual Infrastructure Management and Allocation: Virtual infrastructure management and allocation in cloud is related with framework capacity to manipulated a virtual machines sets. These virtual machines, that interoperates in a virtual environments, such as, virtual networks.

Both, Eucalyptus and Opennebula have capacity to begin, access, manage and kill virtual machines, individually or in groups. In addition, have support to management of virtual networks and SLAs standards. However, the Opennebula have a capacity to scheduling a virtual machine hosted in a external cloud. Currently, the Opennebula have drivers for utilization on frameworks, such as, Amazon EC2, and Eucalyptus, for example.

4) Resource Allocation: Resource Allocation (e.g., memory, hard disk) on clouds is related with framework capacity to manipulate these resources in virtual machines and work with SLAs. The Opennebula uses a scheduler, usually Haizea, to do it. On the other hand, the Eucalyptus uses yours hierarchical architecture to make this.

V. Monitoring in Cloud

A. Monitoring in Eucalyptus

As seen before, the Eucalyptus system enables users and administrators to manipulate an infrastructure of physical and virtual resources through open code solutions for cloud computing systems.

 ISBN: 978-1-61208-183-0

This system is hierarchically structured and has three components of high level which communicate among themselves through communication networks.

In order to achieve that, Eucalyptus utilizes public and private network structures, being the latter heterogeneous and comprised of physical and virtual network devices. The implementation of these virtual networks is achieved through Virtual Distributed Ethernet (VDE), a system which has the ability of creating virtual network according to ethernet standards and provides the use of switches, cables and other virtual network devices for the Eucalyptus.

It is also worth mentioning that the communication between group and instance managers is provided by a private network influencing directly on the communication among the final components of a system. This peculiarity is vital to the load balance within a network, given the adequate structure of a communication network has direct influence on the system's performance altogether.

A virtual network is created from physical devices that host virtual ones which posses as main characteristics isolation capability and migration, which enable the definition of different and flexible use scenarios, cost reduction and effective security policies.

In this respect, it is understood that the effective monitoring of these networks is of crucial importance to framework researches and the own perfecting of it, as well as any other cloud computing system.

The current managing standards such as SNMP protocol, Netconf and managing via Web Services can be applied to Eucalyptus as a whole, although studies which evaluate its performance, efficency and application of these standards have not yet been found in cloud computing system.

The avouchment above highlight the fact that it is still incipient the task of monitoring network resources in Eucalyptus, such notions can only be proved through research, studies and experiments which can evaluate specifically and proficiently these types of solutions.

B. Monitoring in OpenNebula

As observed before, the Opennebula system has a flexible and modular architecture and it is able to manage virtual machines individually or conjoined. This system possesses a core and a scheduler which work with virtual support systems (e.g., Xen, Vmware, among others), storage devices and virtual communication networks, and it is also capable of scheduling external clouds if necessary.

In this regard, virtual network image devices are stored in the so-called network factory, devices which are allocated and deallocated according to momentary communication needs.

Such communication needs may come from within a cloud, which occur between virtual systems initiated and set in motion to cater for applications defined by the core and scheduler in Opennebula.

The monitoring of virtual network devices which serve to the virtual systems are compatible with the solutions already mentioned in this paper, such as SNMP, Netconf and managing through Web Services, given the fact that virtual systems can support these managing solutions, as well as the monitoring of virtual devices.

As the Eucalyptus, it is understood that the virtual monitoring activity is very important to the perfecting of Opennebula, although no research, studies or experiments regarding it have been found, as well as currently in respect to this framework.

VI. Conclusion and Future Work

The cloud computing emerges as an old dream of computer science, called utility computing. From this perspective, customers, companies, government institutions, among others, start to migrate their applications, platforms and infrastructure to the cloud, creating new types of pricing, services, resources use, etc.

This way, virtualization is the most important assumption to achieve the goals addressed by this new paradigm. Specifically, virtualization can enable dynamism and scalability to the cloud, maximizing the potential to services offering and optimizing resources allocation by providers. In the other hand, this flexibility brings an inherent complexity to infrastructure orchestration and management, since constitute an increasingly heterogenous enviroment.

In this work, we investigated the cloud monitoring process in the main cloud plataforms. We can observe that the concerns are more focused on the cloud offering models, i.e., the convergence in a model, which directs the efforts to management tasks, still seems remote. Obviously, we can diagnose easily the well-known management challenges, but the innovations introduced by the cloud require specific researches.

Based on this context, we verify the need for well-defined mechanisms to monitoring process. Such mechanisms should provide the features for low-level metrics measurement (e.g., measure the rate of memory and processor usage in an array of devices that make up the infrastructure for a particular application) and the capacity to diagnose high-level behaviors, such as quality of services from customers point-of-view.

As future works, we intend to investigate the applicability of traditional monitoring mechanisms in order to diagnose the strengths and weaknesses of them in the cloud enviroment. From this point, we aim propose a basic monitoring framework to cover the different models of service provided in the cloud.

References

[1] A. S. Tanenbaum and A. S. Woodhull, *Operating Systems: Design and Implementation*, 2nd ed. Bookman, 2000.

[2] B. Hayes, "Cloud computing," *Commun. ACM*, vol. 51, pp. 9–11, July 2008. [Online]. Available: http://doi.acm.org/10.1145/1364782.1364786

[3] M. R. Head, A. Kochut, C. Schulz, and H. Shaikh, "Virtual hypervisor: Enabling fair and economical resource partitioning in cloud environments," in *NOMS*, 2010, pp. 104–111.

[4] I. Brandic, "Towards self-manageable cloud services," in *Proceedings of the 2009 33rd Annual IEEE International Computer Software and Applications Conference - Volume 02*, ser. COMPSAC '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 128–133. [Online]. Available: http://dx.doi.org/10.1109/COMPSAC.2009.126

[5] G. Reig, J. Alonso, and J. Guitart, "Prediction of job resource requirements for deadline schedulers to manage high-level slas on the cloud," *Network Computing and Applications, IEEE International Symposium on Cluster Computing and the Grid*, vol. 1, pp. 162–167, 2010.

 ISBN: 978-1-61208-183-0

[6] W. Stallings, *SNMP,SNMPv2,SNMPv3 and RMON1 and RMON2*, 3^a ed. Addison-Wesley, 1996.

[7] R. Stephan, P. Ray, and N. Paramesh, "Network management platform based on mobile agents," *Int. J. Netw. Manag.*, vol. 14, no. 1, pp. 59–73, 2004.

[8] M. A. Miller and Atwood, *Managing Internetworks with SNMP; The Definitive Guide to the Simple Network Management Protocal, SNMPV2, RMON, and Rmon2*, 2nd ed. Hungry Minds, Incorporated, 1997.

[9] D. Xiaochao and W. Jinpeng, "Netconf network management model based on web services," in *IT in Medicine Education, 2009. ITIME '09. IEEE International Symposium on*, vol. 1, Aug. 2009, pp. 160 –164.

[10] W. Vambenepe and V. Bullard, "Web services distributed management: Management using web services (muws 1.1) part 1," Available http://docs.oasis-open.org/wsdm/wsdm-muws1-1.1-spec-os-01.htm, Nov, 2011.

[11] Y.-C. Chen and I.-K. Chan, "SNMP getrows: an effective scheme for retrieving management information from MIB tables," *Int. J. Netw. Manag.*, vol. 17, no. 1, pp. 51–67, 2007.

[12] M. T. Rose, "Management information base for network management of tcp/ip-based internets: Mib-ii," 1990. [Online]. Available: http://rfc.net/rfc1158.html

[13] D. Booth, H. Haas, F. McCabe, E. Newcomer, M. Champion, C. Ferris, and D. Orchard, "Web services architecture," 2004. [Online]. Available: http://www.w3.org/TR/ws-arch/, Sep, 2011.

[14] T. Klie, T. U. Braunschweig, A. Belger, T. U. Braunschweig, L. Wolf, and T. U. Braunschweig, "A peer-to-peer registry for network management web services," 2009.

[15] R. Lemos Vianna, M. Almeida, L. Tarouco, and L. Granville, "Investigating web services composition applied to network management," in *Web Services, 2006. ICWS '06. International Conference on*, Sep, 2006, pp. 531 –540.

[16] J. Snell, D. Tidwell, and P. Kulchenko, *Programming Web services with SOAP*. Sebastopol, CA, USA: O'Reilly & Associates, Inc., 2002.

[17] *Understanding Web Services: XML, WSDL, SOAP, and UDDI*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2002.

[18] S. Ran, "A model for web services discovery with qos," *SIGecom Exch.*, vol. 4, pp. 1–10, March 2003. [Online]. Available: http://doi.acm.org/10.1145/844357.844360

[19] D. M. T. Force, "Web services for management specification," Available http://dmtf.org/sites/default/files/standards/documents, Nov, 2011.

[20] Amazon, "Amazon web services (aws)," Available http://aws.amazon.com, Nov, 2011.

[21] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Gener. Comput. Syst.*, vol. 25, pp. 599–616, June 2009. [Online]. Available: http://portal.acm.org/citation.cfm?id=1528937.1529211

[22] I. Foster, Y. Zhao, I. Raicu, and S. Lu, "Cloud computing and grid computing 360-degree compared," *2008 Grid Computing Environments Workshop*, vol. abs/0901.0, no. 5, pp. 1–10, 2008. [Online]. Available: http://arxiv.org/abs/0901.0131

[23] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the clouds: A berkeley view of cloud computing," EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2009-28, Feb 2009. [Online]. Available: http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html

[24] L. M. Vaquero, L. Rodero-merino, J. Caceres, M. Lindner, and T. I. Y. Desarrollo, "A break in the clouds: Towards a cloud definition," *ACM SIGCOMM Computer Communication Review*, pp. 50–55, 2009.

[25] B. Sotomayor, R. S. Montero, I. M. Llorente, and I. Foster, "Virtual infrastructure management in private and hybrid clouds," *IEEE Internet Computing*, vol. 13, pp. 14–22, 2009.

[26] "Amazon elastic compute cloud: Ec2," http://aws.amazon.com/documentation/ec2/. Last access Sep. 2011.

[27] "Eucalyptus cloud," http://www.eucalyptus.com/. Last access Sep. 2011.

[28] "Opennebula project," http://www.opennebula.org. Last access Sep. 2011.

[29] P. Marshall, K. Keahey, and T. Freeman, "Elastic site: Using clouds to elastically extend site resources," *IEEE International Symposium on Cluster Computing and the Grid*, pp. 43–52, 2010.

[30] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, L. Youseff, and D. Zagorodnov, "The eucalyptus open-source cloud-computing system," in *Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid*, ser. CCGRID '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 124–131. [Online]. Available: http://dx.doi.org/10.1109/CCGRID.2009.93

[31] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauery, I. Pratt, and A. Warfield, "Xen and the art of virtualization," in *Proceedings of SOSP'03*, 2003.

[32] The KVM Project Home Page, "KVM-kernel based virtual machine," Available http://www.linux-kvm.org, Nov, 2011.

[33] THE VMWARE Project Home page, "The vmware project," RFC 1832. Available http://www.vmware.com, Nov, 2011.

[34] B. Sotomayor, R. S. Montero, I. M. Llorente, and I. Foster, "Capacity leasing in cloud systems using the opennebula engine," *Cloud Computing and its Applications, CCA´08*, pp. 1–5, 2008.

 ISBN: 978-1-61208-183-0

The Space-Time Information in the Access Management

Jaroslav Kadlec, Radimír Vrba, David Jaroš, Radek Kuchta
Department of Microelectronics
FEEC, Brno University of Technology
Brno, Czech Republic
kadlecja | vrbar | jarosd | kuchtar@feec.vutbr.cz

***Abstract*—This paper deals with the possibility to employ the user's time and space position information in the access management. Using the time and space information as new factors for authentication process is discussed in this paper. We have also considered the issues of indoor localization and possible application scenarios where these two additional authentication factors can be applied. We have developed the Multi-Factor Authentication Device (MAD I) together with active infrastructure, which is required for indoor users' localization, to demonstrate the new main functions and advantages of adding time and space position to the user's authentication factors. The main advantage of the MAD I is that the device helps the AAA system verify the user's location in both main usages, i.e., indoor and outdoor environment.**

Keywords-location-based authentication; active infrastructure; wireless communication.

I. INTRODUCTION

Authentication and authorization are required almost everywhere in today's world. People must be identified when they download emails, read newspapers over the Internet, fill out forms for the government, access company private information, etc. When servers communicate with each other, they have to create trusted connection. Before a connection is created, it is necessary to identify the servers. There are different ways how to identify a user and a server.

The techniques that are used for user´s identity verification can be divided into three main groups along the subject of verification as refers [1].

- **A user knows something –** the user has to know private information, which is not known by anybody else. The password verification technique is one of the most common techniques in this group.
- **A user is somebody –** this group covers techniques that are related to human user authentication. The techniques verify biometric properties of a human´s body. The fingerprint reading technique can be mentioned here.
- **A user has something –** the user brings up a unique thing (token) as subject of credential. For example, the unique thing can be Radio Frequency Identification (RFID) transponder or a hardware key.

When a user or a server needs to authenticate a server, the most common way is using certificates. In this scenario, a trusted authority issues a certificate that is used for asymmetric cryptography.

Especially, the scenario with the user's credentials is sometimes insufficient and some extra information is required for many situations and systems. The information should be the user's certificate, biometric identification or current position.

The main topic of the paper is focused on the possibility of employing the user's space-time information in AAA (Authentication Authorization Accounting) systems [2]. We assume that the space-time information will be used especially for user's identity verification.

The sharp growth in information, technologies and especially information systems require monitored and effective access control. A user has to approve his identity at first. Based on this step, an access management will assign the rights for the user. An accounting system that will create and store records about the user's activities should also be a part of the system. For example, the records can be used as input information for future system development or for an audit. The above mentioned functions are provided by the systems that are commonly called AAA (Authentication Authorization Accounting) systems. The blocks of the main AAA system features are shown in Figure 1. A user is authenticated at first. In the next step rights are assigned to the user. The records are created during a whole session and stored in a database.

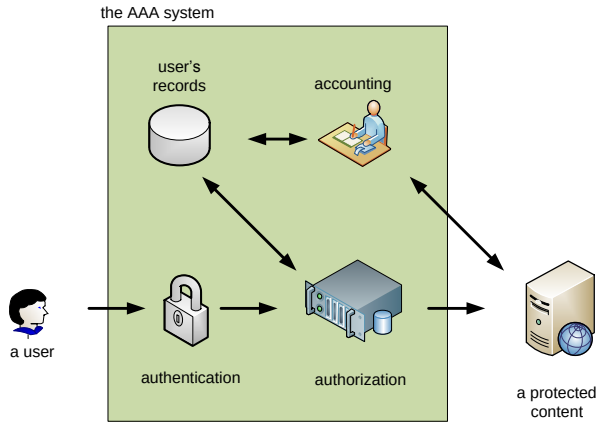


Figure 1. The main features of the AAA system

As a new factor of the user´s identity verification, his/her space-time information is discussed in this paper. That means "a user is on a known place in known time".

The rest of the paper is organized as follows: the space-time information introduction, the main aspects, and possible use in the AAA systems are described in the second section. The third section presents a possible application scenario.

 ISBN: 978-1-61208-183-0

The fourth section introduces an authentication technique with active infrastructure and with the Multi-Factor Authentication Device (MAD I).

II. THE SPACE-TIME INFORMATION

The AAA systems that could work with the space-time information can prove useful in the following fields.

For example, a doctor shouldn't manipulate a patient's private information outside the hospital area, as referred to in [3].

Another example can be found in the financial sector. The user's position verification should be a part of user authentication process before s/he gets access to the bank account.

The SSO principle (Single-Sign-On) [4] could also make use of the space-time information. The user does not need to perform authentication to various systems when they are accessed from approved places (from his/her home or office).

The position information can be interpreted relatively or absolutely [5]. The relative position is determined as proximity to the object the position of which is exactly known. The objects with known position are called anchor points. This interpretation is used in GSM (Global System for Mobile Communications: originally from Groupe Special Mobile). Second, the position information can be interpreted absolutely. The absolute position information utilizes the coordinates in two or three dimensions. This way is employed for example in GNSS (Global Navigation Satellite System) systems.

The space-time information can be assigned into all three main processes of the AAA system, as described in Figure 2. For the authentication the user's space-time information should be verified in conjunction with verification of other authentication factors. The process which performs verification of two or more factors is commonly called multi-factor authentication or strong authentication. Depending on the user's space-time information the user will get different access rights in the system. For example, when the user accesses from the office, s/he will get different rights than when doing so from a public internet café. The user's space-time information could also be used for choosing charging rate for services.

The space-time information is very sensitive private information. Generally, similar information is to be handled very carefully. As shown in [6], the user's space-time information could be abused in various ways.

The space-time information needs in the AAA systems are related especially to mobile users. If a user meets the space-time condition in the verification time, s/he will get access based on submitted credentials. The user has been verified and has access to the system, but he can change his position and move out from the approved area. This problem can be solved by periodically evaluating the space-time information.

The more suitable solution is described in [7]. Direction and speed of the user's movements is additional information used.

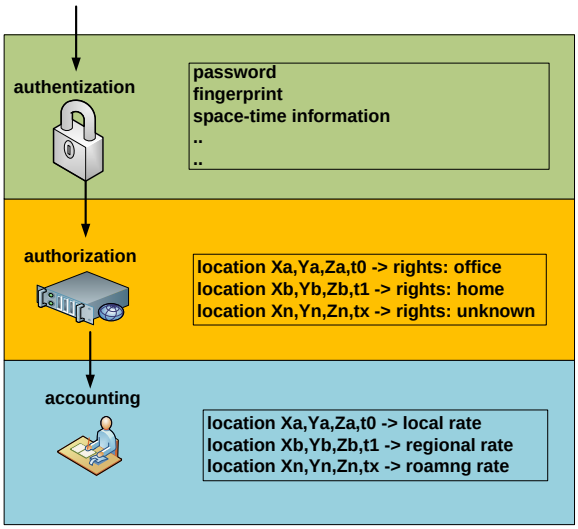


Figure 2. The space-time information in AAA

III. APPLICATION SCENARIO

We have adopted the application scenario as shown in Figure 3. The user wants to get access to the protected domain content (resources, services, etc.). The MAD I is connected to the user's terminal. The request for protected content from the user is redirected to domain controller, which performs access management. The user is requested to give in his/her credentials. If the user has connected MAD I, it provides the space-time information and fingerprint data. The methods for fingerprint processing in general provide the same hash for the same fingerprint, otherwise a fingerprint reader cannot be used in the identity verification. The position information and fingerprint are encrypted by AES (Advantage Encryption System) [8]. The user adds his login and the data are sent to a domain controller. The domain controller will settle the user's authentication depending on received credentials. If the identity is verified, the user's role in the domain is defined. The RBAC (Role Based Access Control) is used for the system [9].

An area management represents a database, which stores the definition of the user's areas. The areas are defined in two ways. A simpler way is to define one point and the distance from it (radius). Thus we get a circle from where the user will get the access. The definition of the net of triangles is more complex (leading to convex combination). This way brings along more difficulties in defining, storing and evaluating but gives us an advantage in the definition area of any shape. Defined areas are stored within IDs and can be used by any users. The defined area can mean different roles (rights) for different users. The user can cooperate with the administration desk to define a new area. The pairs of the area's ID - roles are stored in a user's profile in the Active Directory. Appropriate areas are requested by the domain controller from the management of areas. The domain controller contains API for evaluating the position information (if a user is or is not in an evaluated position). The order in which the area's IDs are stored in a user's

 ISBN: 978-1-61208-183-0

profiles defines the priority of the areas. The last added ID in the list has the highest priority. This right solves the overlapping problem.

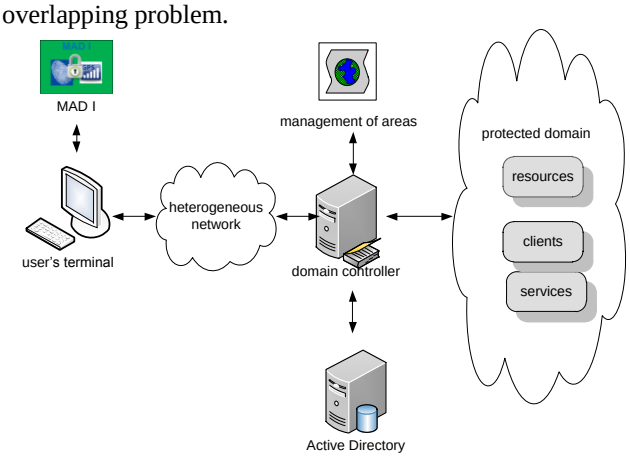


Figure 3. The application scenario with MAD I

An API in the domain controller evaluates mutual relationship between the user's position and the areas defined for its identity.

IV. ACTIVE INFRASTRUCTURE AND MAD I

The Active Infrastructure (AI) is a technology background that is used in the two authentication techniques that are described in two subsequent sections. The key parts of AI are represented by an anchor point, a user's tag and an authenticator. The anchor point is located in position where the users want to be authenticated regarding their position. We assume that the position of the anchor point is exactly known to the authenticator. On the other hand, the user's tag is assigned to the particular user and only with difficulties is it related to its identity. The user's tag can be a part of the user's terminal or autonomy pocket device. The position of the user's tag is determined by proximity between the anchor point and the user's tag. When the user's tag can communicate with the anchor point, it means that it is nearby.

Figure 4 represents the active infrastructure key parts. The anchor point is in known position x_{AP}, y_{AP}, z_{AP}. If the user's tag is in its proximity, it can communicate with the anchor point which means that the position of the anchor point is similar to the position of the user's tag. The similarity between the positions is dependent on the range of transceivers. When the user claims that he is in a position nearby the anchor point, the authenticator asks the anchor point if an appropriate user's tag is in the communication range. It should be noted here that, for example, IQRF [10], Bluetooth [11], or similar wireless communication solutions can be used as wireless technologies.

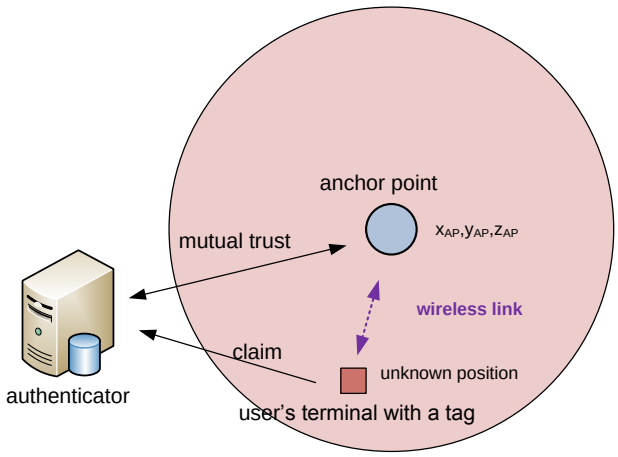


Figure 4. Principle of active infrastructure

Before the first user is authenticated, a mutual binding has to be done. Initial binding has to be executed at the system administration desk over local bus (MAD I is USB enabled). Binding process performs AES key exchange between MAD I and the domain controller or the Active Directory where the key is stored during the binding. The hash of the user's fingerprint is also stored on the server side. This process can also cause the assignment of MAD I to an exact user. The initial binding is described in the following steps.

- First, a secure channel should be established. This is done by Diffie-Hellmann key exchange [12]. Two unknown sides can derive the secret key. This technique is often used for the exchange of symmetrical encryption key.
- When the secured channel is established, the domain controller generates an encryption key for AES. The length of the key is 256 bytes.
- The key is sent via the secured channel created in the first step.
- The MAD I stores the key in secured memory after reception.
- The user is requested to swipe his finger on the fingerprint reader on the MAD I.
- The hash of the user's fingerprint is sent to the domain controller.
- The user's fingerprint hash is stored in the user's profile in the Active Directory.

The MAD I collects principally three authentication factors, i.e., the ownership of certain device, the fingerprint and the user's space-time information, where the user's position is used in the authentication process described.

The MAD I is connected to user's terminal via USB (Universal Serial Bus). The device is designed as a pocket device. The block diagram of the MAD I is shown in Figure 5.

 ISBN: 978-1-61208-183-0

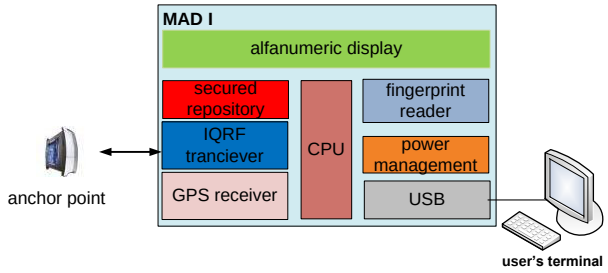


Figure 5. The MAD I block diagram

The position information is provided by the IQRF transceiver. The MAD I has already been assembled with GPS receiver for another way of position determination in other authentication techniques. The fingerprint reader is used for the user‘s authentication employing MAD I. For the security reasons the symmetrical encryption key is stored in the secure data repository. The secure data repository has special features that protect the stored data against unauthorized reading or writing. Alphanumeric display is assembled for communication between the user and MAD I.

The MAD I is a battery-powered pocket device. The power management contains circuits for adjusting power voltages for the other blocks and circuits for battery charging via USB.

V. CONCLUSION AND FUTURE WORK

The paper has introduced quite a new direction in the access management which works with the user’s space-time information. We have enumerated the main aspects of possible applications. Further, we have described possible application scenario and a suitable solution while using the active infrastructure and the MAD I.

We have designed and developed a user’s Multi-Factor Authentication Device, prepared software for this device and started the testing phase of the project. The software for the MAD I device represents only one part of the required software. Another two software pieces had to be prepared. One is on the server side, which allows processing of the received user data and authentication factors and also integrates the position information to the Microsoft Active Directory.

The second part of the software has to be implemented to the user terminal. We are testing the available solutions. One is an extension of the Windows Credential provider. This has required installation to the client computer, an update of local policies and other administrative tasks. Second one can prove useful for public computers. In this case no installation is required. The software will only be executed and will communicate directly with the server and ensure user’s authentication. But this version has some limitations.

All the described methods are in the testing phase. The test results will be followed by other improvements. Also the MAD device is designed for testing purposes only and will be optimized and minimalized in the future.

ACKNOWLEDGMENT

This research has been supported by the ARTEMIS JU in Project No. 120228 AS Nanoelectronics for Mobile Ambient Assisted Living (AAL) Systems and partly by the Czech Ministry of Industry and Trade in project FR-FR-TI3/275 OPS An Open Platform for Smart Cities.

REFERENCES

[1] G. Lenzini, M. S. Bargh, and B. Hulsebosch, "Trust-enhanced Security in Location-based Adaptive Authentication," Electronic Notes in Theoretical Computer Science, vol. 197, pp. 105-119, 2008.

[2] R. He, M. Yuan, J. Hu, H. Zhang, Z. Kan, and J. Ma, "A novel service-oriented AAA architecture," Personal, Indoor and Mobile Radio Communications, 2003. PIMRC 2003. 14th IEEE Proceedings on , vol.3, no., pp. 2833- 2837 vol.3, 7-10 Sept. 2003

[3] E. Bertino and M. Kirkpatrick, "Location-Aware Authentication and Access Control - Concepts and Issues," in 2009 International Conference on Advanced Information Networking and Applications, 2009, pp. 10-15.

[4] D. E. Denning and P. F. MacDoran, "Location-based authentication: Grounding cyberspace for better security," Computer Fraud & Security, vol. 1996, pp. 12-16, 1996.

[5] I. Ray and M. Kumar, "Towards a location-based mandatory access control model," Computers & Security, vol. 25, pp. 36-44, Feb 2006.

[6] B. Schilit, J. Hong, and M. Gruteser, "Wireless location privacy protection," Computer, vol. 36, pp. 135-137, Dec 2003.

[7] P. S. Tikamdas and A. E. Nahas, "Direction-based proximity detection algorithm for location-based services," in Wireless and Optical Communications Networks, 2009. WOCN '09. IFIP International Conference on, 2009, pp. 1-5.

[8] Ch. Lu, Y. Kao, H. Chiang, and Ch. Yang, "Fast implementation of AES cryptographic algorithms in smart cards," in Security Technology, 2003. Proceedings. IEEE 37th Annual 2003 International Carnahan Conference on, 2003, pp. 573-579.

[9] E. Bertino, B. Catania, M. L. Damiani, and P. Perlasca, "GEO-RBAC: A spatially aware RBAC," Acm Transactions on Information and System Security, vol. 10, Feb 2007, pp. 1-39.

[10] MICRORISC. IQRF - wireless technology. <retrieved: 12, 2011> Available: www.iqrf.org

[11] B. SIG, Bluetooth homepage. <retrieved: 12, 2011> Available: www.bluetooth.com

[12] Y. Eun-Jun and Y. Kee-Young, "An Efficient Diffie-Hellman-MAC Key Exchange Scheme," in Innovative Computing, Information and Control (ICICIC), 2009 Fourth International Conference on, 2009, pp. 398-400.

 ISBN: 978-1-61208-183-0

A High-level Network-wide Router Configuration Language

Miroslav Sveda Michal Sekletar Tomas Fidler Ondrej Rysavy
Faculty of Information Technology
Brno University of Technology,
612 66 Brno, Czech Republic
e-mail:{sveda, xsekle00, xfidle01, rysavy}@fit.vutbr.cz

***Abstract*—In this short paper, we discuss the design of a high-level network-wide router configuration language. At its current stage of development, the language enables us to specify basic routing and security configurations. A declarative nature of the language is supposed to be intuitive to network administrators. We have developed an experimental compiler that produces configuration files for Cisco routers. The contribution of the paper consists of the description a language for configuration programming and the demonstration of its capabilities on several examples.**

***Index Terms*—network configuration management; routing configuration; access control lists**

I. Introduction

Configuration languages for network devices enable to define every aspect of their functionality. Network administrators can thus write a network configuration that meets the required functionality for different and often very specific requirements. These languages have a simple declarative form. A network configuration consists of configuration files of all devices in a network. The difficulty in implementation of a correct network configuration stems from the necessity to create several separate configuration files that need to be consistent.

To overcome the difficulty in delivering consistent set of configuration files, device vendors provide tools implementing configuration wizards, web configuration interfaces or configuration generators. These tools may simplify basic configuration tasks but usually do not provide any additional mechanisms to guarantee configuration consistency and correctness. An alternative approach is to use high-level configuration languages.

This idea is behind the design of Nettle language [1], which is a domain-specific high-level language for BGP configurations. The other example is the Flow-based Management Language by Hinrichs et al. [2], which is a declarative policy language supposed for developing configurations for enterprise networks. It covers ACL, VLAN, NAT, policy routing and admission control features. However, the specified configuration is compiled only for the NOX platform. Our motivation is to define a high-level router configuration language that can be compiled to common router configuration languages of devices deployed in present enterprise networks.

In this short paper, we present the design of a high-level network configuration language. Currently, the language allows us to specify a limited set of network configurations, which includes network address assignment, static and dynamic routing and basic security. We also implemented an experimental compiler that produces IOS configuration files for Cisco routers. To be practically usable, the language needs to support other configuration features and to generate configuration for other network devices.

The structure of the paper is as follows. The next section describes a syntax of the language and illustrates its usage on simple examples. Section III briefly describes an implementation of the language compiler and the IOS configuration generator. Finally, section IV concludes the paper by summarizing the current state and discussing the future development.

II. The Language

The language consists of a set of simple declarative statements and an embedded expression language. The statements are grouped in different configuration sections depending on their purposes. Currently, we have defined and implemented rather a small subset of such language, which we present in this paper on a series of examples specifying network devices, routing areas, network areas, network connections, routing options and security policies.

The expressions can be embedded in declarations. In expressions we can refer to declared elements and predefined methods. In the future we plan to extend the language with possibility of defining user methods. It is important to note that these expressions represent the side effect free computations. The proposed expression syntax reassembles the syntax of object-oriented languages. We use dot notation to access fields of objects and call their methods. Usually, an expression is to be evaluated to a collection, a simple value or a structure, which can be inferred by type checking. Types of results have to conform to expected types of surrounding contexts. Declarations define attributes that can be accessed from expressions as shown in several examples in the rest of this section.

Each declaration block specifies a certain part of a network configuration, e.g., routing, address assginement, etc. To be treat as a programming language statement, it can be viewed as a macro definition, which evaluates to a corresponding program block. For instance, a device list from the next subsection can be viewed as the following list definition[1]:

```
var Devices = new[] {
  new Router("Austin", "A", "cisco_2811"),
  new Router("Dallas", "D", "cisco_2811"),
  new Router("Houston", "H", "cisco_2811")
};
```

Another characteristic of the language is that for specifying packet forwarding and filtering, a flow-based description [2] is employed.

[1]We use C# syntax in this example representation.

 ISBN: 978-1-61208-183-0

A. Device List

A device configuration group enumerates all devices in the configured network. A device declaration assigns a specific device type to each router, which tells the compiler what generator should be used for generating a configuration file. The compiler can be extended by custom generators for new platforms and models.

```
Devices {
    Router Austin[A] cisco_2811;
    Router Dallas[D] cisco_2811;
    Router Houston[H] cisco_2811;
}
```

The presented configuration snippet declares three routers appearing in Texas area and specifies their hardware platforms. Together with the full router name we may provide its short name that can be used for referring to the router from other places of the configuration file. For the language of expressions, this declaration defines a collection called `Devices`, which consists of three objects of type router. `Router` class is one of the classes derived from `Device` abstract class. Another derived classes could be `Switch`, `Gateway`, etc.

Currently, a device type specification consists of an enumeration of all device's interfaces, as shown in the following example:

```
Device cisco_2811 {
    PORT    Serial0/0/0 s0/0/0;
    PORT    Serial0/0/1 s0/0/1;
    PORT    FastEthernet0/0 fa0/0;
    PORT    FastEthernet0/1 fa0/1;
}
```

A device type specification is compiled into a plug-in module that is used by the language compiler for generating a device configuration for the specified device type.

B. Area List

The purpose of an area list is to define routing areas. Each routing area consists of routers which run the same instance of a routing protocol. The following is a definition of three different areas:

```
Areas {
    AREA {A, D, H} {A} RIP Texas;
    AREA {A,Tampa,M,T} {A,T} EIGRP Florida;
    AREA {R,S,T} {T} OSPF Washington;
}
```

Each area declaration consists of a list of area routers, a list of border routers, a definition of a routing protocol and a name of the area. A non-empty intersection of sets of border routers denotes routers where the redistribution between routing protocols can be configured. The redistribution options are stipulated in a routing configuration section.

C. Network List

A network list enumerates all destination networks. Each network declaration defines a network address and a network name, as shown in the following example:

```
Networks {
    Intermediate 192.168.1.0/30 Dallas-Austin;
    EndUser 192.168.2.0/24 management;
    EndUser 192.168.3.0/24 servers;
}
```

The network list does not include interconnecting networks except if these networks are significant from the viewpoint of routing or security configurations. The unspecified interconnecting networks are listed in a connection configuration block.

D. Connections

Connections among routers and destination networks are specified in a connection list. Point-to-point and point-to-multipoint connections can be distinguished. Following example contains several kinds of connections:

```
Links {
    A.s0/0/0 -> D.s0/0/0  Austin-Dallas;
    A.fa0/0 -> TERM       management;
    D.fa0/1 -> DEFAULT    ISP;
    Tampa.fa0/0 -> SWITCH Florida-Net;
    Miami.fa0/0 -> SWITCH Florada-Net;
}
```

A connection is specified by its endpoints and its name. An endpoint is either a router interface or one of the following keywords:

- `DEFAULT` denotes that the connection represents a default gateway for the network,
- `TERM` denotes that a router interface is connected to a destination network and there are no other routers in this network, and
- `SWITCH` denotes a router interface connected to a port of a switch.

Note that addresses of interfaces are generated automatically by the compiler. For interconnecting networks these addresses are taken from the pool of addresses that can be defined by the user.

E. Routing

A dynamic routing configuration is implicitly defined by specifying routing areas. In a routing configuration block, static routing, redistribution and other routing related options can be defined to customize the network routing. A following example shows redistribution of routing information from Florida routing instance to Texas routing instance. The redistribution is performed at Austin router, which runs both routing protocols. All routing information on end user networks maintained by the EIGRP in Texas routing domain will be copied to the RIP with a specified metric.

```
Routing {
    REDISTRIBUTE  Florida -> Texas
      END_USER_NETWORK METRIC 5;
}
```

Keyword `END_USER_NETWORK` selects what information is to be redistributed. In the presented example, all destination networks will be redistributed. At this position an arbitrary predicate that selects a set of redistributed networks can be used. For instance, we may write the following configuration:

 ISBN: 978-1-61208-183-0

```
Routing {
  REDISTRIBUTE  Florida -> Texas
    {Networks.Select(n => n.Name.StartsWith("D"))}
    METRIC this.Network.Name.Length;
}
```

The redistribution predicate selects all networks which names begin with letter 'D'. It means that when compiling and generating output for this statement, the expression is evaluated and replaced with a set of networks satisfying this condition. This configuration uses a weird policy for setting metrics. For each network, a metric is set to a value which equals to the length of its name. While this particular example is not very useful in practice, it demonstrates the use of `this` statement that refers to an object in which scope this expression occurs.

A purpose of a static routing configuration is to define preferred paths for network traffic. This is defined separately for each destination network. An example of static configuration is presented for `HoustonNet`. The static routing configuration consists of a subset of network links.

```
Routing {
  STATIC HoustonNet
    { Austin -> Houston,
      Tampa -> Austin,
      Miami -> Tampa }
}
```

It is also possible to apply a predefined algorithm to compute the best paths with respect to given criteria. The following configuration snippet demonstrates this approach:

```
Routing {
  STATIC HoustonNet
    SpanningTree(HoustonNet).Edges.Select
    (e => e.Contains(Austin|Houston|Tampa|Miami))
}
```

For computing the set of links we use `SpanningTree` algorithm, which computes a minimum spanning tree for `HoustonNet`. Then resulting set of links are filtered and only links which begin or end in one of four specified routers are kept in the configuration.

F. Security

Routers implement security policy by filtering traffic according to filtering rules maintained in access control lists (ACL). The configuration language is able to specify a security policy and the compiler generates ACLs and assigns them to appropriate interfaces.

First, a set of interesting flows is enumerated in a flow declaration block:

```
Flows {
  Web tcp any:any -> public:80;
  Mail (s => tcp any:any -> s:25);
}
```

Currently, flows are represented as tuples consisting of five components, namely, protocol type, source address, source port, destination address and destination port. Flows can be parametrized as can be seen in the case of `Mail` flow.

In filtering section, it is specified, which flows are permitted or denied on a particular link. In the following example, only web and mail traffic is permitted. Mail flow is instantiated with `TexasMail` server.

```
Filtering {
  Austin-Dallas {
    allow Web,
    allow Mail(TexasMail),
    implicit deny
  }
}
```

While flow-based security management brings a benefit of simplifying the implementation of security policy, it does not guarantee the correctness and consistency, because one still needs to pair filters with network locations. Along the line of proposals described in [3], [4], [5], [6], we would like to research the possibility to infer security implementations from high-level security policy specifications. Currently, we have attempted to apply techniques for filter consistency verification [7], [8], [9]. We implemented a simple tool, which reports conflicts for the given set of ACL rules. The employed method is based on work reported in [10].

III. Implementation Notes

We have implemented an experimental compiler and a configuration generator in the C++ language. Except the STL library, the compiler depends on the BOOST library, which provides data types and methods for manipulating advanced data structures. The configuration processing consists of the following steps:

1) Parsing an input configuration and generating a *network configuration object model*. This model is a structured description of parsed configuration amenable to further analysis.
2) Evaluation of expressions in the object model. The expressions are replaced with results yielded from their evaluation. After evaluating all expressions we obtain a concrete model.
3) Optional static analysis of the model. For instance, we may run an ACL conflict detection algorithm.
4) Generation of device configurations using plugins for registered device types. Based on the model, the tool generates for every known device its partial configuration by using the corresponding plugin.

Currently, the tool contains plugins only for a few devices and the expression language has a very limited form. In the future, we plan to extend the tool in both directions.

IV. Summary

In this short paper, we presented work in progress, which aims at the definition of a high-level network configuration language and the implementation of its compiler. The compiler produces device configuration files and it is extensible for different vendors and different router models. As it can be seen from the brief language description, the current state provides the basic functionality. The language is able to describe an enterprise network as a collection of devices and routing areas, to generate address assignment and to define basic security policies. The future work is focused on extending

 ISBN: 978-1-61208-183-0

the language with other features, e.g. NAT configuration, VLAN definitions, VPN configuration, policy routing, etc. For a first experimental implementation we decided to implement configuration generator only for CISCO devices. Currently, we are working on the support for other platforms.

The presented approach is directly comparable to Nettle language [1] and the FML [2]. These languages attempt to define a network configuration by specifying which services should be available rather than encoding the network behavior by using low level configuration commands. Nevertheless, there are other methods that simplify the network configuration. From industrial perspective, the major achievement in this area has been made by XML-based network configuration methods and protocols [11]. For instance, Juniper Networks introduced a Network Configuration Protocol called NETCONF, which was standardized by IETF as RFC4741. The protocol provides mechanisms to install, manipulate, and delete the configuration of network devices. The aim of Network Description Language (NDL) [12] is to simplify a description of networks and configurations by creating the ontology for computer networks based on the Resource Description Framework (RDF).

Our approach goes beyond merely introducing a new language for describing network configurations. Rather, we would like to construct network configurations by using a high-level configuration programming language. For this we set foundations in the presented language, which supports declarative statements with embedded expressions increasing the expressiveness and minimizing the need to repeatedly write routine configuration statements. There is a similarity to TCL scripting in IOS configuration, which can be employed to automatize certain tasks. Nevertheless, this scripting is rather limited to a single device.

The proposed language contains also concepts known from network configuration management tools. The NetScope toolkit [13], for instance, integrates topology model, traffic, and routing based on flows. It visualizes traffic and enables users to determine effects of configuration changes before they are applied to a real network. For security specifications, our language employs ideas described by Guttman in [14]. In particular, we attempt to generate access control lists from a security policy specifications.

The other line of research has been focused on configuration synthesis. Tools such as ConfigAssure [15] are able to refine or generate configurations for network devices based on a predefined configuration database and given constraints. This approach requires the implementation of advanced reasoning methods that perform model-finding. The goal of our tool is similar, but we employ less sophisticated techniques requiring that a user will provide the intended configuration by programming it in the proposed configuration language.

The presented paper briefly reported the first attempt to tackle the specified goal. We plan to extend the language with more advanced constructs, which would allow us to define a network specification in the modular manner that is typical for programming languages. This means that a network configuration would be split in modules logically representing network areas. These modules would have defined public interfaces through which interconnections are only possible. These are also points where security enforcement on the highest level is to be implemented. Thus it may be possible to hide internal structures of individual modules to simplify the configuration management.

To evaluate the presented approach we need to i) extend our language beyond the currently supported set of rather basic configuration blocks and ii) support more than a single target platform for which the configuration can be generated. Both are topics for the further work.

ACKNOWLEDGMENT

This work was partially supported by research programs MSM 0021630528 and CZ 1.05/1.1.00/02.0070, and the BUT grant FIT-S-11-1.

REFERENCES

[1] A. Voellmy and P. Hudak, “Nettle: A language for configuring routing networks,” in *Domain-Specific Languages*. Springer, 2009, pp. 211–235.

[2] T. L. Hinrichs, N. S. Gude, M. Casado, J. C. Mitchell, and S. Shenker, “Practical declarative network management,” *Proceedings of the 1st ACM workshop on Research on enterprise networking - WREN ’09*, p. 1, 2009.

[3] J. Guttman, “Filtering postures: Local enforcement for global policies,” in *IEEE Symposium on Security and Privacy*. IEEE Comput. Soc. Press, 1997, pp. 120–129.

[4] G. Stone, B. Lundy, and G. Xie, “Network policy languages: a survey and a new approach,” *IEEE Network*, vol. 15, no. 1, pp. 10–21, 2001.

[5] S. Narain, “Network configuration management via model finding,” in *Proceedings of the 19th conference on Large Installation System Administration Conference-Volume 19*. USENIX Association, 2005, p. 15.

[6] X. Ou, S. Govindavajhala, and A. Appel, “MulVAL: A logic-based network security analyzer,” in *Proceedings of the 14th conference on USENIX Security Symposium-Volume 14*. USENIX Association, 2005, pp. 8–8.

[7] E. Lupu and M. Sloman, “Conflict analysis for management policies,” in *Proceedings of IFIP/IEEE International Symposium on Integrated Network Management (IM1997)*, vol. 97, no. May. Citeseer, 1997, pp. 1–14.

[8] A. Couch and M. Gilfix, “Its elementary, dear Watson: applying logic programming to convergent system management processes,,” in *Proc. Lisa XIII*, 1999.

[9] A. X. Liu and M. G. Gouda, “Firewall Policy Queries,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 20, no. 6, pp. 766–777, Jun. 2009.

[10] F. Baboescu and G. Varghese, “Fast and scalable conflict detection for packet classifiers,” *Computer Networks*, vol. 42, no. 6, pp. 717–735, Aug. 2003.

[11] J. Hong, “XML-based configuration management for IP network devices,” *IEEE Communications Magazine*, vol. 42, no. 7, pp. 84–91, Jul. 2004.

[12] J. van der Ham, P. Grosso, R. van der Pol, A. Toonk, and C. de Laat, “Using the network description language in optical networks,” in *Integrated Network Management, 2007. IM’07. 10th IFIP/IEEE International Symposium on*. IEEE, 2007, pp. 199–205.

[13] A. Feldmann and A. Greenbreg, “NetScope: traffic engineering for IP networks,” *IEEE Network,March/April, 200o. 1l* , 2000.

[14] J. D. Guttman and A. L. Herzog, “Rigorous automated network security management,” *International Journal of Information Security*, vol. 4, no. 1-2, pp. 29–48, Dec. 2004.

[15] S. Narain, G. Levin, and S. Malik, “Declarative Infrastructure Configuration Synthesis and Debugging,” *Journal of Network and Systems*, pp. 1–26, 2008.

 ISBN: 978-1-61208-183-0

A New Approach to NGN Evaluation Integrating Simulation and Testbed Methodology

Marcial P Fernandez
Universidade Estadual do Ceará (UECE)
Fortaleza, Brazil
marcial@larces.uece.br

Sebastian Wahle
Fraunhofer FOKUS
Berlin, Germany
sebastian.wahle@fokus.fraunhofer.de

Thomas Magedanz
Technische Universität Berlin
Berlin, Germany
tm@cs.tu-berlin.de

***Abstract*—Since the beginning of Internet, network researchers have been proposing methodologies and tools to facilitate the design and development of new protocols for Internet. Analytical modeling, network simulation, network emulation, and more recently, testbeds, are being used in these researches. However, there are advantages and disadvantages in all these methodologies making difficult to decide on the ideal methodology and tool. In this paper, we propose a new methodology to evaluate Next Generation Networks that permits the integration of design, development and test of a new protocol or network service using different tools. The approach was demonstrated by designing a simple example of a network service.**

Keywords-Next Generation Networking (NGN); Network Evaluation; Network Simulation; Network Emulation.

I. INTRODUCTION

Design, development, and validation of networks protocols and services are important research issues. Generally, for analysis and comparison of different mechanisms and algorithms, five techniques are applied: analytical modeling, network simulation, network emulation, testbed and real-world experiments. Over the past 50 years, these methodologies were used to develop the protocols used on the Internet today. Although these techniques are known for many years, the predominant use of each technique over the years is dependent upon computer capacity. The potentials and limitations of these methods have been widely discussed by Jain [1].

Concerning the techniques being used currently, simulation, emulation and testbed, we can say that the first is the most distant from reality but is the easiest to work with, while the latter is the closest to real world but it is the most difficult for researchers to use. Regarding costs, simulation is cheaper than emulation and testbed. But because it is close to the real world, testbeds are hard to do, expensive, have a fixed topology, fixed environment, and it is difficult to create new impairment scenarios (broken link, routing table errors, high drops). These objectives further evolved towards refinement of experimentally-driven research as a visionary multidisciplinary research, defining the challenges for and taking advantage of experimental facilities, realized by means of iterative cycles of research, oriented towards the design and large-scale experimentation of new and innovative paradigms for the Future Internet - modeled as a complex distributed system.

A good methodology for the development of protocols and network services would be the use of both approaches: simulation and testbed. The first step, simulation, could be used to test and debug the first prototype, taking advantage of the facility of creating scenarios, traffic sources and network failures. In the second step, we could consider carrying out experiments on a testbed, now taking advantage of the reality offered by this methodology. Thus, it becomes very interesting to create a tool that permits the integration of these two methodologies to make the work of network researchers easier.

In this paper, we present a new tool to provide integrated simulation/experimentation environment to permit to develop protocols and services with four main contributions: provide a unifying approach to simulation/experimentation that makes the transition easy from simulation to network testbeds; provide a graphical interface to facilitate the topology creation and traffic definition; provide analysis tools to permit comparison of simulation and experimentation results; offer a layered and modular architecture to permit to evaluate specific parts without modification on the testbed facilities.

The rest of the paper is structured as follows. In Section II, we present some related works, Section III introduces the network research methodology and in Section IV we present our proposed methodology. Section V shows the proposal evaluation and the results and, finally, Section VI concludes the paper.

II. RELATED WORKS

Netbed/Emulab is a network testbed project, aiming to give network researchers an environment to develop, debug, and evaluate networked systems. Emulab project started as a emulation facility at the University of Utah [2], and consists of a cluster of emulation devices running an *ns-2* (Network Simulator Version 2) script [3]. One of the main contribution of this project was the ns-2 to emulation mapping [4]. The following works focused on implementing network

 ISBN: 978-1-61208-183-0

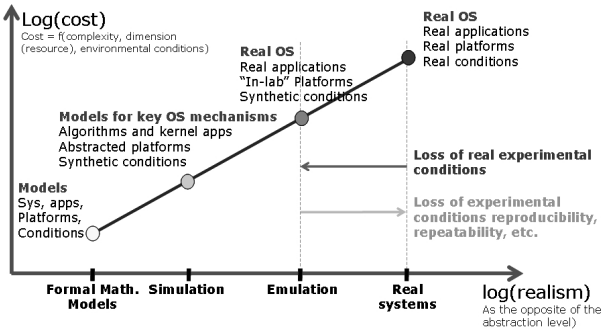


Figure 1. Network research methodology

emulation facilities on PlanetLab testbed [5] [6]. Other contribution of this project was the importance of simulation and emulation integration on network experimentation [7]. The Flexlab is a new framework that combines overlay and emulation testbeds (PlanetLab + Emulab), running an application within the emulation testbed and uses its load to measure the overlay network [8].

PL-VINI [9] is an implementation of VINI on PlanetLab. It runs on each PlanetLab slice providing network resources like link delay, link drop and routing. PL-VINI provides a realistic and controlled environment for evaluating new Internet protocols and services. Some features that could be evaluated in PL-VINI are: routing software, traffic loads and network events. To provide researchers flexibility in designing their experiments, VINI supports arbitrary network topologies on a shared physical infrastructure.

NEPI [10] is a framework proposal that makes the execution of a network experiment possible in different tools, e.g., simulation, emulation, and testbed. NEPI focuses on executing experiments using multiple tools separately and together in order to improve researchers productivity. The tool was implemented in Python and has a script and a GUI interface.

III. Network Research Methodology

The rationale was thus clear: to create a dynamic between elaboration, realization, and validation by means of iterative cycles of experimentation. Nevertheless the "validation by experimentation" objective opens a broad spectrum of experimentation tools (in large sense) ranging from simulation to real system experimentation. Our thesis is that "elaboration" requires validation by means of more abstract tools (not only because their resulting cost is lesser but because such tools produce results verifying all conditions explained here below) followed by progressive addition of realism as part of the experimented system to ultimately reach so called field trials with real systems. Thus, systematic experimentation is a continuum (Figure 1).

1) "Computer Communication/Networking" is characterized by two fundamental dimensions: distribution of a large number of dynamically interacting (non-atomic) components and the variation of their inner properties that in its turn influences these interactions. Thus compared to computer science, the distribution/interaction and the large number of elements composing the system add two fundamental dimensions to computer science "paradigms".

2) On the other hand, one shall characterize the output of experimentation: in order to ensure verifiability, reliability, repeatability, and reproducibility of the experimental results. Ensure these properties implies in provide strict control to the experimental conditions (parametrization, i/o, and running). Verifying the repeatability, reproducibility, and reliability conditions ensures generalization of experimental results, and verifiability of their credibility.

3) Different experimental tools can be used. As stated above their selection is neither arbitrary nor religious: it depends on the experimental objective and maturity of the experimented corpus. Nevertheless, each of them needs to ensure that the conditions defined here above are verified. However it is clear that fulfilling these conditions does not come at the same cost for the same level of abstraction. Validation of a new algorithm would be better conducted on a simulation platform (after formal verification) not only because their resulting cost is lesser but because such tools produce results verifying all conditions explained here above. Emulation experiments can lead to reproducible and repeatable results but only if "conditions" and "executions" can be controlled. Realism can thus be improved compared to simulation (in particular for time-controlled executions of protocol components on real operation system).

A. Simulation

Network simulation is a technique in which a software simulates the behavior of a network and its components (routers, hosts, links, protocols, etc) by calculating the interaction between them using mathematical models. Most network simulators use discrete-event simulation, in which a list of events are processed according to a virtual time, independently of the computer's clock where the simulator software is running. Then, a simulation produces the same result in different computers. Since the beginning of Internet, network simulators have been an invaluable tool for network researchers.

The ns-3 is a discrete-event network simulator, intended to replace the traditional ns-2 simulator [3]. The first release of ns-3 was published in July of 2008 and it has been improved and extended since then. Since it was proposed, ns-3 concept was to be a simulator capable to interact with real world. Some improvements pointed in this direction, e.g., the ns-3 API is a Unix socket-like API, to permit easy migration from simulated code to real-system code, and the Network Simulation Cradle (NSC) allows ns-3 using the Linux TCP/IP stack.

 ISBN: 978-1-61208-183-0

B. Emulation

The Emulation is the technique where a network is simulated in a real hardware and software. The emulation platform implements virtual network topologies and scenarios over real hardware and protocols, i.e., that experiments can be executed in real hardware, use real operating systems and protocols, run their real applications, and obtain actual (not simulated) performance measures. Although emulation is much closer to real environment than simulation, the links should be simulated in order to create delay and communication impairments (noise, drops, etc). Sanaga et.al. [11] shows the difficulties to emulate a network link.

IV. A New Methodology for Network Research and Experimentation

The development of new protocols and services for Internet requires a series of procedures before it can be used in the real world. The first one is to verify whether it works, i.e., the protocol or service performs what we want. Then we must explore the parameter space to find the best configuration to achieve the best trade-off. Thus, if the protocol or service is working, we must verify that it will not kill the network. Finally, we need to perform a couple of experiments to see the overall performance, and scalability. The network research environment must provide:

Reality: the proposal should be tested in real environment.

Configuration: large scale experiments require a lot of configuration.

Instrumentation: need to gather data about the behavior of the experiment to figure out what happened.

Fidelity: did the experiment really capture the effects you are really interested in?

Reproducibility: scientific methodology means that you must publish reproducible results.

The methods currently used to develop protocols and services for Internet are the simulation and the emulation testbed. However, there are advantages in both of them, not found in the other, so instead of comparing both methodologies, we associate them. The first step, simulation, allows easy creation of different scenarios, as well as different types of traffics. As the prototype is running on a computer, we can create a series of experiments that will allow evaluation of the prototype operation in many different situations. The simulation makes easy the creation of impairment in environment, such as link break, link degradation (increased of drop rate), very long delays, routing instability, evaluating the prototype in very adverse situations.

Simulation provides facility to change the source code (we do not need to change any code in multiple remote machines, only on the simulation server) also facilitating the prototype development. The code can be changed and tested very quickly. Another advantage of simulation is the possibility of taking execution snapshots, e.g., we can simply put a *printf* in simulated code. In a testbed, it is often difficult to change the code and create mechanisms for collecting information, making the code debug hard. Since the simulator environment is controlled, it is possible to obtain the repeatability necessary to validate a scientific work. The creation of traffic sources statistically distributed in a controlled environment allows repeatability, which is very difficult to obtain in a testbed due to the difficulty to reproduce the same situation and events in a certain moment. However, simulations tend to be unrealistic. The packets do not pass through a real network, so even if it used sophisticated simulation tools, it continues far from reality in some scenarios. Emulation can provide a more realistic environment because it uses real machines and real operation system. But it is also difficult to emulate the reality, e.g, Sanaga et al [11] shows the difficulties to emulate an Internet path.

Nowadays testbeds are presented as the ideal methodology to develop and test protocols and services for the Internet. As it is an extract from a real network, tests are performed in an actual infrastructure. The possibility of setting up paths and the monitoring tools offer a control degree that allows the creation of specific test situations. But as it is a separated experimental environment, there is no risk to damage the production network. Despite its proximity to the real world, Testbeds are not the ideal tool. Setting up a testbed is complex and may require individual configuration of each resource. The repeatability of an experiment is difficult to be achieved, given the environment unpredictability. Another difficulty is to perform measurements on a testbed compared to simulation, usually performed through measurement points placed on routing devices. Most of testbed provides the collecting of statistics information (e.g., sFlow) or collecting flow traces (e.g., *pcap* format).

As Testbeds could be categorized as an Overlay Emulation, i.e., Testbeds run over production networks. It is difficult to evaluate routing mechanisms, because we use the real network routing environment. This imposes limit to evaluate tests of low level protocols (layers 2 and 3). The implementation of the Click routing engine in testbed performs poorly [9].

Our proposal is to offer an integrated tool to evaluate Internet protocols and services joining simulation and emulation in a testbed. Figure 2 shows a diagram of the proposed solution. We can visualize the two prototype development steps, the first running in a simulation and the second running in a testbed. In the first step, the designer can debug the code in different environments. By the end of this step the code has been debugged and probably, it does not have serious flaws. As we run the experiment in simulation, we can test in a vast range of topologies and network scenarios in order to validate the proposal and explore the parameter space. In the second step, we need to test the code on a testbed, closer to reality. At this point we can make a

 ISBN: 978-1-61208-183-0

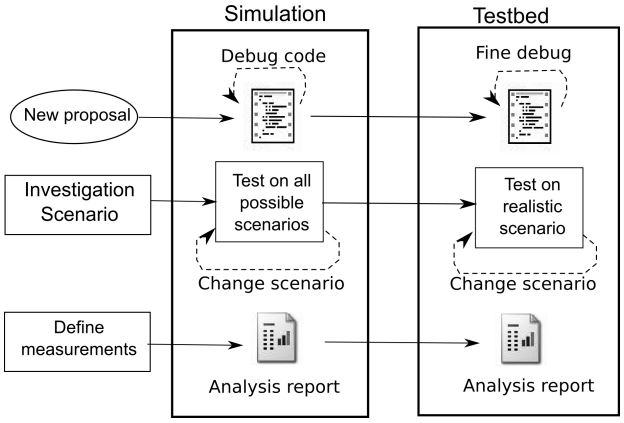


Figure 2. Proposed methodology

fine debug and fine parameters tuning in an (almost) real environment. As we saw, the testbed has less resources to capture information, but now we should be satisfied with some statistical traffic information.

A. Teagle Framework

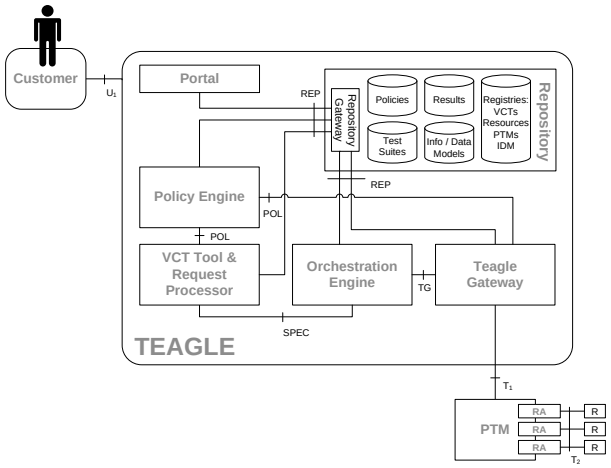


Figure 3. Teagle architecture

Reference [12] defines a federation model and framework that allows users to get access to distributed resources and group them into a virtual environment, which is called Virtual Customer Testbed (VCT). Teagle as a control framework and collection of central federation services helps the user in configuring and reserving a desired set of resources. Resources are offered by participating organizations across Europe.

On the federation layer, our Teagle framework implementation offers several services to the user and other framework entities, such as the registry and a common information model. Teagle allows browsing through the federation offerings, enables the definition of VCTs, and executes their provisioning. Figure 3 shows the Teagle architecture that is detailed in [12].

B. Teagle VCT Tool

The Teagle VCT offers a graphical tool (GUI) that permits the user describe the network topology and parameters to simulation and testbed evaluation. Before starting the experiment description, the user needs to define the kind of experiment: simulation or testbed. Then, Teagle framework will create a simulation script or a testbed setup file. The user can design the test topology using a set of graphical objects interconnected by arrows. Three different components to design an experiment were defined in Teagle VCT: Resource, Connector, and Monitor.

1) Resource: Resource represents a functional unit in a network system. A Resource could be a hardware device, like a Node or Link, or a software entity, like a protocol or a traffic generator. A Resource has Attributes and Events.

- *Attribute* is the configuration parameter of a Resource, which can be defined before the experiment and can be changed while the experiment is running. An example of attribute is the node IP address or the link bandwidth. In some Resource is possible to define the experiment planning, i.e., the sequence of values in experiments that should be performed, e.g., packet length of 100, 500 and 1000 bytes.
- *Events* is the set of timed events that will happen during the experiment. The Event time is based on virtual time independent from the real time. An example of Event is the start and stop time to transmit a traffic or to interrupt a link.

The Figure 4 shows the *Resource Application* configuration and the *Event* definition of a specific Application. The *Application Client* box has a ***cfg*** button that permits the configuration of its attributes like ClientType (Sink or Echo), Packet Length, Data Rate, and Port. The ***cfg*** windows also defines the *Experiment Planning*, i.e., the definition of various experiments will be executed. In this example, we use five different Random Seeds, five different Data Rates, from 10 pkt/s to 50 pkt/s, and three different packet lengths, 500, 1000, and 1500 bytes. It is important to notice that the total quantity of experiments, simulation or testbed, will be the combination of all different attributes, in our example $5 * 5 * 3 = 75$ experiments. The ***rules*** button defines the Events that will happen in this Application, e.g., the data stream starts at 2.0 sec and stops at 14.0 sec.

2) Connector: Connector is a representation of an interconnection from a Resource to another Resource. For example, the Resource Node is connected to Resource Link to build the topology. It is important to notice that the Connector is an abstract component only to permit joining different Resources, i.e., a Link is a Resource not a Connector although we use a Link to connect routers.

 ISBN: 978-1-61208-183-0

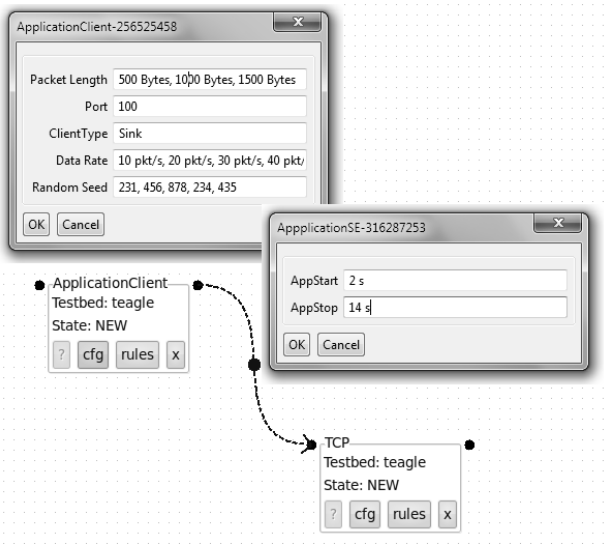


Figure 4. Teagle VCT Application configuration

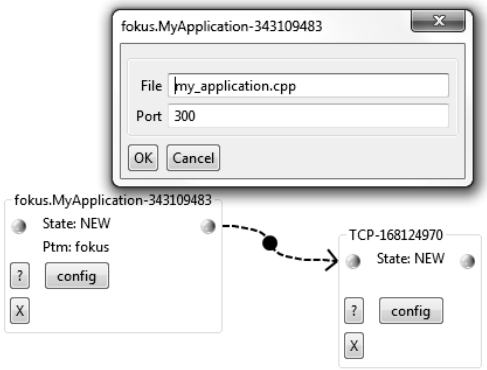


Figure 5. Creating a new application on Teagle VCT

3) Monitor: Monitor is a special Resource used to monitor the experiment and collect information about the experiment. However, Monitor does not interfere in network experiment. We consider using a *pcap* collector that collects useful information at simulation environment and also at testbed environment.

C. Creating a new protocol or service

The key feature of Teagle is the development and test of new protocols and services. For that, you can create a new module that will implement the desired protocol or service. Suppose that the user will design an application level module.

Figure 5 shows the definition of the *MyApplication* module that uses TCP protocol. The configuration sets the protocol port and the filename that contains the C code of the test protocol. The interface with Teagle is based on the Unix Socket, so the protocol implementation should be very similar to actual interface. To facilitate the development, a code template with the interface definition and the suggestion of the most common methods is provided.

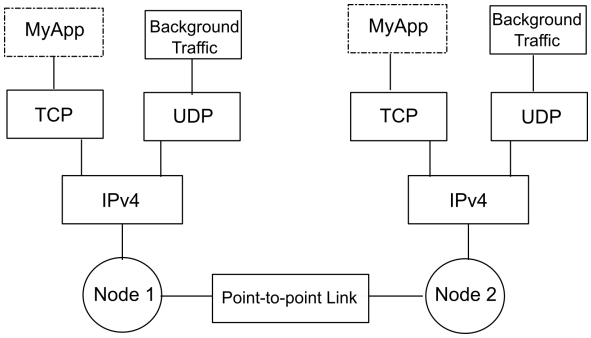


Figure 6. Test topology to validate the proposal

D. Mapping Teagle to ns-3

The Network Simulator version 3 – *ns-3* – was chosen as simulation framework. This version uses a network interface similar to Unix Socket, making easy the Teagle VCT translation to simulation script and Testbed configuration specification. However, the use of ns-3 as a standard tool to Teagle does not invalidate the creation of simulation models in other simulators, if it provides an abstract interface based on Sockets. The Teagle components are similar to ns-3 modules, then the translation is almost direct.

E. Mapping Teagle to Testbed

The Teagle platform aims to coordinate the execution of experiments in a Testbed federation. So naturally, an experiment specification in Teagle can be converted directly to Testbed configuration. However, an application performance validation tests requires more functionality than the standard Testbed platform can offer. Teagle offers the opportunity to create new functionality in a testbed in order to expand the scope of testing to be performed.

F. Analysis of results

Carrying out experiments on two methodologies and tools from a unique specification is not difficult because the components used are similar (nodes, protocols, applications). The great problem is the analysis of the results produced in different environments. However, the ns-3 simulator can produce a file in *pcap* format. In the testbed, it is possible to capture traffic in *pcap* format using tools like *tcpdump*. Comparing both *pcap* files is possible to analyze the results and the conclusions.

V. Proposal Evaluation and Results

Aiming to validate the proposed model, we created a simple test scenario that would allow to run a prototype testing. This experiment does not intend to validate a protocol or service, but it only intends to demonstrate the proposed methodology validation. The testing scenario is

 ISBN: 978-1-61208-183-0

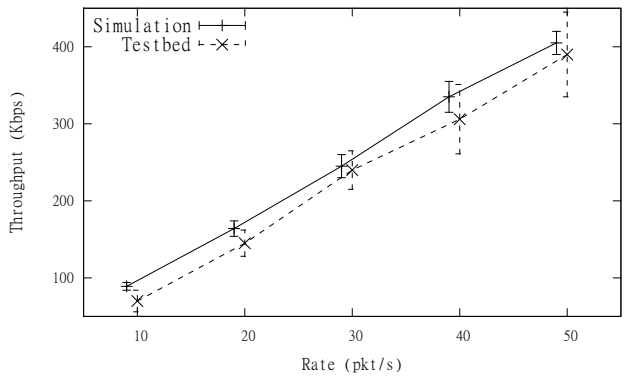


Figure 7. Bandwidth test result

shown in Figure 6, demonstrating a simple application with a background traffic. We have two nodes connected with a link and two applications: one is the proposed protocol running over TCP and a background application over UDP protocol. We run 10 experiments on both environment, simulation and testbed, and confidence interval is calculated.

The result graph is shown in Figure 7. The simulation environment is more controlled than the testbed environment and results tend to be different even with the same parameters. However, although we might have expected that the testbed results are not identical to the simulation, the results are very similar, mostly inside the confidence interval.

VI. Conclusion and Future Works

This paper presented a new methodology for developing and testing protocols and services using simulation and testbed. A single interface for the end user is the Teagle tool, Teagle Simulation and Emulation, enabling it to perform a test in simulation and emulation from the same specification in Teagle-VCT GUI. In both experiments, the technique chosen to collect and evaluate the results of the protocol under test performance and operation was the *pcap* files. They were generated by ns-3 and collected in the testbed using tcpdump software. To demonstrate the feasibility, a prototype model was developed using the Network Simulator ns-3 and the PANLAB testbed. The results were analyzed by comparing the *pcap* files generated in both experiments, which demonstrated the feasibility of the proposed model.

As future work, we wish to improve the collection of information in *pcap* files, which produces large files that require much processing capacity to analyze. One possibility is defining a filter to choose the specific information before the test that we want to collect in Teagle-VCT. It will reduce the amount of information stored. The Teagle-VCT specification translation considered only basic objects, so it becomes necessary to increase the amount of new objects to allow more functionality to the researcher.

References

[1] R. Jain, *The art of computer systems performance analysis: techniques for experimental design, measurement, simulation, and modeling*. Wiley New York, 1991.

[2] B. White, J. Lepreau, L. Stoller, R. Ricci, S. Guruprasad, M. Newbold, M. Hibler, C. Barb, and A. Joglekar, "An integrated experimental environment for distributed systems and networks," in *Proc. of the Fifth Symposium on Operating Systems Design and Implementation*. Boston, MA: USENIX Association, Dec. 2002, pp. 255–270.

[3] T. Henderson, S. Roy, S. Floyd, and G. Riley, "ns-3 project goals," in *Proceeding from the 2006 workshop on ns-2: the IP network simulator*. ACM, 2006, p. 13.

[4] R. Ricci, C. Alfeld, and J. Lepreau, "A solver for the network testbed mapping problem," *ACM SIGCOMM Computer Communication Review*, vol. 33, no. 2, p. 81, 2003.

[5] K. Webb, M. Hibler, R. Ricci, A. Clements, and J. Lepreau, "Implementing the Emulab-PlanetLab portal: Experience and lessons learned," in *Proc. WORLDS*, 2004.

[6] M. Stoller, J. Duerig, S. Guruprasad, T. Stack, K. Webb, and J. Lepreau, "Large-scale virtualization in the emulab network testbed," in *USENIX Annual Technical Conference, Boston, MA*, 2008.

[7] S. Guruprasad, R. Ricci, and J. Lepreau, "Integrated network experimentation using simulation and emulation," in *Testbeds and Research Infrastructures for the Development of Networks and Communities, 2005. Tridentcom 2005. First International Conference on*, 2005, pp. 204–212.

[8] R. Ricci, J. Duerig, P. Sanaga, D. Gebhardt, M. Hibler, K. Atkinson, J. Zhang, S. Kasera, and J. Lepreau, "The Flexlab approach to realistic evaluation of networked systems," in *Proc. of the Fourth Symposium on Networked Systems Design and Implementation (NSDI 2007)*, Cambridge, MA, 2007.

[9] A. Bavier, N. Feamster, M. Huang, L. Peterson, and J. Rexford, "In VINI veritas: realistic and controlled network experimentation," in *Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications*. ACM, 2006, p. 14.

[10] M. Lacage, M. Ferrari, M. Hansen, T. Turletti, and W. Dabbous, "Nepi: using independent simulators, emulators, and testbeds for easy experimentation," *SIGOPS Oper. Syst. Rev.*, vol. 43, no. 4, pp. 60–65, 2010.

[11] P. Sanaga, J. Duerig, R. Ricci, and J. Lepreau, "Modeling and emulation of Internet paths," in *Proceedings of the 6th USENIX symposium on Networked systems design and implementation*. USENIX Association, 2009, pp. 199–212.

[12] S. Wahle, B. Harjoc, K. Campowsky, and T. Magedanz, "Pan-European testbed and experimental facility federation–architecture refinement and implementation," *International Journal of Communication Networks and Distributed Systems*, vol. 5, no. 1, pp. 67–87, 2010.

 ISBN: 978-1-61208-183-0

Impact of Propagation Factors on Routing Efficiency in Wireless Mesh Networks: A Simulation-based Study

Ewa Osekowska, Iwona Pozniak-Koszalka, Andrzej Kasprzak
Department of Systems and Computer Networks
Wroclaw University of Technology
50-370 Wroclaw, Poland
e-mail: ewa.osekowska@bth.se, iwona.pozniak-koszalka@pwr.wroc.pl, andrzej.kasprzak@pwr.wroc.pl

Abstract—**The low installation and maintenance costs, self-healing abilities and the ease of development are some of the qualities that make the multi-hop wireless mesh network a promising alternative to conventional networking in both – rural and urban areas. This paper examines the performance of such a network depending on environmental propagation conditions and the quality of applied routing protocols. This aim is addressed in an empirical way, by performing repetitive multistage network simulations followed by a systematic analysis and a conclusive discussion. This research work resulted in the implementation of an experiment and analysis tools, and a comprehensive assessment of a group of simulated wireless ad-hoc routing protocols.**

Keywords-routing protocol; simulation; wireless mesh network; propagation shadowing

I. INTRODUCTION

The object of these investigations is the wireless mesh network (WMN). It is a multi-hop network that consists of stationary mesh routers, strategically positioned to provide a distributed wireless infrastructure for stationary or mobile mesh clients over a mesh topology [1]. WMN provides more robust, adaptive and flexible wireless Internet connectivity to mobile users compared to conventional local wireless networks (WLAN) or mobile ad hoc networks (MANET). It offers relatively low installation and maintenance costs, self-configuration and self-healing ability, thus ensuring more reliable connection and enlarging the covered area [2].

Routing is a crucial factor influencing connectivity and information exchange across the network. The flexibility, self-configuring and healing, as well as general performance of WMNs is highly dependent on the choice of a routing protocol and the quality of its implementation.

The objective of these investigations is to evaluate the performance of WMNs influenced by propagation factors referred to as signal shadowing. The evaluation is based on network simulations applying a group of commonly used WMN routing protocols, containing the Highly Dynamic Destination-Sequenced Distance-Vector (DSDV), the Ad hoc On-Demand Distance Vector (AODV), the Optimized Link State Routing (OLSR) and the Hybrid Wireless Mesh Protocol (HWMP). The analysis and discussion requires collating the protocols and conducting a comparison by experimentally simulating scenarios of environments with different propagation conditions. The result is an assessment of protocols suitability for the WMN and an evaluation of the overall network performance influenced by propagation conditions.

There have been conducted numerous researches investigating wireless network's performance with respect to the issue of routing [2][3]. A similar comparative evaluation of routing protocols was conducted by Zakrzewska et al. [13]. This work extends previous research in WMN routing by investigating the influence of propagation factors.

The rest of the paper is organized as follows: In Section II, the considered protocols are briefly described. Section III presents the experiment environment and simulation scenarios. Section IV shows the results and discusses the propagation impact. The final remarks appear in Section V.

II. ROUTING IN WMN

The investigated routing protocols are used in 802.11b/g/n standard based WMNs. The selection of DSDV, AODV, OLSR and HWMP was dictated by the intention to investigate a wide spectrum of approaches towards topographical routing. These protocols can be divided into the classical groups of distance vector and link state routing protocols, as well as hybrid protocols that have characteristics of both. The common WMN routing protocols differ also in terms of the events triggering the routing information exchange. Some protocols use a proactive mechanism repeating broadcasts in regular intervals of time. Others exchange the information in reaction to current data transmission and other events.

A. DSDV

DSDV is historically the first of the investigated routing protocols. It operates on ad hoc networks induce inferring a cooperative engagement of mobile hosts without a required intervention of any centralized access point [3]. It specifies each mobile host as a router, which advertises its view of the topology to other mobile hosts within the network [4], by periodically and incrementally broadcasting own routing table entries. DSDV determines the shortest route to a destination, i.e., a route with least intermediate hops.

DSDV construction is based on the basic Bellman Ford (BF) routing mechanisms, as specified by routing internet protocol (RIP), adjusting it to a dynamic and self-starting network mechanism required in ad hoc networks [4]. The

 ISBN: 978-1-61208-183-0

modifications concern, e.g., poor looping properties, such as the counting to infinity problem. Furthermore, in order to damp out fluctuations in route table updates DSDV also includes a sequence number and settling-time data.

There are significant limitations in DSDV protocol, i.e., it provides only a single path between each given source and destination pair [5]. Furthermore, the protocol's performance is highly dependent on selected parameters of periodic update interval, maximum value of the settling time and the number of update intervals. These parameters likely represent a trade-off between the latency of valid routing information and excessive communication overhead [5].

B. AODV

The AODV routing protocol offers the ability of quick adaptation to dynamic link conditions, low processing and memory overhead, low network utilization, and determines unicast routes to destinations within the ad hoc network. Similarly to DSDV, AODV uses destination sequence numbers to ensure the elimination of loops [6], but unlike DSDV, it does not require nodes to maintain routes to destinations that are not active in communication.

The AODV operations require Route Request (RREQ) messages, to be disseminated among a range of network nodes [6]. Despite from RREQ the AODV protocol defines the Route Reply (RREP), and Route Error (RERR) routing messages improving the efficiency of finding routes.

The on-demand character of the protocol implies that as long as the endpoints of a communication connection have valid routes to each other, no routing messages need to be sent. The information is kept in route tables, which (like in DSDV) store entries for all, even short-lived routes. Among the added fields of table entries are the valid destination sequence number flag and the list of precursors [6].

AODV is designed for use in networks, where the nodes can all trust each other, either by use of preconfigured keys, or based on known fact of no malicious nodes. It has been designed to improve the wireless network scalability and performance and eliminate overhead on data traffic.

C. OLSR

OLSR protocol is an adaptation of the wired Link State Routing (LSR) algorithm, specifically designed to serve the needs of mobile ad hoc networks (MANET) [7]. The main adjustments tackle reduction of administrative data exchange and increase the overall protocol performance.

Each router in an OLSR routed network owns a complete representation of the whole network topology and maintains this information periodically by exchanging topology information with other nodes. This makes OLSR a member of the proactive and link state routing algorithms family.

OLSR exchanges information by the means of messages HELLO and Topology Control (TC) [7]. They are used to sense the links between nodes in direct neighborhood. Based on responses from the other nodes each node selects an individual subset of neighbors, which are from then on referred to as Multipoint Relays (MPR). MPR's task is to execute the information exchange called flooding in its part of the topology. Therefore, each of the MPRs sends TC-messages containing local topology information to their respective MPRs, while forwarding received topology information to their non-MPR neighbors [7]. Hence, OLSR ensures a complete distribution of routing information, and limits the flooding data overhead only to MPR nodes. This design aims to lower administrative data exchange and improve scalability to network size and density.

Although OLSR is a young protocol, it is already used as a major WMN routing protocol e.g., by the Freie Funknetze in Berlin, Germany. It is criticized, though, for its large energy consumption due to constant data exchange and large topology databases.

D. HWMP

The HWMP is a mesh routing protocol that combines the flexibility of on-demand routing with proactive topology tree. The reactive and proactive elements of HWMP are combined in order to enable optimal and efficient path selection in mesh networks (with or without infrastructure).

The HWMP protocol uses a set of protocol primitives, generation and processing rules taken from AODV, adapted for Layer-2 address-based routing and link metric awareness. The AODV mode is used for finding on-demand routes in a mesh network, while the optional proactive mode sets up a distance vector tree rooted at a single root mesh node [8].

The control messages in HWMP are the RREQ, RREP, RERR – introduced in AODV, and an additional Root Announcement (RANN) message. The metric cost of the links determines which routes HWMP builds. The needed information is propagated between mesh nodes in the metric fields of RREQ, RREP and RANN messages [8]. The loop free routing is ensured by the use of the sequence numbers.

In the experimental phase of this research, only the on-demand mode of HWMP is enabled, thus it is qualified as a reactive routing protocol.

III. SIMULATION MODELING

The group of chosen routing protocols is compared based on an experiment using ns 2.34 network simulator (licensed for use under version 2 of the GNU General Public License). The choice of this simulator is motivated by its advantages, among which are: open source code, variety of implemented protocols and contributed code [10], as well as the reliability confirmed by the common usage for research purposes [13].

The simulated scenarios represent a structure of a hybrid WMN [9]. This means that all nodes have the mesh routing capabilities. The backbone of the network is formed by more powerful and completely stationary routers covering the topology in shape of a regular square grid. They provide the wireless infrastructure to the mobile users placed in random locations, which also support meshing and improve the internal network coverage [9].

Apart from mobility, router and mobile node properties differ in the matter of transmission capabilities, namely the transmitting power and the receiving threshold. On the sending side of communication, the initial packet signal power is regulated by a transmitting power value [10]. The receiving is limited by a threshold, which is assigned to a wireless node and determines the minimum value of packet's signal power required to succeed with its delivery. If the packet's signal power at the destination node does not reach

 ISBN: 978-1-61208-183-0

the receiving threshold value, it is marked as error and dropped by the Media Access Control (MAC) layer [10].

A. *Radio Signal Propagation*

The signal power fluctuates in a way determined by the phenomenon of radio wave propagation [11], which leads to the next issue of simulating wireless communication, namely radio propagation models. In general, these models predict the received signal power of each packet. There are three propagation models in *ns* [10].

The free space model assumes ideal conditions with a clear line-of-sight between the transmitter and the receiver [10]. This model represents the communication range as a circle around the transmitter. If a receiver is within the circle, it receives all packets. Otherwise, it loses all packets.

The two ray ground reflection model gives a more accurate prediction at a long distance than the free space model, based on considering both – the direct- and ground reflection paths. Still, this model also predicts the received power as a deterministic function of distance representing the communication range as an ideal circle (Fig. 1).

Those are the acceptable and commonly used simplifications of radio wave propagation for most of simulation based research. However, an attempt to investigate realistic conditions requires determining the received power at certain distance by a more complex computation. It is due to multipath propagation effects, which are also known as fading effects. These are taken in consideration in the shadowing propagation model [10]. This model redefines the calculation of the mean received power at a distance, making it dependent on the value called path loss exponent, which also enables a user to manipulate the propagation mechanism in simulations.

The signal power is reduced gradually with raising distance from transmission source, representing the communication range as a fuzzy circle. The diagrams (Fig.. 1, 2) were developed for the *ns* simulator and published by the Institute of Telematics at the Hamburg University of Technology, Germany to demonstrate the differences between propagation models. The upper graph shows the probability of receiving a packet by the middle horizontal line of nodes. The other graph is a 2D area plot representing the probability of receiving packets as grayscale points, where the darker the shade is – the higher the probability.

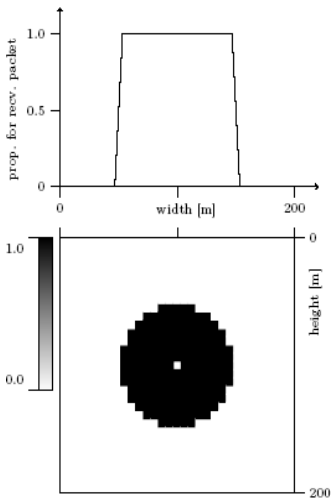


Figure 1. Probability of receiving packet – two ray ground model.

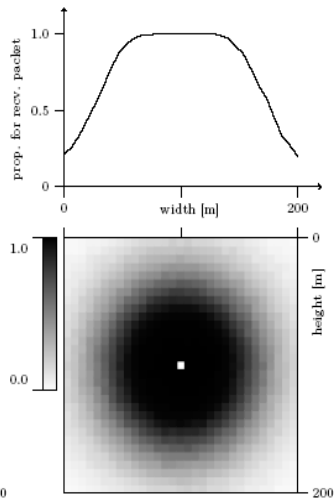


Figure 2. Probability of receiving a a packet – the shadowing model.

The shadowing model (Fig. 2) fulfills the description of IEEE 802.11 physical layer definition, which implies using a medium that has no readily observable boundaries, outside of which stations with conformant physical layers transceivers are known to be unable to receive network frames [11].

Furthermore the shadowing model introduces the shadowing deviation factor, which reflects the variation of the received power at a certain distance by time-varying and asymmetric propagation properties [10][11]. This prevents unrealistic representation of communication range as a circle, which was the case for other propagation models. It is also most probably the only way to simulate the presence of physical obstacles causing the signal power fluctuation in wireless network topologies. The intensity of this fluctuation is controlled by the shadowing deviation parameter [10].

IV. RESULTS DISCUSSION

In this section, the performance of four investigated routing protocols is compared based on the WMN simulations with User Datagram Protocol (UDP) traffic. The simulations were carried out for the hybrid WMN topologies with the following settings:

Topology size	width	- 300m
	length	- 1200m
Amount of nodes	total	- 58
	mobile	- 36
	backbone	- 22
Mobile node speed	maximum	- 5m/s

The experiments were performed in order to observe the influence of two varying propagation parameters:

Shadowing deviation	min. 3 - free space
	max.12 - outdoor, very obstructed
Path loss exponent	min. 2.0 - free space
	max. 4.4 - urban shadowed area

The performance is measured using metrics well describing performance of wireless networks [12][13]. The choice of metrics was dictated by the need of both – precise analysis as well as legible and intuitive representation.

Delivery ratio (DR) – the percentage of successfully delivered packets calculated as the total amount of data packets received at the destinations, divided by the amount of all data packets generated by the sources [12].

End to end delay of data packets (EED) – the average time passing from the moment of sending a data packet to its delivery, measured in milliseconds [13], including all delays such as route discovery latency, interface queuing and retransmissions, as well as propagation and transfer times [12]. The delay for individual hops is not measured.

Normalized routing load (RL) represents the relative content of routing packets. Here it is expressed as the amount of all sent and forwarded routing packets divided by amount of delivered data packets, thus each hop-wise transmission of a routing packet is counted [12].

 ISBN: 978-1-61208-183-0

System throughput (ST) – the aggregate amount of data measured in bytes delivered at all nodes in a given period of time. The unit of the AT is kilobyte per second. In opposition to the received throughput (RT) metric [12], the ST reflects the summary amount of both – data and routing traffic.

A. *Experiment 1. Impact of the shadowing deviation*

The results are collated in diagrams (Fig. 3), where each plot represents the performance of one routing protocol.

The packet DR reaches the maximum, i.e., most desired value of nearly 65% for minimum shadowing deviation, referring to environments with a clear line of sight (e.g., a factory). It holds true for all tested routing protocols. Increasing shadowing deviation lowers the DR. The decrease for the HWMP exceeds 40%; it is smaller for AODV – ca. 35%, whereas for OLSR and DSDV protocols it does not reach 30%. Nevertheless, the DSDV protocol is visibly the least effective – offering more than twice smaller DR in comparison to any other protocol. The favorites are AODV – for the highest DR in this experiment and OLSR – for slightly lower but more stable DR.

The same two routing protocols are leading, taken into account the EED. The HWMP and DSDV plots show similar results, mostly between 4 and 5s. These delays are over two times longer than for AODV protocol. The unquestionably best EED times are reached using OLSR protocol.

The results for OLSR seem implausibly small in comparison to other protocols, however revising the analysis procedure as well as manual analysis of parts of trace files confirms correctness of these EED outcomes based on simulations performed using the ns simulator.

AODV and OLSR produce comparable amount of routing packets sent per one successfully delivered data packet. For AODV it grows from circa 1.5 routing packets for shadowing deviation equal to 3, to almost 4 for maximal deviation. The range for OLSR is smaller but the values are higher – from 2.4 to 4.3. This amount for HWMP is twice smaller; also its increase trend is not as strong. The RL is smallest and nearly stable for DSDV routing protocol. That property corresponds with the proactive character of this routing protocol. Nevertheless, the benefit of a small amount of DSDV routing packets is overweighed by the disadvantage of their large size.

ST reflects the network load and it is not predestined to represent the speed of data transmission. The complete observation requires collating ST versus the DR and RL metrics. The largest ST of almost 1.7Mbps, and thus the greatest network load is generated by the DSDV protocol. This result confronted with low DR puts DSDV in the last place. The correlation with stable and low RL leads to an interesting finding. The amount of successfully transmitted data is small; the number of generated routing packets – smallest of all; the ST on contrary is the biggest. Then the size of the broadcasted routing information must be very large. The second largest ST, circa 1.1Mbps is generated for the HWMP, followed by 600 to 700kbps for the AODV protocol. The smallest result is reached using the OLSR.

This observation points to the conclusion that the optimal network performance for simulated scenarios is reached using OLSR, and the worst for DSDV protocol.

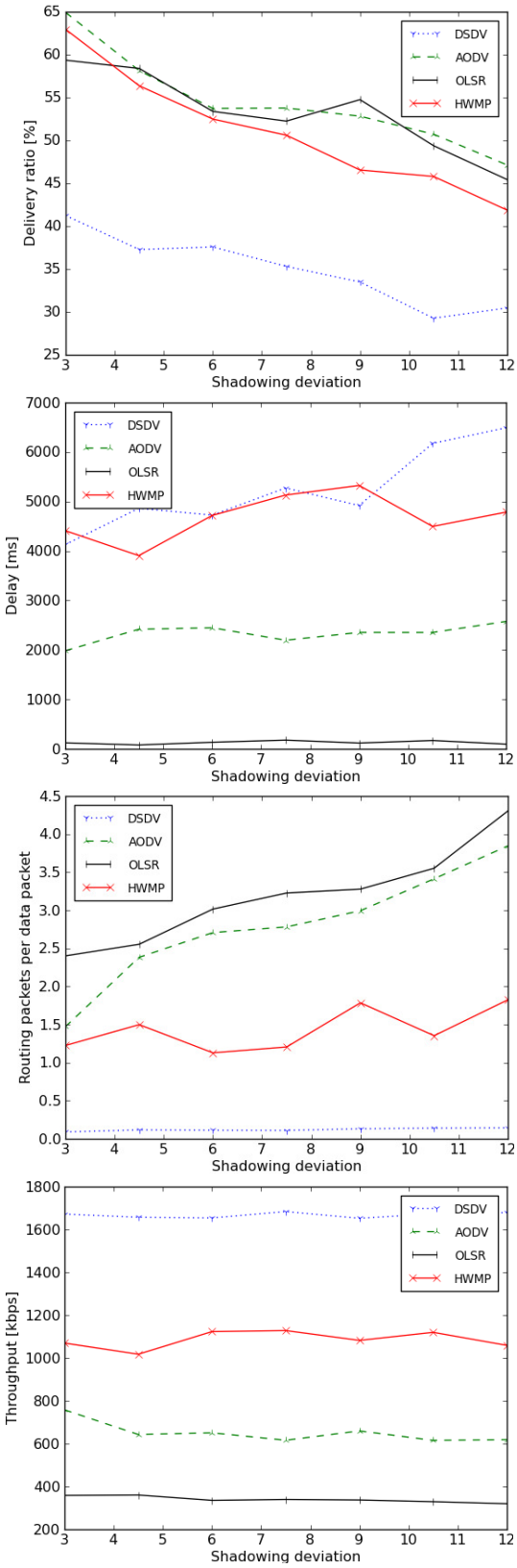


Figure 3. Delivery ratio, packet delay, routing load and system throughput in functions of the shadowing deviation.

 ISBN: 978-1-61208-183-0

B. *Experiment 2. Impact of the path loss exponent*

Unlike the deviation, the path loss exponent does not introduce randomness imitating the presence of encountered obstacles, but affects the transmission range. Higher values of path loss exponent mean faster fading of transmitted signal power with growing distance from the source, and thus shorter transmitting, sensing and receiving ranges.

The plots in the first diagram (Fig. 4) representing the DR for individual routing protocols show very strong decreasing trend (lowering DR approximately to a third). DSDV is shown to be the least effective protocol in the whole experiment. The similarity in DR for AODV, OLSR and HWMP does not allow to select an unambiguous leader. All of them note a rather constant decrease of the DR summing up to the amounts from 53.7% to 56.8%.

The EED diagram (similarly to Fig. 3) shows significant differences between the investigated protocols. The EED function of the path loss exponent has a clearly visible, strong growing trend for HWMP – from 336ms up to nearly 7.4s – respectively for values 2 and 4.4 of the investigated parameter. The increase of EED for AODV routing protocol is also rather constant but considerably less intensive – from approximately 1s to over 3s. Whereas the DSDV plot shows a contradicting trend decreasing from 4.8s to 3.2s, which is a rather unexpected outcome. The other interesting finding concerns the OLSR plot. The EED is diminutive for values from 2 to 3.2 of path loss exponent. For higher values of this shadowing parameter OLSR protocol denotes a rapid EED growth of the order of several seconds.

The relatively smaller EEDs for low path loss exponent and their growth for more intensive signal fading can be logically explained. In perfect circumstances, when the wave propagation is undistorted and the signal power fades slowly, a lot of data is sent directly from its source to the destination. The decreasing range of the source and destination nodes disables the direct transmission and forces the source to send the packet through intermediate nodes. In this case the propagation and transfer times are multiplied by the number of intermediate hops and the total EED grows.

It is harder to explain the behavior of the DSDV protocol. The relatively low DR, especially for low values of path loss exponent, suggest that DSDV protocol, or at least its implementation for *ns* manages UDP traffic routing significantly less effectively than any other protocol.

The proactive character of this protocol is not to blame in this case. The true reason is most likely the way of disseminating routing information in the network. In case of DSDV it is performed by exchanging full routing tables with all of the currently detected one-hop neighbors. The lower the path loss exponent is, the further the nodes' range, and thus the bigger the one-hop neighborhood.

The big network load caused by large routing traffic, can influence the efficiency of data transmission. This effect is amplified by the fact that the routing messages are given higher priority than those carrying data.

The two remaining diagrams substantiate the assumptions made in previous paragraphs in this subsection. The RL is smallest for DSDV, from 0.05 for low values of path loss exponent, up to 0.8 for the strongest shadowing.

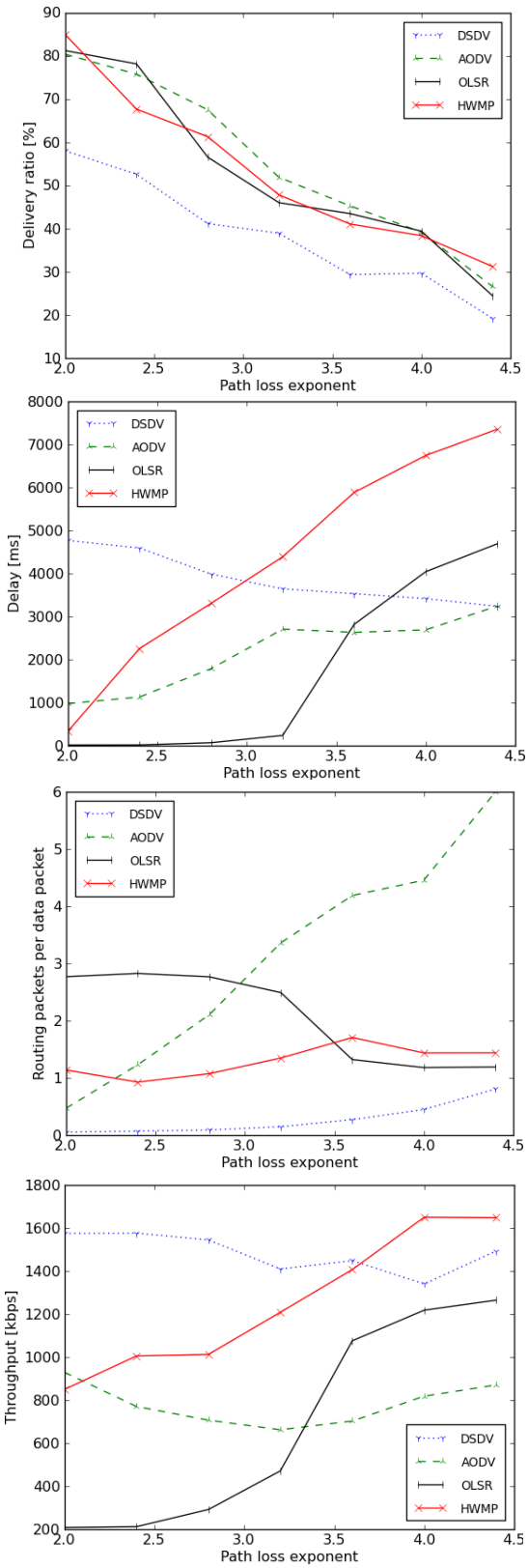


Figure 4. Delivery ratio, packet delay, routing load and system throughput in functions of the path loss exponent.

With more intensive shadowing less and less nodes belong to one-hop neighborhoods. For this reason a lot of broadcasted routing information needs to be forwarded, thus enlarging the RL. The HWMP protocol generates stable amount of approximately 0.9 to 1.4, in the extreme case 1.7, routing messages per one delivered UDP packet. The RL for AODV protocol grows most intensively with the path loss exponent. This happens because the decreasing nodes' range in mobile network causes more changes in topology, and consequently more frequent changes of the neighborhood, thus generating more on-demand routing information for this reactive routing protocol. The unexpected and fairly sudden decrease of RL for OLSR protocol seems somewhat anomalous and, confronted with other diagrams, gives an inkling of problems on implementation level, resulting in anomalies observed for more complex or particularly problematic scenarios.

The ST is the highest for the DSDV protocol, however intensively and almost constantly growing ST for HWMP reaches the same level of approximately 1600kbps for the high path loss exponent values – 4 and 4.4. The level of ST for AODV mostly reaches the values between 700 and 900kbps, which make it the most stable result.

This set of simulations has shown AODV as the most suitable protocol for network topology environments affected with strong shadowing. The DR obtained using AODV is in most cases the highest; the EED is the shortest, and the generated ST – the lowest. For less shadowed environments, like e.g., rural or indoor WMN applications, the OLSR protocol seems to be the right choice. However, in the face of the dubious accuracy of simulation outcomes for OLSR protocol a clear and irrefutable selection of the overall best routing protocol or protocols cannot be made.

V. CONCLUSION AND FUTURE WORKS

This paper presented an investigation of WMN performance based on simulations with varying propagation conditions. The simulation results gave some direct remarks on the WMN performance, which may be used as support for decisions on the choice of a routing protocol or aid in the process of its design and development.

The experiment has shown that the performance, indicated by the predefined metrics, highly depends on the propagation conditions. The investigated protocols perform better in scenarios with low propagation parameter values.

The oldest of the investigated protocols - the proactive, distance vector routing protocol DSDV, performs very well only in terms of the RL. Regarding to the relatively low DR, long EED and high ST, DSDV efficiency is insufficient for use in modern WMNs even in good propagation conditions.

The main drawbacks of the hybrid protocol HWMP are the long EED and large ST. It is, however, still a well and reliably performing routing protocol. Its DR is considerably good. HWMP's strong side is the low and stable RL. Moreover, HWMP protocol, currently developed in IEEE 802.11s standard for WMN may prove to be more efficient, especially with parallel proactive and on-demand modes.

AODV – a reactive distance vector protocol, offers similar DR and low stable network load. The EED is considerably long but still acceptable. The downside is the sensibility to the propagation factors in case of RL. In the experiment with shadowing deviation AODV shows an intensive growth, but the generated RL is still adequate. In case of the path loss exponent experiment the result is the highest of all. Nonetheless, it is stable and robust, when influenced by changeable propagation.

The proactive, link state routing protocol OLSR has shown the best performance. Its EED is multiple times shorter than any other, the ST-indicated network load is low and the DR is among the highest. The drawback is the relatively high RL, which makes it prone to transmission collisions. These are, however, reduced by the MPR based topology information dissemination mechanism. These characteristics make the young OLSR a good routing protocol for areas with all propagation conditions.

In course of investigations several problems, which may create a perspective for future work, were encountered. These are, for example, the lacking compatibility of the routing protocol implementations as well as imprecision of the simulator modules' documentation.

REFERENCES

[1] G. Elias, M. Novaes, G. Cavalcanti, and D. Porto, "Simulation-based performance evaluation of the SNDP protocol for infrastructure WMNs," Proc. 24th IEEE Int. AINA Conf, 2010, pp. 90-97.

[2] N. H. Moleme, M. O. Odhiambo, and A. M. Kurien, "Enhancing video streaming in 802.11 wireless mesh networks using two-layer mechanism solution," Proc. AFRICON '09, 2009, pp. 1-6.

[3] Y. Zhang, J. Luo, and H. Hu, "Wireless mesh networking, Architectures, Protocols and Standards," Auerbach Publications, 2007.

[4] C. E. Perkins and P. Bhagwat, "Highly dynamic destination sequenced distance vector routing (DSDV) for mobile computers," 1994.

[5] V. D. Park and M. S. Corson, "A highly adaptive distributed routing algorithm for mobile wireless networks," Proc. IEEE, INFOCOM '97, vol. 3, 1997, pp. 1405-1413.

[6] C. Perkins, E. Royer, and S. Das, "Ad hoc on-demand distance vector (AODV) routing," RFC 3561, IETF, 2003.

[7] T. Clausen and P. Jacquet, "Optimized link state routing protocol (OLSR)," RFC 3626, IETF, 2003.

[8] The Working Group for WLAN Standards of the IEEE, "HWMP protocol specification," 2006.

[9] I. F. Akyildiz and X. Wang, "A survey on wireless mesh networks," 2005 [Online: September 2011].

[10] K. Fall and K. Varadhan, Eds., "The ns Manual (formerly nsNotes and Documentation)," UC Berkeley, LBL, USC/ISI, and Xerox PARC.California, 2010 [Online: September 2011].

[11] IEEE Standard for information technology: telecommunications and information exchange between systems. Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, 2007. IEEE Std 802.11 - 2007 (Revision of IEEE Std 802.11 - 1999).

[12] C. E. Perkins, E. M. Royer, S. R. Das, and M. K. Marina, "Performance comparison of two on-demand routing protocols for ad hoc networks," IEEE Personal Comm., vol. 8, 2001, pp. 16-28.

[13] A. Zakrzewska, L. Koszalka, and I. Pozniak-Koszalka, "Performance study of routing protocols for wireless mesh networks', Proc. 19th Int. Conf. Systems Engineering ICSENG '08, 2008, pp. 331-336.

 ISBN: 978-1-61208-183-0

Comparison of Heuristic Methods Applied to Optimization of Computer Networks

Tomasz Miksa, Leszek Koszalka, Andrzej Kasprzak
Department of Systems and Computer Networks,
Wroclaw University of Technology,
Wroclaw, Poland
e-mails: tom.miksa@gmail.com, leszek.koszalka@pwr.wroc.pl, andrzej.kssprzak@pwr.wroc.pl

***Abstract*—This paper presents an attempt to solve Capacity and Flow Assignment (CFA) problem, which is NP-complete. Meta-heuristic and heuristic algorithms are invented in order to find not only feasible but also effective solution. A set of test network instances, with provided dual bounds as a reference, is used to: tune algorithms' parameters, conduct experiments and assess results. Final results provide statistical measures derived from experiments and imply which of proposed algorithms provides better solutions. However, the two created algorithms seem to be promising.**

Keywords-computer networks; heuristic algorithm; simulated annealing; flow assignment; simulation

I. INTRODUCTION

The problems connected with design of Wide Area Networks (WAN) are important due to its practical application. Slight difference in solution quality has a big impact on networks maintenance cost. The issues of network design can be grouped in three categories: Flow Assignment (FA), Capacity and Flow Assignment (CFA), and Topology, Capacity and Flow Assignment (TCFA). Different optimization criteria are considered, but in most cases cost and average packet delay are chosen.

This paper deals with CFA for WAN problem with a cost as a criterion. Demands for transfer of data between multiple nodes may change during lifetime of already designed network. Then, in order to reduce upkeep cost, an optimization of routing paths (flow represents routing) and capacity modules installed on links should be made. Furthermore, solution of CFA problem can be used by algorithms that solves more complex problem – TCFA problem [1] [2].

CFA problem is NP-complete [3]. An algorithm of branch and bound was proposed in [1]. Authors of papers [4] [5] [6] applied successfully heuristic approaches. Other heuristic approaches are discussed in [7]. Some strongly polynomial algorithms are described in [8]. Meta-heuristic algorithm MSA and heuristic algorithm MHU are proposed in this paper, in order to solve CFA problem. Both algorithms are original implementation of heuristic approaches described in [1] [9] [10]. Some clues from [11] concerning integer programming are used. What is more, an experimentation system Ultimate Capacity and Flow Assignment (UCFA) was designed and implemented in order to test the proposed algorithms. This system enables to compare results provided by MSA and MHU algorithms against dual bounds or optimum solutions. This is achieved by the use of *sndlib* test instances [12].

The paper is organized as follows. In Section II, problem formulation is presented. The considered algorithms for solving such a problem are described in Section III. In Section IV, neighborhood types that are used by algorithms described in the previous section are presented. In Section V the designed and implemented experimentation system is described. The results of investigations are presented and discussed in Section VI. Finally, conclusions and prospects for future are presented in Section VII.

II. PROBLEM STATEMENT

The CFA problem consists in finding such a multi-commodity flow and capacity modules allocation that satisfies conditions arising from network topology, traffic matrix, etc.

In this paper, the CFA problem is formulated as follows, using notation from [12].

Constants:
D – number of demands,
E – number of edges,
V – number of vertices.

Indices:
$d = 1, 2, \ldots, D$ – demands,
$e = 1, 2, \ldots, E$ – edges,
$v = 1, 2, \ldots, V$ – vertices,
$k = 1, 2, \ldots, K$ – capacity module type,
$p = 1, 2, \ldots, P_d$ – paths for demand.

Indexed constants:
P_d – number of paths for demand d,
h_d – value of demand d,
c_e – capacity of edge e,
M_k – size of the link capacity module of type k,
δ_{edp} - equals 1, if link e belongs to path p of demand d, equals 0, otherwise,
ξ_{ek} - cost of module type k for edge e.

Variables:
x_{dp} – flow allocation vector.

Objective:

$$\min F = \sum_e \sum_k \xi_{ek} \qquad (1)$$

 ISBN: 978-1-61208-183-0

Constraints:

$$\sum_p x_{dp} = h_d, \qquad d = 1,2,\ldots,D, \tag{2}$$

$$\sum_d \sum_p \delta_{edp} x_{dp} \leq \sum_k M_k, e = 1,2,\ldots,E. \tag{3}$$

III. ALGORITHMS

A. MSA algorithm

MSA algorithm is based on meta-heuristic approach known as Simulated Annealing (SA). SA algorithm imitates process of annealing applied in metallurgy. A description of SA can be found in [13]. Key factors for SA algorithm are the way the "neighborhood" is represented and the way the "moves" are made. These factors are closely related to the problem which is to be solved by SA. "Moves", which make a significant difference between "neighbors", should be made in high temperatures. However, when the temperature is close to end temperature, "moves" should be slight. This is the adaptive divide-and-conquer effect [10].

MSA algorithm uses two types of neighborhood:

- Single Random Any Capacity (SRAC),
- Single Random Decrease Capacity (SRDC).

The former is used as default, while the latter one is used when the temperature of current iteration is close to end temperature. SRAC and SRDC are described below.

MSA input parameters are:

- number of possible paths for demand – *KPath*,
- start temperature T_s,
- end temperature T_k,
- temperature interval τ,
- iterations number L.

Best values for above parameters are derived from experiments.

B. MHU algorithm

MHU algorithm is based on a concept presented in [1], which suggests that neighborhood solutions of CFA problem should be browsed in directed way. Direction should be the "excess of unused capacity". In other words, in following iterations, links chosen for modification are not randomly chosen like in MSA algorithm, but according to the amount of unused capacity. Such an approach results in fast increase of solution quality. Kasprzak [1] claims that such approach can lead to optimum solutions.

MHU algorithm starts its operation by installing maximum capacity modules on links. Multi-commodity flow is found, and objective is calculated. If it is not possible to find multi-commodity, the problem is unsolvable. Else, the link that has the biggest excess of unused capacity is chosen. Bandwidth of this link is reduced to previous capacity module from increasing sequence of available capacity modules for given link. If capacity module already installed on this link cannot be decreased (is already first in sequence), next link with biggest excess of unused capacity is chosen. If none of links can be modified, algorithm stops. Once the capacity module installed on chosen link has been modified, multi-commodity flow is calculated. In this way, new neighbor solution is generated. The step described above is repeated given number of times (algorithm's input parameter), or when there are no more links available for modification. In any iteration, solution with best objective is chosen. The result is used as a base solution for the next iteration. In the case when the solution generated in one of iterations is not feasible, because it is not possible to find multi-commodity flow, algorithm proceeds using best, last known, feasible solution.

IV. NEIGHBORHOOD

This section clarifies the concept of neighborhood used for solving CFA problem. Neighbor solution is a new solution that is, in some way, modified when compared to the previous one. In CFA, the problem it consists of:

- link configuration vector.
- demand routing vector.

Link configuration vector consists of elements representing capacity modules installed on following links. Routing of demands is also represented by vector. Following elements are indexes of paths available for given demand. Set of available paths for each demand is created once, before neighboring solution is generated. In order to create this set algorithms like Dijkstra's, Bellman- Ford's or K-ShortestPath's algorithm can be used. Quantity of paths found for each demand can be done arbitrary or empirically.

A. Single Random Any Capacity (SRAC)

In this method of obtaining neighboring solution, operations are made on both link configuration and demand routing vectors. Index of link, that will be modified, is drawn. Then, an index from the list of available capacity modules is drawn. Drawn capacity module is inserted into vector replacing previously installed capacity module in according to drawn link index. For example, in Fig. 1 link that index is 5 was drawn. Capacity module installed on this link in base solution is 128kbps. Neighbor solution has on that link capacity module whose bandwidth is 512 kbps, because index 8 was drawn as a pointer for a list of capacity modules.

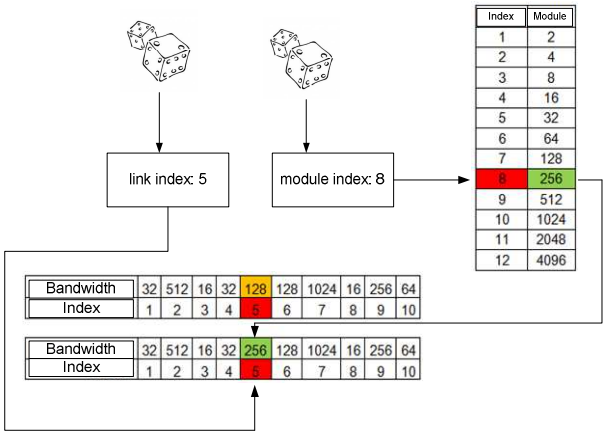


Figure 1. SRAC neighborhood generation scheme.

Other links are left unmodified, according to base solution. Demand routing vector is created anew. For each demand, path's index from list of available paths is drawn.

B. Single Random Decrease Capacity

This method of creation of new neighboring solution is similar to SRAC method, described in previous section. Main difference lies in the way capacity module of drawn link is changed. Instead of random change of capacity module, lower capacity module is used, on condition it exists. For example if drawn link index is 6, only this link is modified. If capacity module installed on that link was 64kbps, and the lower capacity module is 32kbps, then 32kbps capacity module is installed. Demand routing vector is created anew, like it is in SRAC method.

V. EXPERMIENTATION SYSTEM

Experimentation system called Ultimate Capacity and Flow Assignment (UCFA) was created in order to solve CFA problem with a use of MHU and MSA algorithms. System possesses user friendly Graphical User Interface (GUI), which facilitates experimentation. UCFA enables among other things following options:

- Choice of test instances.
- Choice of algorithm.
- Setup of input parameters.
- Multithreaded tests.
- Live progress preview.
- Solution save.
- Results summary save.

Test instances are delivered with the experimentation system. Every instance possesses scheme reflecting nodes and links allocation. All instances are imported from *sndlib* library [12]. This guarantees that instances have always not only at least one feasible solution, but also provided test data is derived from real systems, what makes the simulations more realistic. Furthermore, information about network and problem parameters are available, as well as information about dual bound for problem or best solutions uploaded by other scientists. As a result, the solutions obtained with the examined algorithms can easily be compared to dual bound, or to other known solutions.

In UCFA experimentation system, a user can choose between MSA and MHU algorithms. Each of them has configurable set of parameters. In case of MSA, these parameters are: number of possible paths for demand – *KPath*, start temperature T_s, end temperature T_k, temperature interval τ, iterations number L. In case, of MHU the parameters are: *KPath*, and iterations number L.

Multiple tests can be run in parallel; the only limitation is performance of platform, on which the tests are being run. Each test can be paused and resumed or cancelled. At each stage of performing process the current results can be saved. Solution file format is XML, what facilitates easy integration with other simulators and benchmarks.

Results of tests are appended into summary file, where information on algorithm type, its parameters, solution cost, gap between dual bound and solution, simulation time, etc. are stored. Summary file can easily be used in multiple editors, due to its CSV structure.

VI. INVESTIGATION

Three complex experiments were conducted in order to tune algorithms' efficiency and carry out reliable comparison of MHU and MSA algorithms. These experiments concern:

- influence of *KPath* parameter in MSA algorithm,
- parameters setup in MSA algorithm,
- solution quality gain in MSA and MHU algorithms.

Each experiment was performed with the use of UCFA experimentation system. In the first and in the second experiment, the three network instances were used:

pdh, *dfn-bwin*, *nobel-german*.

Each test was repeated 5 times for a given set of parameters on a given network instance. In the third experiment, the six network instances were used:

dfn-bwin, *nobel-eu*, *nobel-germany*, *pdh*, *norway*, *dfn-gwin*.

In order to evaluate the algorithms, the two indices of performance were introduced:

- gap to dual bound,
- gap to base solution.

Both indices are expressed in percents. Equations (4) and equation (5) show how these measures are calculated.

$$\Delta_{BEST} = \left(\frac{F(current) - dualbound}{dualbound} \cdot 100 \right) \%. \quad (4)$$

$$\Delta_{FIRST} = \left(\frac{F(first) - F(current)}{F(current)} \cdot 100 \right) \%. \quad (5)$$

$F(current)$ denotes the cost of a current solution, while $F(first)$ denotes the cost of the first solution, and *dualbound* is the cost of dual bound.

A. KPath influence in MSA algorithm

This experiment concerns influence of *KPath* parameter onto solution quality. This parameter determines number of different paths for each demand what is in direct correlation with number of possible routing combinations. Ten values of *KPath* values were considered. These values range from 1 to 10. Fig. 2 presents results of the performed experiment.

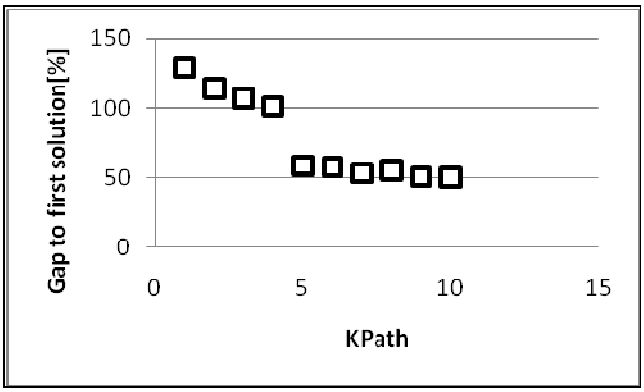


Figure 2. *KPath* influence in MSA algorithm.

 ISBN: 978-1-61208-183-0

One can observe that Δ_{FIRST} value is biggest for *KPath* equal to 1. However, this value must be rejected. It cannot be used, because in case when number of available paths for demand equals 1, then demand can flow along only one path, what means the flow is always same. When flow is always same, solved problem is no longer CFA problem, but only capacity assignment problem! Due to this fact, *KPath* cannot equal to 1.

Having rejected first best value, *KPath* parameters 2, 3 and 4 should be considered. All of them guarantee almost same solution efficiency gain (over 100%). *KPath = 4* seems to be the best choice, because it enhances quantity of possible demand flow combinations, while gap to first solution stays at same high level like when choosing parameters 2 or 3.

KPath values larger than 4 are not to be considered, because it can be easily noticed, that they vastly decrease solution quality. Furthermore, during experiment it was not always possible to find more than 4 different paths for each demand. *Nobel-germany* is an example of such network. If *KPath* parameter was set to value higher than 4, many networks would be unsolvable.

B. *Parameters setup in MSA algorithm*

This experiment concerns the parameters tuning for MSA algorithm. Eight parameter configurations were tested. Table 1 presents names of configurations mapped to parameters setup. It was assumed that each configuration should result in circa 100 000 total number of iterations. This assumption was made because of performance of available test platform. What is more it is assumed that temperature interval is always set to 0.9 and final temperature is of 1. The first assumption followed [10], but the second was done by authors of this paper on the basis of the results of preliminary experiments.

TABLE I. CONSIDERED MSA PARAMETERS CONFIGURATIONS.

Configuration	KPath	Ts	Tk	τ	L
Configuration 1	4	40000	1	0.9	1000
Configuration 2	4	4000	1	0.9	2000
Configuration 3	4	200	1	0.9	2000
Configuration 4	4	13	1	0.9	4000
Configuration 5	4	5	1	0.9	6000
Configuration 6	4	3.5	1	0.9	8000
Configuration 7	4	2.8	1	0.9	10000
Configuration 8	4	1.5	1	0.9	25000

The average gap to first solution Δ_{FIRST} was calculated for each configuration. Results are presented in Fig. 3. One can observe that configuration 8 delivers highest solution quality gain (133%). Configurations 5, 6 and 7 have very similar gain (over 130%). However, Configuration 1 gives only 104% of average improvement. This shows that slight difference in parameters setup results in big difference between qualities of solutions.

.

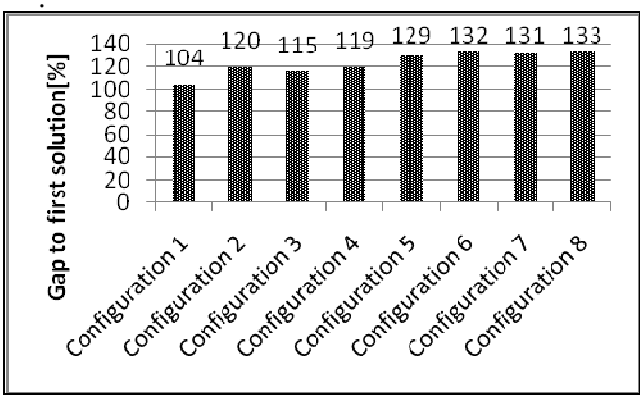


Figure 3. Average gap to first solution for MSA algorithm.

Before one of configurations is chosen as best one, one more calculation of measures (indices) is done. Optimistic cases (maximum gaps) for each configuration are shown in Fig. 4.

For all the configurations, maximum gaps are around 10% better than average gaps. The trend observed in previous chart remains same. Configuration 8 is best again; Configurations 5, 6, 7 have quite similar results to it. Although in this case difference between Configuration 8, and other configurations is higher than previously. Configuration 1 remains worst again.

Taking into consideration result presented above, Configuration 8 is chosen as the best one. Thus, all experiments concerning MSA algorithm are advised to be performed using Configuration 8.

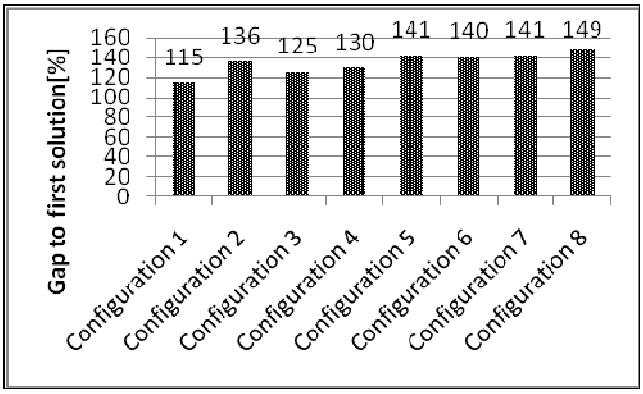


Figure 4. Maximum gap to first solution for MSA algorithm.

C. *Solution quality gain in MSA and MHU algorithms*

This experiment was designed in order to compare solution quality gain over first solution delivered by MSA and MHU algorithms. Parameters used for MSA are taken from previous experiment, e.g., configuration 8 was used. Both algorithms used *KPath* parameter set to 4.

Fig. 5 presents the results obtained with the use of MSA algorithm. Three cases are considered: optimistic, pessimistic

 ISBN: 978-1-61208-183-0

and average. In the worst case, MSA guarantees solution quality gain of 121 %, what is not far from average gain of 138%. In the best case 163% gain is possible. It is worth to mention, that all these values are average values from all tested instances. Hence, it is possible, that in specific conditions these values can even be higher.

The same relation can be observed for results obtained with the use of MHU algorithm (see Fig. 6). Once more three cases are considered: optimistic (maximum), pessimistic (minimum) and the averaged. One can easily notice that solutions obtained with a use of MHU are of better quality than in the case of MSA algorithm. The pessimistic case is circa 50% better than optimistic in MSA algorithm. Average quality gain in MHU algorithm is 239%, but in best case it is possible to get 261% improvement of solution in comparison to the base solution.

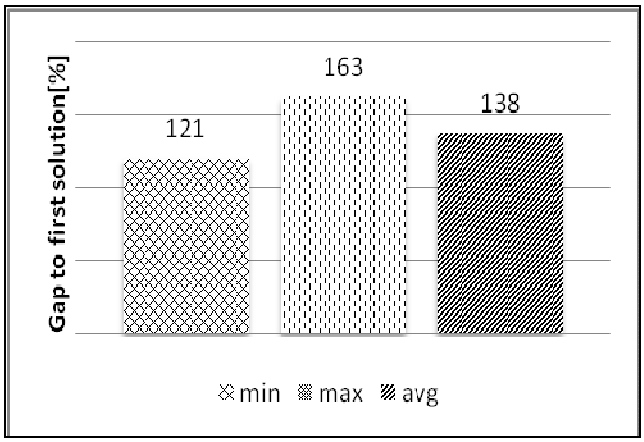


Figure 5. Minimum, maximum and average gap to first solution for MSA.

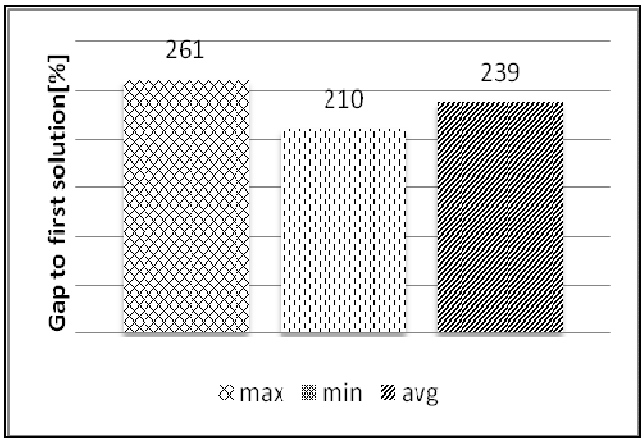


Figure 6. Minimum, maximum and average gap to first solution for MHU.

VII. CONCLUSION AND FUTURE WORKS

In this paper capacity and flow assignment problem, that is NP-complete [3], were discussed. For this purpose, two new algorithms have been proposed by the authors. Moreover, in order to conduct simulations, advanced experimentation system was designed and implemented.

Additional effort to tune parameters of created algorithms was taken. As a result, set of parameters that improve quality of solutions delivered by algorithms is proposed. Finally, both algorithms are examined to check solutions' quality gain over first solution. MSA delivers result which gave average gain of 138%, while MHU algorithm improves base solution 239% in average.

The results of both algorithms are satisfying. We believe that further solution improvement is still possible. Prospects for future are using ideas presented in [14] [15], including new configurations of parameters for MSA, or making simulation along with multistage experiment designs. It seems highly probable, that these two ideas will result in quality improvement of both algorithms.

REFERENCES

[1] A. Kasprzak, Designing of Wide Area Networks, Wroclaw University of Technology Press, (in Polish), 2001.

[2] B. Gendron, T. G. Crainic, and A. Frangioni, Multi-commodity Capacited Network Design, Telecommunications Network Planning, Kluwer, Norwell, MA, 1998, pp. 1-19.

[3] T. Cormen and R. Rivest, Introduction to Algorithms, Warsaw, 2004.

[4] J. Anisiewicz, T. Miksa, and M. Piec, "Cost optimization problem in Wide Area Network design," Proceedings of 10th Polish British Workshop, 2010.

[5] K. Walkowiak, "Ant Algorithm for Flow Assignment in Connection-Oriented Networks", Int. J. Appl. Math. Comput. Sci. 2, 2005, pp. 205-220.

[6] K. Walkowiak, "An Heuristic Algorithm for Non-bifurcated Congestion Problem", Paris, France, Proceedings of 17th IMACS World Congress, 2005..

[7] D. Corne, M. Oates and D. Smith, Telecommunications optimization: heuristic and adaptive techniques, John Wiley & Sons, 2000.

[8] Y. Azar and O. Regev, "Strongly Polynomial Algorithms for the Unsplittable Flow Problem", Proceedings of the 8th Conference on Integer Programming and Combinatorial Optimization, 2001, pp. 15-29.

[9] M. Gerla and L. Kleinrock, "On the Topological Design of Distributed Computer Networks", IEEE Trans Commun., Vol. COM-25, 1977, pp. 48-60..

[10] M. Pioro and D. Mehdi, Routing, Flow, and Capacity Design in Communication and Computer Networks, San Francisco, 2004.

[11] L. A. Wolsey, Integer Programming. Wiley-Interscience, New York, 1998.

[12] S. Orlowski, M. Pioro, and A. Tomaszewski, SNDlib 1.0--Survivable Network Design Library. [Online: 20 May 2011.]

[13] C. D. Gelatt, S. Kirkpatrick, and M. P. Vecchi M. P., "Optimization by Simulated Annealing", Science, New Series, Vol. 220, 1983, pp. 671-680.L. A. Wolsey, Integer Programming. Wiley-Interscience, New York, 1998.

[14] P. Bogalinski, I. Pozniak-Koszalka, L. Koszalka, and A. Kasprzak, "Computer System for Making Efficiency Analysis of Meta-heuristic Algorithms", Proc. of the 2nd ACIIDS conference, LNAI, vol. 5991, Springer, 2010, pp. 225-234.

[15] D. Ohia, L. Koszalka, and A. Kasprzak, "Evolutionary Algorithm for Congestion Problem in Computer Networks", KES 2009, LNAI, vol. 57, 2008, pp. 57-67.

 ISBN: 978-1-61208-183-0

Wireless Home Automation Network Stability Testing

Radek Kuchta, Radovan Novotny, Jaroslav Kadlec, Radimir Vrba
Faculty of Electrical Engineering and Communication,
Brno University of Technology
Brno, Czech Republic
kuchtar | novotnyr |kadlecja | vrbar@feec.vutbr.cz

Vladimir Sulc
MICRORISC, s.r.o.
Jicin, Czech Republic
sulc@microrisc.com

***Abstract* - This paper describes stability testing of the new wireless communication platform for home automation. In the paper is description of two test-cases for wireless communication parameters evaluation to determine the limits of stability and low error rate. The paper also describes used statistical method and discusses testing results together with the main advantages of tested wireless communication platform. This new wireless communication platform was designed and developed especially for home automation and telemetry projects and test case results prove suitability of this wireless communication technology for home and office buildings environment.**

Keywords - Home Automation; IQRF, Wireless communication; Stability testing.

I. INTRODUCTION

With the advance of networking technology and wireless communications, the popularity and the applications of Wireless Sensor Network (WSN) are increasing. Current trends show that the Wireless Sensor Networks will be an integral part of our lives, more than the present-day personal computers [1][2][3].

Usage of Wireless Sensor Networks with low energy demands, low weight and intelligent networking features seems to be the most cost effective solution for many application areas. These devices incorporate wireless transceivers so that communication in short distances over a Radio Frequency (RF) channel is enabled. Wireless Sensor Networks can be used for many applications in various application fields such as automation of the buildings, machines, in the monitoring product quality or conditions at agriculture, medicine, and healthcare.

A general overview of available wireless solution targeted to the small home automation applications and their main parameters and limitations is described in Section II. Following chapter defines test cases and issues of testing of the wireless communication platforms. Statistic tool and evaluation method is described in the Section IV followed by the measured results in Sections V and VII. Conclusion of final measured values and their short assessment is in the last Section VII.

II. STATE OF THE ART

There are available different wireless communication solutions from different vendors on the market place. These solutions support different network topologies. Many of them are based on 802.15.4 [4] standard defining Physical Layer (PHY) and Media Access Layer (MAC) for Low Rate Wireless Personal Area Networks (LR-WPAN). In most cases they work on non-licensed wireless communication bands.

Probably, the most known standardized protocol that works on non-licensed bands is Zigbee [12]. It is a solution based on the IEEE 802.15.4 standard prepared by the Zigbee Alliance [5]. This standard was developed by consortium of industrial companies especially for building automation [6][7]. There are also special applications for industrial control, e.g., [8] [10] on remote access to the system and using small, independent wireless devices, [9][13][14] on building automation and telemetry applications, or an alarm system suitable for pervasive healthcare in rural areas [11]. Among the proprietary solutions, reference can be made to the technology of MiWi launched by Microchip Technology Inc. [15]. MiWi is based on the aforementioned standard but simpler than Zigbee from the implementation point of view. This technology does not support direct cooperation with Zigbee devices. From other solutions available on the market, mention would be made, for example, of the solution promoted by Z-wave alliance [16][17].

These solutions have disadvantage in attempt on being a universal solution targeting every kind of applications. It brings heavier protocols, more difficult and more expensive implementations, lower reliability, and increased network complexity.

Effectiveness of Wireless Sensor Networks (WSN) relies on the communication parameters of interconnected sensors' nodes, which are typically transmitting power, baud-rate, error-rate and their detection range or sensitivity to received signal.

These WSN technologies are determined especially for monitoring environmental and physical conditions, such as temperature, pressure, sound, vibration, humidity, and motion. WSNs applications are often used to perform many critical tasks and sensor networks applications have to meet strict rules and parameters to reliably and error-rate.

A failure of a component or components of a network can result to malfunction in the area of sensing, data processing, and communication. From this point of view it is necessary to evaluate the availability and reliability of application services as two important dependability factors [3].

 ISBN: 978-1-61208-183-0

III. TEST CASE DESCRIPTION AND PROBLEM DEFINITION

Small battery operated wireless sensor nodes are in our network used for automatic inventory system. This application not only expects wireless signal coverage but also need undisrupted service and reliable connectivity. The key aspect of wireless channel is the monitoring and evaluation of the channel quality. Most of the models of radio wave propagation involve questions related to the "free space" radio wave propagation [18].

Radio waves emit from a point source of radio energy, traveling in all directions. Obstacles such as physical and structural components of a building, furniture and fixed or movable structures, or the ground can impact signal propagation paths. Especially ferrous materials, such as steel and iron, can drastically alter signal propagation characteristics, communication distances, link quality, and many other factors [19].

Reflection, diffraction and scattering cause radio signal distortions and give rise to signal fades, as well as additional signal propagation losses. Indoor use of wireless systems creates the necessity for evaluation of indoor radio (RF) propagation. Any obstacles in the pathway would be harmful to RF transmission, radio signals penetrate of obstacles in ways that are very hard to predict. The final composite signal is made up of a number of components from the various sources of scattered and diffracted signal components or reflections from different directions.

To better understand this effect in our test case we at first evaluated the communication characteristics when sensor nodes were placed in various locations and distances. Absorption of RF energy results in loss of signal strength and reduced transmission distances. RF signals from wireless sensor nodes are air radiating from a transmitter and propagating through a medium in all directions. We need to understand the communication distance of individual nodes as well as to evaluate how and where to install the nodes.

The WSN in this application test case is based on the IQRF wireless communication platform for industrial and home automation. This is the technology that was specifically developed for wireless sensor mesh networks by Microrisc company [20]. Typical application scenario of home automation with IQRF communication technology for a smart house is shown in the Fig. 1. The main parts of the platform are covered by Czech and US patents [21][22][23]. For our experimental purposes, the standard IQRF components and development tools have been used. This wireless solution could be used for wireless connectivity necessary for telemetry, remote control, displaying of remotely acquired data, connection of more equipment and building automation. Implementation of IQRF transceiver modules works in non-licensed communication bands, license-free ISM bands 868 MHz in EU, 916 MHz in US, 433 MHz in EU, US and other countries.

Basic features of the IQRF communication platform are especially extra low power consumption (1 µA in the sleep mode and 35 µA in the on-line mode), available networking functions, programmable RF power up to 3.5 mW, SW selectable in steps, up to 170 m communication range, 15 kb/s (optionally 100 kb/s) RF bit rate. A transceiver module is the basic communication component needed for realization of wireless RF connectivity and can work as a node or a network coordinator. The IRQF modules could be integrated into any electronic device via SIM card connector. The low power consumption predetermines these modules for battery powered applications. The transceiver module is equipped with the IQRF operating system supporting functionality for the user application. There are RF functions for transmitting, receiving, network bonding, routing, main parameters configuration, EEPROM access functions, and IIC and SPI communication functions. Data processing, for example, encoding, encryption, checksums, adding headers, is evaluated automatically by IQRF operating system during the communication. The other functions of operating system are three buffers and some other auxiliary functions. IQRF operating system is buffer-oriented and allows sending up to 32 bytes in one packet.

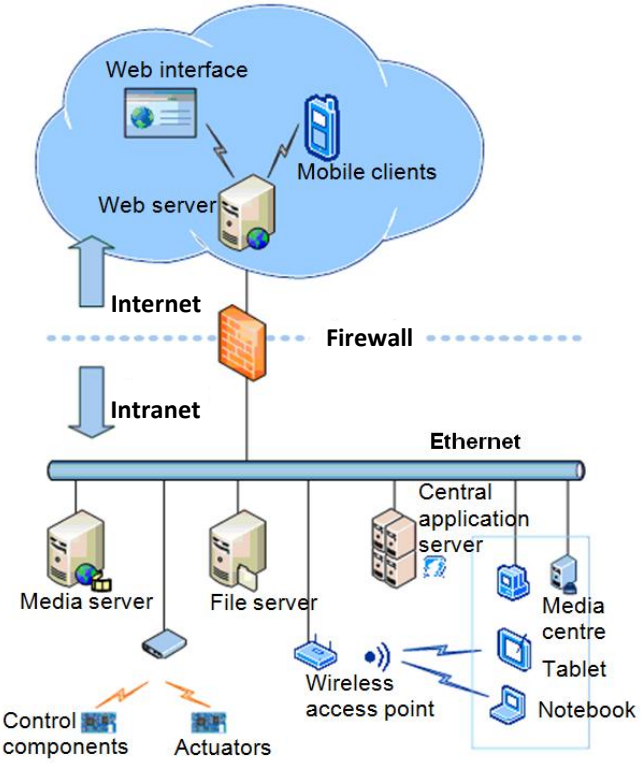


Figure 1. Block diagram of the telemetry and control for smart house

This application test case demonstrates simple data collection from wirelessly connected sensors. The network used for our experiment consists of one coordinator and a set of communication units. This is the basic star network topology where a sensor network is created around a core coordinator. The packet is wirelessly sent by operating system to the coordinator and the quality of the communication is statistically evaluated.

 ISBN: 978-1-61208-183-0

IV. TEST CASES AND USED DATA EVALUATION METHOD

For wireless communication parameters measurement were used two basic tests-case scenarios. The first one examined communications' parameters and wireless technology limits under typical building environment with the set of rooms separated by the plasterboard walls and the second set of measures was done in the long hall without any physical obstacles to test free space signal propagation.

A. Statistical description of the wireless network stability

The binomial distribution $B(n,p)$ with parameters n and p gives the discrete probability distribution of independent observations by the number of observations in the group that represent one of two outcomes. This distribution describes the behavior of a count variable X if the number of samples n is fixed, each sample represents "success" or "failure", each sample is independent, and the probability of "success" p is the same for each outcome.

The binomial distribution gives the approach to dependability evaluation for wireless communication. We expect that in the stable wireless network each type of outcome has a fixed probability and by evaluation of the proportion of individuals in a random sample we could evaluate the stability of the network. They are the sequences of independent transfers in the communication model with two possible outcomes ("success" or "failure").

B. Statistical description of the wireless network stability

We extract samples of a certain size from the ongoing Wireless Sensor Network in our case study related to the stability testing of the channel quality. There are the sequences of independent transfers with two possible outcomes ("success" or "failure") in this experimental situation. The fraction or proportion of "failure" items can be expressed as a decimal or as a percent (when multiplication by 100 is used).

From the statistical point of view, the number of failures is the random variable. Common-causes and special-causes are the two distinct origins of variation in a system. Common-cause variation is the noise within the system and is inherent to the process. It could be removed by making modifications to the process. Special-causes are unusual, not previously observed variation, which is inherently unpredictable. There are only common-causes in the stable system and the statistical monitoring and control could be used for stability evaluation.

Each run that is accomplished is then a realization of a Bernoulli random variable with parameter p. The binomial distribution $B(n,p)$ with parameters n and p gives us the discrete probability distribution of these independent observations. If a random sample of n units of transfer realization is selected and if k is the number of units that are nonconforming, the k follows a binomial distribution with parameters n and p according to following equation

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \forall k; k = 0,1,\dots,n \qquad (1)$$

We expect that in the stable wireless network each type of outcome has a fixed probability and by evaluation of the proportion of individuals in a random sample we can evaluate the stability of the network in setting condition. For the stability evaluation can be used the np-chart as Shewhart control chart with underlying binomial distribution. The sample size is constant and the number nonconforming is plotted against the control limits. The control limits are defined as a

$$n\bar{p} \pm 3\sqrt{n\bar{p}(1-\bar{p})}, \qquad (2)$$

where n is the sample size and $\bar{p}$ is the estimation of the long-term mean. Rational subgroups for our testing are composed of the transfer of packets under essentially the same experimental conditions.

V. STABILITY EVALUATION OF INDOOR RF PROPAGATION (THE TEST CASE OF MORE RF PROPAGATION OBSTACLES)

In this application test case, there were five transmission units in five various rooms each separated by the plasterboard partitions. There are two changed factors in this experiment: eight various levels of transmitting power and the daytime. In each run, there was 80 data transfer execution, which each consists from 500 data frame. The number of failures in the communication was then evaluated. The results from this first experimental test case are summarized in the Fig. 2. There are two factors influencing the results. The mark (a) in the graph highlights the independence of the number of failures on the RF power. For the distance that is higher than 5 meters it is necessary to optimize the RF power value. The mark (b) in the graph highlights the special-causes variation. Experimental results were evaluated by using the np control charts (see Fig. 3 and Fig. 5). An np-chart is a plot of the number of defective items observed in a sample where n is the sample size and p is the probability of observing a defective item when the system is in control without affection of special cause variation. The observed number nonconforming (NP) is plotted against the control limits (UCL – Upper Control Limit, LCL – Lower Control Limit), which are statistically determined.

For the purpose of statistical evaluation of this wireless communication, experiments were used np control charts. The results of this analysis are summarized at the control charts (Fig. 3 and Fig. 5). The fluctuation of the points between the control limits (UCL, LCL) is due to the common cause variation. Any points outside the control limits related to the six standard deviation rule could be attributed to a special-cause variation. There are some cases where special-causes are affecting the results. Out of control points are marked as "1". Overall interpretation of created np control charts for this part of experiment leads to these conclusions:

- The higher distance between the transmitter and receiver is in the relation to the special-cases variation existence and communication failure.

 ISBN: 978-1-61208-183-0

- The higher RF power gives the higher probability for wireless communication without failures.

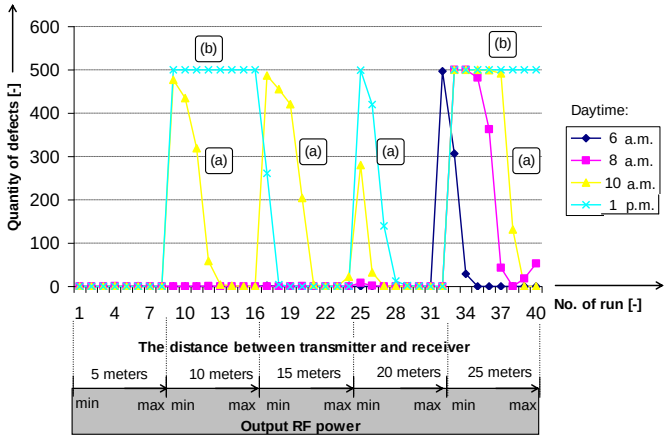


Figure 2. Wireless stability evaluation (the test case of more RF propagation obstacles)

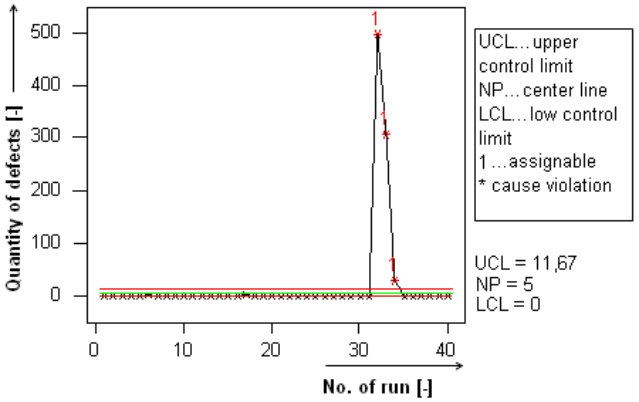


Figure 3. NP chart of wireless stability evaluation for the measurement at 6 p.m. (the test case of more RF obstacles)

VI. STABILITY EVALUATION OF INDOOR RF PROPAGATION (THE CASE OF THE FREE SPACE RF PROPAGATION)

In this part of the experiment, there were five transmission units in five various places in a region, which is free of all objects that might absorb or reflect radio energy. Eight various levels of the RF power and the various daytime are changed in this experiment. In each run there was 80 data transfer execution, which each consists from 500 data frame. The number of failures in the communication system in this configuration was then evaluated. The results from this part of experimental test case are summarized in the Fig. 4.

We could see that there are some communications problems related to the setting of the RF power. The RF power needs the optimization according the distance between the transmitter and receiver. These situations are in the Fig. 4 depicted by mark (a).

Communication with the fifth transmitter unit located in the distance 25 meters for the receiver is affected by special cause variation in this case. This is the limiting distance that the signal is able to penetrate at the building environment in this experiment configuration.

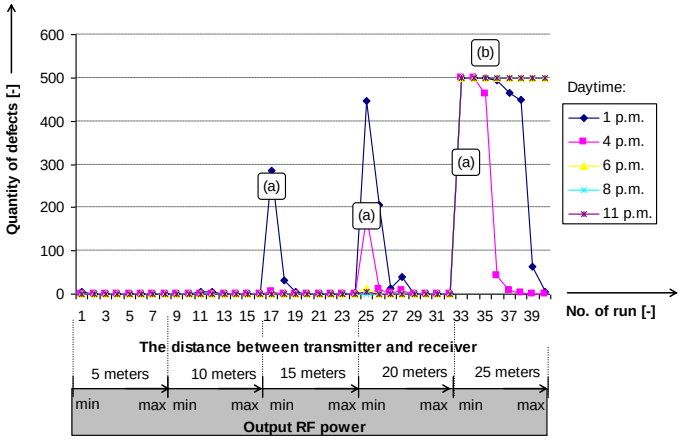


Figure 4. Wireless stability evaluation (the test case of the free space propagation)

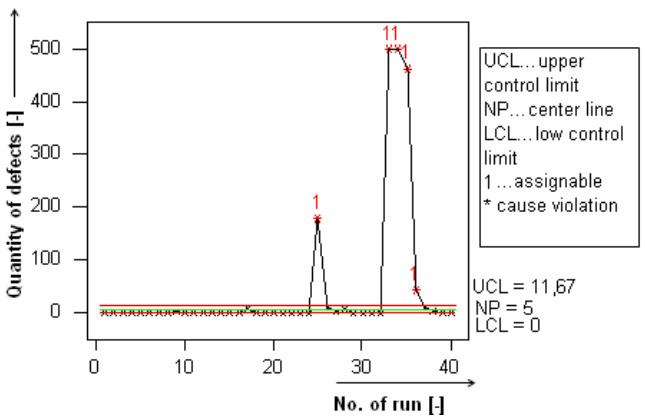


Figure 5. NP chart for wireless stability evaluation (the case of the free space propagation, 4 p.m.)

The mark (b) in the Fig. 4 is related to the communication failure and the special cause variation case.

The number of communication problems in comparison to the case of more RF propagation obstacles is smaller. The requirements for higher RF power are smaller and the overall stability is better. The higher RF power gives the higher probability of wireless communication without failures.

VII. CONCLUSIONS

IQRF is a new wireless communication platform especially designed and developed for specific requirements from home automation and telemetry. One of the main aims was to offer wireless platform to developers of the end user devices that allows rapid development without necessity of stack implementations. As a typical representative of the low-cost wireless communication technology IQRF presents ideal solution for home automation and office or light industry applications. As such, this platform was designed especially for home automation and telemetry applications.

 ISBN: 978-1-61208-183-0

Proposed test cases have proved suitability of this technology to typical application scenarios, test their real communication parameters under buildings environment and determine limits. Based on the statistical result of measured data analysis can be set optimal node distance and output RF power to communication defects ratio. Output RF power influences power consumption and then operation time. Optimal combination of distance and output RF power in specific operation conditions under different environments is therefore highly needed and can significantly improve operation time and minimize communication failures.

Real tests proved wireless communication abilities of IQRF, which fits to the requirements for usage in home automation and telemetry applications and also in the currently developed automatic stochastic system.

ACKNOWLEDGMENT

This research has been supported by ARTEMIS JU in Project No. 100205 POLLUX - Process Oriented Electronic Control Units for Electric Vehicles Developed on a multi-system real-time embedded platform, by the Czech Ministry of Industry and Trade in projects FR-TI3/254 OPT - Open Platform for Telemetry and by the CZ.1.05/1.1.00/02.0068, OP RDI CEITEC - Central European Institute of Technology.

REFERENCES

[1] Akyildiz, F., Weilian, S., Sankarasubramaniam, Y., and Cayirci, E.: A Survey on Sensor Networks, IEEE Communications Magazine, pp. 102-114, August 2002.

[2] Akyildiz, F., Wang, X., and Wang, W.: Wireless mesh networks: a survey, Computer Networks, no. 47, pp. 445-487, January 2005, doi:10.1016/j.comnet.2004.12.001.

[3] Taherkordi, A., Taleghan, A., and Sharifi, M.: Dependability Considerations in Wireless Sensor Networks Applications, Journal of Networks, vol. 1, no. 6, pp. 28-35, November 2006.

[4] Naris, L. and Benedetto, G.: Overview of the IEEE 802.15.4/4a standards for low data rate wireless personal data networks., in 4th Workshop on Positioning, Navigation and Communication 2007 (WPNC 07), 2007, pp. 285-289.

[5] ZigBee: (2009, May) ZigBee Aliance Web Pages. [Online]. HYPERLINK "http://www.zigbee.org" http://www.zigbee.org , Last accessed on 27 December 2011, <retrieved: 1, 2011>.

[6] Evans-Pughe, C.: Bzzzz zzz [ZigBee wireless standard], IEE Review, pp. 28-31, March 2003, 0953-5683.

[7] Gill, K., Yang, H., Yao, F., and Lu, X.: A ZigBee-Based Home Automation System, IEEE TRANSACTIONS ON CONSUMER ELECTRONICS, pp. 422-430, May 2009.

[8] Gill, K., Yang, H., Yao, F., and Lu, X.: A zigbee-based home automation system, Consumer Electronics, IEEE Transactions on, pp. 422-430, March 2009, 0098-3063.

[9] Edgan, D.: The emergence of ZigBee in building automation and industrial control, Computing & Control Engineering Journal, pp. 14-19, April-May 2005.

[10] Zualkernan, A., Al-Ali, R., Jabbar, A., Zabalawi, I., and Wasfy, A.: InfoPods: Zigbee-based remote information monitoring devices for smart-homes, Consumer Electronics, IEEE Transactions on, pp. 1221-1226, August 2009.

[11] Casas, R., Marco, A., Plaza, I., Garrido, Y., and Falco, J.: ZigBee-based alarm system for pervasive healthcare in rural areas, Communications, IET , pp. 208-214, February 2008.

[12] Poole, I.: What exactly is... ZigBee?, Communications Engineer , pp. 44-45, August-September 2004.

[13] Ciardiello, T.: Wireless communications for industrial control and monitoring, Computing & Control Engineering Journal , pp. 12-13, April-May 2005.

[14] Gomez, C. and Paradells, J.: Wireless Home Automation Networks: A Survey of Architectures and Technologies, IEEE COMMUNICATIONS MAGAZINE, vol. 48, no. 6, pp. 92-101, June 2010.

[15] Flowers, D. and Yang, Y.: MiWi Wireless Networking Protocol Stack, 2010. [Online]. HYPERLINK "http://www.newark.com/pdfs/techarticles/microchip/AN1066_MiWi AppNote.pdf", Last accessed on 27 December 2011, <retrieved: 1, 2011>.

[16] Gomez, C. and Paradells, J.: Wireless home automation networks: A survey of architectures and technologies, Communications Magazine, IEEE , pp. 92-101, June 2010.

[17] Walko, J.: Home Control, Computing & Control Engineering Journal, pp. 16-19, October-November 2009.

[18] Rappaport, T.: Wireless Communications: Principles and Practice. Prentice-Hall, Englewood Cliffs, NJ: IEEE Press (The Institute of Electrical And Electronics Engineers, Inc.), 1996, ISBN: 0-7803-1167-1.

[19] Sun, Z. and Akyildiz, F.: Channel Modeling and Analysis for Wireless Networks in Underground Mines and Road Tunnels, IEEE Transactions on Communications, vol. 58, no. 6, pp. 1758-1768, June 2010, ISSN: 0090-6778.

[20] Microrisc: (2009, May) Microrisc Web Page. [Online]. HYPERLINK "http://www.microrisc.cz/new/weben/index.php" http://www.microrisc.cz/new/weben/index.php, Last accessed on 27 December 2011, <retrieved: 1, 2011>.

[21] Šulc, V.: Czech Republic Patent - A method of accessing the peripherals of a communication device in a wireless network of those communication devices, a communication device to implement that method and a method of creating generic network communication, PUV 18679, 2008.

[22] Šulc, V.: US Patent - Method of coding and/or decoding binary data for wireless transmission, particularly for radio transmitted data, and equipment for implementing this method., 7167111, 2007.

[23] Šulc, V., Kuchta, R., Vrba, R.: IQMESH implementation in IQRF wireless communication platform, In 2009 Second International Conference on Advances in Mesh Networks, pp. Pages 62-65, 2009, ISBN 978-0-7695-3667-5.

 ISBN: 978-1-61208-183-0

Band-Pass Filters for Direct Sampling Receivers

Pavel Zahradnik, Boris Šimák and Michal Kopp
Department of Telecommunication Engineering
Czech Technical University in Prague
Prague, Czech Republic
zahradni, simak, koppmich@fel.cvut.cz

Miroslav Vlček
Department of Applied Mathematics
Czech Technical University in Prague
Prague, Czech Republic
vlcek@fd.cvut.cz

***Abstract*—A robust analytical design procedure for high performance digital equiripple band-pass finite impulse response filters for direct sampling receivers is introduced. The filters are optimal in Chebyshev sense. The underlaying generating function of the equiripple approximation is the Zolotarev polynomial. The closed form solution provides a straightforward evaluation of the filter degree and of the impulse response coefficients from the filter specification. One example is included. The robustness of the design procedure is emphasized.**

***Keywords*-FIR filter; band-pass filter; equiripple approximation; Zolotarev polynomial; direct sampling; digital receiver;**

I. INTRODUCTION

Direct sampling receivers are based on the sampling and processing of the amplified radio-frequency (RF) signal incomming from the aerial. The selectivity in the RF signal is obtained using narrow-band band-pass (BP) digital filters. Because of the high ratio between the pass-band frequency and the bandwidth of the filters, high performance digital filters are required. Such filters can be used in the receivers with direct intermediate frequency (IF) sampling and in frequency analyzers as well. Because of the inherent stability and because of the linear phase the digital finite impulse response (FIR) filters are preferred. A filter is optimal in terms of its length provided its frequency response exhibits an equiripple (ER) behavior. In [1] we have introduced an analytical design procedure for the digital ER notch FIR filters. Here, we present an analytical design procedure for the ER BP FIR filters. The proposed design procedure is based on Zolotarev polynomials [2]-[5]. We present here the closed form solution for the design of ER BP FIR filters. It includes the degree equation and formulas for the robust evaluation of the impulse response coefficients of the ER BP FIR filter.

II. ZERO PHASE TRANSFER FUNCTION

We assume the impulse response $h(k)$ with odd length $N = 2n + 1$ with even symmetry

$$a(0) = h(n)\ ,\ a(k) = 2h(n-k) = 2h(n+k)\,,\ k = 1 \ldots n\,. \tag{1}$$

The transfer function of the filter is

$$\begin{aligned} H(z) &= \sum_{k=0}^{2n} h(k)\, z^{-k} \\ &= z^{-n}\left[h(n) + 2\sum_{k=1}^{n} h(n \pm k)\,\frac{1}{2}\left(z^{k} + z^{-k}\right)\right] \\ &= z^{-n}\sum_{k=0}^{n} a(k)\, T_k(w) = z^{-n} Q(w) \end{aligned} \tag{2}$$

where

$$T_k(w) = \cos(k \arccos(w)) \tag{3}$$

is Chebyshev polynomials of the first kind. The function

$$Q(w) = \sum_{k=0}^{n} a(k)\, T_k(w) \tag{4}$$

represents a polynomial in the variable

$$w = \frac{1}{2}(z + z^{-1}) \tag{5}$$

which on the unit circle $z = e^{j\omega T}$ reduces to the real valued zero phase transfer function (ZPTF) $Q(w)$ of the real argument

$$w = \cos(\omega T)\ . \tag{6}$$

III. GENERATING POLYNOMIAL

An approximation of the frequency response of a filter is based on the generating function. The generating function of an ER BP FIR filter is the Zolotarev polynomial $Z_{p,q}(w, \kappa)$ which approximates a constant value in equiripple Chebyshev sense in two disjoint intervals $(-1, w_1)$ and $(w_2, 1)$ as shown in Fig. 1. The lobe with the maximal value $y_m = Z_{p,q}(w_m, \kappa)$ is located inside the interval (w_1, w_2). The notation $Z_{p,q}(w, \kappa)$ emphasizes the fact that the integer value p counts the number of zeros right from the maximum w_m and the integer value q corresponds to the number of zeros left from the maximum w_m. The real value $0 \leq \kappa \leq 1$ which is in fact the Jacobi elliptical modulus affects the maximum value y_m and the width $w_2 - w_1$ of this lobe (Fig. 1). For increasing κ the value y_m increases and the lobe broadens. E. I. Zolotarev

 ISBN: 978-1-61208-183-0

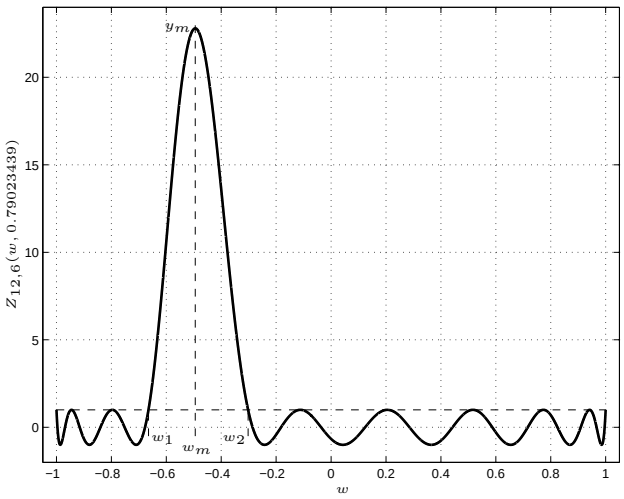


Fig. 1. Zolotarev polynomial $Z_{12,6}(w, 0.79023439)$.

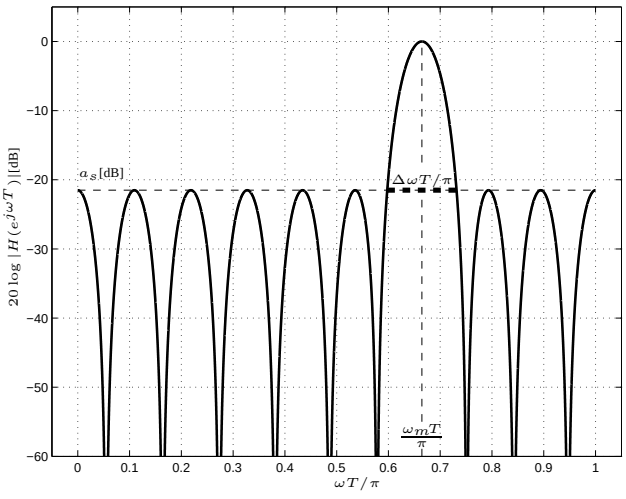


Fig. 2. Amplitude frequency response $20 \log |H(e^{j\omega T})|$ [dB] corresponding to the Zolotarev polynomial from Fig. 1.

(1847-1878) derived the general solution of this approximation problem in terms of Jacobi elliptic functions [3]-[5]

$$Z_{p,q}(w,\kappa) = \frac{(-1)^p}{2} \times \left[\left(\frac{H(u - \frac{p}{n}\,\mathbf{K}(\kappa))}{H(u + \frac{p}{n}\,\mathbf{K}(\kappa))} \right)^n + \left(\frac{H(u + \frac{p}{n}\,\mathbf{K}(\kappa))}{H(u - \frac{p}{n}\,\mathbf{K}(\kappa))} \right)^n \right] . \quad (7)$$

The factor $(-1)^p/2$ appears in (7) as the Zolotarev polynomial alternates $(p+1)-$times in the interval $(w_2, 1)$. The variable u is expressed by the incomplete elliptical integral of the first kind

$$u = F\left(\mathrm{sn}\left(\frac{p}{n}\,\mathbf{K}(\kappa)|\kappa\right) \sqrt{\frac{1+w}{w + 2\,\mathrm{sn}^2\left(\frac{p}{n}\mathbf{K}(\kappa)|\kappa\right) - 1}} \,|\kappa \right) . \quad (8)$$

The function $H\left(u \pm (p/n)\,\mathbf{K}(\kappa)\right)$ is the Jacobi Eta function, $\mathrm{sn}(u|\kappa)$, $\mathrm{cn}(u|\kappa)$, $\mathrm{dn}(u|\kappa)$ are Jacobi elliptic functions, $\mathbf{K}(\kappa)$ is the quarter-period given by the complete elliptic integral of the first kind and $F(\phi|\kappa)$ is the incomplete elliptic integral of the first kind. The degree of the Zolotarev polynomial is $n = p+q$. A comprehensive treatise of Zolotarev polynomials was published in [5]. It includes the analytical solution of the coefficients of Zolotarev polynomials, the algebraic evaluation of the Jacobi Zeta function $\mathrm{Z}(\frac{p}{n}\mathbf{K}(\kappa)|\,\kappa)$ and of the elliptic integral of the third kind $\Pi(\sigma_m, \frac{p}{n}\mathbf{K}(\kappa)|\,\kappa)$. The Jacobi Zeta function and the elliptic integral of the third kind are connected by the formula

$$\Pi(u, u_0|\,\kappa) = \frac{1}{2} \ln \frac{\Theta(u-u_0)}{\Theta(u+u_0)} + u\mathrm{Z}(u_0|\,\kappa) \quad (9)$$

where

$$u_0 = \frac{p}{p+q}\,\mathbf{K}(\kappa) \quad (10)$$

and $\Theta(w)$ is the Jacobi Theta function [4]. The position w_m of the maximum value $y_m = Z_{p,q}(w_m, \kappa)$ is

$$w_m = w_1 + 2\,\frac{\mathrm{sn}\left(\frac{p}{n}\mathbf{K}(\kappa)|\kappa\right)\mathrm{cn}\left(\frac{p}{n}\mathbf{K}(\kappa)|\kappa\right)}{\mathrm{dn}\left(\frac{p}{n}\mathbf{K}(\kappa)|\kappa\right)} \mathrm{Z}\left(\frac{p}{n}\mathbf{K}(\kappa)|\kappa\right) \quad (11)$$

where the edges of the lobe are

$$w_1 = 1 - 2\,\mathrm{sn}^2\left(\frac{p}{n}\mathbf{K}(\kappa)|\kappa\right) \quad (12)$$

$$w_2 = 2\,\mathrm{sn}^2\left(\frac{q}{n}\mathbf{K}(\kappa)|\kappa\right) - 1 \ . \quad (13)$$

The relation for the maximum value y_m

$$y_m = \cosh 2n \left(\sigma_m \mathrm{Z}(\frac{p}{n}\mathbf{K}(\kappa)|\kappa) - \Pi(\sigma_m, \frac{p}{n}\mathbf{K}(\kappa)|\kappa) \right) \quad (14)$$

is useful in the normalization of Zolotarev polynomials. The degree of the Zolotarev polynomial $Z_{p,q}(w,\kappa)$ is expressed by the degree equation

$$n \geq \frac{\ln(y_m + \sqrt{y_m^2 - 1})}{2\sigma_m \mathrm{Z}(\frac{p}{n}\mathbf{K}(\kappa)|\kappa) - 2\Pi(\sigma_m, \frac{p}{n}\mathbf{K}(\kappa)|\kappa)} \ . \quad (15)$$

The auxiliary parameter σ_m is given by the formula

$$\sigma_m = F\left(\arcsin\left(\frac{1}{\kappa\ \mathrm{sn}\left(\frac{p}{n}\,\mathbf{K}(\kappa)|\kappa\right)} \sqrt{\frac{w_m - w_s}{w_m + 1}} \right) |\kappa \right) \quad (16)$$

where $F(\Phi|\kappa)$ is the incomplete elliptical integral of the first kind. The Zolotarev polynomial $Z_{p,q}(w,\kappa)$ satisfies the differential equation

$$(1-w^2)(w-w_1)(w-w_2)\left(\frac{dZ_{p,q}(w,\kappa)}{dw}\right)^2 = n^2\left(1 - Z_{p,q}^2(w,\kappa)\right)(w-w_m)^2 \ . \quad (17)$$

The differential equation expresses the fact that the derivative $dZ_{p,q}(w,\kappa)/dw$ does not vanish at the points $w = \pm 1$, w_1, w_2 where $Z_{p,q}(w,\kappa) = \pm 1$ for which the right hand side of eq. (17) vanishes, and that $w = w_m$ is a turning point corresponding to the local extrema at which $Z_{p,q}(w,\kappa) \neq \pm 1$.

 ISBN: 978-1-61208-183-0

Based on the differential equation (17) we have derived the recursive algorithm for the evaluation of the impulse response $h(k)$ corresponding to the Zolotarev polynomial $Z_{p,q}(w,\kappa)$ based on its expansion into Chebyshev polynomials of the first kind

$$Z_{p,q}(w,\kappa) = \sum_{k=0}^{n} a(k) T_k(w) \ . \qquad (18)$$

The corresponding recursive algorithm is summarized in Table I.

IV. Design Procedure

There are two goals in the design of any filter. The first one is to obtain the minimal filter degree n (or minimal filter length N) satisfying the filter specification while the second one is to evaluate the impulse response $h(k)$ of the filter. The ER BP FIR filter is specified by the pass-band frequency $\omega_m T$ and by the bandwidth $\Delta\omega T$ for the attenuation a_s [dB] in the stop-bands (Fig. 2). The proposed design procedure consists of several steps as follows:

1) Specify the pass-band frequency $\omega_m T$ (or f_m), width of the pass-band $\Delta\omega T$ (or Δf) and the attenuation in the stop-bands a_s [dB] (Fig. 2). For the non-normalized frequencies f_m and Δf specify additionally the sampling frequency f_s.
2) Evaluate the normalized frequencies

$$\omega_m T = \pi \frac{f_m}{\frac{f_s}{2}} \ , \ \Delta\omega T = \pi \frac{\Delta f}{\frac{f_s}{2}} \ . \qquad (19)$$

if the filter is specified by the non-normalized ones.
3) Calculate the band edges

$$\omega_2 T = \omega_m T - \frac{\Delta\omega T}{2} \ , \ \omega_1 T = \omega_m T + \frac{\Delta\omega T}{2} \ . \qquad (20)$$

4) Evaluate the Jacobi elliptic modulus κ

$$\kappa = \sqrt{1 - \frac{1}{\tan^2(\varphi_1)\tan^2(\varphi_2)}} \qquad (21)$$

for the auxiliary parameters φ_1 and φ_2

$$\varphi_1 = \frac{\omega_1 T}{2} \ , \ \varphi_2 = \frac{\pi - \omega_2 T}{2} \ . \qquad (22)$$

5) Calculate the rational values p/n and q/n

$$\frac{p}{n}\mathbf{K}(\kappa) = F(\varphi_1|\kappa) \ , \ \frac{q}{n}\mathbf{K}(\kappa) = F(\varphi_2|\kappa) \ . \qquad (23)$$

6) Determine the required maximum value y_m

$$y_m = \frac{2}{10^{0.05 a_s[dB]}} \ . \qquad (24)$$

7) Calculate and round up the minimum degree n required to satisfy the filter specification using the degree equation (15). For the algebraic evaluation of the Jacobi Zeta function $\mathrm{Z}(\frac{p}{n}\mathbf{K}(\kappa)|\,\kappa)$ and of the elliptic integral of the third kind $\Pi(\sigma_m, \frac{p}{n}\mathbf{K}(\kappa)|\,\kappa)$ in the degree equation (15) use the algebraical procedures [5].
8) Calculate the integer values p and q of the Zolotarev polynomial $Z_{p,q}(w,\kappa)$

$$p = \left[n \frac{F(\varphi_1|\kappa)}{\mathbf{K}(\kappa)} \right] \ , \ q = \left[n \frac{F(\varphi_2|\kappa)}{\mathbf{K}(\kappa)} \right] \ . \qquad (25)$$

The brackets [ ] in (25) stand for rounding.
9) For the values p, q, κ and y_m evaluate the impulse reponse $h(k)$ algebraically using the procedure summarized in Tab. I.

V. Example of the Design

Design the ER BP FIR filter specified by the pass-band frequency $f_m = 10.7$ MHz and by the bandwidth $\Delta f = 50$ kHz for minimal attenuation in the stop-band $a_s = -80$ dB. The specified sampling frequency is $f_s = 30$ MHz.

From the filter specification we get $\omega_m T/\pi = 0.71\overline{3}$ and $\Delta\omega T/\pi = 0.00\overline{3}$ (19). Further we get $\kappa = 0.16239149$ (21), $n = 2026$ (15), $p = 1445$ and $q = 581$ (25). The filter length is $N = 4053$ coefficients. The actual attenuation in the stop-bands is $a_{s\,act} = -80.13$ dB. The amplitude frequency response of the ER BP FIR filter is shown in Fig. 3. A detailed view of its passband is shown in Fig. 4.

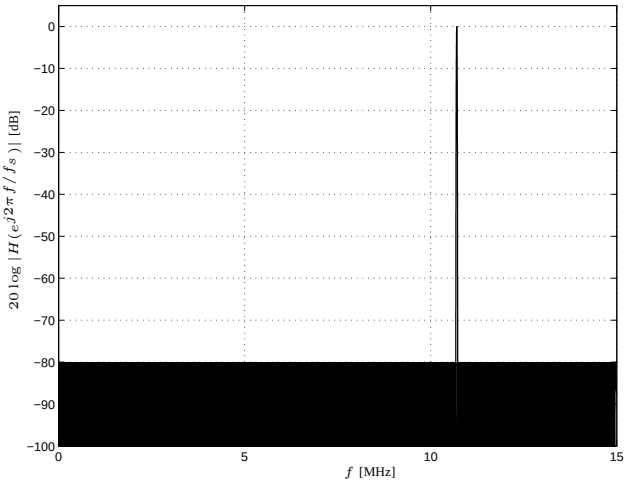


Fig. 3. Amplitude frequency response $20\log|H(e^{j2\pi f/f_s})|$ [dB].

VI. Robustness of the Design Procedure

In order to demonstrate the robustness of the presented design procedure, let us design the filter which was specified in our example, however, with modified bandwidth which is now specified by $\Delta f = 5$ kHz.
We get $\kappa = 0.05166139$ (21), $n = 20248$ (15), $p = 14444$ and $q = 5804$ (25). The filter length is $N = 40497$ coefficients. The actual attenuation in the stop-bands is $a_{s\,act} = -80.04$ dB. The amplitude frequency response of the ER BP FIR filter is shown in Fig. 5. A detailed

 ISBN: 978-1-61208-183-0

TABLE I
ALGORITHM FOR THE EVALUATION OF THE IMPULSE RESPONSE $h(k)$.

given

$p,\ q,\ \kappa,\ y_m$

initialisation

$$n = p + q$$

$$w_1 = 1 - 2\,\mathrm{sn}^2\left(\frac{p}{n}\mathbf{K}(\kappa)|\kappa\right)$$

$$w_2 = 2\,\mathrm{sn}^2\left(\frac{q}{n}\mathbf{K}(\kappa)|\kappa\right) - 1$$

$$w_a = \frac{w_1 + w_2}{2}$$

$$w_m = w_1 + 2\,\frac{\mathrm{sn}\left(\frac{p}{n}\mathbf{K}(\kappa)|\kappa\right)\,\mathrm{cn}\left(\frac{p}{n}\mathbf{K}(\kappa)|\kappa\right)}{\mathrm{dn}\left(\frac{p}{n}\mathbf{K}(\kappa)|\kappa\right)}\mathrm{Z}\left(\frac{p}{n}\mathbf{K}(\kappa)|\kappa\right)$$

$$\alpha(n) = 1$$

$$\alpha(n+1) = \alpha(n+2) = \alpha(n+3) = \alpha(n+4) = \alpha(n+5) = 0$$

body

(*for* $m = n + 2$ *to* 3)

$$8c(1) = n^2 - (m+3)^2$$

$$4c(2) = (2m+5)(m+2)(w_m - w_a) + 3w_m[n^2 - (m+2)^2]$$

$$2c(3) = \frac{3}{4}[n^2 - (m+1)^2] + 3w_m[n^2 w_m - (m+1)^2 w_a] - (m+1)(m+2)(w_1 w_2 - w_m w_a)$$

$$c(4) = \frac{3}{2}(n^2 - m^2) + m^2(w_m - w_a) + w_m(n^2 w_m^2 - m^2 w_1 w_2)$$

$$2c(5) = \frac{3}{4}[n^2 - (m-1)^2] + 3w_m[n^2 w_m - (m-1)^2 w_a] - (m-1)(m-2)(w_1 w_2 - w_m w_a)$$

$$4c(6) = (2m-5)(m-2)(w_m - w_a) + 3w_m[n^2 - (m-2)^2]$$

$$8c(7) = n^2 - (m-3)^2$$

$$\alpha(m-3) = \frac{1}{c(7)}\sum_{\mu=1}^{6} c(\mu)\alpha(m+4-\mu)$$

(*end loop on* m)

normalisation

$$s(n) = \frac{\alpha(0)}{2} + \sum_{m=1}^{n}\alpha(m)$$

$$a(0) = (-1)^p\frac{\alpha(0)}{2s(n)}$$

(*for* $m = 1$ *to* n)

$$a(m) = (-1)^p\frac{\alpha(m)}{s(n)}$$

(*end loop on* m)

impulse response

$$h(n) = \frac{y_m - a(0)}{y_m + 1}$$

(*for* $m = 1$ *to* n)

$$h(n \pm m) = -\frac{a(m)}{2(y_m + 1)}$$

(*end loop on* m)

 ISBN: 978-1-61208-183-0

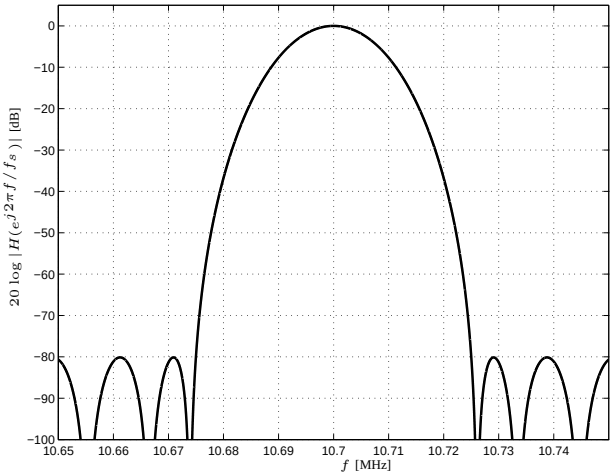


Fig. 4. Detailed view of the pass-band.

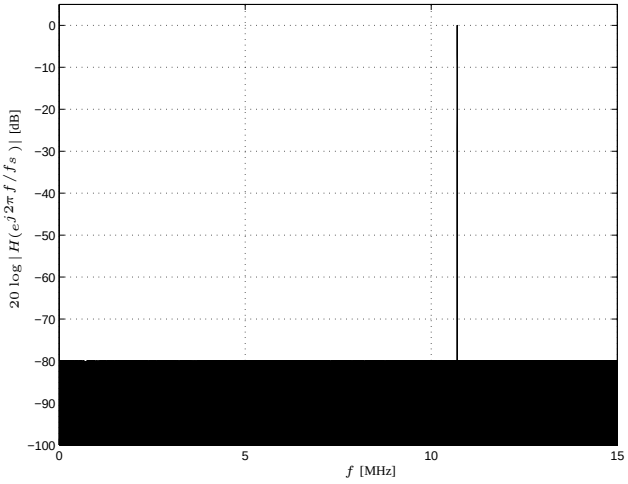


Fig. 5. Amplitude frequency response $20\log|H(e^{j2\pi f/f_s})|$ [dB].

view of its passband is shown in Fig. 6. In order to point out the robustness of our design procedure, let us do comparative designs of both filters presented above by the established numerical Parks-McClellan procedure [6] which is implemented e.g. in the Matlab function *firgr*. The filter with the length of 4053 coefficients from our example can be designed using the function call [h,err,res]=firgr(4052,[0 10.675 10.7 10.725 15]/15,[0 0 1 0 0],'n' 'n' 's' 'n' 'n') easily. Note that the Matlab function $firgr$ returns filters with different normalization of the ripples of the corresponding frequency response. Except for the difference in the normalization, the obtained results are identical. On the other hand the design of the filter with the length of 40497 coefficients specified in this section cannot be designed using the function call [h,err,res]=firgr(40496,[0 10.6975 10.7 10.7025 15]/15,[0 0 1 0 0],'n' 'n' 's' 'n' 'n') as it collapses (Matlab R2010b) because of numerical problems. We assume that this failure is caused by the failed numerical evaluation of the densely located roots (isoextremal values) of the optimized function. To our knowledge, our design

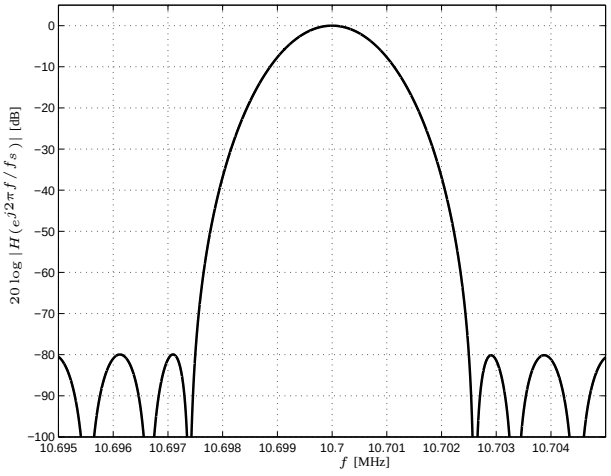


Fig. 6. Detailed view of the pass-band.

procedure presented here has no parallel in the design of high performance ER BP FIR filters with the length beyond several thousands of coefficients. Further note that in the Parks-McClellan design procedure the length of the filter is an input argument, not the result of the design.

VII. Implementation of the Filter

There are various ways for implementing of FIR filters in real time available. For high order filters, the digital signal processors (DSP) and the field programmable gate arrays (FPGA) dominate. We prefer the filter implementation using digital DSPs over the implementation using FPGAs mainly because it is a less time consuming process. The band-pass FIR filter with the length of 4053 coefficients from our example requires 122 billions multiply-and-accumulate (GMACs) operations per second. The adequate DSP is the eight-core DSP TMS320C6678 [7] which provides 320GMACs operations per second in the 16-bit fixed point arithmetics. For the real-time implementation of the filter we use the DSP Evaluation Module TMDXEVM6678 (Fig. 7). The implementation of the filter in the IEEE-754 compliant single precision floating point arithmetics would require a two chip solution based on the DSP TMS320C6678 which provides 160 billions floating point operations (GFLOPs) per second per chip, or a single chip solution based on the 32-core DSP TMS320TCI6609 [8]. The implementation of the filter with the length of 40497 coefficients from previous section requires 1215 GMACs and consequently its implementation would require a multi-chip solution, e.g. three chip solution based on the 32-core DSP TMS320TCI6609.

VIII. Conclusions

We have presented an analytical design of high performance digital equiripple band-pass finite impulse response filters. In contrast to the established numerical design procedures the proposed design method is based on the generating polynomial

 ISBN: 978-1-61208-183-0

Fig. 7. Evaluation Module TMDXEVM6678.

and provides a formula for the degree of the filter and formulas for the evaluation of the coefficients of the impulse response of the filter. The demonstrated robustness is another advantage of the proposed design method.

ACKNOWLEDGEMENTS

This activity was thankfully supported by the grant No. VG20102015053, Ministry of the Interior, Czech Republic and by Texas Instruments, Inc.

REFERENCES

[1] P. Zahradnik, M. Vlček, Fast Analytical Design Algorithms for FIR Notch Filters, *IEEE Transactions on Circuits and Systems*. March 2004 , Vol. 51, No. 3, pp. 608 - 623.

[2] N. I. Achieser, Über einige Funktionen, die in gegebenen Intervallen am wenigsten von Null abweichen, *Bull. de la Soc. Phys. Math. de Kazan*, Vol. 3, pp. 1 - 69, 1928.

[3] D. F. Lawden *Elliptic Functions and Applications* Springer-Verlag, New York Inc., 1989.

[4] M. Abramowitz, I. Stegun, *Handbook of Mathematical Function*, Dover Publication, New York Inc., 1972.

[5] M. Vlček, R. Unbehauen, Zolotarev Polynomials and Optimal FIR Filters, *IEEE Transactions on Signal Processing*, Vol.47, No.3, pp. 717-730, March 1999.

[6] J. H. Mc Clellan, T. W. Parks, L. R. Rabiner, FIR Linear Phase Filter Design Program, *Programs for Digital Signal Processing*. newblock New York: IEEE Press, 1979.

[7] www.ti.com/product/tms320c6678

[8] www.ti.com/lit/ml/sprt619/sprt619.pdf

 ISBN: 978-1-61208-183-0

Comparison of the Fully-Differential and Single-Ended Solutions of the Frequency Filter with Current Followers and Adjustable Current Amplifier

Jan Jerabek and Kamil Vrba
Department of Telecommunications, Faculty of Electrical Engineering and Communication
Brno University of Technology
Purkynova 118, 612 00 Brno, Czech Republic
Emails: jerabekj@feec.vutbr.cz, vrbak@feec.vutbr.cz

***Abstract*—Two solutions of universal and adjustable current-mode filters are presented in this contribution. The first of them is able to process single-ended (S-E) signals in communications and the other can operate fully-differentially (F-D) and therefore is well applicable for balanced transmission lines. Both circuits have adjustable quality factor and both are analyzed in this contribution. Their simulation results are compared to each other. Main contribution of this paper is the presentation of two novel solutions and their mutual comparison.**

Keywords-adjustable amplifier; DACA; fully-differential.

I. INTRODUCTION

F-D structures [1]–[10], usually used on balanced communication lines, have several benefits when compared to the single-ended (S-E) circuits. It is, for instance, higher dynamic range of the signals, high attenuation of common-mode signal, better power supply rejection ratio, and lower harmonic distortion. F-D structures also have some disadvantages. They are, in particular, larger area needed on the chip, which is related to greater power consumption, and sometimes the design of F-D structures is more complex with respect to S-E topologies.

The basics of the design of simple F-D structures with a high Common Mode Rejection Ratio (CMRR) (by coupling two S-E structures) were described in [1]. Transconductance elements such as the Balanced Operational Transconductance Amplifier (BOTA) [2] are very often present in F-D filters. Differential-input buffered and transconductance amplifiers (DBTA) [3] can also be applied, for instance. The Fully Differential Current Feedback Operational Amplifier (FDCFOA) operating in the voltage mode and having various internal structures is also quite common [4]; for example fully-differential current conveyors of the second generation (FDCCII) [5]–[7] or fully-differential current followers (FD-CF) [9], [10]. The structures traditionally work in the voltage-mode (VM); however, recent research is also focused on the current-mode (CM) filters. Various conceptions of simple F-D circuits capable of processing current-mode signals can be found in [8], while the methodology for the F-D filter design with various target requirements was presented in [11].

Recently, current followers with non-unity gain [12] or current amplifiers [13], [14] have been presented and should be suitable for high-frequency applications. In [9], [15], [16], the Digitally Adjustable Current Amplifier (DACA) has been presented.

The newly designed structure of the universal filter working in the current mode is compared with its F-D equivalent in this contribution. Both solutions provide the possibility of digital adjustment of the quality factor. Multiple-output current follower (MO-CF) [17], [18], its fully-differential equivalent, Fully-Differential Current Follower (FD-CF), and DACA are used as active elements. The main aim of this work is to compare these F-D and S-E solutions, because this approach is not so common.

Contribution is organized as follows: Section II provides short description of active elements; Section III includes designed filters and Section IV summarizes simulation results.

II. ACTIVE ELEMENTS DEFINITIONS

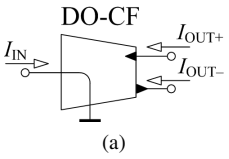
The S-E and F-D structures presented in this contribution operate with three types of active element. One is a simple current active follower with dual or multiple outputs (DO-CF, MO-CF) [17]. As an example, the DO-CF schematic symbol is shown in Fig. 1a, and its simple 3rd-level simulation model suitable for AC analysis is shown in Fig. 1b. This model covers only input and output impedances. Ideally, the current transfer from an input to an output is unity, with inverted or non-inverted phase of the signal.

The F-D equivalent of the DO-CF circuit is the Fully-Differential Current Follower (FD-CF), which is suitable for fully-differential signal processing. It has at least four outputs, two with positive current transfer and two with negative current transfer from the input nodes. The FD-CF schematic symbol is shown in Fig. 2a, a simple 3rd-order AC simulation model is shown in Fig. 2b. The ideal FD-CF is described by

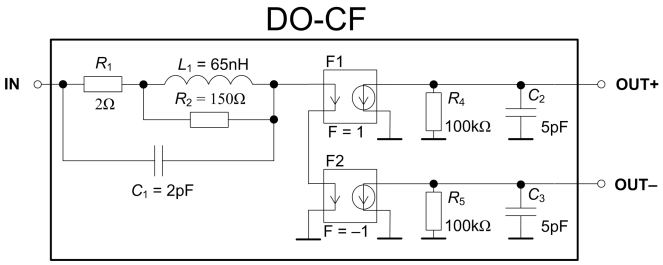
$$I_{\mathrm{OUT1+}} = I_{\mathrm{OUT2+}} = (1/2)(I_{\mathrm{IN+}} - I_{\mathrm{IN-}}), \qquad (1)$$

$$I_{\mathrm{OUT1-}} = I_{\mathrm{OUT2-}} = -(1/2)(I_{\mathrm{IN+}} - I_{\mathrm{IN-}}). \qquad (2)$$

 ISBN: 978-1-61208-183-0

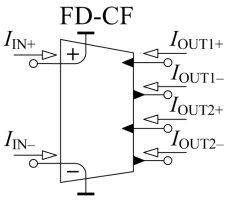


(a)

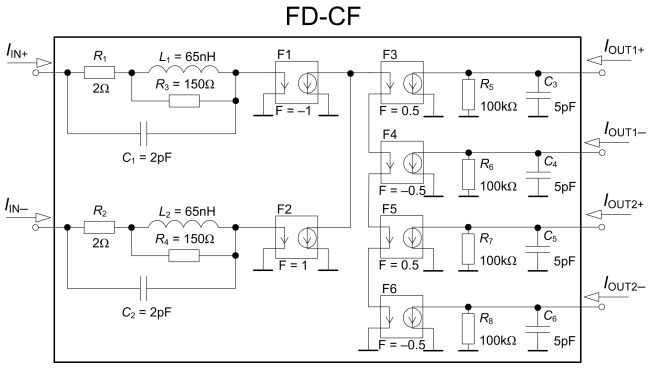


(b)

Figure 1. Dual-Output Current Follower (DO-CF): (a) schematic symbol (b) 3rd-order AC simulation model



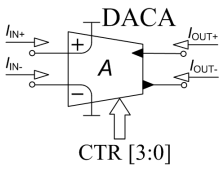
(a)



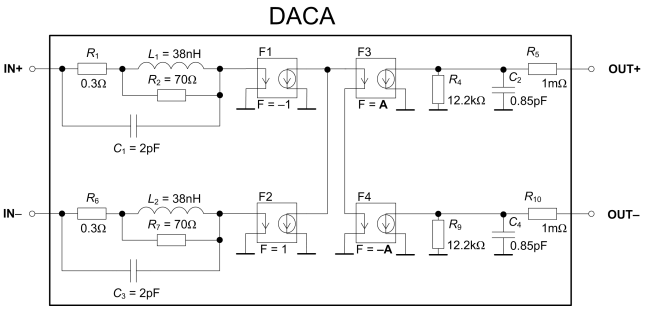
(b)

Figure 2. Fully-Differential Current Follower (FD-CF): (a) schematic symbol (b) 3rd-order AC simulation model

A Digitally Adjustable Current Amplifier (DACA) (Fig. 3a) is the other active element. The key feature of DACA is that current gain (A) is adjustable and can be controlled by three-bit digital bus. The DACA circuit was lately developed in cooperation with ON Semiconductor in the CMOS 0.35 μm technology. We have several samples from the second test batch available and they are currently undergoing the first tests. The DACA 3rd-level AC simulation model is depicted in Fig. 3b. The current transfers of the DACA element are given by the relations



(a)



(b)

Figure 3. Digitally Adjustable Current Amplifier (DACA): (a) schematic symbol (b) 3rd-order AC simulation model

$$I_{\mathrm{ID}} = I_{\mathrm{IN+}} - I_{\mathrm{IN-}}\,, \tag{3}$$

$$I_{\mathrm{OD}} = I_{\mathrm{OUT+}} - I_{\mathrm{OUT-}}, \quad I_{\mathrm{OD}} = 2AI_{\mathrm{ID}}\,, \tag{4}$$

$$I_{\mathrm{OUT+}} = A(I_{\mathrm{IN+}} - I_{\mathrm{IN-}}), \tag{5}$$

$$I_{\mathrm{OUT-}} = -A(I_{\mathrm{IN+}} - I_{\mathrm{IN-}}). \tag{6}$$

where I_{ID} represents the differential input current, I_{OD} is the differential output current, and A stands for the adjustable current gain of DACA element. It is clear that the differential gain is twice higher than the single-ended gain. A can be adjusted from 1 to 8 in steps of 1.

Measurement results for the DACA features are not yet available; therefore the DACA is modeled only partially and the model does not cover all parameters. Only input and output impedances are modeled, similarly to DO-CF and FD-CF elements.

III. Designed S-E and F-D Filter

Universal filter with current-only active elements was designed in both the single-ended (Fig. 4) and the fully-differential (Fig. 5) variant. Independent adjusting of the quality factor for every filtering function is possible by adjustable current gain of DACA in both variants.

The denominator of all transfer functions is for the S-E filter equal to:

$$D(\boldsymbol{s}) = 1 + \boldsymbol{s}C_2R_2A + \boldsymbol{s}^2C_1C_2R_1R_2\,. \tag{7}$$

Provided transfer functions are:

$$\frac{\boldsymbol{I}_{\mathrm{LP}}}{\boldsymbol{I}_{\mathrm{IN}}} = -\frac{\boldsymbol{I}_{\mathrm{iLP}}}{\boldsymbol{I}_{\mathrm{IN}}} = \frac{1}{D(\boldsymbol{s})}\,, \tag{8}$$

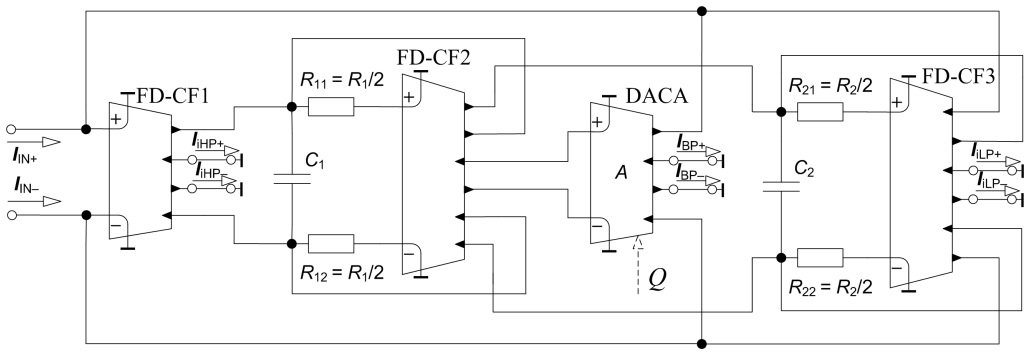


Figure 5. Fully-differential universal and adjustable frequency filter with three FD-CF and one DACA elements working in the current mode

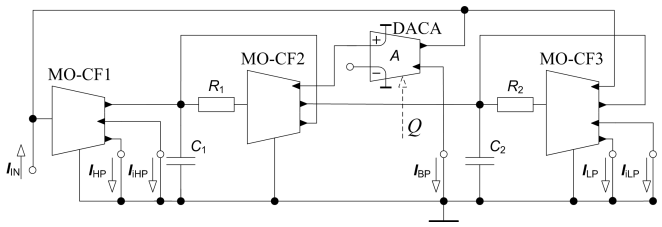


Figure 4. Single-ended universal and adjustable frequency filter with three MO-CF and one DACA elements working in the current mode

$$\frac{\boldsymbol{I}_{\mathrm{BP}}}{\boldsymbol{I}_{\mathrm{VST}}} = \frac{\boldsymbol{s}C_2R_2A}{D(\boldsymbol{s})}\,, \tag{9}$$

$$\frac{\boldsymbol{I}_{\mathrm{HP}}}{\boldsymbol{I}_{\mathrm{IN}}} = -\frac{\boldsymbol{I}_{\mathrm{iHP}}}{\boldsymbol{I}_{\mathrm{IN}}} = \frac{\boldsymbol{s}^2C_1C_2R_1R_2}{D(\boldsymbol{s})}\,, \tag{10}$$

$$\frac{\boldsymbol{I}_{\mathrm{LP}}+\boldsymbol{I}_{\mathrm{HP}}}{\boldsymbol{I}_{\mathrm{IN}}} = -\frac{\boldsymbol{I}_{\mathrm{iLP}}+\boldsymbol{I}_{\mathrm{iHP}}}{\boldsymbol{I}_{\mathrm{IN}}} = \frac{1+\boldsymbol{s}^2C_1C_2R_1R_2}{D(\boldsymbol{s})}\,, \tag{11}$$

$$\frac{\boldsymbol{I}_{\mathrm{iLP}}+\boldsymbol{I}_{\mathrm{BP}}+\boldsymbol{I}_{\mathrm{iHP}}}{\boldsymbol{I}_{\mathrm{IN}}} = -\frac{1-\boldsymbol{s}C_2R_2A+\boldsymbol{s}^2C_1C_2R_1R_2}{D(\boldsymbol{s})}\,. \tag{12}$$

Relations for angular frequency and quality factor can be easily derived:

$$\omega_0 = \sqrt{\frac{1}{R_1R_2C_1C_2}}\,, \tag{13}$$

$$Q = \frac{1}{A}\sqrt{\frac{R_1C_1}{R_2C_2}}\,. \tag{14}$$

It is obvious that the quality factor of filters from Fig. 4 and Fig. 5 can be controlled by DACA gain A with an inverse proportion. The F-D filter is designed so as to have almost the same transfer functions as the S-E filter thanks to appropriately modified values of passive elements as shown in Fig. 5. In order to obtain particular transfer functions for the F-D filter, A in each of the equations has to be replaced by $2A$ because of the differential gain of DACA, which is twice higher than the S-E gain, as demonstrated by eqs. (3)–(6).

Table I
VALUES OF PASSIVE COMPONENTS

Variant [–]	C_1 [pF]	C_2 [pF]	R_1 [kΩ]	R_2 [Ω]
S-E	430	68	2	390
F-D	430	68	4	200

IV. SIMULATION RESULTS

To verify the theoretical presumptions, the behavior of both the S-E and the F-D filters has been analyzed by Spice simulations. The chosen or calculated values are summarized in Table I. Theoretical pole frequency is 1 MHz in each case, theoretical quality factor is $Q = \{0.9;\ 1.3;\ 2.9;\ 5.7\}$, obtained by gain values $A = \{8;\ 5;\ 2;\ 1\}$. It is clear that resistor values are changed in the case of the F-D filter from Fig. 5, because they are placed in lengthwise branches. Therefore, $R_{11} = R_{12} = 2\ \mathrm{k\Omega}$ and $R_{21} = R_{21} = 100\ \Omega$. Floating capacitor C_1 (and C_2, of course) could be replaced by two grounded capacitors in the particular solution. These capacitors would be 860 pF in the case of C_1 and 136 pF in the case of C_2.

Simulation results comparing the S-E and the F-D filter are shown in Fig. 6. All simulations were done with simple models shown in Fig. 1b, Fig. 2b and Fig. 3b. The graph in Fig. 6a contains magnitude responses of inverting low-pass, band-pass, inverting high-pass and inverting band-stop filters, Fig. 6b shows an example of quality factor adjustment in the case of band-pass filter, and Fig. 6c includes all characteristics of all-pass filter.

The differences between the S-E and the F-D solutions are clearly visible in the low-frequency area, particulary in the case of iHP and BP functions. The F-D filter provides a slightly higher low-frequency attenuation than the S-E solution. In the current mode, low-frequency attenuation is dependent on the output impedances of active elements, but in this particular case, the difference is caused mainly by unequal values of resistors. The theoretical values of the quality factor of BP filters are included in Fig. 6b, the

 ISBN: 978-1-61208-183-0

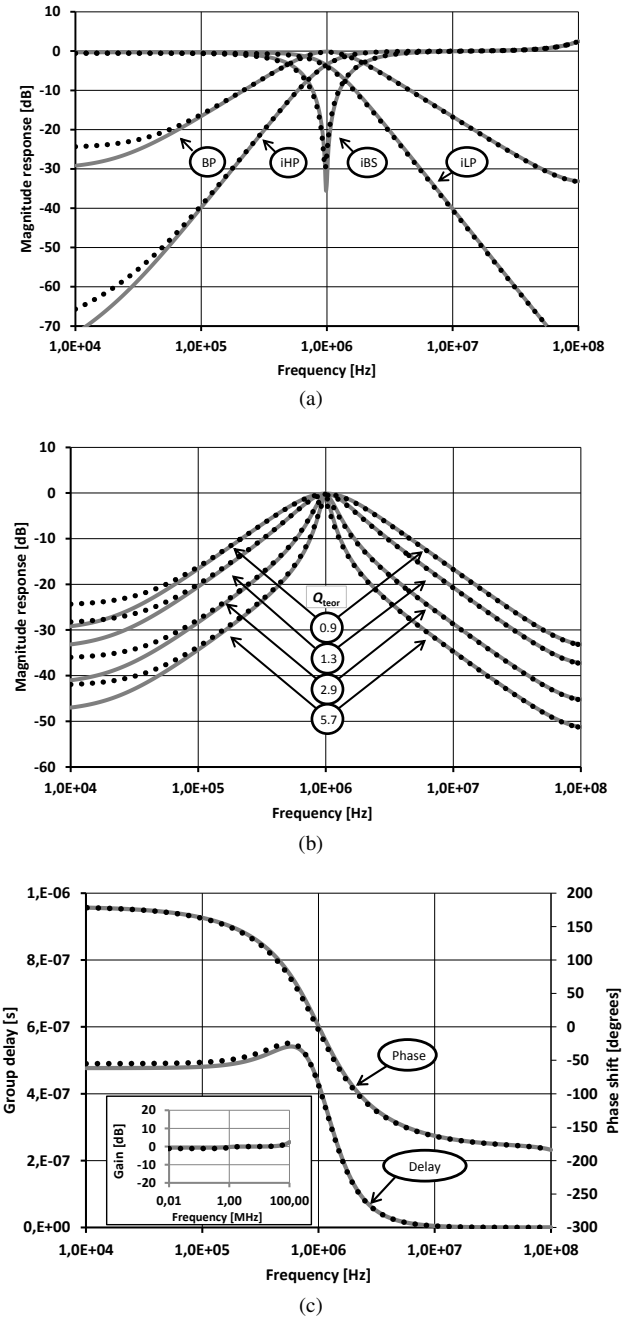


Figure 6. Simulation results of universal and adjustable filters with four current-only active elements; F-D filter (solid line) compared to S-E filter (dotted line). (a) magnitude response of iLP, iHP, BP and iBS functions (b) adjustement of quality factor in case of BP filter (c) iAP - magnitude and phase response, group delay

simulation results for S-E are 4.3, 2.5, 1.2 and 0.8, and the simulation results for F-D are 4.9, 2.6, 1.2, 0.8.

V. Conclusion

Both S-E and F-D filters have several benefits when compared to each other. Simulation results that were shown in this contribution showed that both solutions provide comparable features and therefore both of them can find good applications in communications and transmission systems.

Acknowledgment

This work was supported in part by the Czech Science Foundation, project 102/09/1681 and by the Czech Ministry of Education, program MSM 0021630513.

References

[1] O. Casas and R. Pallas-Areny. "Basics of Analog Differential Filters", *IEEE Transaction on Instrumentation and Measurement*, Vol. 45, No. 1, pp. 275–279, 1996.

[2] M. O. Shaker, S. A. Mahmoud, and A. M. Soliman, "New CMOS Fully-Differential Transconductor and Application to a Fully-Differential Gm-C Filter". *ETRI Journal*, Vol. 28, No. 2, pp. 175–181, 2006.

[3] N. Herencsar, J. Koton, and K. Vrba, "Differential-Input Buffered and Transconductance Amplifier (DBTA)-Based New Trans-Admittance- and Voltage-Mode First-Order All-Pass Filters".*In Proceedings of the 6th International Conference on Electrical and Electronics Engineering - ELECO' 09*. Turkey: EMO Yayinlari, pp. 256–259, 2009.

[4] S. A. Mahmoud, "Low Voltage Fully Differential CMOS Current Feedback Operational Amplifier", *Proc 47th IEEE Midwest Int Symposium Circuits and Systems*, Vol. 1, pp. 49–52, 2004.

[5] E. A. Soliman and S. A. Mahmoud, "New CMOS fully differential current conveyor and its application in realizing sixth order complex filter". *IEEE International Symposium Circuits and Systems*, pp. 57–60, 2009.

[6] E. A. Sobhy and A. M. Soliman, "Realizations of fully differential voltage second generation current conveyor with an Application". *International Journal of Circuit Theory and Applications*, Vol. 38, No. 5, pp. 441–452, 2010.

[7] C. M. Chang, B. M. Al-Hashimi, C. L. Wang, and C. W. Hung, "Single fully differential current conveyor biquad filters", *IEEE Proc. of Circuits, Devices and Systems*, No. 5, pp. 394–398, 2003.

[8] R. H. Zele, D. J. Allstot, and T. S. Fiez, "Fully balanced CMOS current-mode circuits". *IEEE Journal of Solid-State Circuits*, Vol. 28, No. 5, pp. 569–575, 1993.

[9] J. Jerabek, R. Sotner, and K. Vrba, "Fully-differential current amplifier and its application to universal and adjustable filter". *In 2010 Int Conf on Applied Electronics*. Pilsen: University of West Bohemia, Czech Republic, pp. 141–144, 2010.

[10] A. Hussain and N. Tasadduq, "Realizations of CMOS Fully Differential Current Followers/Amplifiers", *IEEE International Symposium on Circuits and Systems (ISCAS 2009)*, Taipei, Taiwan, pp.1881–1384, 2009.

[11] M. Massarotto, O. Casas, V. Ferrari, and R. Pallas-Areny, "Improved Fully Differential Analog Filters". *IEEE Transaction on Instrumentation and Measurement*, Vol. 56, No. 6, pp. 2464–2469, 2007.

[12] W. Tangsrirat and T. Pukkalanun, "Digitally programmable current follower and its applications," *Int. J. Electron. Commun. (AEU)*, vol. 63, no. 5, pp. 416–422, 2009.

 ISBN: 978-1-61208-183-0

[13] H. A. Alzaher, ”A CMOS digitally programmable universal current mode filter,” *IEEE Trans. Circuits and Systems II*, vol. 55, no. 8, pp. 758–762, 2008.

[14] B. Sedighi and M. S. Bakhtiar, ”Variable gain current mirror for highspeed appications,” *IEICE Electronics Express*, vol. 4, no. 8, pp. 277–281, 2007.

[15] J. Koton, N. Herencsar, J. Jerabek, and K. Vrba, ”Fully Differential Current-Mode Band-Pass Filter: Two Design Solutions”. *In Proc. 33rd Int Conf on Telecom and Signal Procesing, TSP 2010*. Baden, Austria, pp. 1–4, 2010.

[16] J. Jerabek, K. Vrba, and M. Jelinek, ”Universal Fully-Differential Adjustable Filter with Current Conveyors and Current Amplifier in Comparison with Single-Ended Solution,” *In 2011 Int Conf on Applied Electronics*. Pilsen: University of West Bohemia, Czech Republic, pp. 189–192, 2011.

[17] J. Jerabek, R. Sotner, and K. Vrba, ”Current-Mode Filters with Single-Ended and Fully-Differential Nth- order Synthetic Elements”. *In Proceedings of the 34th International Conference on Telecommunications and Signal Processing TSP 2011*. Budapest: Hungary, pp. 269–273, 2011.

[18] J. Jerabek and K. Vrba, ”Design of Fully-Differential Filters with nth-order Synthetic Elements and Comparison with Single- Ended Solution”. *In Proc of the 2011 Int Conf on Computer and Communication Devices (ICCCD 2011)*. Bali, Indonesia, pp. V1-48 – V1-52, 2011.

 ISBN: 978-1-61208-183-0

Precision Full-wave Rectifier Using Current Conveyors and Two Diodes

Jaroslav Koton, Norbert Herencsar, Kamil Vrba
Dept. of Telecommunications
Brno University of Technology
Purkynova 118, 612 00 Brno, Czech Republic
{*koton, herencsn, vrbak*}*@feec.vutbr.cz*

***Abstract*—In this paper, precision full-wave rectifier employing two current conveyors and only two diodes is presented. The proposed structure operates in mixed- or voltage-mode. To compare the behavior of the proposed structure, the frequency dependent RMS error and DC transient value for different values of input voltage amplitudes are evaluated.**

***Keywords*-Precision full-wave rectifier; current conveyor; voltage conveyor; measurements.**

I. INTRODUCTION

In applications such as ac voltmeters and ammeters, signal-polarity detectors, averaging circuits, peak-value detector rectification function is of great importance [1]. Because of the threshold voltage of the diodes, simple passive rectifiers operate inaccurately, if low-voltage signals are analyzed. Therefore, precision rectifiers employing active elements have to be used.

Probably, the most known precision rectifiers are based on operational amplifiers (opamps) [1]. However, because of the finite slew-rate and effects caused by diode commutation, these circuits operate well only at low frequencies [2], [3]. This problem can be overcome by the use of current conveyors (CCs), where the diodes are connected to the high-impedance current outputs of the active elements. In [4]–[7] the same precision full-wave rectifier is analyzed (Fig. 2b). It uses two second-generation CCs and four diodes. To further extend the frequency range the voltage [4], [7] or current [6], [7] biasing scheme can be used. Another precision full-wave rectifier is presented in [8] that is based on the standard opamp rectifier shown in Fig. 2a. Here, the OPA_1 is replaced by the operational conveyor and later by second-generation CC [3]. A full-wave rectifiers using second-generation and dual-X CCs are presented in [9] and [10], respectively, where the required diodes are suitably replaced by NMOS transistors. The use of fully differential operational transconductance amplifiers (BOTA) operating in weak inversion region for the design of precision full-wave rectifiers is presented in [11], which is based on the idea discussed in [12], where simple transconductance amplifiers (OTA) are used. Here, the transconductance of OTA is controlled by the current derived from the input signal to be rectified. In another group of precision rectifiers, a transistor connected to the current output of an active element operates as a switch. For this purpose the current conveyor [13] or transconductance amplifiers [14]–[16] are used.

In this paper, precision full-wave rectifier employing two second-generation CCs and two diodes is presented. It is of minimal configuration and operates in the voltage- or mixed-mode. Voltage or current biasing scheme can be also used to extend the frequency operation range. The behavior of the circuit is compared to the known conveyor based solution presented in [4]–[7]. Simulation results are given that show the feasibility of the newly designed circuit to rectify signals up to 1 MHz and beyond with no or little distortion.

II. CURRENT CONVEYORS

In 1968, the current conveyors were presented for the first time [22], however they did not found any significant usage since the operational amplifiers were more attractive at that time. Current conveyors received considerable attention after the second (CCII) [23] and later third (CCIII) [24] generation current conveyors were designed. These elements are now advantageously used in applications, where the wide bandwidth or current output response is necessary. Nowadays, different types of current conveyors are described that mostly base on the CCII, e.g., CCCII [25], DVCC [26], or ECCII [27]. The behavior of a four-terminal CCII (Fig. 1) is described by following equations:

$$v_{\mathrm{X}} = v_{\mathrm{Y}}, \quad i_{\mathrm{Y}} = 0, \quad i_{\mathrm{Z+}} = i_{\mathrm{X}}, \quad i_{\mathrm{Z-}} = -i_{\mathrm{X}}. \tag{1}$$

III. NEW PRECISION FULL-WAVE RECTIFIER

The standard op amp based circuit from Fig. 2a [1] is a connection of an inverting half-wave rectifier (OPA_1) and summing amplifier (OPA_2). For desired full-wave rectification following conditions have to be fulfilled:

$$R_1 = R_2, \quad R_4 = 2R_3, \tag{2a}$$

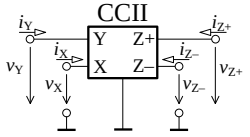


Figure 1. Circuit symbol of the four-terminal CCII

 ISBN: 978-1-61208-183-0

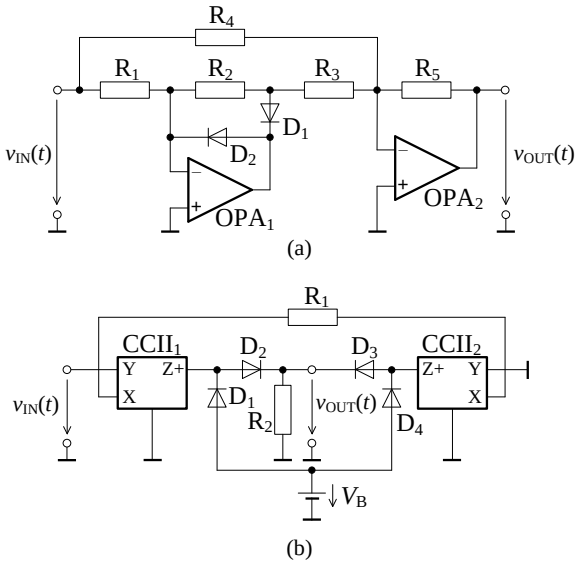


Figure 2. Voltage-mode (a) standard op amp based [1], (b) known conveyor based full-wave rectifier from [4]-[7]

or

$$R_2 = 2R_1, \quad R_3 = R_4, \tag{2b}$$

which generally means that the half-wave rectified signal must be amplified two times higher than the original input signal, either more commonly by the summing amplifier (according to (2a)) or by the half-wave rectifier (according to (2b)).

A well known circuit topology of the full-wave rectifier using two second-generation current conveyors and four diodes is shown in Fig. 2b [4]-[7]. Both CCIIs form a differential voltage-to-current converter. During the positive and negative input cycle the output currents make only the diodes D_2, D_4 and D_1, D_3 active, respectively. On the resistor R_2 the output current is converted back to voltage.

The newly proposed structure of the full-wave rectifier is shown in Fig. 3. Basically, it uses only one current conveyor ($CCII_1$) and two diodes. The second current conveyor operates as a current follower, where the resistors R_2 and R_3 connected to the Z-terminals convert the current back to voltage. Therefore, this full-wave rectifier can operate in the voltage- or mixed-mode.

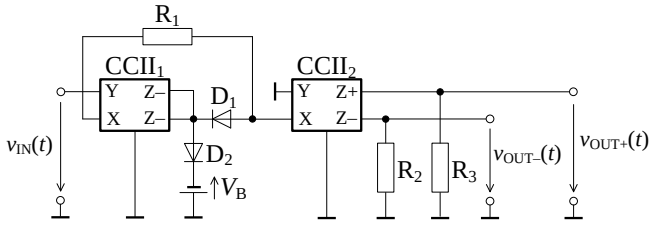


Figure 3. Proposed voltage-mode precision full-wave rectifier

The output voltages of the rectifier from Fig. 3 can be expressed as:

$$v_{\mathrm{OUT}+}(t) = \frac{R_3}{R_1}|v_{\mathrm{IN}}(t)|, \quad v_{\mathrm{OUT}-}(t) = -\frac{R_2}{R_1}|v_{\mathrm{IN}}(t)|, \tag{3}$$

Since the input voltage $v_{\mathrm{IN}}(t)$ is directly connected to the Y-terminal of the $CCII_1$, the input impedance of the rectifier is infinitely high in theory. Operated in mixed-mode, the current response is directly sensed at the Z-terminal having in theory infinitely high output impedance.

IV. DC and RMS Error Analyses

To evaluate and compare the accuracy of the voltage-mode full-wave rectifiers from Fig. 2 and Fig. 3, the DC value transfer p_{DC} and RMS error p_{RMS} have been analyzed [28]:

$$p_{\mathrm{DC}} = \frac{\int\limits_T y_{\mathrm{R}}(t)\,\mathrm{d}t}{\int\limits_T y_{\mathrm{ID}}(t)\,\mathrm{d}t}, \tag{4a}$$

$$p_{\mathrm{RMS}} = \sqrt{\frac{\int\limits_T \left[y_{\mathrm{R}}(t) - y_{\mathrm{ID}}(t)\right]^2\,\mathrm{d}t}{\int\limits_T y_{\mathrm{ID}}^2(t)\,\mathrm{d}t}}. \tag{4b}$$

where the $y_{\mathrm{R}}(t)$ and $y_{\mathrm{ID}}(t)$ represent the actual and ideally rectified signal and T is the period of the input signal. The ideal behavior of the rectifier is characterized by the values $p_{\mathrm{RMS}} = 0$ and $p_{\mathrm{DC}} = 1$.

V. Simulation Results

The behavior of the proposed voltage-mode full-wave rectifier has been compared with the circuit solutions from Fig. 2. As active elements the universal current conveyor UCC-N1B have been used. The current and voltage transfer bandwidths of the UCC are about 35 MHz [29]. Therefore, in the standard op amp based rectifier the AD8656 has been used [30]. The diodes are general purpose 1N4148 and all

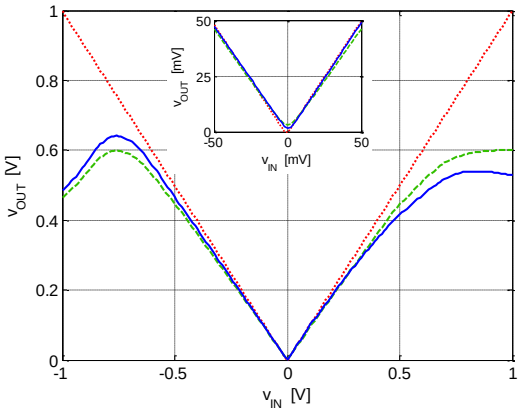


Figure 4. DC transfer of the full-wave rectifiers from Fig. 2a (dotted line), Fig. 2b (dashed line), and Fig. 3 (solid line)

 ISBN: 978-1-61208-183-0

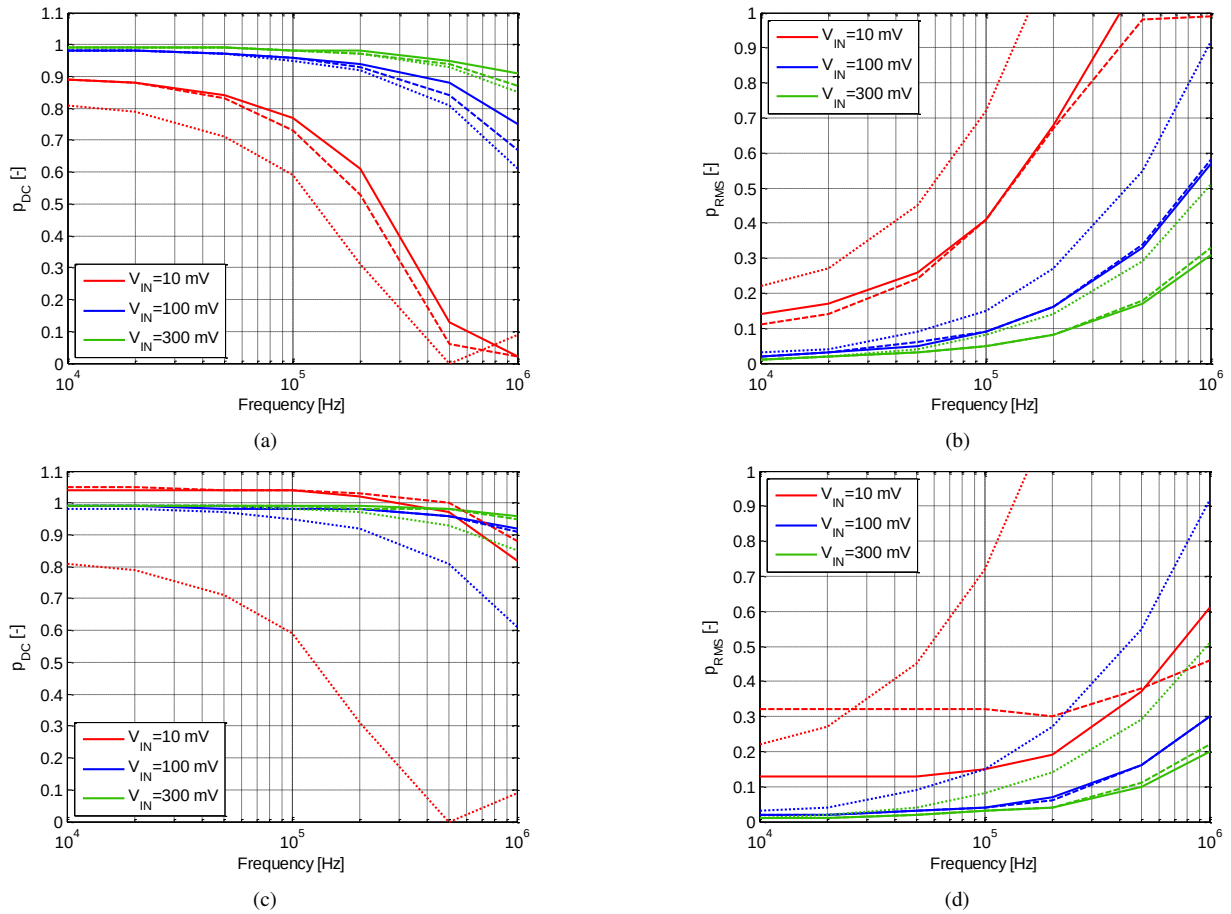


Figure 5. (a), (c) DC value transfer for $V_{\mathrm{B}} = 0$ V and $V_{\mathrm{B}} = 0.65$ V, (b), (d) RMS error for $V_{\mathrm{B}} = 0$ V and $V_{\mathrm{B}} = 0.65$ V of the rectifiers from Fig. 2a (dotted line), Fig. 2b (dashed line), and Fig. 3 (solid line), for input voltage amplitudes 10 mV, 100 mV, and 300 mV

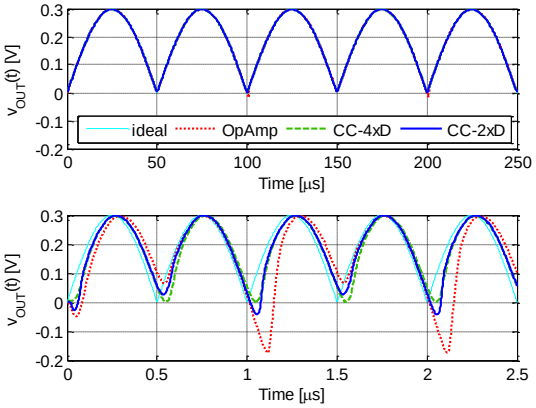


Figure 6. Transient simulation results of the full-wave rectifiers from Fig. 2a (dotted line), Fig. 2b (dashed line), and Fig. 3 (solid line) for frequencies 10 kHz and 1 MHz

resistors are 1 kΩ (in Fig. 2a $R_3 = 500$ Ω). In Fig. 4, the DC transfer characteristics are shown. Due to high voltage gain of the op amps, the DC error of the circuit from Fig. 2a is minimized and the DC transfer is almost ideal (in Fig. 4 dotted line). The non-unity voltage and current transfers of the current conveyors cause higher DC error of the conveyor based full-wave rectifiers (solid and dashed line).

Using (4) the behavior of the proposed rectifier has been analyzed in frequency domain and compared to the circuit solution from Fig. 2. The simulation results of the frequency dependent RMS error and DC value transfer for chosen values of amplitudes v_{IN} are shown in Fig. 5a and Fig. 5b, where the dotted, dashed, and solid lines stand for the circuits from Fig. 2a, Fig. 2b, and Fig. 3, respectively. If the frequency increases and/or amplitude decreases distortions occur and the p_{DC} decreases below one and p_{RMS} increases. From Fig. 5a, it is evident that the best results are achieved with the new minimal configuration rectifier. For an appropriate value of the bias voltage (here $V_{\mathrm{B}} = 0.65$ V), the conveyor based precision rectifiers can operate at higher frequencies (Fig. 5c, Fig. 5d).

The complete the simulations, for frequencies 10 kHz and 1 MHz and input amplitude $V_{\mathrm{IN}} = 300$ mV the time-domain performance of the analyzed rectifiers is shown in Fig. 6 ($V_{\mathrm{B}} = 0.65$ V). Also, one can observe, that at low

 ISBN: 978-1-61208-183-0

frequencies the behavior of the rectifiers is nearly the same. Once the frequency of the processed signal rises deviations occur.

VI. Conclusion

In this paper, the performance of conveyor based precision full-wave rectifier has been analyzed and compared to the standard opamp based topology. A new minimal configuration conveyor based rectifier has been presented, that employs current conveyors and two diodes. The rectifier can work in the voltage- or mixed-mode. Simulation were performed that prove the feasibility of the proposed conveyor based full-wave rectifier.

Acknowledgment

The paper has been supported by projects: GACR 102/10/P561, GACR 102/09/1681, MSM0021630513, and FEKT–S–11–15.

References

[1] U. Tietze, CH. Schenk, and E. Gramm, *Electronic Circuits-Handbook for Design and Application*, Springer, 2008.

[2] C. Toumazou and F. J. Lidgey, Fast current-mode precision rectifier, *Electron. Wireless World*, vol. 93, no. 1621, pp. 1115-1118, 1987.

[3] S. J. G. Gift and B. Maundy, Versatile precision full-wave rectifiers for instrumentation and measurements, *IEEE Trans Instrum. Meas.*, vol. 56, no. 5, pp. 1703-1710, 2007.

[4] C. Toumazou F. J. Lidgey, and S. Chattong, High frequency current conveyor precision full-wave rectifier, *Electronics Letters*, vol. 30, no. 10, pp. 745-746, 1994.

[5] A. A. Khan, EL-ELA, M. A. El-Ela, and M. A. Al-Turaigi, Current-mode precision rectification, *Int. J. Electron.*, vol. 79, no. 6, pp. 853-859, 1995.

[6] B. Wilson and V. Mannama, Current-mode rectifier with improved precision, *Electronics Letters*, vol. 31, no. 4, pp. 247-248, 1995.

[7] D. Stiurica, Truly temperature independent current conveyor precision rectifier, *Electronics Letters*, vol. 31, no. 16, pp. 1302-1303, 1995.

[8] S. J. G. Gift, A high-performance full-wave rectifier circuit, *Int. J. Electron.*, vol. 87, no. 8, pp. 925-930, 2000.

[9] E. Yuce, S. Minaei, and O. Cicekoglu, Full-wave rectifier realization using only two CCII+s and NMOS transistors, *Int. J. Electron.*, vol. 93, no. 8, pp. 533-541, 2006.

[10] S. Minaei and E. Yuce, A New Full-Wave Rectifier Circuit Employing Single Dual-X Current Conveyor, *Int. J. Electron.*, vol. 95, no. 8, pp. 777-784, 2008.

[11] C. Chanaproma, and K.Daoden, A CMOS fully differential operational transconductance amplifier operating in subthreshold region and its application, in *Proc. IEEE 2nd Int. Conf. Signal Proc. Systems - ICSPS 2010*, pp. V2-73–V2-77, 2010.

[12] C. Jongkunstidchai, C. Fongsmut, K. Kumwachara, and W. Surakampontorn, Full-wave rectifiers based on operational transconductance amplifiers, *Int. J. Electron. Commun.* vol. 61, pp. 195-201, 2007.

[13] S. Maheshwari, Current controlled precision rectifier circuits, *J. Circuits, Systems, and Computers*, vol. 16, no. 1, pp. 129-138, 2007.

[14] N. Minhaj, Transconductance element-based non-inverting and inverting precision full-wave rectifier circuits, in *Proc. IEEE Int. Conf. Advantages in Computing, Control, and Telecommunication Technologies*, pp. 442-445, 2009.

[15] N. Minhaj, OTA-based non-inverting and inverting precision full-wave rectifier circuits without diodes, *Int. J. Recent Trends in Engineering*, vol. 1, no. 3, pp. 72-75, 2009.

[16] N. Minhaj, Electronically controlled precision full-wave rectifier circuits, in *Proc. IEEE Int. Conf. Advantages in Recent Technologies in Communication and Computing*, pp. 240-243, 2009.

[17] G. Ferri and N. C. Guerrini, *Low-voltage low-power CMOS current conveyors*, Cluwer Academic Publishers, 2003.

[18] S. Takagi, Analog circuit designs in the last decade and their trends toward the 21st century, *IEICE Trans. Fundamentals*, vol. E84-A, no. 1, pp. 68-79, 2001.

[19] S. Minaei, O. K.Sayin, and H. Kuntman, A new CMOS electronically tunable current conveyor and its application to current-mode filters, *IEEE Trans. Circuits Systems I*, vol. 53, no. 7, pp. 1448-1457, 2006.

[20] N. Herencsar, J. Koton, K. Vrba, and O. Cicekoglu, Single UCC-N1B 0520 device as a modified CFOA and its application to voltage- and current-mode universal filters, in *Proc. Applied Electronics - APPEL 2009*, Pilsen, Czech Republic, pp. 127-130, 2009.

[21] N. Herencsar, J. Koton, and K. Vrba, Single CCTA-based voltage- and current-mode universal biquadratic filters employing minimum components, *Int. J. Comp. Elect. Engineering*, vol. 1, no. 3, pp. 316-319, 2009.

[22] K. C. Smith and A. Sedra, The current conveyor: a new circuit building block, *IEEE Proc.*, vol. 56, pp. 1368-1369, 1968.

[23] A. Sedra and K. C. Smith, A second-generation current conveyor and its application, *IEEE Trans. Circ. Th.*, vol. 17, pp. 132-134, 1970.

[24] A. Fabre, Third-generation current conveyor: a new helpful active element, *Electronics Letters*, vol. 31, no. 5, pp. 338-339, 1995.

[25] A. Fabre, O. Saaid, F. Wiest, and C. Boucheron, High frequency applications based on a new current controlled conveyor, *IEEE Trans. Circuits Syst.-I*, vol. 43, no. 2, pp. 82-90, 1996.

[26] H. O. Elwan and A. M. Soliman, Novel CMOS differential voltage current conveyor and its applications, *IEE Proc. Circuits, Devices, Systems*, vol. 144, no. 3, pp. 195-200, 1997.

[27] W. Surakampontorn and K. Kumwachara, CMOS-based electronically tunable current conveyor, *Electronics Letters*, vol. 28, no. 14, pp. 1316-1317, 1992.

[28] D. Biolek, V. Biolkova, and Z. Kolka, AC analysis of operational rectifiers via conventional circuit simulators, *WSEAS Transactions on Circuits and Systems*, vol. 3, no. 10, pp. 2291-2295, 2004.

[29] R. Sponar and K. Vrba, Measurements and behavioral modeling of modern conveyors, *Int. J. Comp. Science Net. Secur. IJCSNS*, vol. 6, mo. 3A, pp. 57-65, 2006.

[30] Datasheet AD8656, Analog Devices, Rev. A, 06/2005.

 ISBN: 978-1-61208-183-0

Single GCFDITA and Grounded Passive Elements Based General Topology for Analog Signal Processing Applications

Norbert Herencsar, Jaroslav Koton, Kamil Vrba
Department of Telecommunications
Brno University of Technology
Purkynova 118, 612 00 Brno, Czech Republic
{*herencsn, koton, vrbak*}*@feec.vutbr.cz*

Abhirup Lahiri
36-B, J and K Pocket
Delhi-110095
India
lahiriabhirup@yahoo.com

***Abstract*—In this paper, a novel topology using generalized current follower differential input transconductance amplifier (GCFDITA), which is suitable to realize several current-mode (CM) analog functions is presented. The topology is created using a single GCFDITA, a maximum of two grounded passive elements and can realize CM amplifier, integrator, first-order low-pass, high-pass, and all-pass filtering functions. The workability of one of the GCFDITA variants, namely the current inverter differential input transconductance amplifier (CIDITA) is proved by SPICE simulation results, based on the commercially available ICs OPA860 by Texas Instruments, and is in good accordance with theoretical predictions.**

Keywords-current-mode circuit; GCFDITA; CIDITA; general topology; all-pass filter.

I. INTRODUCTION

Analog frequency filters are widely used as anti-aliasing video filters in the analog sections of high-speed data communication systems defined by ITU BT 601 standard, for signal processing in wireless LANs described by IEEE 802.11 standard, measurement systems, etc. [1]–[3]. One of the most often used versatile modern current-mode (CM) active building blocks (ABBs) is the current differencing transconductance amplifier (CDTA) [4]. In the current technical literature, numerous publications providing several simple circuit solutions using CDTA can be found. These include CM biquadratic filters [4], [5], first-order all-pass filters [6], [7], higher-order filters [1], [8], full-wave CM precision rectifiers [9], or sinusoidal oscillators [10]. However, many of the aforementioned circuits only partially utilize the input p or n terminals of the CDTA. In such cases, the input current differencing unit (CDU) is reduced to a current follower (CF) or current inverter (CI). This fact has been the first time noticed in [11] and later in [12], and which led to the evolution of a generalized current follower transconductance amplifier (GCFTA) [13]. Depending on the value of the conveyance coefficients, six types of CFTA variants can be defined. In general, these are the CFTA and the inverted CFTA (ICFTA) [13]–[19]. In one of more recent article [20], Biolek and Biolkova introduced a modified version of GCFTA with buffered voltage outputs and transconductance of conventional CDTA or GCFTA changed to differential input transconductance amplifier. Limiting this work to the circuits with only current inputs and current outputs, in this paper, we provide a new ABB called generalized current follower differential input transconductance amplifier (GCFDITA), which as compared to [20] does not have a buffered voltage output terminal. This modification leads to simpler internal structure, and subsequently, less power consumption of its applications. An important advantage of GCFDITA is also that it can be easily created using commercially available ICs OPA860 [21].

II. PROPOSED CIRCUIT

A generalized current follower differential input transconductance amplifier (GCFDITA) has positive or negative current follower input that transfers the input current at terminal f to the z terminal and a balanced-output differential input transconductance amplifier (BO-DITA) stage, which is used to convert the difference voltage between the z and v terminals to balanced output currents. The transconductance parameter g_m corresponds for the positive output and $-g_m$ for the negative output. The circuit symbol of GCFDITA is shown in Figure 1. In general, the equations characterizing an ideal GCFDITA are:

$$V_f = 0,\ I_z = aI_f,\ I_v = 0,\ I_{x+} = -I_{x-} = g_m(V_z - V_v), \tag{1}$$

where $a \in \{1, -1\}$. Depending on the values of a, two variants of GCFDITA are possible, namely current follower differential input transconductance amplifier (CFDITA) for

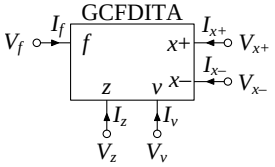


Figure 1. Circuit symbol of GCFDITA

 ISBN: 978-1-61208-183-0

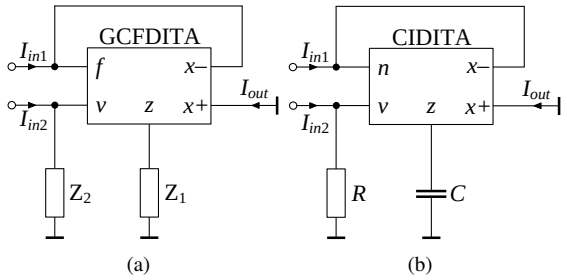


Figure 2. (a) Proposed circuit topology for realizing CM analog functions, (b) derived all-pass filter

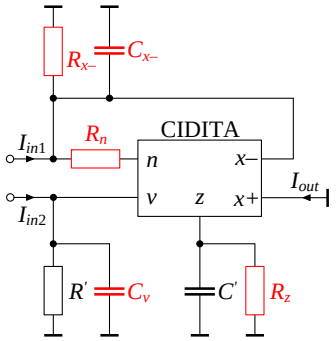


Figure 3. The all-pass filter in Figure 2(b) including dominant parasitics

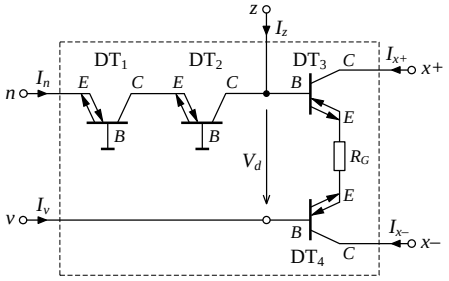


Figure 4. CIDITA implementation by ICs OPA860

$a = 1$ (let us label the input terminal as p) and current inverter differential input transconductance amplifier (CIDITA) for $a = -1$ (its input terminal is labeled as n).

The proposed general circuit topology used to realized several CM analog functions is shown in Figure 2(a). Recently, a similar general topology for realizing multiple voltage-mode (VM) analog functions using differential voltage current conveyor (DVCC) has been proposed in [22]. However, it is worth mention that the circuit in [22] does not provide a low-pass and high-pass filtering functions that can also be realized by here proposed general topology.

Considering the used ABB to be a CIDITA (i.e. $a = -1$) and doing routine circuit analysis using (1), the output current can be expressed as follows:

$$I_{out} = \frac{g_m(I_1 Z_1 - I_2 Z_2)}{g_m Z_1 + 1}. \qquad (2)$$

By appropriately choosing different impedances (Z_1 and Z_2 - as combinations of resistor and capacitor) and current inputs (I_1 and I_2), various CM analog functions can be derived. For example, with $I_1 = I_2 = I_{in}$, $Z_1 = 1/sC$ and $Z_2 = R$, the transfer function (TF) becomes:

$$T(s) = \frac{I_{out}}{I_{in}} = -\frac{g_m(sCR - 1)}{sC + g_m}. \qquad (3)$$

The above TF represents an all-pass (AP) filter under the condition $g_m R = 1$. If it is satisfied, the phase response of the filter is given as:

$$\varphi(\omega) = -2\tan^{-1}(\omega CR) = -2\tan^{-1}\left(\frac{\omega C}{g_m}\right). \qquad (4)$$

It is worth noting that here proposed AP circuit uses all grounded passive elements, a feature which is absent in previously reported CDTA based APF in [6].

III. Non-idealities of the CIDITA

For a complete analysis of the AP filter, it is important to take into account the non-idealities of CIDITA shown in Figure 3:

- $I_z = -\alpha I_n, I_{x+} = \beta_1 g_m V_d, I_{x-} = -\beta_2 g_m V_d$, where $V_d = V_z - V_v$, $\alpha = 1-\varepsilon_1$, $\beta_1 = 1-\varepsilon_2$ and $\beta_2 = 1-\varepsilon_3$. The parameters ε_1, ε_2 and ε_3 $(|\varepsilon_1|, |\varepsilon_2|, |\varepsilon_3| \ll 1)$ denote the current tracking error of the current inverting stage and of BO-DITA, respectively.
- The non-zero parasitic input resistance at terminal n of the CIDITA is represented by R_n, which typical value in case of implementation by ICs OPA860 in Figure 4 is 10.5 Ω.
- The parasitic resistance R_z and parasitic capacitance C_z appear between the high output impedance z terminal of the CIDITA and ground and their typical values are 455 kΩ||2.1 pF. The parasitic capacitance C_z is absorbed into external capacitor C as it appears in shunt with it and in Figure 3 labeled as C'.
- The parasitic resistance R_v and parasitic capacitance C_v appear between the high input impedance v terminal of the CIDITA and ground and their typical values are equal to z terminal parasitics. The parasitic resistance R_v is absorbed into external resistor R as it appears in shunt with it and labeled as R'.
- The parasitic impedances appearing between the high-impedance x terminals of the CIDITA and ground. For the circuit in Figure 3, these impedances are modeled at terminal $x-$ by R_{x-} and C_{x-} that represent the parasitic resistance and parasitic capacitance, respectively, and their typical values are 54 kΩ||2 pF.

Considering the aforementioned tracking errors and parasitic capacitances, the ideal TF (3) turns to:

$$T(s) = \frac{I_{out}}{I_{in}} = -\frac{\beta_1 g_m[s(C' - \alpha C_v) - \frac{\alpha}{R}]}{(sC' + \alpha\beta_2 g_m)(sC_v + \frac{1}{R})}, \qquad (5)$$

 ISBN: 978-1-61208-183-0

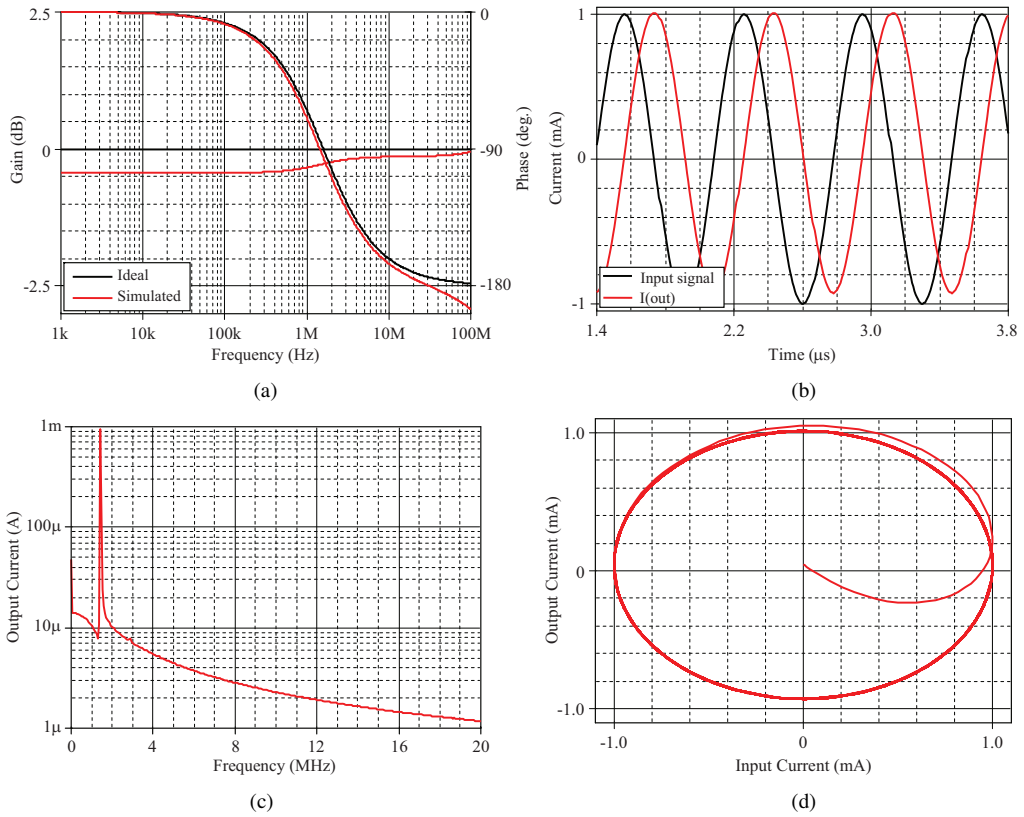


Figure 5. (a) Simulated gain and phase characteristics of the proposed CM all-pass filter, (b) time-domain responses at 1.44 MHz, (c) Fourier spectrum of the output signal, (d) Lissajous pattern showing -90° phase shift at pole frequency

where $C' = C + C_z$. Assuming that the external capacitor C is much greater than C_v and C_z, the effects of these parasitics can be omitted and (5) changes to:

$$T(s) = \frac{I_{out}}{I_{in}} = -\frac{\beta_1 g_m (sCR - \alpha)}{sC + \alpha\beta_2 g_m}. \quad (6)$$

Now, the zero ω_z and pole ω_p frequencies are not equal and can be expressed as:

$$\omega_z = \frac{\alpha}{CR}, \qquad \omega_p = \frac{\alpha\beta_2 g_m}{C}. \quad (7)$$

It should be noted that, the angular pole frequency ω_p differs from the corresponding angular zero frequency ω_z and thus this mismatch affects both the magnitude and phase response of the circuit. Hence, to receive $\omega_z = \omega_p$ precise matching condition $g_m R = 1$ and careful design of CIDITA with values of α, β_1 and β_2 close to unity are needed.

IV. Simulation Results

In order to verify the workability of the designed AP filter shown in Figure 2(b), it has been simulated using SPICE software. Figure 4 shows the implementation of the CIDITA using commercially available ICs, e.g., the OPA860 by Texas Instruments [21]. The DC power supply voltages of OPA860 SPICE macromodels were equal to ± 5 V. The OPA860 contains the so-called 'diamond' transistor (DT) and fast voltage buffer (VB). In the output stage, in order to increase the linearity of collector currents versus input voltage V_d, the DT3 and DT4 are complemented with degeneration resistor $R_G \gg 1/g_{mT}$, added in series to the emitters, where the g_{mT} is the DT transconductance. Then the total transconductance decreases to the approximate value $1/R_G$. In simulations the passive element values were selected as follows: $C = 100$ pF and $R = R_G = 1$ kΩ, and hence, the above mentioned matching condition $g_m R = 1$ ($R = R_G$) is fulfilled. In this case, a 90° phase shift is at pole frequency $f_p \cong 1.59$ MHz. Figure 5(a) shows the ideal and simulated gain and phase responses of the proposed filter, from which the obtained f_p is 1.44 MHz. Time-domain simulation result of the proposed filter is shown in Figure 5(b) in which a sinusoidal input current signal with 1 mA peak value at 1.44 MHz is applied to the filter. The total harmonic distortion at this frequency is found as 0.102%. The Fourier spectrum of the output signal, showing a high selectivity for the applied signal frequency, is shown in Figure 5(c). The Lissajous pattern for the circuit as -90° phase shifter is shown in Figure 5(d). The total power dissipation of the circuit is found to be 204 μW. From the simulation results it can be seen that the final solution is in good agreement with the theory.

 ISBN: 978-1-61208-183-0

V. Conclusion

The paper presented a novel general topology suitable to realize several current-mode analog functions. It is created using single versatile ABB, namely the generalized current follower differential input transconductance amplifier (GCFDITA) and maximum of two grounded passive components. As an example, first-order all-pass filter has been derived and SPICE simulation results of the proposed filter have been provided. It is expected that GCFDITA would prove to be a versatile ABB for the general design of active filters and sinusoidal oscillators. It is worth mention that a similar general circuit is proposed for voltage-mode functions in [22], however, in [22] the low-pass and high-pass filtering functions are not provided.

Acknowledgement

The research described in the paper was supported by the following projects: MSM0021630513, GACR P102/11/P489, GACR P102/09/1681, and FEKT–S–11–15.

References

[1] A. Uygur and H. Kuntman, "Seventh-order elliptic video filter with 0.1 dB pass band ripple employing CMOS CDTAs," *Int. J. Electron. Commun. (AEU)*, vol. 61, pp. 320–328, 2007.

[2] T. Y. Lo, C. S. Kao, and C. C. Hung, "A Gm-C continuous-time analog filter for IEEE 802.11 a/b/g/n wireless LANs," *Analog Integrated Circuits and Signal Processing*, vol. 58, pp. 197–204, 2009.

[3] G. Ferri and N. C. Guerrini, *Low-voltage low-power CMOS current conveyors*, London: Kluwer Acad. Publ., 2003.

[4] D. Biolek, "CDTA - building block for current-mode analog signal processing," In *Proc. of the 16th European Conf. on Circuit Theory and Design - ECCTD'03*, Krakow, Poland, pp. 397–400, 2003.

[5] W. Tangsrirat, T. Dumawipata, and W. Surakampontorn, "Multiple input single-output current-mode multifunction filter using current differencing transconductance amplifiers," *Int. J. Electron. Commun. (AEU)*, vol. 61, no. 4, pp. 209–214, 2007.

[6] A. U. Keskin and D. Biolek, "Current mode quadrature oscillator using current differencing transconductance amplifiers (CDTA)," *IEE Proc.-Circuits Devices Syst.*, vol. 153, no. 3, pp. 214–218, 2006.

[7] A. Lahiri and A. Chowdhury, "A novel first-order current-mode all-pass filter using CDTA," *Radioengineering*, vol. 18, pp. 300–305, 2009.

[8] T. Dostal, "All-pass filters in current mode," *Radioengineering*, vol. 14, no. 3, pp. 48–53, 2005.

[9] J. Koton, N. Herencsar, S. Minaei, and K. Vrba, "Precision full-wave current-mode rectifier using current differencing transconductance amplifier," In *Proc. of the Int. Conf. on Computer and Communication Device - ICCCD'11*, Bali Island, Indonesia, vol. 1, pp. 71–74, 2011.

[10] A. Lahiri, "Novel voltage/current-mode quadrature oscillator using current differencing transconductance amplifier," *Analog Integr. Circ. Signal Proc.*, 2009, vol. 61, pp. 199–203.

[11] N. Herencsar, J. Koton, I. Lattenberg, and K. Vrba, "Signal-flow graphs for current-mode universal filter design using current follower transconductance amplifiers (CFTAs)," In *Proc. of the Int. Conf. on Applied Electronics - APPEL'08*, Pilsen, Czech Republic, pp. 113–116, 2008.

[12] D. Biolek, R. Senani, V. Biolkova, and Z. Kolka, "Active elements for analog signal processing: classification, review, and new proposals," *Radioengineeing*, vol. 17, no. 4, pp. 15–32, 2008.

[13] N. Herencsar, J. Koton, K. Vrba, I. Lattenberg, and J. Misurec, "Generalized design method for voltage-controlled current-mode multifunction filters," In *Proc. of the 16th Telecommunications forum - TELFOR'08*, Belgrade, Serbia, pp. 400–403, 2008.

[14] N. Herencsar, K. Vrba, J. Koton, and A. Lahiri, "Realisations of single-resistance-controlled quadrature oscillators using generalised current follower transconductance amplifier and unity-gain voltage-follower," *International Journal of Electronics*, vol. 97, no. 8, pp. 897–906, 2010.

[15] N. Herencsar, J. Koton, K. Vrba, and I. Lattenberg, "Novel SIMO type current-mode universal filter using CFTAs and CMIs," In *Proc. of the 31st Int. Conf. on Telecommunications and Signal Processing - TSP'08*, Paradfurdo, Hungary, pp. 107–110, 2008.

[16] N. Herencsar, J. Koton, and K. Vrba, "Realization of current-mode KHN-equivalent biquad using current follower transconductance amplifiers (CFTAs)," *IEICE Trans. Fundamentals*, vol. E93-A, no. 10, pp. 1816–1819, 2010.

[17] W. Tangsrirat, "Novel current-mode and voltage-mode universal biquad filters using single CFTA," *Indian Journal of Engineering & Materials Sciences*, vol. 17, pp. 99–104, 2010.

[18] N. Herencsar, J. Koton, K. Vrba, and O. Cicekoglu, "New active-C grounded positive inductance simulator based on CFTAs," *Proc. of the 33th Int. Conf. on Telecommunications and Signal Processing - TSP'10*, Baden near Vienna, Austria, pp. 35–37, 2010.

[19] R. Sotner, J. Jerabek, N. Herencsar, T. Dostal, and K. Vrba, "Additional approach to the conception of current follower and amplifier with controllable features," In *Proc. of the 2011 34th Int. Conf. on Telecommunications and Signal Processing - TSP'11*, Budapest, Hungary, pp. 279–283, 2011.

[20] D. Biolek and V. Biolkova, "Modified buffered transconductance amplifier for analog signal processing," In *Proc. of the Int. Conf. Radioelektronika'09*, Bratislava, Slovak Republic, pp. 191–194, 2009.

[21] Datasheet OPA860 - Wide Bandwidth Operational Transconductance Amplifier (OTA) and Buffer. Texas Instruments, SBOS331C–June 2005–Rev. August 2008.

[22] S. Maheshwari, "Analogue signal processing applications using a new circuit topology," *IET Circuits Devices Syst.*, vol. 3, no. 3, pp. 106–115, 2009.

 ISBN: 978-1-61208-183-0

How Many Cores Does Parallel BGP Need in a High-Speed Router

Yaping Liu, Shuo Zhang, Zexin Lu and Baosheng Wang
School of Computer Science, National University of Defense Technology
Changsha, Hunan, P.R.China
e-mail: {ypliu, zhangshuo, lzx, wbs }@nudt.edu.cn

***Abstract*—The performance problem of BGP has raised great concerns both in industry and research. With rapid expansion of Internet, how to improve the performance of BGP to support more BGP neighbors in a high-speed router is a practical urgent problem. In this paper, we presented a Minimal Cores Computing (MCC) algorithm based on multi-root tree model to compute the minimal cores for parallel BGP in the context of the multi-cores platform. The algorithm is an approximation algorithm as the problem is a nonlinear programming problem. Simulation results show that MCC can get reasonable good speedup with minimal number of cores. MCC can give direction to the design of the control node in a core router.**

Keywords-parallel BGP; speedup; multi-cores; performance; router.

I. INTRODUCTION

With rapid expansion of Internet, BGP (Border Gateway Protocol) [1] confronted serious performance problem. Feldmann [2] pointed out that with the increase of BGP neighbors, the router can not deal with the update packets immediately, thus the routing update time will increase. If the neighbors are beyond 250, the router will hardly maintain these neighbors even if it does not process any other packets. According to Agarwal [3]'s analysis, rapid route changes have made BGP process consume over 60% of CPU cycles.

To overcome the low-efficiency of BGP, some mechanisms are used to improve its scalability, such as confederation and route reflector. However, these mechanisms require more complex network configuration, which may lead to configure mistakes. The performance problem also exists.

At present, the main solution is to use BGP distributed computing, which focuses on the corporation among BGP instances on different control nodes in one router. However, with the development of multi-cores CPU, we can create parallel BGP on multi-cores CPU, which focuses on the model and implementation in one node. Comparing with dsitributed BGP, parallel BGP has much lower communication cost. And one multi-cores control node is cheaper than multiple control nodes. Our research belongs to parallel BGP on a multi-cores CPU.

We present a multi-root tree model (MR-PBGP) for parallel BGP, which is an integrated model with neighbor-based division and data division. According to the model, the problem of compute the minimal cores for a BGP router with appointed number of neighbors is a nonlinear programming problem. Thus, we presented an approximation algorithm named MCC (Minimal Cores Computing algorithm) and simulated it using Matlab. Simulation results show that the algorithm can get a sound result giving appointed number of neighbors, the probability of comparing all routing information to select an optimal route, the probability of routing updates with the same prefix from different neighbors, the probability of EBGP, and the performance ratio to the ideal situation or an appointed parallel speedup.

To the best of our knowledge, there is no previous work on determining the minimal cores for parallel BGP similar to ours.

The rest of the paper is organized as follows. Section II reviews the related work. Section III describes our problem and presents MCC algorithm. Section IV simulates MCC in the Matlab environment. Section V draws conclusions for this research.

II. RELATED WORK

The research about BGP protocol parallelism mainly centers on BGP distributed computing. It catches high-end router manufacturers' attentions especially [4-6].

Markus Hidell proposed a distributed BGP protocol model based on data division [7]. Its main idea was to divide all of the network prefixes into several disjoint subsets, and assign different network prefixes to different BGP protocol entities to do the parallel processing. However, its main drawback is that with the increase of neighbors, session management (SM) will become a bottleneck. This model is suitable for the network environment that has relatively small number of neighbors.

Kun Wu et al. [8,9] proposed a distributed BGP route processing model based on a tree structure. Their study was based on the following two assumptions. First, BGP can select the optimal route without the requirement of the whole route information for the same network prefixes. Secondly, they mainly centers on how to improve the performance of the process of optimal route selection of BGP.

Xiaozhe Zhang [10] proposed an agent-based distributed parallel implementation BGP model. The model introduced the team work idea of the agent technology, extended BGP protocol to be a BGP entity independently running on each control node.

Although the above mentioned methods are the parallel processing techniques of protocols, they are not suitable for multi-cores platform.

 ISBN: 978-1-61208-183-0

III. MCC ALGORITHM

To simply the problem, no routing policies are configured. Suppose that we use t_1 to represent the average time in creating a new entry, t_2 to represent the average time in selecting the optimal route for one prefix, t_3 to represent the average time of unpacking the optimal route for one prefix, t_4 to represent the average time of advertising the optimal route for one prefix, if its neighbor is an ebgp neighbor. Moreover, n is the number of neighbors, x is the set of routing prefixes, p_1 is the probability of comparing all routing information to select an optimal route, p_2 is the probability of routing updates with the same prefix from different neighbors, p_3 is the probability of EBGP, and finally, λ is the average arrival rate of routing updates.

A typical serial model of BGP can be considered as a queue system of M/M/1. Its average service time S_o satisfies (1).

$$s_o = t_1 + t_2 + t_3 + p_3 n t_4 \tag{1}$$

Theoretically speaking, t_3 and t_4 are constants. The value of t_1 is related to the size of x. Since the cost of allocating memory is far greater than the cost of searching and inserting in a structure of tree, the value of t_1 mainly depends on the cost of allocating memory, hence it can also be treated as a constant. The value of t_2 is proportional to the number of neighbors, from which a router receives routing updates for one prefix.

Let $k_i = \frac{t_i}{t_1}(i \geq 2)$, then the average service rate satisfies (2). The value of k_2 is a liner function of n. Moreover, k_{2a}, k_{2b}, k_3, k_4, and C are constants.

$$\mu_o = \frac{1}{s_0} = \frac{1}{1 + k_2 + k_3 + p_3 n k_4} \times \frac{1}{t_1}$$
$$k_2 = k_{2a} n + k_{2b} \tag{2}$$
$$C = \frac{1}{\lambda t_1}$$

If $\lambda < \mu_0$, the average stay time of one bgp message in the router is t_s, which, by the queue theory, equals (3).

$$t_s = \frac{1}{\mu_o - \lambda} \tag{3}$$

To decrease the value of t_s, we propose a multi-root tree model and call it MR-PBGP as illustrated in Fig.1. This model can not only reduce the arrival rate but also cut down neighbors for each thread, so that k_2, $p_3 n$, and λ are decreased.

In Fig. 1, *w* represents the number of roots, *h* represents the height of the tree, and x_i represents sons of the ith-level node (not leaf node). Leaf nodes represent bgp neighbors. The first-level threads are nodes that are fathers of leaf nodes. Every first-level thread creates several BGP sessions and handles BGP routing information from its neighbors. Each first-level thread may have multiple fathers depending on *w*. Different father node deals with different scope of routing prefixes showing ideas of data division. If the selection of optimal route for one prefix requires all routing information from every neighbor, the routing information of that prefix will be sent directly to the corresponding master-thread by the first-level thread. Thus, in Fig.1, there are direct lines from the first-level nodes to root nodes.

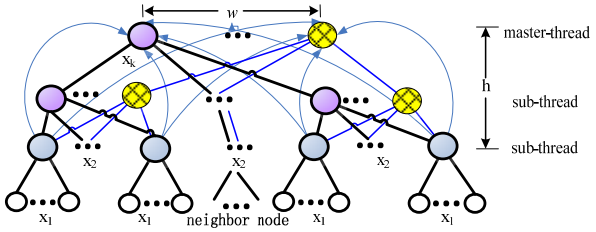


Figure 1. MR-PBGP

The problem of compute the minimal cores according to MR-PBGP can be described as following (4).

$$\lambda_0 = \lambda \times \frac{1}{n}$$
$$\lambda_1 = \lambda \times \frac{x_1}{n} = \lambda_0 \times 1 \times x_1$$
$$s_1 = (1 + k_3 + (1 - p_1 p_2) k_2(x_1) + p_3 x_1 k_4) t_1$$
$$\mu_1 = \frac{1}{s_1}$$
$$t_{s1} = \frac{1}{\mu_1 - \lambda_1}$$
$$\lambda_2 = \frac{1}{w} \times \frac{\lambda_1}{x_1} \times \left(\frac{p_2(1-p_1)(x_1+1)}{2} + (1-p_2) \times x_1\right) \times x_2$$
$$s_2 = (1 + k_2(x_2)) \times t_1$$
$$\mu_2 = \frac{1}{s_2}$$
$$t_{s2} = \frac{1}{\mu_2 - \lambda_2}$$
……
$$\lambda_k = \frac{\lambda_{k-1}}{x_{k-1}} \times \left(\frac{p_2 - p_1 p_2}{1 - p_1 p_2} \frac{x_{k-1}+1}{2} + \frac{1-p_2}{1-p_1 p_2} \times x_{k-1}\right) \times x_k$$
$$s_k = (1 + k_2(x_k)) \times t_1$$
$$\mu_k = \frac{1}{s_k}$$
$$t_{sk} = \frac{1}{\mu_k - \lambda_k} \qquad (k<h)$$
$$\lambda_h = \frac{\lambda_{k-1}}{x_{k-1}} \times \left(\frac{p_2 - p_1 p_2}{1 - p_1 p_2} \frac{x_{k-1}+1}{2} + \frac{1-p_2}{1-p_1 p_2} \times x_{k-1}\right) \times x_h + \frac{1}{w} \lambda p_1 p_2$$
$$s_h = \left[\ 1 + (1 - p_1 p_2) \times k_2(x_h) + p_1 p_2 \times k_2\left(\frac{n}{2}\right)\ \right] \times t_1$$
$$\mu_h = \frac{1}{s_h}$$
$$t_{sh} = \frac{1}{\mu_h - \lambda_h}$$
$$\lambda_{ideal} = \lambda \times \frac{V}{n}$$

 ISBN: 978-1-61208-183-0

$$s_{ideal} = (1 + k_3 + (1 - p_1 p_2)k_2(V) + p_3 V k_4)t_1$$

$$\mu_{ideal} = \frac{1}{s_{ideal}}$$

$$t_{ideal} = \frac{1}{\mu_{ideal} - \lambda_{ideal}}$$

$$\min_{x_1, x_2, \ldots, x_h} m = w \times (1 + \sum_{k=3}^{h} \prod_{i=k}^{h} x_i) + \prod_{i=2}^{h} x_i \quad (4)$$

Subject to:

$$\max(t_{s1}, t_{s2}, \ldots, t_{sh}) = At_{ideal}, A > 1 \; is \; a \; cons\tan t$$

$$x_1 \times x_2 \times \ldots \times x_h = n$$

$$x_1, x_2 \ldots, x_k \geq 2, w \geq 1, h \geq 2 \;\; are \; \text{int} \, ergers$$

Algorithm: **MCC**

Inputs: p_1,p_2,p_3,C,k_{2a},k_{2b},k_3,k_4,n,V,A; ***Outputs*****:** m, x_i , j, and w

1) Solve the following equation and set the value of x_1

$t_{s1} = At_{ideal}, x_1 = \lceil result \rceil$

2) Compute the possible max height of the tree H

$H = \log_2^{n/x_1} + 1$

3) j=2; /*search the optimal values*/

while (j < H) **do**
  if (j= =2) **then**
    $x_2 = \lceil n / x_1 \rceil$;calculate w according to $t_{s1} = t_{sh}$;
    if (w<1) **then**
      break;
    end if
    else then
      Calculate m according to (4); j++;
      continue;
    end else
  end if
  else then
    optimizer(w,j, x_i);
  end else
end while

optimizer(w,j, x_i)

w--;
if (w < 1) **then**
  return;
end if
else then
  compute the optimal integer values of x_i(i=2,…)
  according to $t_{s1} = t_{s2} = \ldots = t_{sh}$;
  if (no feasible solutions) **then**
    j++; **return**;
  end if
  else then
    calculate m according to (4) and update w,j, x_i;
    optimizer(w,j, x_i);
  end else
end else

Figure 2. Pseudo Codes of MCC

In formula (4), V is the value of the number of the first-level thread's neighbors calculated by the ideal MR-PBGP model with no constraint on cores. It can be proved that the ideal MR-PBGP model is a multi-root binary tree with V being greater than 2 [11]. The value of V can be calculated by the following (5), in which v is the minimal value of x_1 according to $s_h \leq t_{s1}$.

$$V = \max(\lceil v \rceil, 2) \quad (5)$$

And usually, the value of V is equal to 2. The above problem is a nonlinear programming. Thus, we presents an approximation algorithm named MCC as illustrated in Fig. 2.

The complexity of MCC depends on the cost of solving equations and searching times. We use Newton iterative method to solve polynomial equations. Its complexity is O(zM), in which M is max iterative number and z is highest-degree of a polynomial. In MCC, the highest-degree of equations is 5, so that the complexity of solving equations is O(M). The max searching times satisfies following (6), in which w is the solution of equations with j=2 in MCC.

$$\sum_{j=2}^{H} w(j-1) = \frac{H(H-1)w}{2} < w(\log_2^n)^2 \quad (6)$$

As w is considered to be a constant in this case, the complexity of MCC is $O((\log_2 n)^2 M)$.

IV. SIMULATION

We simulated MCC using MatLab. We use Quagga BGP [12] as the sample of typical BGP running on Lenovo with 4-cores Intel Xeon E5405 CPU, 4G Memory, and two 1G Ethernet interface. Tested by Spirent AX4000 [13], C, k_2, k_3 and k_4 were obtained as follows:

$C = 2.93, k_3 = 0.32, k_4 = 0.15$

$k_2 = k_{2a} \times x + k_{2b} (k_{2a} = 0.03243, k_{2b} = 0.58122)$

(x is neighbors)

At first, we chose p_1 as 0.1, p_2 as 0.7, p_3 as 0.2, which is nearer to current real network [14]. The neighbors change from 64 to 2048. As the serial typical BGP do not accord with conditions of queue theory in those cases, parallel speedup should be infinite. Thus, we chose performance ratio with that of the ideal parallel model changing from 0.1 to 0.7.

Fig. 3 shows the simulation results of cores computed by MCC. It shows that the number of cores is increased with the increase of neighbors and performance ratio. The cores change from 3 to 10 with neighbors changing from 64 to 2048 in the case of performance ratio equaling 0.1. And the cores change from 9 to 42 in the case of performance ratio equaling 0.7. The increase speed of cores is higher with high performance ratio than that with low performance ratio. The phenomena also show that if a router wants to support large neighbors with performance near to the ideal optimal value, a large number of cores are needed. However, if a router wants to support large neighbors with acceptable performance, only several cores are needed. For example, if we chose performance ratio to 0.1, only 10 cores are needed to support 2048 BGP neighbors with its parallel speedup being infinite.

 ISBN: 978-1-61208-183-0

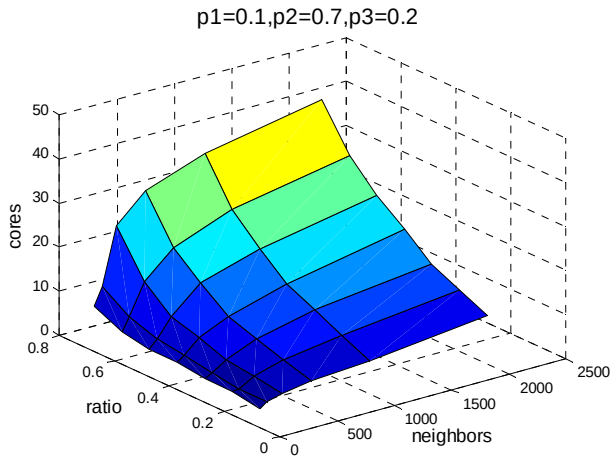


Figure 3. Cores computed by MCC

Fig. 4 shows the simulation results of *w* computed by MCC. The value of *w* represents the width of MR-PBGP tree. The results show that if the value of performance ratio is lower, *w* often equals to 1 which means one master thread in the model. With the increase of performance ratio, *w* also increases. The reason is that the cost time of first-level thread becomes lower so that the master thread may be bottleneck. Thus, *w* should increase to reach load balance between threads. In our simulation, the value of *w* does not be beyond 4 in most cases.

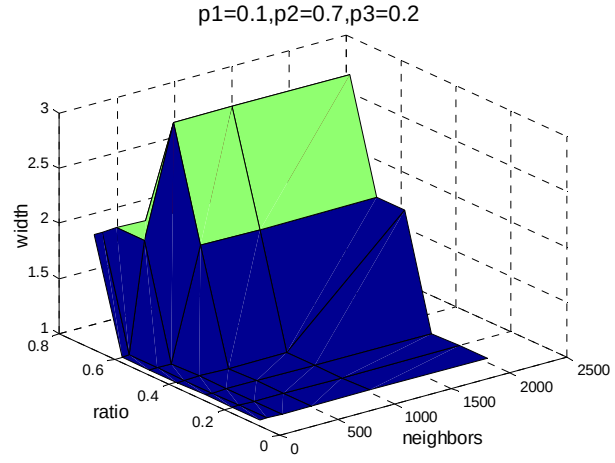


Figure 4. w computed by MCC

Fig. 5 shows the simulation results of *h* computed by MCC. The value of *h* represents the height of MR-PBGP tree. The results show that in most cases, the height of the tree in the model is 2. And with the increase of performance ratio and neighbors, *h* may increase. The reason is that the cost time of master thread can decrease more rapidly by increase of *w* than by increase of *h*. Thus, the value of *h* becomes high only with high performance ratio and large neighbors. The value of *h* usually does not be beyond 4.

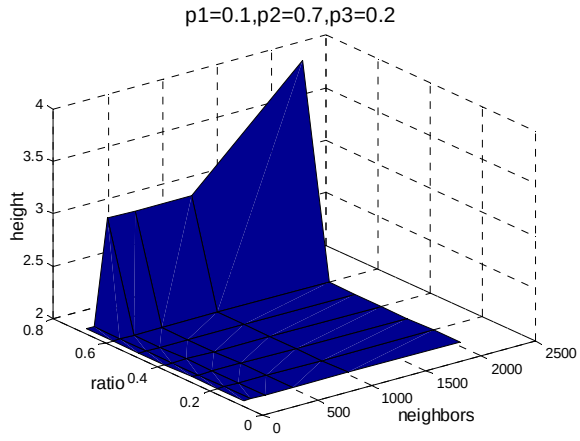


Figure 5. h computed by MCC

Then we research on the relationship among cores, p_1, p_2, and p_3 giving appointed neighbors and performance ratio. We chose neighbors as 2048, and performance ratio as 0.1.

Fig. 6 represents the simulation results about the change of cores, *w* and *h* with the change of p_1. The result shows that the cores, width and height change little with the increase of p_1 excluding p_1 equaling 1. The value of core number is around 10. The value of *h* is 2 or 3. And in most cases, the value of *w* is 1. The reason is that though the value of x_1 increases with the increase of p_1, the bottleneck of system mainly lies on the first-level thread in most cases. Thus, the core number changes little. If p_1 is equal to zero, the value of *V* (according to (5)) is equal to 2. But if p_1 is not zero, *V* increases quickly (For example, if p_1 is equal to 0.1, *V reaches 34*). The value of x_1 computed by MCC is much lower in the case of p_1 equal to zero. But for higher values of p_1, the bottleneck of the system is master thread *so that w* increases rapidly which leads to cores' increasing rapidly. Thus, the value of cores is firstly slightly decreased and increased for higher values of p_1.

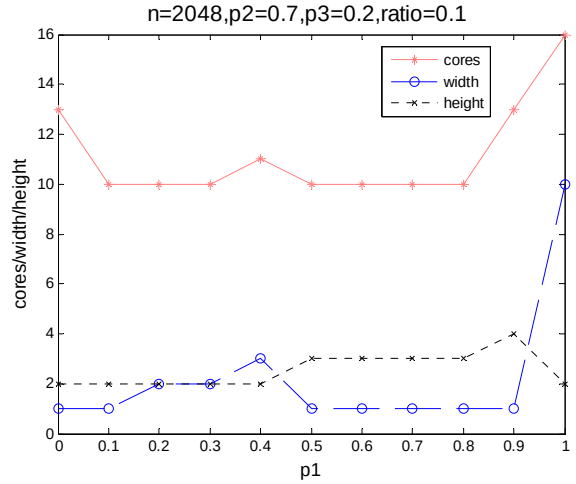


Figure 6. m/w/h with p1

Fig. 7 represents the simulation results about the change of cores, *w* and *h* with the change of p_2. The results show that the width (*w*=1) and height (*h*=2) keep constant with the increase of p_2. The value of core number decreases with the increase of p_2.The reason is that the cost of master thread decreases with the increase of p_2. We can decrease cores by increasing the value of x_1.

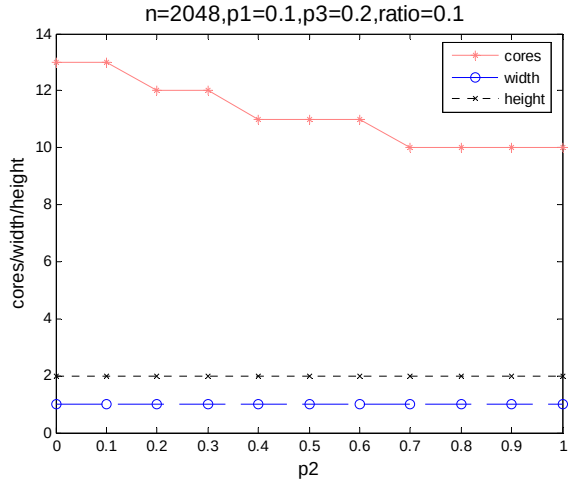


Figure 7. m/w/h with p_2

Fig. 8 represents the simulation results about the change of cores, *w* and *h* with the change of p_3. The results show that the width (*w*=1) and height (*h*=2) keep constant with the increase of p_3. The value of core number increases with the increase of p_3.The reason is that the cost of first-level thread increases with the increase of p_3. To reach load balance, the value of x_1 should be decreased so that cores increase.

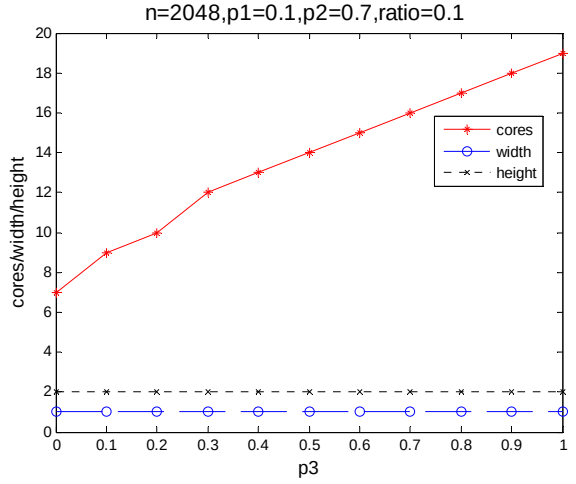


Figure 8. m/w/h with p_3

In a word, if we chose an acceptable low value of performance ratio, the core number computed by MCC is small with the condition of supporting large neighbors. The current common high-end multi-cores CPU can satisfy the condition. For example, 8-cores CPU can support 2048 BGP neighbors with acceptable parallel speedup. Moreover, the tree is two- level structure in this situation with *w* equaling to1 in most cases. However, if a router wants to be designed to support large neighbors with its parallel speedup near to its ideal value, a large number of cores are needed. In this situation, the current highest-performance CPU are needed with 64-cores or 128-cores. And the model will appear as a multi-root and multi-level tree.

If we want to implement parallel BGP in a high-end router, there are three steps to do it. At first, we need testing the values of C and k_i according to the supported interface rate, CPU, and original implementation of typical serial BGP. Secondly, we can determine the cores by algorithm MCC with requirement of supporting number of neighbors, and other parameters. At last, we can implement parallel BGP with multiple threads according to the model of MR-PBGP and the parameters's values computed by MCC. It is our current work to implement a real prototype.

V. CONCLUSION AND FUTURE WORKS

Our main contribution of this paper is presenting a minimal cores computing algorithm named MCC to compute the minimal cores for parallel BGP oriented the multi-cores platform. Simulation results showed that current common high-end multi-cores CPU can be used in a high-speed router with an acceptable good speedup supporting large BGP neighbors. The model and the algorithm are very usable to the design of high-speed routers. However, in real networks, BGP may get dynamic payload from different neighbors. We are implementing a real prototype for parallel BGP in a high-end router. We will test the performance of the prototype and research on how to reach the optimal status under a dynamic running environment.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China under Grant No. 61070200, the National Key Technology R&D Program under Grant No. 2008BAH37B03-03, and the National High-Tech Research and Development Plan of China under Grant No.2008AA01A325.

REFERENCES

[1] Y. Rekhter. A Border Gateway Protocol 4 (BGP-4), 2006. RFC 4271.

[2] Feldmann, H.W. Kong, O. Maennel, and A. Tudor. Measuring bgp pass-through times. Barakat C and Pratt I, editors, In Passive Active Measurement Workshop, vol. 3015, pp. 267-277. France: Springer-Verlag GmbH, 2004.

[3] S. Agarwal, C.N. Chuah, S. Bhattacharyya, and C. Diot. Impact of bgp dynamics on router cpu utilization. Barakat C and Pratt I, editors, In Passive Active Measurement Workshop, volume 3015, pp. 278-288. France: Springer-Verlag GmbH, 2004.

[4] D. A. Ball, R. E. Bennett, et al. Distributed software architecture for implementing BGP, Patent Application Publication, no. US 2005/0074003, April, 2005.

[5] Juniper Networks, Inc. T640 routing node and TX matrixTM platform: architecture. White Paper (Part Number 350031-001), http://www.juniper.net, Dec, 2004.

 ISBN: 978-1-61208-183-0

[6] Cisco Systems, Inc. Next generation networks and the cisco carrier routing system. White Paper, http://www.cisco.com, Dec, 2004.

[7] Markus Hidell, Peter Sjödin, Tomas Klockar, and Lenka Carr-Motyckova. A Modularized Control Plane for BGP, IEEE Proceeding PDCS07, 2007.

[8] K. Wu, J. Wu, and K. Xu. A tree-based distributed model for BGP route processing. In: HPCC 2006, LNCS 4208, Berlin: Springer-Verlag, pp. 119-128, 2006.

[9] K. Xu and H. He. BGP parallel computing model based on the iteration tree. Journal of China Universities of Posts and Telcommunications, 15(Suppl.), pp. 1-8, September, 2008.

[10] X. Zhang, P. Zhu, and X. Lu. Fully-distributed and highly-parallelized implement- tation model of bgp4 based on clustered routers. In: ICN 2005, LNCS 3421, Berlin: Springer-Verlag, pp. 433-441, 2005.

[11] Y. Liu, S. Zhang, G. Zhang. Parallel BGP: Analysis, Model and Experiment. Technical Report, School of Computer Science, National University of Defense Technology, Sep, 2011.

[12] Quagga project. Available from http://ww.quagga.net/download/, Feb, 2011.

[13] Spirent Adtech AX4000 broadband test system. Available from http://www. smartechconsulting.com/Adtech-Spirent-AX4000-AX-4000-16-Slot-Rack-mount-Chassis, Oct, 2010.

[14] CIDR Report [EB/ OL] [2008-2-10]. http :/ / www. cidr-report . org, Sep, 2009.

 ISBN: 978-1-61208-183-0

Multi-level Machine Learning Traffic Classification System

Géza Szabó*, János Szüle†, Zoltán Turányi*, Gergely Pongrácz*
*TrafficLab, Ericsson Research, Budapest, Hungary,
E-mail: [{geza.szabo, zoltan.turanyi, gergely.pongracz}@ericsson.com]
†Complex Networks Laboratory, Eötvös Lóránd University,
E-mail: [szule@complex.elte.hu]

***Abstract*—In this paper, we propose a novel framework for traffic classification that employs machine learning techniques and uses only packet header information. The framework consists of a number of key components. First, we use an efficient combination of clustering and classification algorithms to make the identification system robust in various network conditions. Second, we introduce traffic granularity levels and propagate information between the levels to increase accuracy and accelerate classification. Third, we use customized constraints based on connection patterns to efficiently utilize state-of-the-art clustering algorithms. The components of the framework are evaluated step-by-step to examine their contribution to the performance of the whole system.**

***Keywords*-traffic classification; machine learning; packet header**

I. INTRODUCTION

In-depth understanding of the Internet traffic is a challenging task for researchers and a necessary requirement for Internet Service Providers (ISP). Usually, Deep Packet Inspection (DPI) is used by ISPs to profile networked traffic. Using the results ISPs may apply different charging policies, traffic shaping, and offer differentiated QoS guarantees to selected users or applications (where legally possible). Deep Packet Inspection usually extracts information from both the packet headers and the payload. In some cases, this approach is not feasible due to, e.g., processing constraint or when the payload is encrypted.

Our goal is to classify traffic based solely on packet header information, such as packet size, arrival time, addresses, protocols and ports. The following requirements have to be fulfilled by our system:

- It should be robust: the characteristics of the network, such as speed or load should not impact accuracy
- It should be fast: classification results shall be provided after as few packets of a flow as possible
- It should be accurate: results should have high true positive (TP) with minimal false positive (FP) ratio

In current state-of-the-art, traffic classification engines, which rely only on packet header information, the effects of network environment changes influence the performance of the identification methods (e.g., [1], [2]). This results in reduced accuracy when the model trained in one network is used for testing in a different one. To become robust in such scenarios, our proposed method incorporates unsupervised learning for the basic clustering of the input flows and supervised clustering to automatically deduce the resulting classes. In this way we achieved a method that performs well under changing network conditions.

Another disadvantage of current state-of-the-art methods is that they can provide information about a flow only after its full processing (e.g., [3], [4]). They cannot conclude the processing of the data flow even if a certain confidence is reached in the middle of it. In the proposed framework data collection happens on several granularity levels and the results of one level are fed to a lower granularity level. Therefore result generation can be considered at several checkpoints during the flow to provide information the soonest possible.

Constraint clustering is a state-of-the-art [5],[6] technique to improve clustering. We propose to use this technique with constraints based on connectivity patterns to further increase classification accuracy.

The main contributions of the paper are as follows:

- The evaluation of various clustering and classification algorithms and an efficient combination of them
- The introduction of traffic granularity levels and a proposal to efficiently utilize them
- The efficient utilization of constraint-based clustering algorithms

This paper is organized as follows. Section II overviews the related work and introduces the terms used in the paper. In Section III, the data used for evaluation purposes is described. Section IV compares clustering and classification algorithms and proposes a combination of them. In Section V, the granularity levels of the traffic and its effective use are discussed. In Section VI, some preprocessing steps are introduced to exploit the advantages of constraint-based clustering algorithms. Finally, the paper is concluded in Section VII.

II. RELATED WORK AND TAXONOMY

In the following bullets, we define the terms used in current state-of-the-art papers about machine learning (ML).

- *Feature*: An attribute of the studied objects (e.g., the average bitrate of a flow), the input to machine learning

 ISBN: 978-1-61208-183-0

	Features (measured)					Test result		
Flow ID	avg IAT	psize dev	sum byte	time len	Label	Classification	Clustering (hard)	Clustering (soft)
1	41	54	53	74	P2P	P2P	1	1(80%), 2(15%)
2	64	6	62	45	P2P	P2P	1	1(75%),3(10%)
3	48	80	27	83	E-mail	P2P	2	2(95%)
4	48	83	35	78	VoIP	VoIP	3	3(45%),2(9%)

Fig. 1. Example input for ML algorithms derived from network traffic

algorithms. The algorithms aim at segment the space defined by the features as dimensions.

- *Label*: The goal of ML is to learn to categorize objects based on features. The labels are the name of the categories, hence the label is the result of the testing phase.
- *Training*: The first phase of ML algorithms, when the set of input samples are evaluated (using their features) and models are created.
- *Testing*: The second phase when the models are utilized and tested on unknown traffic to find which model describes them the best. The input to this phase is the models and the features of an unknown object (e.g., flow).
- *Accuracy*: In the test phase, what fraction of the tested objects get the proper label. Labeled test data is needed to measure the accuracy.
- *Classification* is a type of ML algorithm. Label information is used during training (along with the features) that is why it is called *supervised learning*.
- *Clustering* is another type of ML algorithm, also called *unsupervised learning*. This method automatically assigns points into clusters based solely on the features. The label information is not needed during clustering thus it makes possible to deal with new unknown applications. After clustering the label to cluster mapping function must still be defined. One approach is e.g., the most labels in the specific cluster.

Figure 1 shows an example input for ML algorithms derived from network traffic.

There are a large number of publications in the clustering and packet classification area. Most papers usually focus on either clustering [3], [7] or classification [8], [2], [4], [9] but not on their combination. In [10], authors introduced hybrid clustering method which first uses k-means and k-neighrest neighbor clustering to deal with the issue of applications clustered in overlapping clusters, thus improving accuracy and improve performance. In our work, the combination of clustering and classification is used to exploit the different robustness of the methods in case of network parameter changes.

The majority of the publications deal with algorithms working on flow level [1], [2], [8], [3], [7], [11], [12], [4], [13], [14]. Papers introducing methodologies working on packet level information also exist [15], [16], [17]. The flow level information based methods can only identify the flows after the complete processing of the flow. The packet level information based methods can deduce a hint for a traffic flow after a few packets, but they neglect the case when the flow changes traffic characteristics during its lifetime e.g., a VoIP flow starts with signaling and later used for the transferring of the voice. In mobile environments where the available resources of a user is dependent on the load of the mobile cell and the channel resources are reserved according the traffic needs the information in the first few packets of the flow may not sufficient for a robust decision. Our proposed solution operates simultaneously on packet, flow slice and flow levels to achieve a robust and accurate decision as early as possible.

TABLE I
COMPOSITION OF MERGED TRAINING DATA

Protocol	flow%	Protocol	flow%
BitTorrent	61.11	RTP	0.02
DNS	4.50	RTSP	0.02
DirectConnect	0.06	SIP	0.94
FTP	0.01	SMTP	0.01
Gnutella	6.87	SSH	0.01
HTTP	19.82	Source-engine	0.53
ICMP	4.05	UPnP	0.05
IGMP	0.01	WAP	0.22
IMAP	0.03	Windows	1.40
POP3	0.18	XMPP	0.01
PPStream	0.16		

III. INPUT DATA

Later, in the paper, the following data is used for evaluation purposes. We constructed the training and testing data in the same way as it was done in [8]. The *training data* of the system we used a one day long measurement from an European FTTH network, a 2G and a 3G network measurement from Asia and a measurement from a North-American 3G network each of them measured in 2011. We aimed at choosing measurements from networks with very different access technologies and geolocations to make the traffic characteristics varied. Flows are created from the network packet data, where a flow is defined as the packets traveling in both directions of a 5-tuple identifier, i.e., protocol, srcIP, srcPort, dstIP, dstPort with a 1 min timeout. Flows are labeled with a DPI tool developed in Ericsson [18]. The flows are randomly chosen into the training and testing data set with 1/100 probability from those flows where the protocol is recognized by the DPI tool and contained at least 3 packets. From Section IV-D we merge all the training and testing data from the several networks sets into one training and testing dataset containing 50 million flows each. Table I shows the composition of the merged training data.

IV. CLUSTERING VS. CLASSIFICATION

We found that clustering and classification methods perform differently when we use them to identify traffic on unknown networks. In this section we examine an algorithm that mixes these two types of algorithms.

 ISBN: 978-1-61208-183-0

TABLE II
MEASURED ACCURACY OF CLUSTERING METHODS

Method	Tested on same network	Cross-check on other networks
Expectation Maximalization (EM) [3], [7]	85%	65%
K-Means [7]	84%	62%
Cobweb hierarchic clustering [22]	70%	42%
Shared Nearest Neighbor Clustering [23]	95% (20% of the flows are clustered)	93% (12% of the flows)
Autoclass [24]	79%	55%
Constrained clustering [5]	88%	48%
Average	78.5%	60.8%

TABLE III
MEASURED ACCURACY OF CLASSIFICATION METHODS

Method	Tested on same network	Cross-check on other networks
SVM [13], [17], [14], [25]	89%	61%
Logistic Regression	89%	59%
Naive Bayes (complete pdf estimation) [8], [2]	74%	58%
Naive Bayes Simple (mix normal distributions) [8], [2]	70%	57%
Random Forrest [9]	93%	54%
Multilayer Perception [26]	85%	47%
C4.5 [2]	90%	45%
Bayes Net [26]	89%	43%
Average	85%	53.1%

We made experiments with several tools [19], [20], [21], algorithms and with several parameter settings. The features we used are the total set of features, which were mentioned in the related work in Section II (e.g., [2], [11]) and the feature reduction in [12] were applied on them. The results of the classification experiments are collected in Table II and III. The first column shows the case when the training data and the testing data are from the same network, the second column shows the case when the testing data is from a different network than the training data (similar experiment as in [1]). In both columns, we show the result of those scenarios and parameter settings, which give maximum accuracy. The accuracy measures the ratio of correctly classified flow number in terms of the protocol.

In case of clustering methods, the mapping of a specific cluster to an application is a majority decision, e.g., if in the training phase `Cluster_5` contained 100 Bittorrent flows and 10 HTTP flows than during the testing phase if a flow happen to fall into `Cluster_5` it is considered Bittorrent.

We found that clustering is more robust to network parameter changes thus the accuracy drops less when the test set is measured in a different network than the training set comparing to the classification algorithms. On the other hand, classification algorithms can learn a specific network more accurately, thus trained and tested on the flows of the same network, the achieved accuracy is usually higher than the one in the clustering case. Algorithms mainly differ in learning speed and in the number of parameters which has to be set (same conclusion in [8]).

In the following section, we propose a method to combine the advantages of both clustering and classification algorithms.

A. Refinement of clustering with classification

In current state-of-the-art, solutions either standalone supervised (e.g., [8], [2], etc.), or unsupervised methods (e.g., [3], [7], etc.) are used. They either perform well on one specific network but significantly worse on others or they provide more balanced, but less accurate results. We also note that in case of the usage of unsupervised methods, the mapping function has to be defined manually.

Below, we propose a method incorporating unsupervised learning for the basic clustering of the input flows and supervised clustering to automatically deduce the resulting label. Our method is divided into two main phases.

1) Training phase: The input of the training phase is the labeled raw traffic. The output of the system is the clustering and classification models. First, flow descriptors (features) are calculated from the raw traffic, e.g., average payload size, deviation of payload size, etc. Next, an automatic unsupervised clustering is performed, and the resulting clustering model is stored. Finally, the result of the clustering is added to the features of the raw traffic as an additional feature and this extended feature set is fed to an automatic supervised classification system. The resulting classification models are also stored. See Figure 2 for details.

2) Testing phase: The input of the testing phase is the unknown raw traffic. Features are calculated for each flow as in the training phase and are tested on the clustering model. The number of the resulting cluster is added to the feature set, which is then tested on the classification model. The output of the system is a list of traffic types with a confidence level. The classification method also works as a cluster to application mapping function. See Figure 3 for details.

B. Combination of clustering and classification methods

There are two possible ways of combining the clustering and classification methods:

Classification with clustering information: The result of clustering (with the cluster to application mapping completed) is fed to the classification algorithm as a new feature. In this case, the feature expressiveness is chosen arbitrary by the classification method. The advantage of this approach is that it is easy to implement. On the other hand, the clustering information may be neglected or considered with low importance by the classification method thus the clustering cannot always improve the overall accuracy.

Model refinement with per cluster based classification: After the clustering step, a separate classification model is built for the set of flows of each cluster (see Figure 4). The advantage of this approach is that the clustering results are considered always with high importance. The classification methods can construct simple models because the clusters contain a limited number of flow types. As a result, the

 ISBN: 978-1-61208-183-0

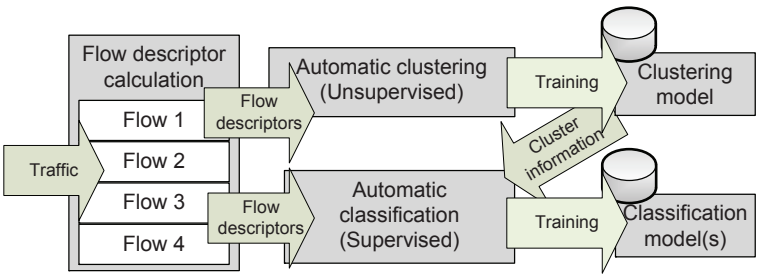


Fig. 2. Refinement of clustering with classification – Training phase

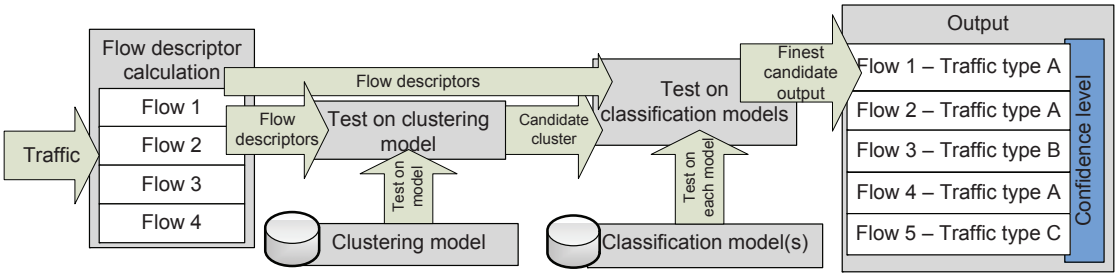


Fig. 3. Refinement of clustering with classification – Testing phase

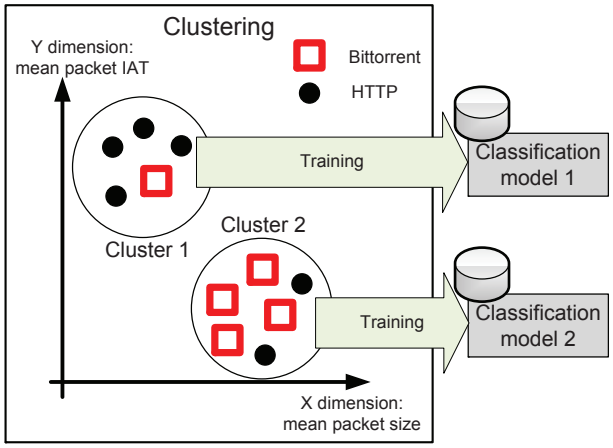


Fig. 4. Per cluster based classification

impact of the overfitting of the classification model is decreased. This approach showed significant improvement over the classification with clustering information scenario (but results in a more verbose model).

C. Preferred implementation

We selected the constrained clustering [5] algorithm (see Section VI for further details) and the SVM [25] classification algorithm to perform the experiments as these algorithms were the most robust for the case when the training and testing data were from different networks.

The focus of the ML-algorithms is slightly different in the clustering and classification case. Clustering calculates Euclidean distances. SVM is a Kernel-based algorithm that projects data into high dimension feature space where the instances are separated using hyperplanes. An important task regarding SVM is to choose an appropriate kernel function. To extend the linear models that constraint clustering can learn, SVM implementations can be tuned to use Gaussian or polynomial kernels. With such kernels it is possible to model non-linear, but exponential dependence of variables thus the clustering and classification models can complement each others capability with linear and non-linear modeling features.

D. Evaluation

Table IV shows that both types of combination improves the performance of both the same network and cross-check case. The increase of the cross-check case improves significantly comparing to the standalone cases (see Tables II, III). In case of the per cluster based classification, the increase is even more significant in both cases than the classification with clustering information case thus we will use it in the rest of the paper.

TABLE IV
MEASURED ACCURACY OF THE COMBINATION OF CLUSTERING AND CLASSIFICATION METHODS

Method	Tested on same network	Cross-check on other networks
Classification with clustering information	89%	72%
Per cluster based classification	93%	75%

We also made measurements of the basic clustering (Figure 5 'Clustering with majority decision' column), classification (Figure 5 'Classification' column), trivia combination

 ISBN: 978-1-61208-183-0

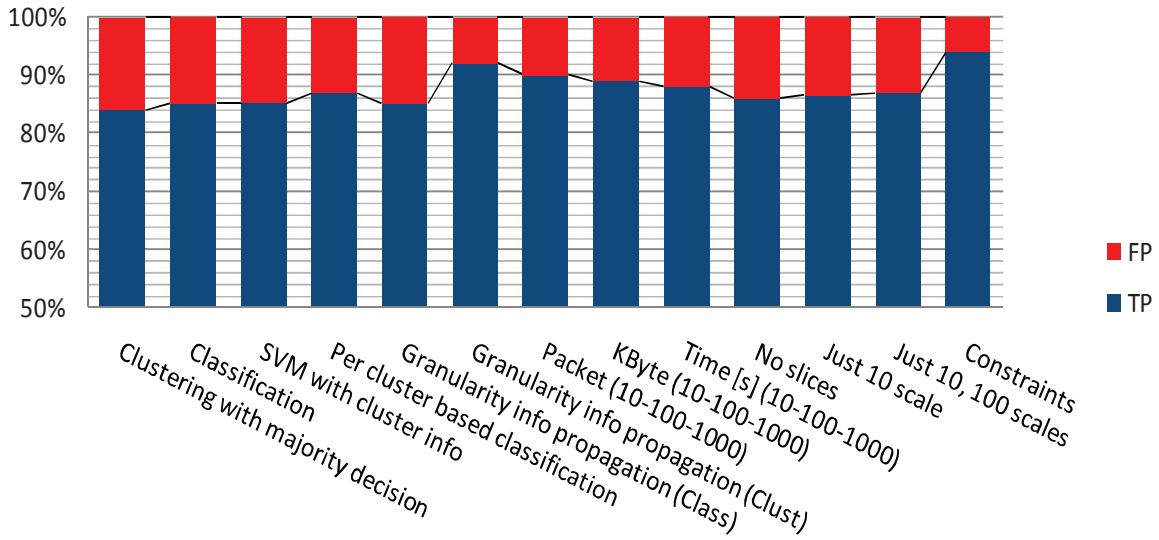


Fig. 5. The summary of the accuracy in case of the application of the proposed improvements

(Figure 5 'SVM with cluster info' column) and per cluster based classification (Figure 5 'Per cluster based classification' column) case on the merged training and testing data (see Section III). TP hit occurs when the label in the original flow equals the hint given for the specific factor. The classification with clustering information case improve its accuracy slightly, but the per cluster based classification overperforms all of them.

V. Granularity levels

Current state-of-the-art packet header-based traffic classification methods can provide information about the flow after its full processing (e.g., [3], [4], etc.). They either collect information at packet or flow level but they cannot propagate the information to other levels.

A. Usage of granularity levels

ML-based traffic classification systems use a set of features. Features can be calculated on several granularity levels. We use multiple granularity levels in our system (see Figure 6) as follows. Collecting traffic description information on *packet level* introduces a limitation on the derived descriptors (features). Practically, only the packet inter-arrival time, packet size and the direction of the packet is available. On the other hand, due to the large number of packets this granularity level provides a sample-rich input.

The most straightforward descriptors on the *flow level* are, e.g., the number of transmitted packets, the sum of bytes transmitted, the distribution of the packet inter-arrival times and packets sizes (or a certain derivative, such as minimum, maximum, average, standard deviation, median, quantiles, etc.). More complex statistical descriptors can also be used, e.g., further moments, autocorrelation, spectrum, H-parameter, recurrence plot-statistics, etc.

Flow characteristics can change over time. The same flow can be used for multiple purposes during its lifetime. This behavior results in misleading conclusions if one views only the statistics calculated for the overall flow without paying attention to the evolution of statistics during the life of the flow. A somewhat finer level, *"slices"*, can be defined as part of a flow divided into multiple pieces, e.g., comprising a certain number of packets, bytes or a given time period. Flow slices can be constructed on several aggregation levels, e.g., based on 10, 100, 1000 packets. The flows can have different characteristics on the different aggregation levels. In this way the scaling property of the traffic can be captured in a similar way as the Hurst-parameter does [27]. It is also possible to segment a flow into slices using some algorithmic determination of slice boundaries, e.g., using TCP flags, significant changes in bitrate, etc. The statistical descriptors can be the same as in case of flow granularity. This approach has the potential of grabbing the temporal changes in the flow during its lifetime and, e.g., remove the inactive periods, which distort the statistical descriptors otherwise.

As we noted, features captured on a lower granularity level (e.g., flow) are richer, but with low number of samples, whereas features captured on a high granularity level (e.g., packet) are simpler, but with high number of samples. It would be desirable if we could combine information from both sources. Another aspect is that high granularity descriptors allow to make a quicker decision, that is, after fewer packets of the flow. An ideal system would provide a quick (potentially lower accuracy decision) fast and would keep refining it as more of the flow is processed. To quickly establish a result, we keep modeling on multiple levels in parallel and propagate information between the levels for higher accuracy.

 ISBN: 978-1-61208-183-0

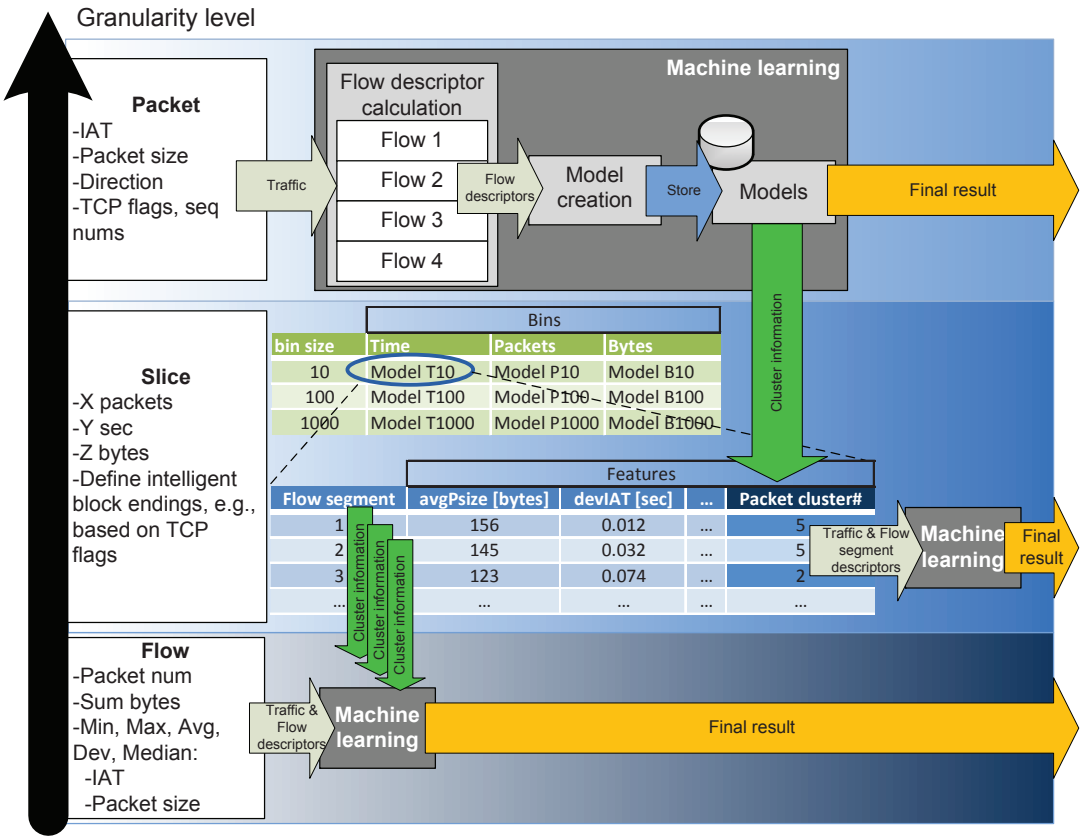


Fig. 6. Usage of granularity levels

B. Propagation of granularity information from one level to an other

The system can provide results as soon as enough information is gained on a granularity level to achieve a required confidence level. This means that, for example, if only 5 packets are enough to provide classification with high level of confidence then further processing is not needed. If the confidence level is still low then the results of a specific granularity level is passed to a lower granularity level. The lower level can then make use of this unreliable, but still potentially indicative information.

There are two possible ways to propagate granularity information from one level to an other:

Propagate final result of a specific level: The result of the refinement, thus the result of the classification, a specific protocol, functionality value, etc. (with a text to number mapping) are fed to the next level classification algorithm as additional features.

Propagate clustering information of a specific level: Each resulting cluster (without cluster to application mapping) is fed to the next level clustering as an additional feature (see Figure 6). Cluster numbers are normalized, aggregated results of several features. They mean that due to some features some flows are similar to each other. This information does not introduce any error to the system. See also Figure 6 for further details.

We should note that it would be possible to use a per cluster based classification like solution as it is proposed in Section IV-B, thus flows in each cluster would generate a separate model on the next granularity level. We did not make experiments with such a setup as the resulting cluster on, e.g., flow level would have a very limited number of flows for training to create a meaningful model. Nevertheless, in a system that handles much more flows, this approach can be feasible and may perform well.

C. Preferred implementation

On packet level the inter-arrival time, packet size, direction (uplink, downlink) and the TCP flags in case of TCP packet can be stored for the first, e.g., 10 packets. This means $10 * (3 + F)$ features to be stored, where F is the number of relevant flags.

On the slice level, we consider, e.g., 10 second long slices. In this case, the first 10 seconds of the flow constitute the first slice. Statistical descriptors are calculated for each slice and all of these features are used as features to the ML-algorithm. Statistics of the next 10 seconds of the flow are also stored, and so on. It is also possible to define a fix number of slices, e.g., 10 and only maintain the statistics

 ISBN: 978-1-61208-183-0

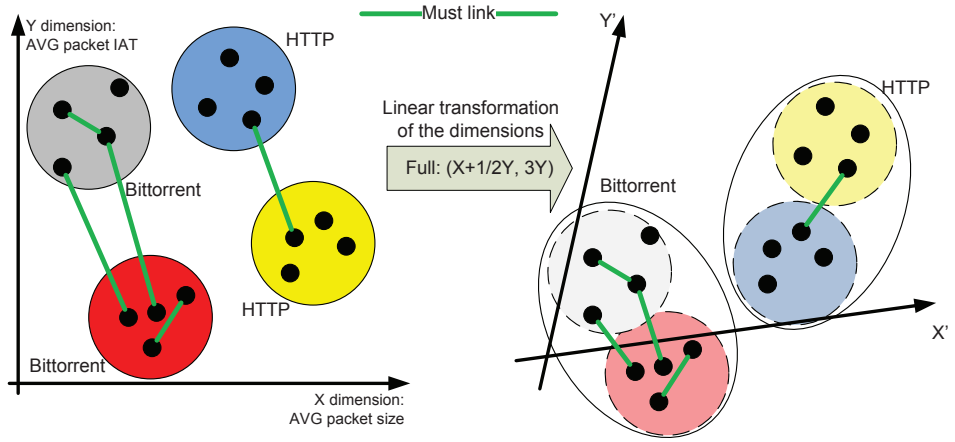


Fig. 7. Constraints clustering mechanism

descriptor for such many slices and cumulate statistics in a circular fashion. Thus the first set of descriptors would hold activity from seconds 0-10, 100-110, 200-210 and so on. The memory consumption of the slices can be limited with the above technique.

D. Evaluation

Figure 5 'Granularity info propagation (Clust)' column shows that propagation of clustering information outperforms the propagation of classification info ('Granularity info propagation (Class)' column). The granularity information introduced provides a further 6% gain comparing to the 'Per cluster based classification' case.

We also made experiments to check which slice dimension (packet/byte/time) contributes with the most information to the final result. In these cases the features were calculated on all scales (10, 100, 1000) and were propagated downward. It is interesting to see that if we have to choose only one of the slice dimensions the highest TP ratio can be achieved with the packet based bin definition.

Another experiment we made that we removed all the slice level information and packet level data was propagated directly to the complete flow level statistics (see 'No slices' column in Figure 5) in the first phase. Later, we extended the information with slice level information considering only the 10 size bins in packet, kbyte and time dimensions as well in the second phase and the 10, 100 size bins in the third. Practically when all the information is propagated from the slice to flow level that means 9 cluster numbers. When only the time bins are propagated that means 3 cluster numbers (for the time 10, 100, 1000 scales) and when e.g., the 10, 100 scales are propagated in all the dimensions it means 6 cluster numbers (the packet, kbyte, time triplet once in the 10 size scaled bin and an other triple for the 100 size scaled bin). We found that providing more and more information by the cluster values of the slices the TP ratio increases by 1-2% step by step (see 'No slices', 'Just 10 scale', 'Just 10, 100 scales' columns in Figure 5). Note that during the granularity level info propagation phase the next level feature set is extended after the feature selection phase therefore they are considered by the clustering methods for sure.

The introduced method can correctly recognize 83% of the flows on the packet level, 8% on segment level and 3% on flow level.

VI. Constrained Clustering

Constraint clustering is a state-of-the-art [5],[6] technique to improve clustering. The key idea is to describe constraints, which tell which instances *must* or *must not* be in the same cluster. Then the feature-space is transformed to fulfill the constraints as much as possible (see Figure 7). It is important to note that constraint clustering can improve the feature selection deficiencies as well (it improves features in a similar way as in Principle Component Analysis [28]).

A. Introduced constraints

In our system, we introduced constraints providing information about the flow instances from independent traffic classification methods. Note that we only propose *must* constraints. It is important that constraints must not introduce error to the system. To achieve this we use only simple and strong heuristics. The introduced *must* constraints are always defined between flows with the same label.

We propose to use the following three constraints (defined for flows being around the same time)

- Constraint Type 1 (red, row/col (1,3); (1,5); (2,3); (2,5)): Flows originating from different srcIPs going to the same dstIP (if we know they are both P2P, we can be sure they are the same app client (factor #1), as well, such as Azureus or uTorrent)
- Constraint Type 2 (orange, row/col (1,3); (1,4); (3,3); (3,4); (4,3); (4,4)): Flows originating from the same srcIP from the same srcPorts (and same for dst) (flows

 ISBN: 978-1-61208-183-0

Flow ID	proto	srcIP	srcPort	dstIP	dstPort	label	avg througput [Mbps]	Constraints
1	TCP	A	B	C	D	P2P	1	1--2
2	TCP	E	F	C	G	P2P	1	
3	TCP	A	B	H	I	P2P	2	1--3
4	TCP	A	B	J	K	P2P	1	1--4
5	TCP	A	L	M	N	P2P	0.2	5--6
6	TCP	A	L	M	N	P2P	5	

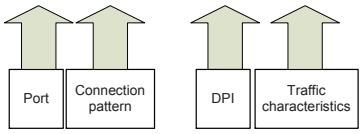


Fig. 8. Introduced constraints

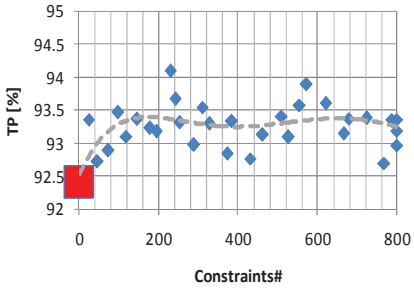


Fig. 9. TP ratio in the function of used constraints /Leftmost square: TP ratio without constraints, dot line: trendline/

from the same IP:port share both application and client program (factors #1 and #3), as well)

- Constraint Type 3 (yellow, row/col (5,3); (5,8); (6,3); (6,8)): Flows with significantly different traffic characteristics with the same user IP address (different characteristics imply different network conditions, factor #5)

B. Evaluation

Figure 9 shows the TP ratio in the function of used constraints. In general, a huge number of constraints could be collected, e.g., for each of the flows which take part in the definition of a type 2 constraint can be constructed a constraint. As increasing the number of constraints increases the run time of the constraint clustering algorithm a lot, the constraints are sampled in practice. The two extreme cases are easy to interpret: it is better to use constraints than not, and it is also very clear that increasing the number of constraints does not imply the increase of TP ratio directly. In the right corner of Figure 9 the TP ratio shows a big variance around the application of 600-1000 constraints. What can be learned from these experiments is that it is advisable to add more and more constraints iteratively during the model construction phase and evaluate whether it increased the overall accuracy or not. It is possible to achieve even 2% gain in TP ratio with a limited number of constraints. The detailed study of the variance of the accuracy in the function of the introduced constraints can be the focus of a further work.

VII. Conclusion

In this paper, we introduced several steps to improve the current state-of-the-art in traffic classification engines relying entirely on packet header data. To become robust our proposed method incorporates clustering and classification methods. This way our method performs well even under changing network conditions. In this step we gained 3% TP ratio compared to standalone clustering or classification methods.

In the second step, we proposed to perform the data collection on several granularity levels and the results of one level to be fed to a lower granularity level. In this step a further 5% is gained relative to the previous step.

Third, we introduced constraint clustering based on connectivity patterns. This resulted in a 2% increase of accuracy.

The overall accuracy of the system on a mix of real world network traffic is 94% which is a 9-10% increase in accuracy comparing to state-of-the-art algorithms.

Acknowledgment

János Szüle thanks the partial support of the EU FP7 OpenLab project (Grant No.287581), the National Development Agency (TAMOP 4.2.1/B-09/1/KMR-2010-0003) and the National Science Foundation OTKA 7779 and 80177.

References

[1] M. Pietrzyk, J.-L. Costeux, G. Urvoy-Keller, and T. En-Najjary, "Challenging Statistical Classification for Operational Usage: The ADSL Case," in *IMC '09: Proceedings of the 9th ACM SIGCOMM conference on Internet Measurement Conference*. New York, NY, USA: ACM, 2009, pp. 122–135.

[2] A. W. Moore and D. Zuev, "Internet Traffic Classification Using Bayesian Analysis Techniques," in *Proc. SIGMETRICS*, Banff, Alberta, Canada, June 2005.

[3] A. McGregor, M. Hall, P. Lorier, and A. Brunskill, "Flow Clustering Using Machine Learning Techniques," in *Proc. PAM*, Antibes Juan-les-Pins, France, April 2004.

[4] F. Palmieri and U. Fiore, "A Nonlinear, Recurrence-based Approach to Traffic Classification," *Comput. Netw.*, vol. 53, pp. 761–773, April 2009.

[5] M. Bilenko, S. Basu, and R. J. Mooney, "Integrating Constraints and Metric Learning in Semi-supervised Clustering," in *ICML '04: Proceedings of the twenty-first international conference on Machine learning*. New York, NY, USA: ACM, 2004, p. 11.

[6] S. Basu, "Repository of Information on Semi-supervised Clustering," retrieved: Oct, 2011. [Online]. Available: http://www.cs.utexas.edu/users/ml/risc/code/

[7] J. Erman, M. Arlitt, and A. Mahanti, "Traffic Classification Using Clustering Algorithms," in *Proc. MineNet '06*, New York, NY, USA, 2006.

 ISBN: 978-1-61208-183-0

[8] N. Williams, S. Zander, and G. Armitage, "A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification," *SIGCOMM Comput. Commun. Rev.*, vol. 36, no. 5, pp. 5–16, 2006.

[9] L. Jun, Z. Shunyi, X. Ye, and S. Yanfei, "Identifying Skype Traffic by Random Forest," in *WiCom '07: Proceedings of the 3rd International Conference on Wireless Communications, Networking and Mobile Computing*, Sept. 2007, pp. 2841 – 2844.

[10] R. Bar-Yanai, M. Langberg, D. Peleg, and L. Roditty, "Real-time Classification for Encrypted Traffic," in *Proc. SEA*, 2010, pp. 373–385.

[11] A. W. Moore, M. L. Crogan, and D. Zuev, "Discriminators for Use in Flow-based Classification," Tech. Rep., 2005.

[12] J. H. Plasberg and W. B. Kleijn, "Feature Selection Under a Complexity Constraint," *Trans. Multi.*, vol. 11, no. 3, pp. 565–571, 2009.

[13] A. Este, F. Gringoli, and L. Salgarelli, "Support Vector Machines for TCP Traffic Classification," *Comput. Netw.*, vol. 53, pp. 2476–2490, September 2009. [Online]. Available: http://dl.acm.org/citation.cfm?id=1576850.1576885

[14] A. Yang, S. Jiang, and H. Deng, "A P2P Network Traffic Classification Method Using SVM," in *Young Computer Scientists, 2008. ICYCS 2008. The 9th International Conference for*, nov. 2008, pp. 398 –403.

[15] L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, and K. Salamatian, "Traffic Classification On The Fly," *SIGCOMM Comput. Commun. Rev.*, vol. 36, no. 2, pp. 23–26, 2006.

[16] E. Hjelmvik and W. John, "Statistical Protocol IDentification with SPID: Preliminary Results," in *SNCW 2009: 6th Swedish National Computer Networking Workshop*, May 2009.

[17] G. G. Sena and P. Belzarena, "Early Traffic Classification Using Support Vector Machines," in *Proceedings of the 5th International Latin American Networking Conference*, ser. LANC '09. New York, NY, USA: ACM, 2009, pp. 60–66. [Online]. Available: http://doi.acm.org/10.1145/1636682.1636693

[18] "Ericsson Network Performance Partnership Tools," retrieved: Oct, 2011. [Online]. Available: http://www.ericsson.com/ourportfolio/telecom-operators/network-performance-partnership?nav=marketcategory002

[19] "Weka 3: Data Mining Software in Java," retrieved: Oct, 2011. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/

[20] "Rapidminer," retrieved: Oct, 2011. [Online]. Available: http://www.rapidminer.com/

[21] S. Zander, T. Nguyen, and G. Armitage, "Automated Traffic Classification and Application Identification Using Machine Learning," in *Proc. IEEE LCN*, Sydney, Australia, November 2005.

[22] D. H. Fisher, "Knowledge Acquisition via Incremental Conceptual Clustering," *Machine Learning*, vol. 2, pp. 139–172, 1987, 10.1007/BF00114265. [Online]. Available: http://dx.doi.org/10.1007/BF00114265

[23] M. S. L. Ertoz and V. Kumar, "A New Shared Nearest Neighbor Clustering Algorithm and Its Applications," in *Workshop on Clustering High Dimensional Data and its Applications at 2nd SIAM International Conference on Data Mining*, 2002.

[24] R. H. John Stutz and P. Cheeseman, "Bayesian Classification Theory," NASA Ames Research Center, 1991, Tech. Rep.

[25] C. Cortes and V. Vapnik, "Support-Vector Networks," *Mach. Learn.*, vol. 20, pp. 273–297, September 1995.

[26] T. Auld, A. W. Moore, and S. F. Gull, "Bayesian Neural Networks for Internet Traffic Classification," *IEEE Transaction on Neural Networks*, vol. 18, pp. 223–239, 2007.

[27] W. E. Leland, M. S. Taqqu, W. W., and D. V. Wilson, "On the Self-Similar Nature of Ethernet Traffic (Extended Version)," *IEEE/ACM Transactions on Networking*, vol. 2, pp. 1–15, 1994.

[28] K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space," *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, Sixth Series*, vol. 2, pp. 559–572, 1901.

 ISBN: 978-1-61208-183-0

TCP SYN Protection: An Evaluation

Pascal Anelli, Fanilo Harivelo
LIM
Université de la Réunion
La Réunion, FRANCE
e-mail: pascal.anelli@univ-reunion.fr,
fanilo.harivelo@univ-reunion.fr

Richard Lorion
LE2P
Université de la Réunion
La Réunion, FRANCE
e-mail: lorion@univ-reunion.fr

***Abstract*—The drop of initial TCP control packets can dramatically penalize flow performance. More the flow is small, more the penalty is important. This paper studies an Active Queue Management (AQM) aiming to protect TCP SYN and SYN-ACK from losses, and evaluates the improvements for short TCP flows and the impacts on long-lived TCP flows. This AQM is an extension based on Random Early Detection (RED). Evaluations are performed in the ns-2 simulator. Results demonstrate the effectiveness of the idea supported in a decrease in the transfer delay of short flows and indiscernible effects on large flows.**

***Index Terms*—TCP; Short-lived flow; AQM; RED; Connection establishment.**

I. INTRODUCTION

Each of us has already noticed that sometimes a web page takes a significant time to display. Asking for refresh permits the page to be loaded. One possible reason for this behavior could be the loss of SYN or SYN-ACK TCP segments as explained in [1]. These TCP segments are exchanged in the three-way handshake procedure used in the connection establishment phase [2]. Congestion is the main cause of these losses. Network reacts by dropping packets in a router's queue. TCP relies on Round Trip Time (RTT) values to tackle packet losses. As no RTT estimate is available at connection opening, the retransmission timeout (RTO) is set by default to 3 seconds [3]. Recently, the IETF working group tcpm (*TCP Maintenance and Minor Extensions*) has proposed decreasing the default RTO value to 1 second [4]. The rationale behind this change is that the RTT of more than 97.5% of the connections observed in a large scale analysis were less than 1 second [5]. However, retransmission rates within the three-way handshake are measured roughly at 2%. This shows that a solution to avoid packet loss in the connection establishment phase will benefit a non-negligible set of connections. While setting the initial RTO value to 1 second provides interesting results, its deployment requires end host modification. Furthermore, one second is inappropriate for brief exchanges. We offer here another option. Rather than dealing with retransmission concerns, i.e., after the loss, we act upstream. Our idea consists in preventing initial packet losses, avoiding in this way, the RTO triggering.

Loss issues at the connection establishment phase are not limited to TCP flow. All connection-oriented services and protocols, such as DCCP and SCTP, are affected. However, short-lived TCP flows are the most impacted. Those are flows made up of few packets. The applications generate short flows in order to ensure an interactive feature. As presented in [6] and [7], Internet traffic is mostly populated by short-lived TCP flows, mainly generated by Web applications. Theses flows suffer more from the initial drops than long-lived flows. Indeed, the connection establishment phase is a process that takes a significant time compared to the duration of the connection. The RTO expiration and corresponding backoff time, due to SYN or SYN-ACK packet loss, add a delay that is significant for short-lived flows. The RTO penalty is perceptible at the service level as the main performance metric is the latency. On the other hand, when a long-lived flow experiences an initial drop, it is equivalent to a shift in the starting of data transfer. The performance metric for this kind of flow is the goodput.

In this paper, we aim to evaluate the benefits of protecting the packets used in connection establishment. The evaluation is made in the context of TCP. The idea is to protect the TCP segments with the SYN flag set (referred to as SYN segment or SYN packet in the following) from losses. The term protection means keeping the SYN segment even if the queue is full. This action is performed on the router's queue. SYN packets are never dropped whenever data packets are present in the queue. Queue management is done in a push-out fashion. If the queue is full on a SYN packet arrival, the last enqueued data packet is pushed out of the queue and dropped. The motivation to act at the router level is that the congestion and the choice of, which packet must be lost, are made at the router level. Furthermore, as the problem affects all connection-oriented services, the solution at this network level deals with this problem globally. With the proposed solution, only data packets are intended to be lost. These packets will be recovered in a better fashion than TCP SYN segment, i.e., either by fast retransmit [8] or by RTO adjusted relatively to the RTT estimations. No additional change other than SYN packet protection is done on routers. Provided service offered by network remains best effort.

The main contribution of this work is the demonstration that the proposed scheme is able to significantly improve the performance of flows by the protection of the segments exchanged in the initial phase. The proposition does not involve a complex identification scheme or per-flow state management. Improvements should be obtained without penalizing long

 ISBN: 978-1-61208-183-0

flows. We propose an implementation of the idea to prove its effectiveness. We then study the integration of the proposed scheme with an existing Active Queue Management (AQM) mechanism, namely, Random Early Detection (RED). This extension of RED is developed and analyzed.

The next section presents the related work that deal with connection establishment and packet losses due to congestion. The proposed idea is described in Section III. In Section IV, we evaluate the performance experienced by short and long TCP flows with the implementation of the proposed AQM. We conclude by our findings.

II. Related work

In the literature, there are overall three types of approaches that address the problem of losses:

1) Preventive actions to reduce the loss rate. Indeed, Besides wasted bandwidth, the retransmission of lost packets introduces extra delay. AQM aims to deal with this congestion issue.
2) Marking packets rather than dropping them; and
3) Responsive actions to improve the retransmission procedure. Solutions based on this approach consider modification of TCP settings.

As mentioned previously, our motivation to act at the router level is that the congestion occurs at this level. For more than a decade, the research community has developed Active Queue Management (AQM) in order to prevent packets being dropped and to maintain high throughput and low delay. In [9], it is recommended to deploy RED [10]. RED monitors the queue length and adopts a packet drop policy based on probabilities, which increase with the level of congestion. However, RED fails to improve the performance of short flows and to provide fairness with unresponsive flows. Choke [11] has been proposed as a solution for this problem. It approximates max-min fairness for the flows that pass through a congested route. It draws a packet at random from the FIFO buffer when a packet arrives and compares it with the arriving packet. If they both belong to the same flow, they are both dropped. Choke can also be considered as a solution for the short-lived flows by considering that it corrects unfairness problems between short and long-lived flows. However, Choke doesn?t prevent the SYN lost when the queue is full. Another way in the router context is the use of DiffServ architecture [12]. In [6], a proposition relies on DiffServ to protect retransmissions and the first packets against loss. After loss detection, the segments are sent with higher priority. This solution is inappropriate for best effort network.

These previous works handle the problem of packet losses, but they do not specifically focus on connection packets. In [13], the authors recognize this problem of lost packets belonging to the connection establishment phase and their simulations show how the response time can be significantly increased by just avoiding the loss of the SYN packet. They show that setting Explicit Congestion Notification (ECN) bits [14] in IP header of TCP control packets while leaving the treatment of the initial TCP SYN packet unchanged, can significantly improve system performance. But, as authors mentioned, this method has a limited scope due to poor usage of ECN on servers and in routers.

Another way to improve the performance in case of lost SYN packets is based on the modification of TCP settings in the operating system, such as defining a smaller value [4] to the initial retransmission value. In [1], the authors investigate the possibility of setting the initial retransmission time to a value smaller than 3 seconds. However, this will then apply to every TCP connection and possibly introduce unnecessary retransmissions and could even cause TCP to fail in certain cases of extreme delay. So they implement an application layer tool to keep a copy of sent packets belonging to the connection opening phase. In case the corresponding acknowledgement does not arrive within a given and a configurable time, the packet is retransmitted. The designed application can be used only for specific ports, such as 80 and not for all TCP connections as opposed to the approach of RTO decrease.

III. Description

Our solution to initial drops in the connection establishment phase is to protect SYN packets within the network. As those losses appear in congestion situations, the proposition takes place on routers. Indeed, a congested router drops packets when its queue fills up (or is about to be filled).

Two types of approaches are possible: scheduling and active queue management. In scheduling, router's buffer is partitioned into separate queues. Each queue holds the packets of one flow or a category of flow. A scheduling mechanism determines which packet to serve next; it is used primarily to manage the allocation of bandwidth (and provide fair sharing) among flows but it can also apply to traffic protection or isolation. This is an interesting option for the isolation of SYN packets from the other traffic. However, algorithmic complexity and scaling issues of scheduling make its deployment on Internet routers difficult.

On the other hand, active queue management, which is concerned with managing the length of packet queues by dropping packets when necessary or appropriate, has a simpler design. A single queue contains all the packets. The deployment of RED, that falls within this class, on Internet routers is highly recommended. RED possesses interesting and useful features; such as its ability to avoid global synchronization, its ability to keep buffer occupancies small and ensure low delays, and its lack of bias against bursty traffic. Our proposition is, then, compared to AQM mechanisms. As the comparison holds on the effectiveness of the SYN packet protection, RED is extended with this additional feature. This new variant of RED will be referred to as REDFavor hereinafter.

With REDFavor, the router serves as a shield for SYN packets against losses. A congestion episode manifests itself by the filling up of the queue. Any new arriving packet is discarded. In normal operation, the router performs this dropping with no regard to the packet type. REDFavor reacts in a different manner if the new packet is a SYN packet. The router makes sure that no SYN packet is rejected if at least

 ISBN: 978-1-61208-183-0

Algorithm 1 REDFavor algorithm

```
1: function enqueue(p)
2: # A new packet p arrives
3: if the SYN flag is set on p then
4:    # p must be protected
5:    if the queue is full or p is an early drop packet then
6:       if only protected packets in the queue then
7:          p is drop
8:          return
9:       else
10:         # Push out
11:         the last standard packet is dropped
12:      end if
13:   end if
14:   p is enqueued in front of all standard packets
15: else
16:   # p is a standard packet
17:   Fall back to RED
18: end if
```

one standard packet is present in the queue. A standard packet is dequeued and dropped in a push-out fashion, as presented in Algorithm 1, to release space for the SYN packet. This latter is, then, enqueued. However, if all packets in the queue are SYN packets, the arriving SYN packet will be dropped as there is no possibility of making room for it.

Thus, although SYN packets are protected during congestion periods at the expense of standard packets, they can still encounter losses. That happens when the queue contains only SYN packets. To lower this potential risk, SYN packets accumulation must be avoided. One response to this point consists in limiting their waiting time in the queue as much as possible. Then, a new enqueued SYN packet is positioned in front of all standard packets and at the tail of already enqueued SYN packets. Thus, it is prioritized in transmission over standard packets.

The exposed protection mechanism can be considered as an isolation or separation of SYN packets from standard ones. This separation relies on SYN flag identification. This flag acts as a priority bit that triggers special and privileged treatment for corresponding packets. A transport layer signalisation is handed over to network-level entities to solve a transport layer issue. Such cooperation can be seen as a cross-layer approach. This operation does not involve complex scheme or per-flow state management. A simple check on the SYN flag suffices; this ensures the scalability of deployment on real networks.

Nevertheless, some questions may arise with the use of isolation and prioritization of SYN packets. Indeed, both types of packets are competing for transmission. One possible issue might be the starvation of standard packets in bandwidth sharing. The problem is not relevant in non-congestion periods. Intuitively, in case of congestion, starvation should not happen as the number and size of SYN packets are relatively small (40 bytes) compared to standard packets. Another potential problem relates to the impact of protection on RED operations and properties. In fact, SYN packets are not checked against RED filters on their arrival. They can be seen as unresponsive and may raise or reinforce congestion. However, the same assertion about the number and size of SYN packets still holds. We think that the effects on RED performance are negligible or minor. These assumptions are validated by simulation results in the evaluation section.

We do not claim that the combination of the proposition and RED is the best one nor gives the best performance. However, this choice highly facilitates analysis and evaluation of the presented solution.

IV. EVALUATION

This section presents the performance simulation results of REDFavor using ns-2 simulator. We look at 2 points:

- Latency of short-lived flows, that expresses the improvements brought by SYN packets protection,
- The counterpart of the observed improvements on long-lived flows.

We compare the performance of the SYN protection with RED and Choke. The simulation is designed to demonstrate the improvements in latency of the proposed SYN protection in a single bottleneck scenario. We adopt the model of web traffic developed in [15]. In this model, a pool of clients request web objects from a pool of servers. Pools are interconnected by a pair of routers and a bottleneck link. This link has a bandwidth of 10 Mbits/s as shown by Figure 1.

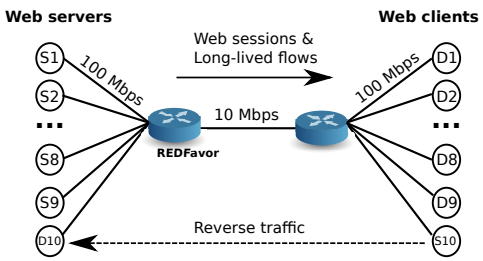


Fig. 1. Simulation model

In detail, we used TCP flows with RTT varying from 5 ms to 100 ms. A client makes a web session. A web session is composed by a sequence of web pages. A page is constituted by several objects. Each object is downloaded through a TCP connection. Then, each object will result in a new flow. The parameters used to obtain the simulated web traffic are the average of the number of pages per session, the average of the number of objects per page, the inter-session time, the inter-page time and the inter-object time, as seen in Table I. Exp(x) means the exponential distribution with mean x. Settings for the object size and the number of objects per page in Table I, are similar to those used by [16]. Consequently, we consider a Pareto object size denoted as P(1,1.2,12) where 1 is the minimum possible object size (in packets), 1.2 is the shape parameter, and 12 is the mean object size (in packets). The Pareto distribution shows high variability. It represents an accurate model of flow size distributions as

 ISBN: 978-1-61208-183-0

TABLE I
SIMULATION PARAMETERS

Parameter	Value
Number of pages per session	Exp(240)
Number of objects per page	Exp(3)
Intersession (s)	Exp(0.5)
Interpage	Exp(5)
Interobject	Exp(0.1)
Object size	P(1, 1.2, 12)

empirically observed on the Internet. The settings for the remaining parameters applied to the web traffic model lead to usage of around 70% of the bottleneck link capacity. The web traffic load is generated by 135 web sessions taking place on each of the 9 web servers. Besides, each web server sends a long-lived TCP to one web client. A flow is generated in the reverse direction to mitigate potential synchronization between flows. REDFavor is applied only on the congested link. The simulation model is illustrated in Figure 1. Data are collected after a 100 second warm-up period. The simulation duration is set to 500 seconds.

A. Web traffic

This subsection evaluates the efficiency of REDFavor to improve the performance of short lived TCP flows. As mentioned earlier, the short flows are the most affected by the loss of initial TCP control packets, in terms of latency. The efficiency of SYN packet protection can be appreciated by a decrease in transfer delay.

Figure 2 shows the cumulative distribution of request completion time. The request completion time of a flow is the time interval starting when the first packet leaves that server and ending when the last packet is received by the corresponding client. RED experiences fewer sessions that terminate their requests within 3 seconds. A noticeable peak appears in 3 seconds with RED. This corresponds to the occurrence of initial RTO. The same observations are reported by [17]. REDFavor eliminates these earlier timeouts. It behaves and leads to the same results as Choke.

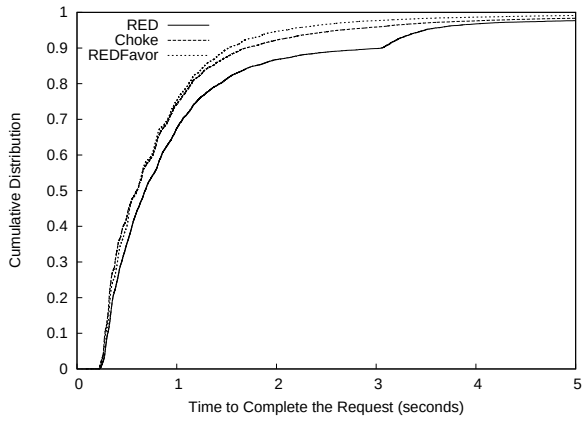


Fig. 2. CDF of completion times

Figure 3 presents the mean completion time as a function of the flow size. REDFavor ensures a lower transfer delay for both short flows and long flows. It performs like Choke with short flows and falls back to RED behaviour on large flows. These results prove that short flows benefit substantially from SYN packets protection offered by REDFavor while long flows are not penalized further than they would with RED.

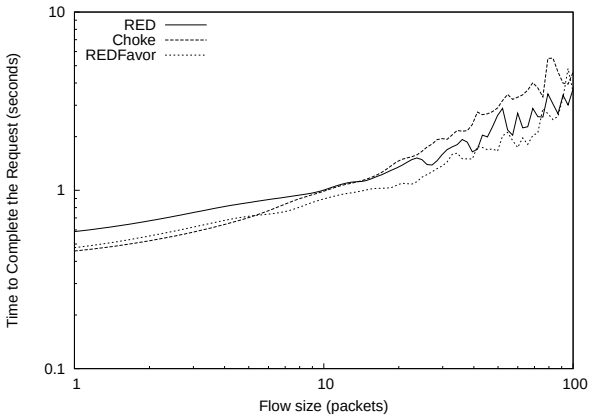


Fig. 3. Mean completion time

Drop protection is evaluated in Figure 4. This figure shows the distribution of dropped packet numbers on the congested link. We note that for flow sizes less than 10 packets, REDFavor has nearly the same behaviour as RED, then, it follows Choke. Quantitative results in Table II show that not a single SYN packet is dropped with REDFavor. Lesser standard packets are even lost compared to the two other schemes. So, protection is obtained at the expense of the loss of some standard packets, i.e., those with a higher packet number. This criticism should be moderate as the drop rate decreases. The obtained results show that REDFavor achieves the initial goal of SYN packet protection.

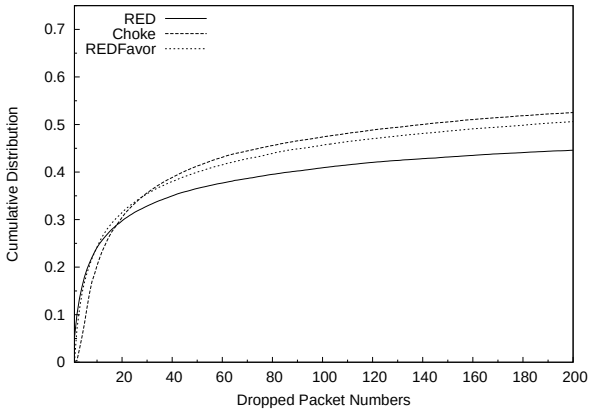


Fig. 4. Distributions of dropped packet numbers on the congested link.

B. Long-lived TCP flow

This subsection studies the impact of the proposition on long-lived flows. Let us remember that the throughput is the

TABLE II
DROPPED PACKETS COUNT

	Choke	RED	REDFavor
Dropped packets count	55566	36758	33150
Drop rate	0.082	0.143	0.073
Dropped SYN packets count	196	3391	0

TABLE III
BANDWIDTH USAGE

	Choke	RED	REDFavor
Bandwidth usage (%)	83.34	96.87	96.71

metric that matters for this type of traffic.

Figure 5 shows the normalized rate obtained by each of the 9 long-lived flows between a couple of web client and web server. With REDFavor, all large flows get the same throughput as in RED. This efficient use of bandwidth is confirmed by the quantitative results presented in Table III. We can note that Choke's improvements for short flows are obtained by a decrease in bandwidth share for long flows.

The impacts of SYN packets protection on long-lived flows are negligible. As stated previously, the "unresponsive" character of SYN packets (and short flows), has no impact on overall performance. Indeed, due to their small size, their participation in congestion occurrence is largely limited.

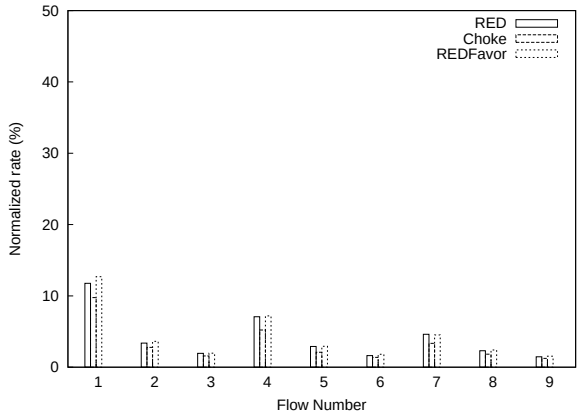


Fig. 5. Bandwidth by the long flows

V. CONCLUSION

This paper proposed a scheme consisting in protecting SYN packets and by the way, leverages the penalty of short flows. The proposition is relatively simple as it is implemented as an AQM scheme on a router's queue. Evaluations validate the idea while showing that noted improvements are not resulting in substantial impact on long-lived flows. A non-permanent (performed at the flow startup) and targeted action significantly improves short flow performance without a significant decrease in the throughput of large flows. Benefits are higher than costs.

The simplicity of the proposition constitutes its main advantage. Operations are solely based on the content of the packet's SYN flag. Deployment of the solution is transparent to end hosts as it involves routers (specifically, queue management), only. As REDFavor works independently, its deployment can be done in incremental manner, i.e., only on routers with heavily loaded links. When the congestion is not present, our AQM has no effect.

In operational aspect, a special attention should be paid to security concerns as the proposition relies on SYN packets identification. For example, it is vulnerable to SYN flood attacks. However, solutions to those security issues, such as firewalling, packet filtering or Intrusion Detection and Prevention Systems, exist and are fully functional. These solutions mitigate those security threats.

ACKNOWLEDGMENT

The authors would like to thank Emmanuel Lochin for earlier discussions on the subject.

REFERENCES

[1] D. Damjanovic, P. Gschwandtner, and M. Welzl, "Why is this web page coming up so slow? investigating the loss of SYN packet (work in progress)," in *IFIP NETWORKING*, Aachen, Germany, May 2009.
[2] J. Postel, "Transmission Control Protocol," RFC 793, Internet Engineering Task Force, September 1981.
[3] R. Braden, "Requirements for internet hosts," RFC 1122, Internet Engineering Task Force, October 1989.
[4] V. Paxson, M. Allman, J. Chu, and M. Sargent, "Computing TCP's retransmission timer," RFC 6298, Internet Engineering Task Force, June 2011.
[5] J. Chu, "Tuning TCP parameters for the 21st century," IETF 75 - Stockholm, Sweden, July 2009. [Online]. Available: http://www.ietf.org/proceedings/75/slides/tcpm-1.pdf
[6] M. Mellia, I. Stoica, and H. Zhang, "TCP-aware packet marking in networks with diffserv support," *Elsevier Computer Networks*, vol. 42, no. 1, pp. 81–100, 2003.
[7] C. Labovitz, S. Iekel-Johnson, D. McPherson, J. Oberheide, F. Jahanian, and M. Karir, "Atlas internet observatory 2009 annual report," in *47th NANOG*, 2009.
[8] M. Allman, V. Paxson, and E. Blanton, "TCP Congestion Control," RFC 5681, Internet Engineering Task Force, September 2009.
[9] B. Braden, D. Clark, J. Crowcroft, B. Davie, S. Deering, D. Estrin, S. Floyd, V. Jacobson, G. Minshall, C. Partridge, L. Peterson, K. Ramakrishnan, S. Shenker, J. Wroclawski, and L. Zhang, "Recommendations on Queue Management and Congestion Avoidance in the Internet," RFC 2309, Internet Engineering Task Force, April 1998.
[10] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance," *IEEE/ACM Transactions on Networking*, vol. 1, no. 4, pp. 397–413, August 1993.
[11] R. Pan, B. Prabhakar, and K. Psounis, "Choke: A stateless aqm scheme for approximating fair bandwidth allocation," in *IEEE INFOCOM*. IEEE, 2000.
[12] S. Blake, D. Blak, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An architecture for differentiated service," RFC 2475, Internet Engineering Task Force, December 1998.
[13] A. Kuzmanovic, "The power of explicit congestion notification," in *ACM SIGCOMM*, Philadelphia, August 2005, pp. 61–72.
[14] K. Ramakrishnan, S. Floyd, and D. Black, "The addition of explicit congestion notification (ECN) to IP," RFC 3168, Internet Engineering Task Force, September 2001.
[15] A. Feldmann, A. Gilbert, P. Huang, and W. Willinger, "Dynamics of IP traffic: A study of the role of variability and the impact of control," in *ACM SIGCOMM*, Vancouver, British Columbia, September 1999.
[16] I. Rai, E. Biersack, and G. Urvoy-Keller, "Size-based scheduling to improve the performance of short TCP flows," *IEEE Network*, vol. 19, no. 1, pp. 12–17, January 2005.
[17] D. Ciullo, M. Mellia, and M. Meo, "Two schemes to reduce latency in short lived TCP flows," *IEEE Communications Letters*, vol. 13, no. 10, October 2009.

 ISBN: 978-1-61208-183-0

Extension of the Shared Regional PACS Center MeDiMed to Smaller Healthcare Institutions

Karel Slavicek, Michal Javornik, Otto Dostal
Institute of Computer Science
Masaryk University
Brno, Czech Republic
Email: karel@ics.muni.cz, javor@ics.muni.cz, otto@ics.muni.cz

***Abstract*—Masaryk University in Brno is operating a regional Picture Archiving and Communications System serving to mostly all hospitals in Brno metropolis and a lot of remote healthcare institutions. The system known under name MeDiMed is utilized by most of the regional hospitals. The last MeDiMed enhancements, which open this system for small healthcare institutions and private doctor's offices, is discussed in this paper.**

Keywords*-*PACS; DICOM; shared solution.

I. INTRODUCTION

The PACS - Picture Archiving and Communications System - [1] is a currently used procedure and methodology for processing medical multimedia data obtained from picture acquisition machines like computer tomography, ultrasound, x-ray, etc. Multimedia medicine data obtained from these machines - in PACS terminology called modalities - are stored in central PACS server. The PACS server then provides these multimedia data to viewing stations. Viewing stations serve to radiologists for analyzing the multimedia data. This approach offers much more capabilities than former film medium. Viewing stations allow image transformation, combination of images from more modalities etc. National Electrical Manufacturers Association - NEMA - [2] has developed a standard DICOM [3] - Digital Imaging and Communications in Medicine - for communications between modalities, PACS servers and viewing stations. DICOM version 3.0 is the currently used by mostly all modalities and PACS servers. The structure of PACS is presented on the Figure 1. In depth background of PACS principle and DICOM protocol is discussed in [4] and [5].

This paper is organized as follows: The overview of MeDiMed project is presented in Section II. Section III describes describes the underlying networking infrastructure. Sections IV and V discuss the service reliability and data redundancy and the overall impact of the MeDiMed project to the healthcare institutions. Currently, we are improving the MeDiMed system to better support small healthcare institutions and private doctor's offices. This effort is discussed in Section VI. The conclusion and further work is presented in Section VII.

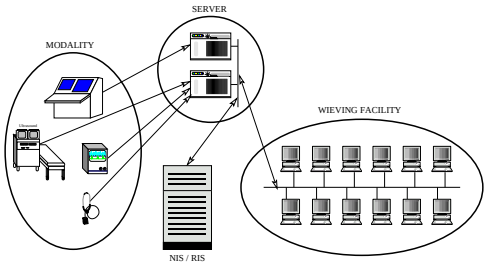


Figure 1. Common structure of PACS system. Modalities serve for acquisition of medicine multimedia data. These data are stored in PACS server and examined and analyzed in viewing stations.

II. MEDIMED

The Shared Regional PACS project MeDiMed started as a collaborative effort among Brno hospitals to process medical multimedia data. Masaryk University is the co-coordinator of this project ensuring that the demands and requirements of radiology departments are met, overseeing the changing legislative standards and the practical limitations of technology. Masaryk University, in cooperation with CESNET (the Czech national research and education network operator), also provides the necessary network infrastructure.

The system serves for transmitting, archiving, and sharing medical image data originating from various medical modalities (computer tomography, magnetic resonance, ultrasound, mammography, etc.) from hospitals. The central PACS serves as a metropolitan communications node as well as a long term archive of patients' image studies. More detailed information about regional PACS archive maintained under the MeDiMed project is described in [6], [7], [8], [9] and [10]. The MeDiMed demand for dedicated communication channels has induced some exploration of backbone optical transport network [11].

Outsourcing of the hospitals' archiving and communications technology permits cooperation among hospitals and the use of existing patient multimedia data. The Shared Regional PACS is more than just a computer network. Gradually, it changes the thinking of medical specialists and gets them to cooperate and share data about patients in

 ISBN: 978-1-61208-183-0

electronic form. It builds a network of medical specialists. The impact of this work is not only in patient care but also in the education of medical specialists. The data stored in the shared archive can be converted to anonymous study or survey (i.e., personal data of patients is replaced by fictitious data) and used for educational purposes.

The realization of the project facilitates fast communication among individual hospitals, allows decision consultations, and brings various other advantages due to direct connections via optic networks. The Masaryk Memorial Cancer Institute is an important node in the regional network of hospitals. It specializes, in a complex way, in oncology diagnostics, treatment and prevention as well as respiratory diseases and gynecology [12]. In general, the proposed MeDiMed project is clearly designed to support society-wide healthcare programs in the Czech Republic as well as programmed implemented by other countries. The system is also supposed to serve as a learning tool for medical students of the Masaryk University as well as physicians in hospitals. The system works in the context of the existing legislative system and will also reflect its changes, especially in the field of data security. The long-term goal will be to adapt legislative standards to the needs of the medical practice (an obligation to provide information, data security issues, etc.).

Nowadays, some type of modalities are commonly used not only by large hospitals but also by small healthcare institutions or even private doctor's offices. This fact brings a new demand for PACS systems. There are ICT departments in large hospitals with enough staff for servicing PACS systems powerful enough server farms, dedicated computer rooms, etc. In small institutions, there is no ICT department so the PACS system should be more robust and reliable. Development of PACS system tailored for small healthcare institutions and private doctor's offices is goal of our current project.

The new goal for the MeDiMed project is to offer PACS system to small institutions. Small healthcare institutions and private doctor's offices usually have limited Internet connectivity and data network availability in general. They are typically located near patients and data communication is not their priority. ICT staff in such institutions is also very limited if it exists at all. For this reason the solution used by large hospitals is not suitable for small institutions. Eventhough the basic principles used in large hospitals can be preserved also in this case.

III. The basic networking principles of MeDiMed

Medicine picture data like X-Ray, CT, US, MR, etc. cannot be used without additional information like picture data description or evaluation, diagnosis, may be reference to history of patients health, previous treatments and other information relevant to patients health. All medicine images have to be equipped with patients identity as well. We are dealing with sensitive information about the patient. The patients privacy must not be compromised.

We have to provide high level of security for medicine picture data maintained by regional PACS archive. We have to protect the data at three stages: data stored on servers of regional PACS, data transported over network between this archive and user, and users access to these data. Security of data stored in regional PACS servers is provided by usage of dedicated hardware for this application and by strict limitation of access (both physical and network based) to this equipment. Security of data transported over network is provided by usage of dedicated fibre optics lines when available and by employing of strong cryptography (IPSEC with AES-256 encryption algorithm) on all lines, which are shared with other data communication traffic.

The main principle of hospital to MeDiMed connection is usage of two firewalls. One of them is in from of MeDiMed PACS servers and is under control of MeDiMed staff. The second one is hospitals firewall and is controlled by hospital staff.

It allows us, as administrators of the application, to control the access to central resources and allows the administrators of the hospital's network to control the access to the hospital's network. That way everybody has under control access to the network he is responsible for. This principle holds for all types of connections (dedicated fiber optics, IPSEC tunnel, or any other) between MeDiMed and the hospital. The communication infrastructure principle is easy to see from Figure 2.

Since Regional PACS system is used on a regular production basis it should provide users with reliable and safe services. Because we are dealing with very sensitive information, we strongly relay on data storage and transport security. Regional PACS archive is running on dedicated network infrastructure with mostly no interaction with another data networks. In case that it is necessary to use public data networks for servicing remote hospitals, we are using strong cryptography to secure medicine data transport.

IPSEC VPN server is connected to the PACS firewall in the same way like local hospitals connected via dedicate fibre optics pairs. The VPN server provides only secure data transport channel. Users connected via IPSEC tunnel then have to follow the same packet filtering like locally connected users.

IV. Server redundancy

For all applications provided to MeDiMed users, we have to offer reliable enough service. Reliability of a service is a quite complicated thing. MeDiMed uses a client-server model. Clients are modalities and viewing station and servers are PACS servers and other servers used to store and retrieve medical images. Under the term reliable enough service we understand the situation when the client can in a reasonable time store or retrieve the medicine image.

 ISBN: 978-1-61208-183-0

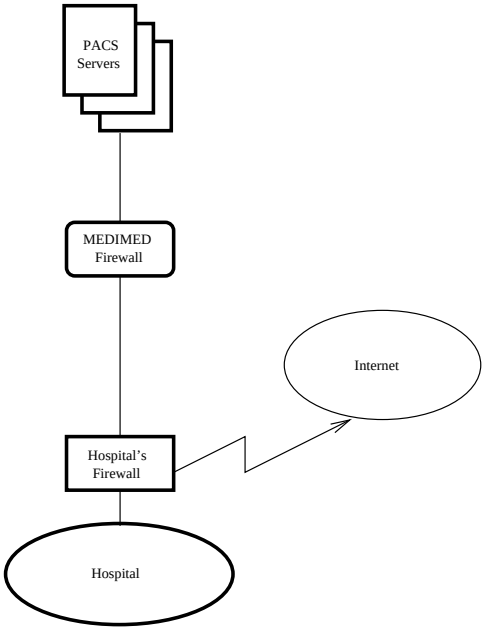


Figure 2. Common principle of hospital to MeDiMed connection. There are two firewalls on the path form modalities and viewing stations to the PACS server. One of them is controlled by hospital and the second on by university.

For better reliability, the key services of Regional PACS system are operated in two distinct locations. PACS systems from different vendors may be used on primary and backup sides if requested. This way, the Regional PACS is able to survive failure of any single fibre optics line, server, storage, electricity (though it is backup-ed via UPS and motor-generator) in one location or even vendor of PACS software.

There is a set of PACS servers used for a routine storage of medicine images. PACS server for this type of service runs on dedicated hardware in both university sites used by MeDiMed. PACS server installations on these sites are standalone and up to some extent independent. One site serves as a primary and until it fails all images are stored into this server. The second site serves as a backup and all data are automatically copied from master to the backup. The backup server is all the time available for retrieving of the medicine images in a read-only mode. This way the overall performance of the system can be improved. If the master site fails we can manually switch the backup site to read-write mode and the former master site to read-only. In many cases the primary and the backup PACS servers are from different vendors. To keep bidirectional synchronization of PACS servers is more than complicated. This manual switching of primary and backup PACS servers provides good enough service with regards to number of failures. Moreover, modalities have some local cache; so, that they can keep images for several days. Older images are available for reading on both primary and backup site.

V. PACS COMMUNICATION SERVERS

A distinct set of PACS servers are so-called communications PACS servers. That means PACS servers used for interchange of medicine images between healthcare institutions. Communications PACS subsystem allows medicine specialist to share the picture data for diagnosis consultations, second reading or even load balancing of radiologists.

Many PACS installations are only limited to the scope of a particular radiology department. An effective use of that technology means image distribution at least throughout the whole healthcare institution. However, the most promising approach to exploiting the PACS technology is to use it at the regional or national level and to support the associated medical processes this way. That means not only basic support of daily routines in radiology departments but also the support of distant consultations, digital long-term archiving or development of shared knowledge databases for research and teaching in this particular area.

Current ICT, as well as existing and developing standards, enable physicians in the region to deliver some services through the computer network. It means that medical specialists from distant specialized departments can consult urgent cases or make decisions. It is a concept of expert centers based on the practices of telemedicine. Image studies of every patient can be referred to a distant expert center for a primary diagnostic or second opinion. This way a much higher quality diagnosis can be assured.

Another important application of the shared regional PACS servers is education. Interesting cases rid of patients personal data and used for both education and research.

The shared regional collaborative environment is more than just a set of computer network applications. Gradually, it changes the thinking of medical specialists and enables them to cooperate and share data about patients in electronic form. It builds a network of medical specialists. The implementation of the system has increased the speed of communication among individual hospitals, allowed decision consultations, and brought various other advantages due to dedicated network connections.

VI. INSTANTPACS PROJECT

The new goal for the MeDiMed project is to offer PACS system to small institutions. Small healthcare institutions and private doctor's offices usually have limited Internet connectivity and data network availability in general. They are typically located near patients and data communication is not their priority. ICT staff in such institutions is also very limited if it exists at all. For this reason the solution used by large hospitals is not suitable for small institutions. Eventhough the basic principles used in large hospitals can be preserved also in this case.

 ISBN: 978-1-61208-183-0

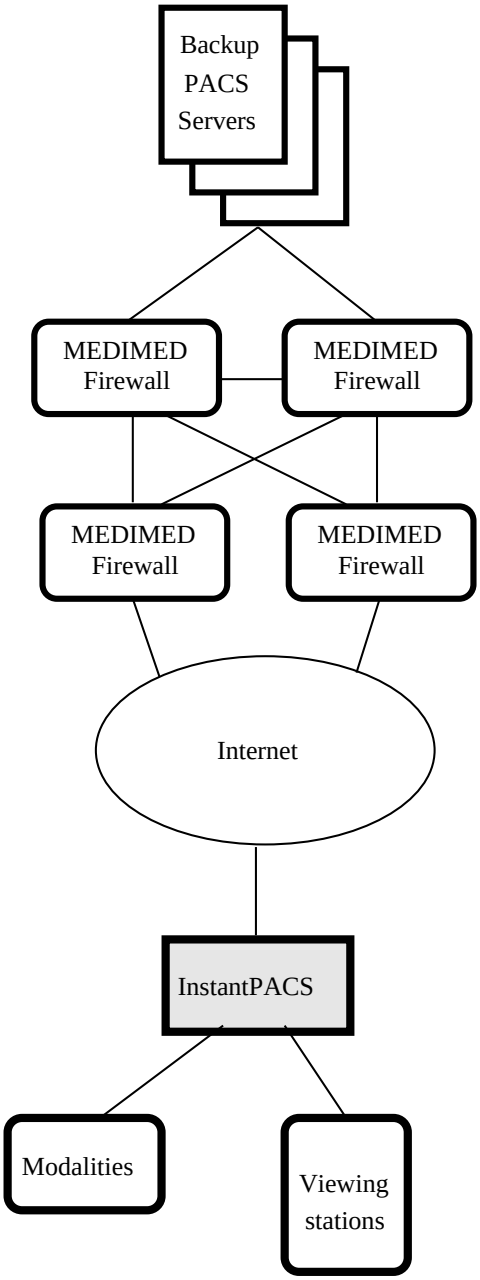


Figure 3. Common principle of InstantPACS communication with the centralized PACS servers.

The aim of the InstantPACS project is to develop a maintenance-free PACS system suitable for small and mid-sized healthcare institutions. This PACS system should offer a user amenity obvious in hospitals including e.g. automatic backup of medicine data. The most important properties are user friendliness, maintenance free operations and pricing acceptable for private doctor's offices. The project is an integral part of the MeDiMed shared regional PACS server overlay project.

As small healthcare institutions and private doctor's offices are being more and more equipped with diagnostics devices, like CT X-ray ultrasound, etc., we expect demand for medicine picture data processing capabilities and services. Our intention is to offer PACS services also to these new perspective medicine users. The specific property of PACS or any ICT services in small healthcare institution is lack of technical staff capable to solve issues on place. For this reason we are developing an "all-in-one" device, which will serve as local PACS server for the healthcare institution and provide backup and communication services. This device will be fully remotely controllable and from point of view of users will not ask for any local maintenance.

Typical user of the InstantPACS will be private doctor's office, which is typically equipped with an ultrasound and one or two more modalities like CT. Currently, modalities in private doctor's office are (from point of view of data communication) isolated devices and data transport is usually performed on USB sticks or the data are processed locally on the modality's console. To offer medical picture processing comfort usual in large hospitals we need to interconnect modalities and viewing stations in the doctor's office. Once the data will be transported from the modality outside we need to provide at least the following services:

- Transport data to the viewing station
- Backup the data on any external device
- Long term archive of the data
- Prevent any unauthorized access to the data while the data are in our device
- Allow to share data between authorized users

The InstantPACS server will be used in a very similar manner like PACS systems in large hospitals. Of course there are some technology discrepancies given by different server placement possibilities in large hospitals (dedicated computer room with air conditioning enough space, etc.) and private doctor's offices (one shared room for treatments and server hosting, room temperature, etc). These worst environmental conditions have introduced some InstantPACS server hardening demands. Backup of medicine picture data from Instant PACS server will be performed on two backup PACS servers located at Masaryk University. The data communication will be performed over Internet via two IPSEC tunnels as shown on Figure 3.

The main properties of the InstantPACS server located at doctor's premises are the following:

- Small dimensions
- General environmental conditions
- No demand for regular local maintenance
- Easy to use
- Expenses corresponding to small size of user's company

The key requirement is no or as small as possible demand for regular local maintenance of the system. Users of the

 ISBN: 978-1-61208-183-0

InstantPACS are expected to have no or very little experience with management of servers operating systems, etc. On the other hand we expect rather large number of users. This leads to demand for maintenance performed by systems user. All critical events and states should be automatically detected and reported. Daily routine maintenance of the system should be practicable in an intuitive way. For example, introduction of new modality (which is typically performed by trained ICT staff in large hospitals) should be performed by general InstantPACS user (medicine doctor).

The hardware platform used for InstantPACS is based on off-the-shelf components. However, it is not like a general PC-like station. It has dedicated memory for system software and configuration and redundant disk subsystem for storage of medicine picture data. It contains also embedded ethernet switch for simple connection of few modalities in a typical private doctor's office. Currently there are two versions available with active and passive cooling system. IPSEC tunnels for backup data encryption are terminated directly in the InstantPACS so no additional equipment is needed.

There is yet one important property of the system to discuss: data protection under marginal or special cases like disk replacement in the InstantPACS or even theft of the whole system. PACS systems are typically located in areas with limited and controlled access in large hospitals. This way enough physical protection of the media containing sensitive patients data is enforced. In case of failure of a disk in raid containing patients data the failed disk is usually physically destroyed so that the data can't leak to unauthorized persons. We expect to use more service or implementation companies who will install the system so it is very difficult to enforce data protection in case of disk replacement. We intend to protect sensitive patients data even in the case when unauthorized person can gain physical access to the InstantPACS device. The work on this topics is still in progress.

VII. Conclusion

Medicine modalities are becoming more widely deployed in medicine public and more commonly used even by private doctor's offices. Development and deployment of PACS systems should follow this trend and offer proper services to new smaller users of medicine modalities. This can improve the healthcare and potentially save life of patients. The introduction of PACS system to small healthcare institutions will bring both some comfort and order into processing of medicine picture data necessary for proper medical examination. The InstantPACS project intends to bring PACS environment commonly used in large hospitals to all medicine users.

Acknowledgment

This work is supported by Czech Technology Agency fund project number TA01010268 - "Maintenance-free PACS system for small and mid-sized healthcare institutions".

References

[1] O. Dostal, M. Petrenko, and M. Filka, "Picture archiving and communication systems," in *Telecommunications and Signal Processing - TSP 2000. Brno*. cwVUT Brno, 2000, pp. 25–27.

[2] "Nema homepage," Online, http://www.nema.org, Retrieved November, 2011.

[3] "Dicom 3.0 standard," Online, http://medical.nema.org/standard.html, Retrieved November, 2011.

[4] K.J.Dreyer, D.S.Hirschorn, J.H.Thrall, and A. Mehta, *PACS A Guide to the Digital Revolution*. USA: Springer Science +Business Media, 2006.

[5] H. K. Huang, *PACS and Imaging Informatics: Basic Principles and Applications*. Hoboken, NJ: Wiley, 2004.

[6] O. Dostal, M. Javornik, and K. Slavicek, "Medimed-regional centre for archiving and interhospital exchange of medicine multimedia data,," in *IASTED International Conference on Communications, Internet, and Information Technology, Scottsdale, Arizona, USA, IASTED, 2003*, M. H. Hamza, Ed. IASTED/ACTA Press, 2003, pp. 609–614.

[7] O. Dostal, M. Javornik, K. Slavicek, M. Petrenko, and P. Andres, "Development of regional centre for medical multimedia data processing," in *IASTED International Conference on Communications, Internet, and Information Technology, November 22 - 24, 2004, St. Thomas, US Virgin Islands*, M. H. Hamza, Ed. IASTED/ACTA Press, 2004, pp. 632–636.

[8] O. Dostal, M. Javornik, and K. Slavicek, "PKI utilisation for PACS users authentication," in *ICN/ICONS/MCL*. IEEE Computer Society, 2006, p. 84. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/ICNICONSMCL.2006.170

[9] O. Dostal and K. Slavicek, "Wireless technology in medicine applications," in *PWC*, ser. IFIP, R. Bestak, B. Simak, and E. Kozlowska, Eds., vol. 245. Springer, 2007, pp. 316–324. [Online]. Available: http://dx.doi.org/10.1007/978-0-387-74159-8_30

[10] O. Dostal, M. Javornik, and K. Slavicek, "Opportunity of current ICT in the processing of medical image information," in *IASTED International Conference on Advances in Computer Science and Technology, January 23-25, 2006, Puerto Vallarta, Mexico*, S. Sahni, Ed. IASTED/ACTA Press, 2006, pp. 193–195.

[11] K. Slavicek and V. Novak, "Introduction of alien wavelength into cesnet dwdm backbone," in *Sixth International Conference on Information, Communications and Signal Processing*. IEEE, 2007, pp. 1–5.

[12] O. Dostal, M. Petrenko, and P. Ventruba, "First application of picture archiving and communication system (pacs) in gynecologic endoscopy," in *ISGE - 6th Regional Meeting*. ISGE, 2002, pp. 53–56.

 ISBN: 978-1-61208-183-0

Multi-Tenancy Authorization System with Federated Identity for Cloud-Based Environments Using Shibboleth

Marcos A. P. Leandro, Tiago J. Nascimento, Daniel R. dos Santos, Carla M. Westphall, Carlos B. Westphall
Post Graduation Program in Computer Science (PPGCC)
Federal University of Santa Catarina (UFSC-INE) - Florianópolis, Brazil
{ marcosleandro, tiagojn, danielrs, carlamw, westphal }@inf.ufsc.br

Abstract— **The services provided in clouds may represent an increase in the efficiency and effectiveness in the operations of the enterprise business, improving the cost-effectiveness related to services and resources consumption. However, there is concern about the privacy of data, since such data are outside the client's domain. For these services to be effectively enjoyed by organizations it is necessary to provide access control. The objective of this work is to provide identity management, based on digital identity federation, with authentication and authorization mechanisms for access control in cloud computing environments to independent, trusted third-parties.**

Keywords-cloud computing; identity management; multi-tenancy; federation; Shibboleth; access control; authentication; authorization.

I. INTRODUCTION

Cloud computing enables the use of services and resources on demand. It uses existing technologies such as virtualization, web services, encryption, utility computing and the Internet [1] [2].

The services provided in clouds may represent an increase in the efficiency and effectiveness in the operations of the business enterprise, improving cost-effectiveness in relation to the consumption of resources and services. Cloud computing systems have many superiorities in comparison to those of existing traditional service provisions, such as reduced upfront investment, expected performance, high availability, infinite scalability, tremendous fault-tolerance capability and so on [3]. Enterprises such as Salesforce.com and Google build and offer a cloud service, while many companies and government entities consider building private cloud data centers or integrating cloud services into their infrastructure [4].

However, there is concern about the privacy of data, since such data are outside the domain of the client. That is, on the one hand we have the advantages of the services available, but, on the other hand, there is concern about security. For these services to be effectively enjoyed by organizations is necessary to provide access control.

The success of cloud computing depends on the evolution of the customer mechanisms of Identity and Access Management (IAM) to service providers. IAM plays an important role in controlling and billing user access to the shared resources in the cloud [5]. IAM must evolve for the cloud to become a trusted computing platform [6]. For consumer organizations using the services offered in the cloud it is necessary to implement a safe and reliable IAM model [1] [7] [8].

IAM systems need to be protected by federations, which are groups of organizations that establish trust among themselves to cooperate safely in business. Identities used in this context are called "federated identity". The user can be authenticated in an organization and can use the services of another organization of the federation without the need to repeat the process of authentication (Single Sign-On). Some technologies implement federated identity, such as the SAML (Security Assertion Markup Language) and Shibboleth system [5] [9].

The aim of this paper is to propose a multi-tenancy authorization system using Shibboleth [10] for cloud-based environments. The main idea is to demonstrate how an organization can use Shibboleth to implement in practice a system of access control in a cloud computing environment, without a trusted third-party.

The following sections are organized as follows: Section II describes related work; Section III introduces the basic concepts of cloud computing; Section IV describes the concepts of identity management and presents the architecture and operation of the Shibboleth; Section V presents the proposed multi-tenancy authorization system; Section VI presents the scenario of implementation of the proposed system and how it was implemented; Section VII presents the results and Section X presents the conclusions and future work.

II. RELATED WORK

In [11], an architecture for a new approach to the problem identified as "Mutual Protection for Cloud Computing (MPCC)" is presented. The main concept underlying MPCC is based on the philosophy of Reverse Access Control, where customers control and attempt to enforce the means by which the cloud providers control authorization and authentication within this dynamic environment, and the cloud provider ensures that the customer organization does not violate the security of the overall cloud structure itself. This work only provides a theoretical framework.

In [12], an approach for IDM is proposed, which is independent of Trusted Third Party (TTP) and has the ability to use identity data on untrusted hosts. The approach is based on the use of predicates over encrypted data and multi-party

 ISBN: 978-1-61208-183-0

computing for negotiating the use of a cloud service. It uses active bundle—which is a middleware agent that includes PII data, privacy policies, a virtual machine that enforces the policies, and has a set of protection mechanisms to protect itself. An active bundle interacts on behalf of a user to authenticate to the cloud service using the user's privacy policies. A prototype using the technology of Java agents on the JADE environment was developed.

In [13], an entity-centric approach for IDM in the cloud is proposed. The approach is based on: (1) active bundles—similarly to [12]; (2) anonymous identification to mediate interaction between the entity and cloud service by using the entity's privacy policies. Angin et al. [13] proposed the cryptographic mechanisms used in [12] without any kind of implementation or validation.

In comparison with the related work, the infrastructure obtained to provide identity management and access control aims to: (1) be an independent third party, (2) authenticate cloud services using the user's privacy policies, providing minimal information to the SP, (3) ensure mutual protection of both clients and providers. This paper highlights the use of a specific tool, Shibboleth, which provides support to the tasks of authentication, authorization and identity federation. Beyond these objectives, the main contribution of our work is the implementation in cloud and the scenario presented.

III. CLOUD COMPUTING

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [14].

In this business model, a customer only pays for services used and as use, without prior commitment, enabling cost reductions in IT deployment and a scalability of far greater resources, which are abstracted to users in order to appear unlimited, and presented through a simple interface that hides the inner workings [15].

From the provider side, the services to be provided are automatically prepared and managed in a multi-tenant model, where a physical server can simultaneously respond to multiple users through virtualization technologies of computing resources.

There are three types of service models that may be offered by cloud computing [14]:

1) Software as a Service (SaaS): providing applications running in the cloud, where the customer has virtually no access control or management of the internal infrastructure;

2) Platform as a Service (PaaS): providing a set of tools that support certain technologies of development and all the necessary environment for deploying applications created by the customer, who is able to control and manage them within the limits of its application;

3) Infrastructure as a Service (IaaS): providing basic computing resources such as processing, storage and network bandwidth where the client can run any operating system or software and maintain as much control as possible.

IV. IDENTITY MANAGEMENT

Digital identity is the "representation of an entity (or group of entities) in the form of one or more elements of information (attributes) that enable the entity to be recognized only within a context" [9].

Identity Management (IdM) is a set of functions and capabilities, such as administration, management and maintenance, discovery, information exchange, policy enforcement and authentication, used to ensure identity information, thus assuring security. An identity management system (IMS) provides tools for managing individual identities in a digital environment [9].

Some specialized features for IMS includes a Single Sign-On (SSO), where a user does not need to be signed on several times to call various applications, and can reuse the authenticated status of a previous application in the same session [16].

An IMS consists of protocols and software components that address the identities of individuals throughout the life cycle of their identities. It involves three main types of entity: the user, the Identity Provider (IdP) and Service Provider (SP). IdPs are responsible for issuing and managing user identities and issue credentials. SPs (also known as relying parties) are entities that provide services to users based on their identities (attributes) [17].

A. Functions of Identity Management Systems

Following are the main functions of an IMS:

- *Provisioning:* the practice of provisioning of identities within an organization addresses the provisioning and deprovisioning of several types of user accounts (e.g. end user, the application administrator, IT administrator, supervisor, developer, etc.) [8].
- *Authentication:* is the process of ensuring that the individual is who he claims to be, and is identified through various mechanisms, such as login, password, biometrics, token, etc. [16].
- *Authorization:* a common need in security is to provide different access levels (e.g. deny/allow) for different parts or operations within a computing system. This need is called authorization [16].
- *Federation:* it is a group of organizations or SPs that establish a circle of trust that allows the sharing of information of user identities to each other [17].

B. Shibboleth

The OASIS SAML standard defines an XML-based framework for describing and exchanging security information between on-line business partners. This security information is expressed in the form of portable SAML assertions that applications working across security domain boundaries can trust. The OASIS SAML standard defines precise syntax and rules for requesting, creating, communicating, and using these SAML assertions [18].

 ISBN: 978-1-61208-183-0

The Shibboleth [10] is an authentication and authorization infrastructure based on SAML that uses the concept of federated identity. With it you can create a safe structure that simplifies the management of identities and provides the user with a SSO for different organizations belonging to the same federation, and who share their identity information in order to do so. The Shibboleth system is divided into two entities: the IdP and SP (Figure 1).

The IdP is the element responsible for authenticating users. It maintains and controls their credentials and attributes, disseminating this information to requests from entrusted organizations. It is composed of four components:

1) Handle Service (HS): authenticates users along with the authentication mechanism and creates a handle token (the SAML assertion that carries the credentials) to the user. Allows an organization to choose the authentication mechanism.

2) Attribute Authority (AA): AA handles requests for SP attributes, applying privacy policies on the release of these attributes (Attribute Release Policies - ARP). Allows the user to specify who can access them. Allows the organization to decide which directory service is used.

3) Directory Service: (external to Shibboleth) local storage of user attributes.

4) Authentication Mechanism: (external to Shibboleth) allows users to authenticate with the central service with only a login/password pair.

The SP Shibboleth is where the resources are stored, that are accessed by the user. It enforces access control on resources based on information sent by the IdP. A single SP may be composed of several applications, but will still be treated as a single entity by an IdP. It has three main components:

1) Assertion Consumer Service (ACS): responsible for receiving messages (SAML) to establish a secure environment.

2) Attribute Requester (AR): responsible for obtaining and passing user attributes to RM.

3) Resource Manager (RM): intercepts requests for resources and makes decisions to control access based on user attributes.

The WAYF ("Where Are You From", also called the Discovery Service) is an optional feature on the Shibboleth system, responsible for allowing an association between a user and organization. When trying to access a resource, the user is forwarded to an interface that asks you to choose the institution to which it belongs. After choosing the institution, the user is redirected to start the authentication process. The WAYF service can be distributed as part of a SP or as part of the third code operated by a federation. In cases where it is used with SPs offering resources for registered users in several IdPs it becomes quite useful.

The flow of operation of Shibboleth is represented in Figure 1.

In Step 1, the user navigates to the SP to access a protected resource. In Steps 2 and 3, Shibboleth redirects the user to the WAYF page, where he should inform his IdP. In Step 4, the user enters his IdP, and Step 5 redirects the user to the site, which is the component HS of the IdP. In Steps 6 and 7, the user enters his authentication data and in Step 8 the HS authenticate the user. The HS creates a handle to identify the user and sends it also to the AA. Step 9 sends that user authentication handle to AA and to ACS. The handle is checked by the ACS and transferred to the AR, and in Step 10 a session is established. In Step 11 the AR uses the handle to request user attributes to the IdP. Step 12 checks whether the IdP can release the attributes and in Step 13 the AA responds with the attribute values. In Step 14 the SP receives the attributes and passes them to the RM, which loads the resource in Step 15 to present to the user.

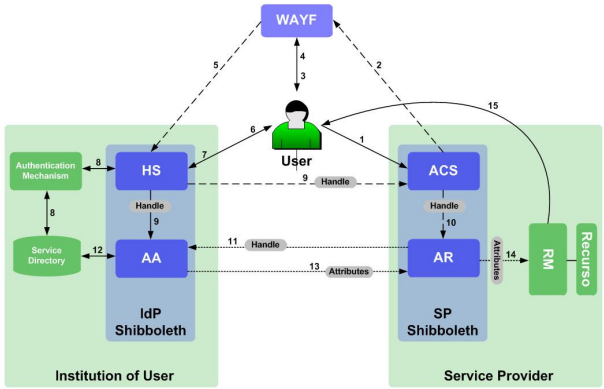


Figure 1. Operation of Shibboleth.

V. FEDERATED MULTI-TENANCY AUTHORIZATION SYSTEM ON CLOUD

According to [5], in order to ensure access control in open environments such as cloud computing, IdM can be implemented in several different types of configuration Figure 2. Firstly, IdM can be implemented in-house. In this configuration, identities are issued and managed by the user companies. Also, IdM itself can be delivered as an outsourced service, which other companies and consumers use. This is called Identity as a Service (IDaaS). There are several commercial offerings in the market. In this configuration, identities are issued and managed by user companies and/or IDaaS providers. In a “managed” hosting case, an IDaaS provider maintains a complete set of employee data that a user company outsources. In other cases, IDaaS providers only maintain pseudonyms of employees, which user companies map to real employee identities. Lastly, each cloud SP may independently implement a set of IdM functions. This configuration requires user companies to maintain a different set of identities for each of the relying parties.

In this work, it was decided to use the first case configuration (in-house), where the client company has complete control and responsibility for the digital identities of its users.

 ISBN: 978-1-61208-183-0

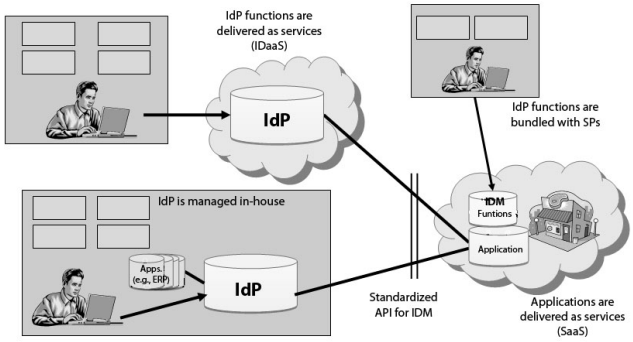


Figure 2. Configurations of IDM systems on cloud computing environments [5].

This work presents an authorization mechanism to be used by an academic institution to offer and use the services offered in the cloud.

The part of the management system responsible for the authentication of identity will be located in the client organization, and communication with the SP in the cloud (Cloud Service Provider, CSP) will be made through identity federation. By establishing trust between the parties, the CSP will request the authentication of users to the IdP located in the client. Thus, the user's data remain under the care of his own company, enhancing privacy and preventing loss of information.

The access system performs authorization or access control in the environment. The CSP should be able to interpret and separately allow access in accordance with the privileges of each user. The institution has a responsibility to provide the user attributes for the deployed application SP in the cloud.

The authorization system should be able to accept multiple clients, such as a multi-tenancy. The concept of multi-tenancy [19] states that an application is used equally across a series of users, each receiving comparable or equitable levels of responsiveness and bandwidth through the use of the Tenant Load Balancer.

VI. SCENARIO

The setting is an academic federation sharing services in the cloud (Figure 3).

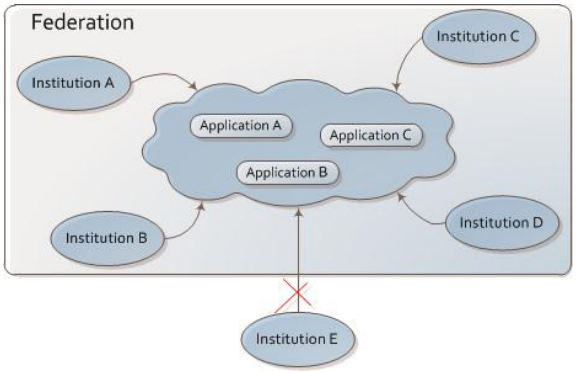


Figure 3. Academic Federation sharing services in the cloud.

A service is provided by an academic institution in a CSP, and shared with other institutions. In order to share services is necessary that an institution is affiliated to the federation.

For an institution to join the federation it must have configured an IdP that meets the requirements imposed by the federation. Since these requirements are expressed in the form of access and privacy policies defined by SAML.

Once affiliated with the federation, the institution will be able to authenticate its own users, according to the authentication and authorization system described in the previous session, since authorization is the responsibility of the SP.

A. Implementation of the Proposed Scenario

For testing and demonstration, a SP was primarily implemented in the cloud. Resulting in the deployment of an Apache server on a virtual machine hired by the Amazon Web Services cloud provider—as illustrated in Figure 4. In this server, beyond the installation of the Shibboleth SP, an application was chosen to serve as an example of the resource to be offered as a service: the software development and collaborative editing of documents DokuWiki [20]. The concept of a lazy session was used to allow users to access the wiki anonymously for reading, and only having to authenticate when permission was needed to edit documents.

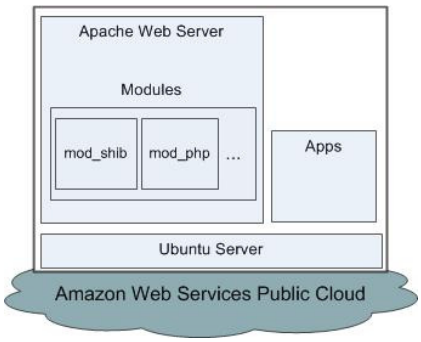


Figure 4. Cloud Service Provider Diagram.

Authorization within the Shibboleth SP can be accomplished in three ways:

- via directives in the .htaccess Apache file, where instructions "require" may include specific users, groups, etc.;
- via the <AccessControl> element that provides several possibilities for more complex use cases of access control [10];
- via the application, which is free to create internal rules according to the attributes available.

The SP was configured with authorization via application, to differentiate between common users and administrators of Dokuwiki.

Before releasing access to users, it was necessary to specify which attributes, among those released by the IdP, the application would be using and how they would be used. This step is application specific, and Figure 5 shows the contents of the file /etc/dokuwiki/local.php, which combines

 ISBN: 978-1-61208-183-0

the attributes of the IDP "Shib-inetOrgPerson-cn " and "Shib-eduPersonPrincipalName" to the attributes of the application "var_remote_user" and "var_name", respectively. Other combinations are also possible.

```
$conf['auth']['shib']['var_remote_user'] =
'Shib-inetOrgPerson-cn';

$conf['auth']['shib']['var_name'] = 'Shib-
eduPerson-eduPersonPrincipalName';
```

Figure 5. Contents of the file /etc/dokuwiki/local.php

Later, a cloud IdP was installed (Figure 6), only to illustrate that each institution has its own IdP control, without regard to whether it is local or cloud.

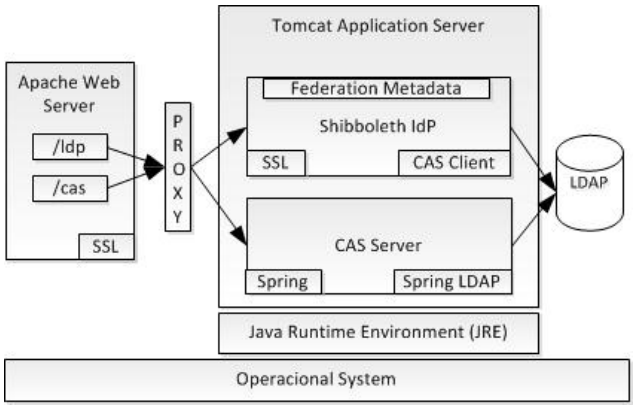


Figure 6. IdP detailed diagram.

The authentication mechanism is external to Shibboleth, and for this purpose we used the JASIG CAS Server [21] that performs user authentication through login and password and provides SSO via a web interface, and then passes the authenticated users to Shibboleth. The CAS has been configured to search for users in a Lightweight Directory Access Protocol (LDAP). To use this directory OpenLDAP [22] was installed in another virtual machine, also running on Amazon's cloud.

To demonstrate the use of SP for more than one client, another IdP was implemented, also in cloud, similar to the first. From this point, the concept of multi-tenancy is necessary, since the service provided by SP will be shared by multiple clients. To support this task Shibboleth provides a WAYF component, which is responsible for allowing the association between a user and an organization. This mechanism was set up by the SP to manage the institutions belonging to the federation.

VII. USE CASES: ANALYSIS AND TEST RESULTS WITHIN SCENARIO

The result of the deployment of IdPs and SPs is shown in Figure 7.

In this resulting structure, each IdP is represented in a private cloud, and the SP is in a public cloud.

Once the scenario was implemented, some tests were performed. The results highlighted two main use cases:

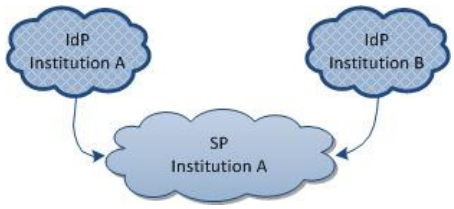


Figure 7. Resulting deployment scenario

A. *Read access to documents*

In the case of read-only access, the service offered allows anonymous use. To perform this type of access the user simply types the URL of the desired service in the Web Browser. In this case, the service is also available for external access to the federation.

B. *Access for editing documents*

In order for a user to have permission to alter documents, authentication is required. In this case, the user will require authentication through the "Login" button that directs them to a URL protected by Shibboleth. The Mod_shib process verifies that the user does not have an open session and forwards it to the Discovery Service of the Federation (WAYF), where the user chooses their institution and is forwarded to the URL of the IdP of the institution chosen, by an HTTP redirect. On the Web page of the IdP, the user enters their credentials (username and password) and if successful, cookies are registered and a handle—a SAML assertion with data from the Authentication—is created.

This handle is used by the SP to request user attributes from the IdP, which analyzes the request based on previously established rules of release. If the handle is valid and the request is accepted, another cookie is created and the user is finally redirected to the SP (the Web application that was originally accessed), and their attributes are sent by the Shibboleth module to this application as values of the environment variables, so that it can use them any way it chooses. With these attributes, the application uses internal authorization rules to determine whether the user has administrative rights on the system.

VIII. CONCLUSIONS AND FUTURE WORK

The adoption of secure IdM in a cloud environment addresses the issues of identity provisioning, authentication, authorization and federation. The use of federations in IdM plays a vital role in enabling organizations to authenticate their users cloud services using any chosen IdP [7].

The work of Albeshri and Caelli [11] only provides a theoretical framework. Ranchal et al. [12] developed a means to protect privacy in a cloud environment and a prototype using the technology of Java agents on the JADE environment. An active bundle was used, that is a container with a payload of sensitive data, metadata, and a virtual machine. Angin et al. [13] proposed the cryptographic mechanisms used in [12] without any kind of implementation or validation.

The focus of this work was aimed at an alternative solution to a IDaaS, which is a solution where the activities

 ISBN: 978-1-61208-183-0

concern outsourced IdM, and therefore, the data and sensitive information of users are outside the domain of the organization, since they are controlled and maintained by a third party.

The infrastructure obtained to provide identity management and access control aims to: (1) be an independent third party, (2) authenticate cloud services using the user's privacy policies, providing minimal information to the SP, (3) ensure mutual protection of both clients and providers. This paper highlights the use of a specific tool, Shibboleth, which provides support to the tasks of authentication, authorization and identity federation.

Shibboleth was very flexible with regards to its use in a cloud environment, allowing a service to be provided reliably and securely. In addition, Shibboleth is based on SAML, which means it is compatible with international standards, thus ensuring interoperability.

With the settings applied to the scenario, it became possible to offer a service allowing public access in the case of read-only access, while at the same time requiring credentials where the user must be logged in order to change documents.

As future work, we propose an alternative authorization method, where the user, once authenticated, carries the access policy, and the SP should be able to interpret these rules. Thus, the authorization process will no longer be performed at the application level.

We also suggest expanding the scenario to represent new forms of communication, and thus create new use cases for testing. For example, (i) provide a service deployed on a new SP where this service is provided by another institution; (ii) provide more than one service in the same SP.

A further example would be to use pseudonyms in the CSP domain, which should ensure the nature of the individual user without the need to expose their real information.

REFERENCES

[1] B. Grobauer, T. Walloschek, and E. Stocker, “Understanding Cloud Computing Vulnerabilities,” IEEE Secur. Priv., vol. 9, no. 2, pp. 50–57, Mar.-Apr. 2011, doi: 10.1109/MSP.2010.115.

[2] F. Maggi, and S. Zanero, “Is the Future Web more Insecure? Distractions and Solutions of New-Old Security Issues and Measures,” Proc. 2nd Worldwide Cybersecurity Summit (WCS 11), 1-2 June 2011, pp. 1–9.

[3] M. Zhou, R. Zhang, D. Zeng, and W. Qian, “Services in the Cloud Computing Era: A survey,” 4th Intl. Univ. Communication Symposium (IUCS 10), pp. 40–46, doi: 10.1109/IUCS.2010.5666772.

[4] “Identity Federation in a Hybrid Cloud Computing Environment Solution Guide,” JUNIPER Networks, accessed in Oct. 2011. Online at: http://www.juniper.fr/us/en/local/pdf/implementation-guides/8010035-en.pdf.

[5] E. Bertino, and K. Takahashi, Identity Management - Concepts, Technologies, and Systems. ARTECH HOUSE, 2011.

[6] E. Olden, "Architecting a Cloud-Scale Identity Fabric," Computer, vol. 44, no. 3, Mar. 2011, pp. 52–59, doi: 10.1109/MC.2011.60.

[7] “Security Guidance for Critical Areas of Focus in Cloud Computing,” CSA, accessed in May 2011. Online at: http://www.cloudsecurityalliance.org.

[8] “Domain 12: Guidance for Identity and Access Management V2.1.,” Cloud Security Alliance. - CSA, accessed in Sep. 2011. Online at: https://cloudsecurityalliance.org/guidance/csaguide-dom12-v2.10.pdf.

[9] D. W. Chadwick, Federated identity management. Foundations of Security Analysis and Design V, Springer-Verlag: Berlin, Heidelberg 2009 pp. 96–120, doi: 10.1007/978-3-642-03829-7_3.

[10] “Shibboleth 2 Wiki,” SHIBBOLETH, accessed in Sep. 2011. Online at: https://wiki.shibboleth.net/confluence/display/SHIB2/Home.

[11] A. Albeshri, and W. Caelli, “Mutual Protection in a Cloud Computing environment,” Proc. 12th IEEE Intl. Conf. on High Performance Computing and Communications (HPCC 10), pp. 641-646, doi:10.1109/HPCC.2010.87.

[12] R. Ranchal, B. Bhargava, A. Kim, M. Kang, L. B. Othmane, L. Lilien, and M. Linderman, “Protection of Identity Information in Cloud Computing without Trusted Third Party,” Proc. 29th IEEE Intl. Symp. on Reliable Distributed Systems (SRDS 10), pp. 368–372, doi: 10.1109/SRDS.2010.57.

[13] P. Angin, B. Bhargava, R. Ranchal, N. Singh, L. B. Othmane, L. Lilien, and M. Linderman, “An Entity-Centric Approach for Privacy and Identity Management in Cloud Computing,” Proc. 29th IEEE Intl. Symp. on Reliable Distributed Systems (SRDS 10), pp. 177–183, doi:10.1109/SRDS.2010.28.

[14] “Definition of Cloud Computing,” NIST, accessed in May 2011. Online at: http://csrc.nist.gov/publications/drafts/800-145/Draft-SP-800-145_cloud-definition.pdf

[15] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz et al., “A View of Cloud Computing,” Communications of the ACM. Association for Computing Machinery, vol. 53, no. 4, 2010. p. 50–58.

[16] A. Belapurkar, A. Chakrabarti, H. Ponnapalli, N. Varadarajan, S. Padmanabhuni, and S. Sundarrajan, Distributed Systems Security: Issues, Processes and Solutions, John Wiley & Sons, 2009.

[17] A. Bhargav-Spantzel, J. Camenisch, T. Gross, and D. Sommer, “User Centricity: A Taxonomy and Open Issues,” Journal of Computer Security - The Second ACM Workshop on Digital Identity Management - DIM 2006, vol. 15, iss. 5, IOS Press Amsterdam, October 2007, pp. 493–527.

[18] “Security Assertion Markup Language (SAML) V2.0 Technical Overview,” OASIS, accessed in Sep. 2011. Online at: http://www.oasis-open.org/committees/download.php/27819/sstc-saml-tech-overview-2.0-cd-02.pdf.

[19] R. PalsonKennedy, and T. V. Gopal, “Assessing the Risks and Opportunities of Cloud Computing – Defining Identity Management Systems and Maturity Models,” Proc. Trendz in Information Sciences & Computing (TISC 10), pp. 138–142, doi: 10.1109/TISC.2010.5714625.

[20] “Dokuwiki Features,” DOKUWIKI, accessed in Sep. 2011. Online at: http://www.dokuwiki.org/dokuwiki.

[21] “JASIG CAS”, JASIG, accessed in Sep. 2011. Online at: http://www.jasig.org/cas.

[22] “OpenLDAP Foundation,” OpenLDAP, accessed in Sep. 2011. Online at: http://www.openldap.org.

 ISBN: 978-1-61208-183-0

A SPIT Avoidance Workflow for SIP-Provider

Nicolas Rüger, Sebastian Hübner, Bettina Schnor
Institute of Computer Science
University of Potsdam
Potsdam, Germany
{rueger,huebners,schnor}@cs.uni-potsdam.de

***Abstract*—Voice-over-IP (VoIP) replaces traditional telephony network infrastructures in growing numbers. Along with this infrastructure change, Spam over Internet Telephony (SPIT) is likely to spread massively, similiar to spam in e-mail infrastructures. Thus, it is necessary to develope appropriate countermeasures. Since the request for a call, usually indicated by the ringing of the phone, might already be a disturbance or annoyance of the called party, traditional content-based preventive and defensive measures, used to avoid e-mail spam, are not applicable anymore. This paper presents a new SPIT Avoidance Workflow for the detection and prevention of unsolicited calls and its implementation. The majority of the presented preventive security measures are applied on side of the provider using the Session Initiation Protocol (SIP). The workflow therefore is implemented for the widely used Kamailio SIP Server that has become a de facto standard in the field of SIP telephony because of its high-performance and robustness. Further, this paper gives an evaluation of the overhead introduced by the different workflow modules. The measurements of the prototype confirm that it is possible to filter and rate call attempts in a larger scale.**

Keywords-Session Initiation Protocol (SIP); Voice over IP (VoIP); Spam over Internet Telephony (SPIT); Security

I. INTRODUCTION

Spam over Internet Telephony (SPIT) might become a serious problem due to the continuous infrastructure change of traditional telephone networks to Voice-over-IP (VoIP) infrastructures. The Session Initiation Protocol (SIP) [14] has prevailed for signalling in VoIP infrastructures. Signalling includes establishment, modification and termination of a media session between the communication endpoints. During the signalling all necessary data for a call, like identities of the communication partners, represented by Uniform Resource Identifier (URI), are exchanged.

So far, a concept is missing, which should combine different approaches for the avoidance of SPIT, and thereby enables a reliable SPIT detection and prevention, without ignoring the requirements of existing communication infrastructures.

In [9] Liske et al. already point out that the known and adapted methods for the prevention of spam are mostly inappropriate for SPIT and therefore fail. SPIT disturbs the called user already by ringing the phone. Therefore, a telephone call needs to be filtered before any voice content is received, while e-mail spam can be analyzed and filtered after receiving the content. Hence, known filter methods are not applicable for VoIP traffic.

When Internet telephony replaces the traditional telephony, we will face VoIP infrastructures, which are managed by different providers. Here, we focus on SIP based infrastructures. Proprietary protocols (e.g., Skype [8]) are not considered. Our approach is aimed at a VoIP infrastructure that offers services comparable to the traditional telephony. Skype uses a restrictive approach with the introduction of buddylists that limit communication to the number of known contacts. This violates the principle of traditional telephony where it is possible to call everybody.

In this paper, we present a mechanism for the detection and avoidance of SPIT that we implemented on side of the provider by extending the widely accepted and used software *Kamailio* [12]. For that purpose, we have analyzed, extended and combined different approaches and results from real VoIP traffic analysis [5] in our overall concept of a SPIT Avoidance Workflow.

This paper is structured as follows: In Section II, we discuss several approaches related to the detection and avoidance of SPIT. The SPIT Avoidance Workflow is presented in Section III. In Section IV, our prototype implementation is described. The results of a detailed investigation of the overhead introduced by the different modules of the SPIT Avoidance Workflow is given in Section V.

II. RELATED WORK

There exist different approaches that relate to the detection and prevention of SPIT including consequences on suspicious call attempts.

A. Authentication

As SPIT detection is necessarily to be done during the call initiation, it has to focus on the available information, e.g., the caller's identity. Especially the appliance of actions with long term effects, e.g., blacklisting a certain user, desires for reliability about the user's identity. Therefore, authentication is an essential requirement for most approaches regarding SPIT prevention as mentioned by Hansen et al. in [7] and Liske et al. [9].

In [13], Mueller and Massoth further describe a basic approach that validates the existence of a calling user during

 ISBN: 978-1-61208-183-0

the initiation of calls. Therefore, at least non-existing faked identites can not be used to initiate malicious calls.

B. Filter Mechanisms

While the known traditional content filter methods, used to detect spam e-mails, are not useful for identifying SPIT, some adapted filter methods will apply for VoIP traffic. Therefore, new filter methods have to be applied during the call initiation for certain attributes, like the caller's identity, as there is no content to analyze. In [7] Hansen et al. introduced a concept for several SPIT filter mechanisms, e.g., whitelists including a web of trust, statistical blacklists or greylists. But, an implementation of the described mechanisms and a performance evaluation was not done.

C. Micro Payment

The use of a payment mechanism initiated by the called party in case of uncertainty about the caller's trustworthiness is introduced in [10], along with the necessary SIP extensions for the micro payment. The use of micro payment seems to be an effective method to prevent SPIT, as the initiater of spit is not willing to pay any amount, due to fact that these calls are initiated en masse.

D. Reputation

In [9], Liske et al. present a way of building a reputation system on base of a micro payment system. The payment requests and their corresponding responses are analyzed to calculate a caller's reputation. Thus SPIT detection by reputation benefits from SPIT prevention by payment. The authors explain that only a single header extension of the underlaying protocol is necessary in order to pass the reputation to the callee. Hence, the approach can be easily integrated in a SIP network that already integrates payment functionality.

In [1], Balasubramaniyan et al. introduce a defense mechanism that is based on the duration of calls between certain users. Therefore, a network of relations is spanned between single users of a VoIP-System. The characteristics of every user, regarding the call duration, is observed during a longer time period. Based on the resulting history of this behaviour analysis, a rating for every user is generated. The rating reflects the reputation of the rated user within the VoIP-System.

E. Behavioral Analysis

In [17], Sengar et al. present two approaches based on the anomaly detection of the distributions of selected call features (inter-arrival time between calls and call duration). The first approach is to detect individual SPIT call and has similarities to some modules presented in this paper. The second approach is designed to detect groups of (potentially collaborating) VoIP spam calls, e.g., a botnet used for sending SPIT. The authors analyze the call behaviour and compare it to theoretical reference pattern to detect unsusual call behaviour. Overhead measurements with prototypes are not given.

F. Consequences on suspicious call attempts

Once a call attempt is suspected by the provider to be a SPIT call, consequences need to follow. According to legal regulations [2], the provider must not drop a call. Therefore, the call is forwarded to the callee. The callee needs to handle the call attempt appropriately. For this purpose several actions may be taken, e.g., reject the call, answer the call, forward the call to a mailbox. In [7] Hansen et al. explain different options like voice menus or announcements of alternative reachability for the handling of an unsolicited call where they emphasize the use of mailboxes.

To sum up, a powerful Anti-SPIT solution has to combine different approaches for an effective detection and prevention mechanism. Just the combination of different approaches for avoiding SPIT will lead to a successful solution as Mueller and Massoth mention in [13]. Especially, the interactions between methods applied by the provider and methods implemented by the client are important.

III. SPIT Avoidance Workflow

We propose a SPIT Avoidance Workflow with a modular structure. The workflow is applied at the provider's side during the call initiation. The modular structure of our overall approach allows easy customizing and experimental combinations of single modules. Furthermore, the solution is easily extendable and can be changed to fit future requirements. The overall architecture of the SPIT Avoidance Workflow is shown in Figure 1. Details for the boxes *Check Filterlits* and *SPIT-Estimation* are shown in Figure 2 and Figure 3.

In addition to the implemented workflow, we will describe an approach for some consequence modules that need to be applied on side of the client in order to evaluate results given by the provider.

A. Check Syntax Module

Certain fields of an incoming message (e.g. caller address, callee address, etc.) are checked, whether or not the fields are filled with valid values, according to the RFC 3261 [14].

If the check is not passed, the response *484 Address Incomplete* or *400 Bad Request* is sent and the call attempt is canceled as it can not be processed. Otherwise, the Filterlist Modules are applied. The list of syntax checks is optionally extendable.

B. Filterlist Modules

Filterlists provide an effective means to categorize incoming messages during the call initiation. Certain fields of an

 ISBN: 978-1-61208-183-0

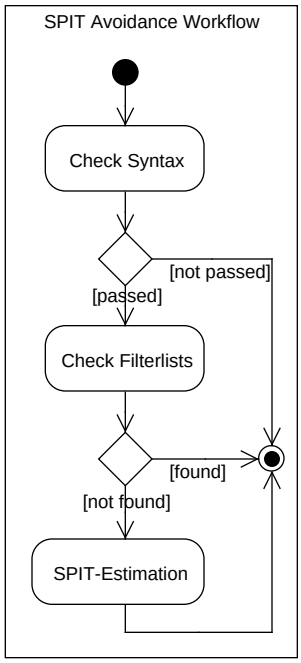


Figure 1. SPIT Avoidance Workflow

incoming message (e.g., the caller's identity in the *From* header) are compared to lists with previously stored values. If a match is present, a corresponding action is executed.

Simple concepts like the black- or whitelisting of users are already well-known from e-mail infrastructures. By using such lists, it is important to be aware about the priority of the interpreted lists, e.g., personal vs. global lists or manually vs. automatically managed lists. Implementing our approach, we set a higher priority to personal and to manually managed lists. In our opinion, the personal settings like the manually and therefore intentionally set entries, should be preferred.

Therefore, we introduce global and callee specific lists for black- and whitelisting. In addition, we propose a global delay list that is to indicate suspicious callers whose calls are delayed during the initiation. This idea is based on the approach that such timeouts will lead callers, that initiated calls en masse, to hang up the phone before the call attempt is actually forwarded to the callee.

The detailed part of the workflow for the Filterlist Modules is shown in Figure 2. A blacklisted user, e.g., receives *403 Forbidden* and the call attempt is canceled as the callee decided so in advance by putting the caller on the blacklist. Once a caller can not be categorized by any filterlist, the Estimation Modules will apply.

C. Estimation Modules

The Estimation Modules realize the main concept of SPIT detection. The modules focus on undesired calls that have

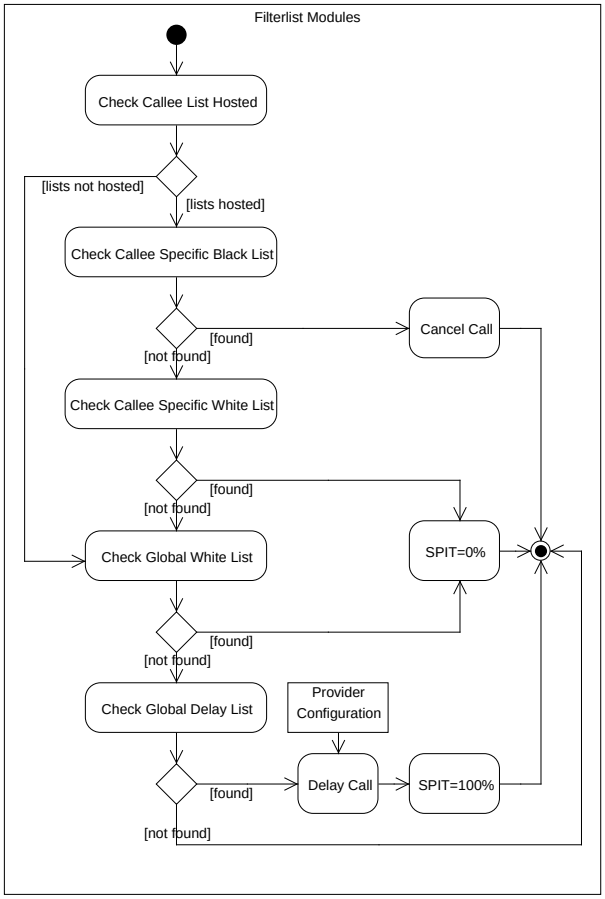


Figure 2. Filterlist Modules

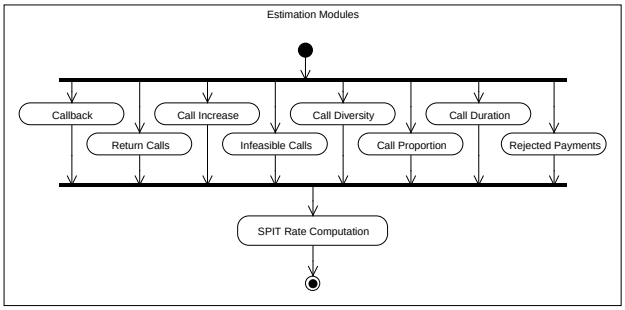


Figure 3. Estimation Modules

 ISBN: 978-1-61208-183-0

been initiated in a larger scale. Both criteria characterize SPIT. Various methods that are managed modular, evaluate the caller's behaviour in the past. The approach is based on the assumption that it is possible to identify users with criminal or malicious intentions because of their specific behaviour related to different criteria. Parameters, such as the number of calls initiated by the user, the duration of calls, the number of different contacts during a certain period of time etc. are analyzed by the provider. Based on this data, the probability of SPIT related to a special criterion is estimated by each method and therefore realized by a single module in each case (see Figure 3).

The detailed mathematical formulas of the *SPIT Estimation Modules* are given in [15].

1) Callback: The module Callback returns a low SPIT probability if a caller was contacted by the callee already before. The idea given by Seifert in [16] is that a user who returns someone's call is a legitimate and desired contact. Therefore, also past call attempts of the callee of the ongoing call are rated.

2) Return Calls: This module returns a low SPIT probability if a user gets called back frequently. This approach is based on the consideration that a user who is never called back by others, is probably to be classified as an unwanted dialog partner.

We now propose that only those interactions between users should be considered, where the relevant past calls have been successful. These calls must have been confirmed by both partners and there must have been a conversation with submitted voice content. Both parties need to confirm at least one call request from the other party. This ensures that both participants really wanted to interact with each other. Such a interaction is rated as a successful *returned call*.

3) Call Increase: The module Call Increase recognizes increasing call numbers of a single user in comparison to himself during a short time period. In [16] Seifert describes a mechanism our module is based on. The module is designed to identify accounts that already exist for a longer time and have now been compromised. The number of outgoing calls of the current caller, in several elapsed intervals, is set into relation to his initiated calls during the last interval. If the number of calls has risen too high, the module returns a high SPIT probability. Balasubramaniyan et al. in [1] and Sengar et al. in [17] assume that a significant divergence from the normal call behaviour is a signal for outgoing SPIT.

4) Infeasible Calls: Based on an approach in [16], our module Infeasible Calls evaluates the number of calls that have not been successful because a recipient with the selected SIP address was nonexistent. Therefore, the module controls which response codes the caller has received on his call attempts in the past. If, too often in the past, the caller of the current call tried to attain identities that did not exist, the module returns a high SPIT probability.

Therefore, we do now propose to extend the message *404 Not Found* by a SIP conforming string extension. In the communication between various providers it should be distinguished between *404 Not Registered* and *404 Not Existing*, while the client still only receives *404 Not Found*.

Thus, it is possible for the providers, to separate infeasible calls from the calls to currently not available users. In addition the approach eliminates the chance of creating a file of existent addresses, as these could possibly be used for sending SPIT in the future.

5) Call Diversity: This module rates the number of calls to different SIP URIs as proposed by Liske et al. in [10]. For that purpose, the module compares the number of made calls to the number of diverse dialed numbers.

We propose that callers who call too many different SIP URIs are suspected to spit.

6) Call Proportion: The module Call Proportion assesses the number of calls made by a user, compared to the average number of calls of all users in an observation period. If in the past the caller of the current call has initiated significantly more calls than other users in average, our module returns a high SPIT probability. This again bases on the assumption made by Balasubramaniyan et al. in [1] and Sengar et al. [17] that a strong divergence from the normal call behaviour is probably SPIT and has been mentioned by Seifert in [16] as well.

7) Call Duration: This module considers the duration of current caller's past calls as Liske et al. mention in [9]. A caller who initiates many very short calls is suspected to be an unwanted caller. Therefore, we compare the number of short calls (e.g., $\leq$ 5sec.) to the overall number of phone calls, the user has initiated.

8) Rejected Payments: The module evaluates the number of calls of the caller, which have not been successfully concluded because the caller did not agree to pay the fees that were caused by his call. In [9], Liske et al. suppose that a caller is suspected to spit if he did reject such requests too often. Thereto, the caller is rated with a high SPIT probability by the module.

Through *SPIT Rate Computation*, the results of the individual modules are combined. The result is represented within a single value for the SPIT probability. In [9], Liske et al. already mention a related reputation system.

The detailed mathematical formulas of the *SPIT Rate Computation* are given in [15].

Finally, the *Consequence Modules*, not shown in the figures due to their client side's deployment, generate an appropriate reaction to the determined *SPIT Rate*. In [7], Hansen et al. already emphasize the importance of mailbox mechanisms to prevent SPIT. Our proposal is to provide certain capabilities to the client, to evaluate the *SPIT Rate* that has been forwarded by the provider.

Thus, our Consequence Modules decide by *time of day*

 ISBN: 978-1-61208-183-0

and by transmitted *SPIT Rate* about how to respond to the call. Possible consequences are: signal an incoming call (e.g., by ringing or vibrating the phone), forward the call to a mailbox, request micro-payment from the caller or reject the call. The Consequence Modules are not part of our prototype implementation as they should be located within the client's VoIP (soft-)phone.

IV. Implementation

The implementation presented in this paper contains the *Syntax Check Module*, the *Filterlist Modules* and the *Estimation Modules* including the *SPIT Rate Computation* as described in Section III.

Our implementation extends the functionality of the *Kamailio SIP Server* [12]. It is complying with the Kamailio project guidelines and compatible with the existing code. Kamailio is de-facto standard in the field of telephony via SIP due to its open source code, modularity, world-wide usage, high-performance, robustness and its active developer community.

The Estimation Modules access the information in the database, previously collected by a *Call Trace Update Functionality*. Therefore, the Call Trace Update Functionality logs all relevant data (e.g., start and end time, caller, callee, etc.) of every call. Based on this information, the behaviour of the caller gets evaluated and expressed as module-specific SPIT probability.

As a final result, the *SPIT Rate* is calculated using the SPIT probabilties from all SPIT Estimation Modules. Different strategies can be applied to include the single SPIT probabilities in the computation, as described by Seifert in [16]. It is possible to incorporate the ratings of all modules in the same degree in the reported value as we did for our prototype implementation. For future use, we propose a weighted accumulation to underline the greater importance of certain modules. As the number of modules that contribute to the SPIT rating may vary, the calculation should be implemented dynamically. If, for example not enough information about the caller is available, not all input parameters for all modules are applicable.

We assume that the relevant data for the single modules is available when it needs to be analyzed. In case no relevant data is found during the evaluation of the available data, the corresponding module is not able to compute a result and can not be involved in the *SPIT Rate Computation*. Our implementation considers this dynamic computation.

The final result of the computation is added as an additional header field *Spit Rate* to the initial *Invite* message. Thus it is transmitted to the client as a compressed value. This extension is conformant with the RFC 3261 [14].

For our implementation, we have chosen a MySQL database [3] as backend because the link is well supported by Kamailio and the database is Open Source like Kamailio itself. The entity relationship model is shown in Figure 4.

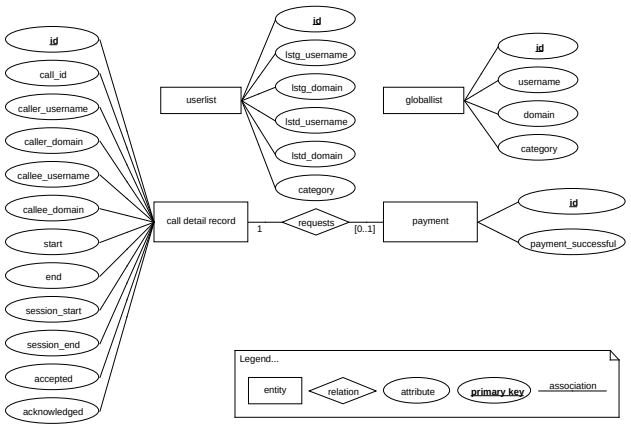


Figure 4. Database Entity Relationship Model

Within the shown database structure, all relevant data needed by the modules described in Section III is stored.

We have implemented our SPIT Avoidance Workflow by extending the Kamailio routing logic in *kamailio.cfg*. We have chosen the interface SQLOps [11] provided by Kamailio. It allows interaction with the database backend directly from the routing logic.

Analysis of different available interfaces arrived at the conclusion that SQLOps shows the best performance. It holds and reuses one database connection that is established on Kamailio startup. The *Perl* interface [4] seemed to be an alternative because it allows the execution of arbitrary Perl scripts, but it etablishes a new database connection for each running script. Therefore, it is not scalable for growing numbers of calls per second.

V. Measurements

The measurements were made to compare the performance of the Kamailio SIP server with and without the extensions for SPIT prevention. Thus, it is possible to evaluate the overhead of our solution and its impact on the number of processed calls per second. Therefore, telephony traffic was simulated, while the SPIT prevention modules have been active or inactive.

A. Testbed and Scenarios

We used three nodes (each with 2 x AMD Opteron 244 CPU, 1.8GHz, 4GB RAM, Gigabit Ethernet Interconnection) to setup one SIP proxy (Kamailio [12], v3.1.1), two user agents (SIPp [6], v3.1) that generated (resp. processed) a various number of SIP calls and a MySQL [3] database (v.14.12 distrib. 5.0.51a) to store the collected connection data. Kamailio has been configured to use 1024MB of memory, to create eight processes and its log level was set to zero.

We measured the default configuration of the proxy in comparison to the behaviour of the proxy with one single

 ISBN: 978-1-61208-183-0

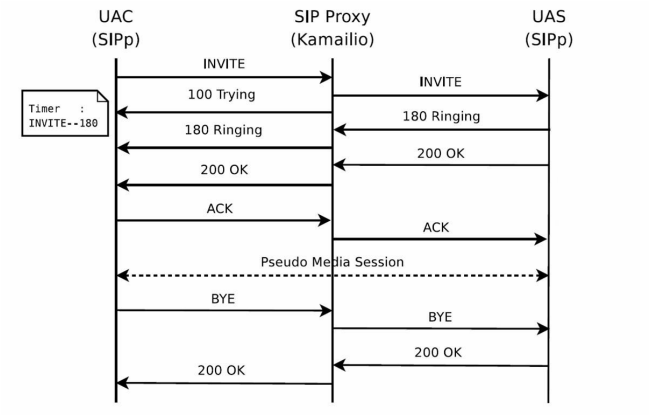


Figure 5. Measurement of Round Trip Time

module switched on in each case. Thereby, we wanted to figure out which module causes the most overhead.

Each call generates a time value. The round-trip time (RTT) represents the delay of a user agent server's response, including Kamailio action (processing the activated modules). We focused on this part of the session initiation as it measures the most expensive part. It represents the time interval from the start of the session initiation to the first ringing (see Figure 5).

For each single module, we measured the RTT in different series. Each series used a constant call frequency (number of calls per second) for 60 seconds. We have chosen different call frequencies for each series in steps of 1 from the range of 1 to 10 calls per second (cps), steps of 10 from the range of 10 to 100cps and steps of 100 from the range of 100 to 1000cps. That sums up to 28 series of different call frequencies per module lasting 60 seconds each (see Table I). The proxy and the user agents have been restarted for each measurement.

RANGE (CPS)	STEPS	NO. OF SERIES
1 - 10	1	10
20 - 100	10	9
200 - 1000	100	9
		$\sum$ 28 a 60 sec.

Table I. Measurements for each module

B. Results

For all modules and each call frequency we calculated the corresponding median values for the RTT.

1) Syntax Check and Filterlist Modules: The module Global List includes the global delay- and global whitelist whereas the module User List includes the callee's personal black and white list. The lists were filled with 50 sample entries each. As shown in Figure 6, the Kamailio SIP server only shows slightly higher response times of maximum (~0.25ms) with the Syntax Check or the Filterlist Modules switched on, than without any modules for SPIT detection.

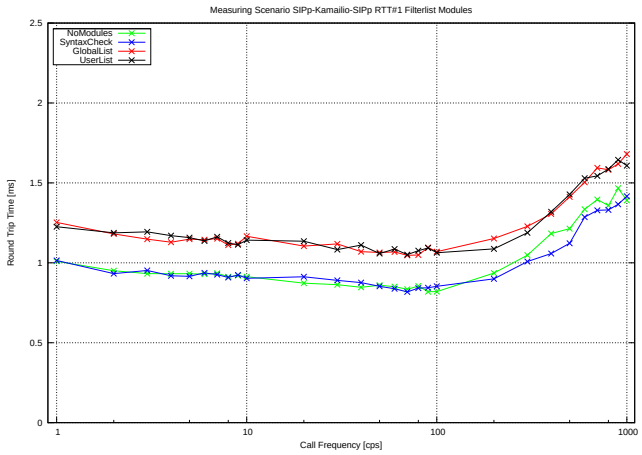


Figure 6. Median RTT with activated Filterlist Modules

Slight deviation of 0.1ms is due to normal measurement fuzziness using the described measurement environment. The reached accuracy of 0.1ms is sufficient to show the impact of the modules.

2) Estimation Modules: Figure 7 shows the RTT for the Estimation Modules, with exception of module Call Increase as it does not fit within the chosen scale, in comparison to the proxy's plain behaviour. The modules Callback, Infeasible Calls and Rejected Payment show just a slightly higher response time of ~0.25 to 0.5ms.

From a call frequency of 100cps the response time for the modules Call Diversity, Return Calls, Call Duration and Call Proportion reaches unacceptable high values, whereas the module Call Increase shows this behaviour already from a call frequency of 10cps.

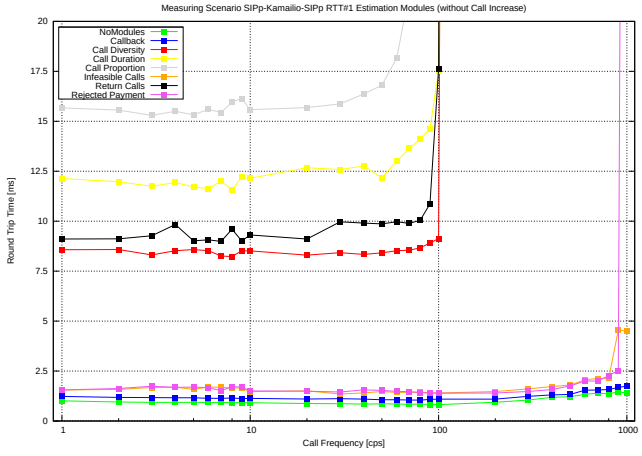


Figure 7. Median RTT with activated Estimation Modules

The measured delays for some modules are resulting from the queries to the database, initiated by the Kamailio routing

 ISBN: 978-1-61208-183-0

logic during the processing of messages. Further tests have shown, that the delays do not occur if the requests on the database are not performed, even though the remaining program code (e.g., decisions, branches, variable assignments, adding the SPIT-header fields etc.) of the SPIT Avoidance Modules kept unchanged. Therefore, the delays result from the database queries only. This conclusion is backed by the measurement results, which show that especially the modules with more frequent requests to the database scale poorly.

3) Call Trace Update Functionality: The Call Trace Update Functionality does not scale well as it accesses the database very often during the call initiation to log all necessary data. Figure 8 shows that the functionality causes high response times already from a call frequency of 10cps.

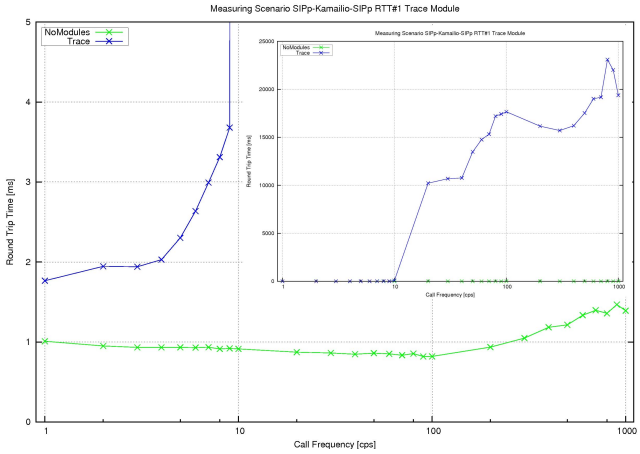


Figure 8. Median RTT with activated Call Trace Update

As the Call Trace Update Functionality provides the basis for the Estimation Modules, some effort needs to be made to improve the performance of this basic functionality.

C. Result Summary and Suggestions for Improvement

In summary, six modules scale satisfying and five modules do not. In addition, the trace mechanism for data collection does not scale very well.

Since we have identified the database queries as the reason for the measured delay, the following actions might lead to better response times and thus, to a higher performance of the SPIT Avoidance Modules that did not scale and finally to a better Kamailio scalability. The selected MySQL database was used in the standard configuration. We have made no improvements for our measurements. Therefore, indexing of database entries, detailed analysis and optimization of the individual database queries, the usage of numeric data types instead of strings, increasing the cache size of the database and the number of possible child processes of Kamailio might lead to a much better performance.

Further, the usage of more powerful hardware as well as separating the database system will improve the overall system performance in addition.

VI. Conclusion and Future Work

We presented a SPIT Avoidance Workflow consisting of filterlists, call detail record analysis, and a rating system based on this analysis. Between modules known from prior works, we also proposed new modules like a global delay list.

We have shown that the workflow can be integrated within the Kamailio routing logic to detect and prevent SPIT. For some of our modules, the measurements with the prototype confirm that it is possible to filter and rate call attempts in a larger scale without increasing the response time or the scalability of the Kamailio SIP Server. Already six from eleven SPIT Avoidance Modules worked very well without any database optimization or any special hardware.

Future work has to be done to optimize the SQL queries and to improve the underlying database structure for a better performance.

References

[1] Vijay Balasubramaniyan, Mustaque Ahamad, and Haesun Park. CallRank: Combating SPIT Using Call Duration, Social Networks and Global Reputation. In *CEAS'07*, 2007.

[2] Translation of the German Criminal Code provided by Prof. Dr. Michael Bohlander Bundesministerium der Justiz. German Criminal Code - Strafgesetzbuch (StGB). http://www.gesetze-im-internet.de/englisch_stgb/german_criminal_code.pdf, 2010.

[3] Oracle Corp. MySQL v.14.12 distrib. 5.0.51a. http://downloads.mysql.com/archives.php?p=mysql-5.0&v=5.0.51a, 2008.

[4] Bastian Friedrich. Kamailio (OpenSER) Perl Module. http://www.kamailio.org/docs/modules/3.1.x/modules_k/perl.html, 2007.

[5] Stefan Gasterstädt and Bettina Schnor. What VoIP-CDR can tell us (not)? Technical Report ISSN 0946-7580, TR-2010-2, Potsdam University, Germany, May 2010.

[6] Richard Gayraud, Olivier Jacques, et al. SIPp: An Open Source Performance Testing Tool for SIP [v3.1]. http://sipp.sourceforge.net, March 17^{th}, 2009.

[7] Markus Hansen, Marit Hansen, Jan Moeller, Thomas Rohwer, Carsten Tolkmit, and Henning Waack. Developing a Legally Compliant Reachability Management System as a Countermeasure against SPIT. In *Third Annual VoIP Security Workshop*, Berlin, Germany, June 2006.

[8] Skype Limited. Skype Technologies (Microsoft). http://www.skype.com, 2011.

 ISBN: 978-1-61208-183-0

[9] Stefan Liske, Klaus Rebensburg, and Bettina Schnor. Implicit Reputation in a Payment Integrated SIP Network. In *Proceedings of the 14th Annual Workshop of HP Software University Association (HP-SUA)*, pages 161–170, Munich, Germany, July 2007.

[10] Stefan Liske, Klaus Rebensburg, and Bettina Schnor. SPIT-Erkennung, -Bekanntgabe und -Abwehr in SIP-Netzwerken. In U. Ultes-Nitsche, editor, *Proceedings of KiVS – NetSec 2007, Workshop „Secure Network Configuration“*, pages 33–38, February 2007.

[11] Daniel-Constantin Mierla. Kamailio (OpenSER) SQLOps Module. http://www.kamailio.org/docs/modules/3.1.x/modules_k/sqlops.html, 2008.

[12] Ramona-Elena Modroiu, Bogdan Andrei Iancu, Daniel-Constantin Mierla, et al. Kamailio (OpenSER) [v3.1.1]. http://www.kamailio.org/, December 02^{th}, 2010.

[13] Juergen Mueller and Michael Massoth. Defense Against Direct Spam Over Internet Telephony by Caller Pre-Validation. In *2010 Sixth Advanced International Conference on Telecommunications (AICT 2010), Barcelona, Spain, 9-15 May 2010*, pages 172–177, Washington, DC, USA, 2010. IEEE Computer Society. Best Paper Award.

[14] Jonathan Rosenberg, Henning Schulzrinne, Gonzalo Camarillo, Alan Johnston, Jon Peterson, Robert Sparks, Mark Handley, and Eve Schooler. SIP: Session Initiation Protocol. RFC 3261 (Proposed Standard), June 2002. Updated by RFCs 3265, 3853, 4320, 4916, 5393, 5621, 5626, 5630, 5922.

[15] Nicolas Rueger. Konzeption und Implementierung providerseitiger Sicherheitsmechanismen zur Erkennung und Abwehr von SPAM over Internet Telephony (SPIT) in VoIP-Netzwerken. Diploma thesis, University of Potsdam, Institute for Computer Science, 2011.

[16] Juergen Maximilian Seifert. Klassifizierung und Implementierung von SPIT (Spam over Internet Telephony) Abwehrmassnahmen in SIP-Netzen. Diploma thesis, University of Potsdam, Institute for Computer Science, 2006.

[17] Hemant Sengar, Xinyuan Wang, and Art Nichols. Thwarting Spam over Internet Telephony (SPIT) Attacks on VoIP Networks. In *Proceedings of the Nineteenth International Workshop on Quality of Service*, IWQoS '11, pages 25:1–25:3, Piscataway, NJ, USA, 2011. IEEE Press.

 ISBN: 978-1-61208-183-0

Using PGP Signatures for Securing SIP Infrastructures

Sebastian Hübner, Nicolas Rüger, Bettina Schnor
Institute of Computer Science
University of Potsdam
Potsdam, Germany
{huebners, rueger, schnor}@cs.uni-potsdam.de

***Abstract*—Because of increasing bandwidth and decreasing costs for the provider, Voice-over-IP is an alternative to the Public Switched Telephone Network for many users. But with the propagation of Voice-over-IP new harassments and threats occur. Assuring the identity of communication partners is significant in this context. Without the authentication of communication partners, the infrastructure is vulnerable to attacks like URI-spoofing, call and registration hijacking. Authenticity is necessary for detecting and avoiding Spam over Internet Telephony (SPIT). Only if the identity of a caller can be verified, a source of SPIT can be exposed and appropriate countermeasures can be taken. In this paper, we present a decentralized approach for authentication in the Session Initiation Protocol (SIP) using PGP signatures. Due to already existing data structures this mechanism can be easily integrated in the SIP without the need of new SIP extensions. Measurements show that our approach results into tolerable overhead.**

***Keywords*-Voice-over-IP (VoIP); Authentication; Pretty Good Privacy (PGP); Session Initiation Protocol (SIP); Signature**

I. Introduction

The Session Initiation Protocol (SIP) [14] is one of the most commonly used protocols in Voice-over-IP communications. SIP handles the signaling, which includes establishment, modification and termination of a media session between two or more communication endpoints. During the signaling the negotiation of call properties is done Further, necessary data for a call, e. g. identities of the communication partners is exchanged. In combination with SIP the Real-time Transport Protocol (RTP) [15] is usually utilized for transferring the media data. Figure 1 shows a common network topology known as the SIP Trapezoid. It consists of the following components: registrar, proxy, User Agent Client (UAC) and User Agent Server (UAS). The UAC sends a request to an UAS. A UAS receives a request and answers with one or more appropriate responses. In SIP, different network entities may be in the role of a UAC or UAS. For example, during the call invitation the caller acts as UAC and the callee as UAS.

Our goal is a secure Voice-over-IP (VoIP) architecture, which guarantees a quality of service comparable to the Public Switched Telephone Network (PSTN) service. It offers the possibility to integrate Pretty Good Privacy (PGP) signatures in the SIP context similar to their more common use in e-mails.

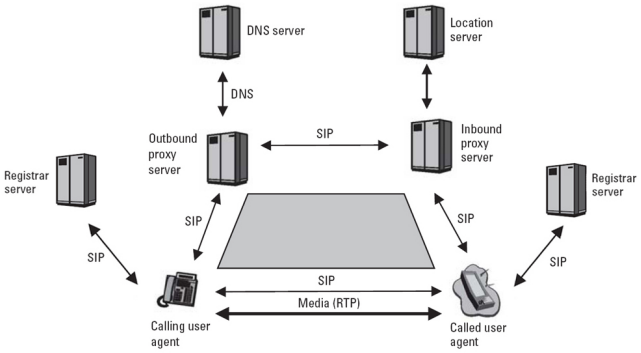


Figure 1. SIP Infrastructure [8]

This paper is structured as follows: In Section II, we define the requirements for a secure SIP infrastructure. In Section III, we discuss related work. In Section IV, we present our approach and its integration into SIP. A security analysis of our concept is given in Section V. In Section VI, we describe the implementation of our protoype and present measurement results in Section VII.

II. Security Requirements in SIP

This section describes security requirements we consider important in a SIP infrastructure:

A. End-to-End Authentication

The assurance about the identities of the involved communication partners is mainly in the interests of the endpoints. They exchange potentially private or personal data during their communication and want to be sure about the recipient's identity. Thus, the decision whether certain information is given or not depends on the authenticity of the communication partners. An authentication mechanism has to realize a direct end-to-end authentication between the endpoints.

B. Mutual Authentication

Caller and callee have to authenticate themselves against each other. Both endpoints of the communication want to

 ISBN: 978-1-61208-183-0

be sure about the other's identity. In SIP, messages are exchanged among different network entities, not only between the endpoints. For example endpoints also communicate with registrar or proxy servers. The SIP standard describes various threats and attacks caused by missing authentication of server components, e.g., Call Hijacking, Registration Hijacking or Impersonation [14] . Their attempt is to make endpoints unreachable to others (DoS). These attacks do not directly affect the authenticity of the communication partners but they have impact on the call's quality. To ensure the quality of calls it is necessary to apply mutual authentication between all UAs in a SIP network. Consider that in SIP different entities are able to act as a UA, not only the endpoints. Any logical entity that creates and sends a request is a UAC, any logical entity that creates and sends responses to a request is a UAS. Thus, requests and responses should be exchanged between mutual authenticated network components only.

C. Authentication During Signaling

Authentication has to be realized during signaling. Once a media stream is established confidential information can be transmitted. Therefore, before a call is accepted by the callee or before the phone even rings, the authenticity of the caller has to be verified. Moreover, it is important to secure all relevant signaling messages during a SIP session. Especially, the termination of a call is a crucial point. In SIP, there are several signaling messages, i. e. *BYE* or *CANCEL* requests to terminate a session. Only authenticated participants should be able to send these requests to avoid an unwanted call termination. To prevent Man-In-The-Middle or reply attacks, the signaling messages have to be protected by a signature field.

D. Technical Requirements

The use of security mechanisms in a given SIP infrastructure has to be practical. The routing of messages may not be hindered or impeded. Additionally, the functionalities of the different SIP components may not be affected.

III. Related Work

There are different approaches that relate to requirements presented in the previous section.

A. SIP Digest Authentication

SIP Digest [14] is based on a challenge/response principle. For its appliance a shared secret between UAS and UAC is necessary. Usually, this is the case between UAs and the registrar or proxy server. The User Agent (UA) authenticates itself against the server by using its associated credentials (user name, password).

In reality, there is no such relationship between two endpoints. The called party cannot hold personal data for any possible caller. But, this would be necessary to verify the origin of an incoming call. Moreover, SIP Digest authentication only allows the authentication of the caller. The initiator of a conversation is not able to authenticate the callee. Therefore, this method is not suitable for mutual authentication. Guillet et al. [5] extend SIP Digest authentication by mutual authentication, but still a shared secret is needed.

Strand and Leister [17] point out some weaknesses and drawbacks of SIP Digest Authentication. It is not suitable for end-to-end or cross-domain authentication. Moreover it is vulnerable to different attacks. The authors focus on a register attack, which is caused by modifying the *Contact* header of SIP message during the registration phase. They suggest extending SIP Digest Authentication by including the contact header value in the digest computation to counter that specific register attack.

B. TLS

RFC 3261 [14] defines the utilization of Transport Layer Security Protocol (TLS) [3] within a SIP network. By using client-side and servers-side certificates a mutual authentication can be achieved. However, TLS realizes a hop-by-hop security. Only the connection to the next node is secured and authenticated. To realize a secure connection between the endpoints via TLS a chain of trust has to be established between all hops on the path from the caller to the callee. The endpoints trust in each other's identities because of an existing trusted relation between them. But there is no direct end-to-end authentication.

SIP provides the SIPS URI Scheme to initiate a hop-by-hop TLS connection. But the last hop between the inbound proxy and the callee is not necessarily included in this trust chain. According to RFC 3261 the security mechanisms on that last hop depends on the policy of the domain.

In spite of this, there exist different approaches to secure SIP infrastructures on the base of TLS. Jiang [7] uses a hop-by-hop TLS connection to exchange a session-key to encrypt the following media streams and a so-called setup-key. The setup-key is valid only for the next call and used for a direct end-to-end authentication. But the concept is based on the trust in the hop-by-hop TLS connection.

In [10], Kong et al. present a solution for securing the localization of communication partners. This is achieved by providing integrity for the contact header of a SIP message by using signatures. For that purpose each endpoint generates a public and a private key. During call initiation the caller creates a signature for the contact header using his private key. Now, the callee is able to verify the identity of the caller and sends a signed *200 OK* response, if authentication was successful. After receiving the response the caller verifies the callee's identity as well. The endpoints exchange their public keys using a hop-by-hop TLS connection outbound and inbound proxy. Again, the last hop is not considered. While the authors focus on the localization of

 ISBN: 978-1-61208-183-0

communication partners, the integrity of other SIP messages and header fields is not verified.

C. S/MIME

S/MIME [13] allows end-to-end encryption. Entire SIP messages are encapsulated within a MIME body. They are signed with the sender's private key and encrypted with the public key of the intended recipient. To allow the routing of encrypted messages their header is duplicated. So, the recipient has to deal with "inner" and "outer" message headers (SIP Tunneling). The "outer" header is used to verify the authenticity of the encapsulated information [14]. But, there are parts of the header, e.g., the via header field, which is legitimately modified during routing. Thus, end-to-end authentication can only be realized for unchangeable parts of the header.

D. PGP

RFC 2543 [6], the previous SIP standard, describes the usage of PGP-based encryption to provide authenticity of SIP messages. RFC 2543 introduces the basic structures and headers for the appliance of PGP in the SIP context. A complete description of security aspects and mechanisms, which are realized by PGP, is not given. This may be a reason why the usage of PGP is described as "incompletely specified". The current RFC 3261 deprecates PGP in favor of S/MIME.

IV. Our Approach: PGP Signatures

Next, we present our approach using PGP signatures in SIP. Our approach fulfills the requirements from Section II. It can be used within SIP infrastructures conform to RFC 3261.

A. Motivation

Although PGP is deprecated in the current RFC 3261, we favor it for the following reasons:
In TLS and S/MIME hierarchical PKIs depending on X.509 certificates are used. Among others, Ellison and Schneier discuss the risks of this approach [4]. They argue that vague Certificate Authority practices to issue certificates cause an imprecise meaning of the word "trust". Furthermore, current events show that a valid certificate does not necessarily mean the owner is trustworthy [1].

Unlike this hierarchical approach PGP, utilizes a "Web of Trust" in which trust is considered private information (cf. IV-C). In [19], Ulrich et al. point out that this trust concept helps to prevent the propagation of faked certificates.

Since PGP is in widespread use for encrypting and signing e-mails, we propose to use the already existing PGP-Keys and trust relationships to secure VoIP communications as well.

B. Concept

In contrast to SIP digest, we do not use message authentication codes, but signatures. Every entity of the SIP infrastructure needs a pair of PGP keys for signing and verifying messages.

After receiving a request the UAS sends an appropriate response to challenge the identity of the UAC. The UAC repeats the initial request and appends a signature for the message body, a subset of the header fields and certain elements from received challenge. The UAS verifies the signature and sends a signed response. In result, the UAC can verify the server's identity. Figures 2 and 3 show the computation of signatures for requests and responses.

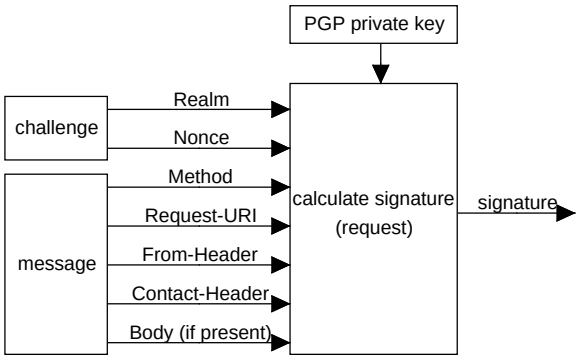


Figure 2. PGP Signature - Request

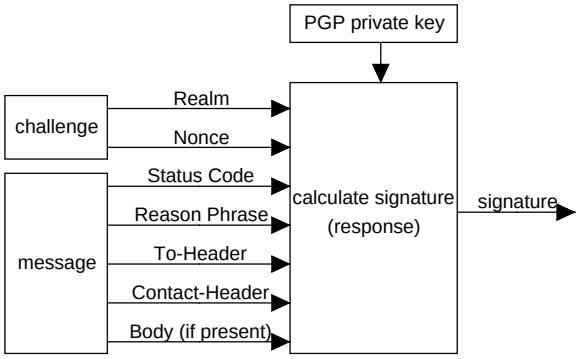


Figure 3. PGP Signature - Response

The verification of a signature is the same process for both UAC and UAS. The signature of a message is calculated by using the sender's private key. So the recipient needs the corresponding public key. Before checking the signature it is crucial to verify the key's associated identity (see Section IV-C). If the key's identity could be verified, the recipient checks whether the signature of the message is correct or

 ISBN: 978-1-61208-183-0

not. Calls should only be established if the *INVITE* request and the referring *200 OK* response are correctly signed and the identity of the corresponding keys is verifiable by the recipient.

After the establishment of a call all SIP messages, which affect the state of the session, have to be signed (see Figure 4).

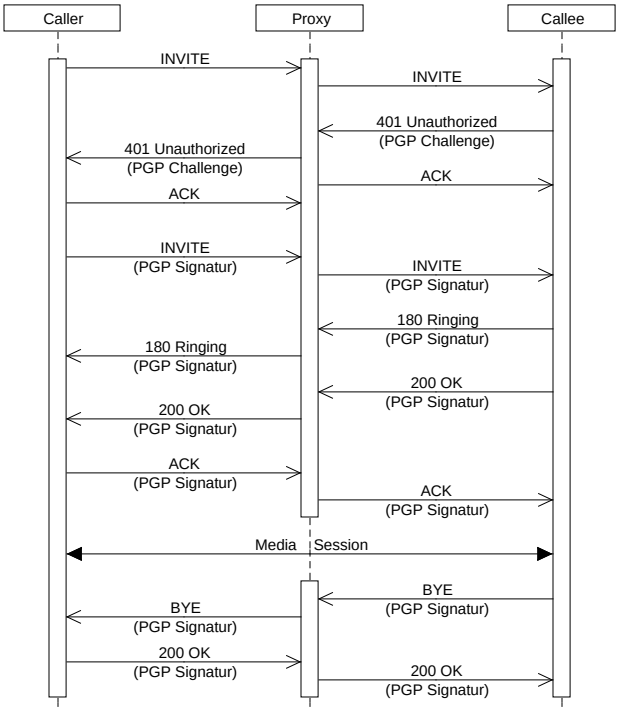


Figure 4. PGP Authentication - Message Sequence

Since a UAC can be challenged by different components of a SIP infrastructure, for example proxy or callee, a request may contain more than one signature. This procedure has to be applied between all components of a SIP infrastructure, which act as UAC and UAS to fulfill the requirements from Section II. In the following, the focus is primarily on the endpoints.

C. Evaluation of the concept

The main feature of this concept is a direct end-to-end authentication. Signed messages are generated and verified by the endpoints. The authenticity of the participants does not depend on the intermediary network entities (cf. TLS). So the last decision on the call establishment is up to the endpoints. This also means that keys with every communication partner have to be exchanged.

The functionality of PGP signatures is based on the binding of the keys and their associated identities. The OpenPGP standard [2] defines two concepts for establishing trust: By signing another key a user claims to be sure of the key-entity binding ("Public-Key Trustworthiness" [19]). By adding a certain trust level it can be determined how much another user is trusted to sign other keys carefully ("Introducer Trustworthiness" [19]). Unlike signatures, the trust level can only be set manually in the local keyring, it is not exported. A key is valid if it is signed directly by the recipient or the validity can be derived from a transitive trust chain ("Web of Trust"). For that the following conditions have to be met: Each key has to be signed by the preceding node and for each key the trust value must be set. Therefore, this chain can only be generated within the local keyring of a user.

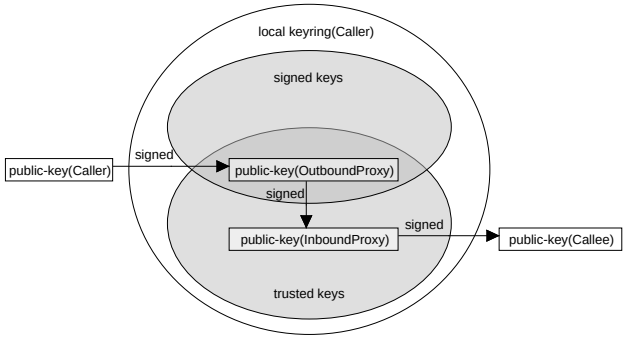


Figure 5. PGP Trustchain in SIP Infrastructure

Figure 5 shows a chain of trusted keys in a SIP infrastructure (see Figure 1) from the caller's view. To verify the callee's key the public keys of the proxies have to be trusted. For establishing a trust chain the key of the Outbound Proxy has to be signed by the caller and the key of the Inbound Proxy has to be signed by the Outbound Proxy. Note that it is also possible to find another path to the callee's key, for example with keys from existing social relationships of the caller. The "Web of Trust" is practical especially for closed groups with signed keys or those users that frequently sign other keys and get signed by them [19].

In case the key's identity cannot be assured by the described concept the recipient will not be able to verify the sender's identity. Consider that it is still possible to check the message's signature with an unknown key. But even though the signature of a SIP message is correct the corresponding key may be falsified. So the endpoint has to decide whether the call should be established or rejected without any further knowledge of the sender's identity. It is also important to note that a signature only assures the integrity and authenticity of elements, which are included in its calculation. So for our concept it is crucial to choose the parts of a message, which are necessary to verify the sender's identity.

 ISBN: 978-1-61208-183-0

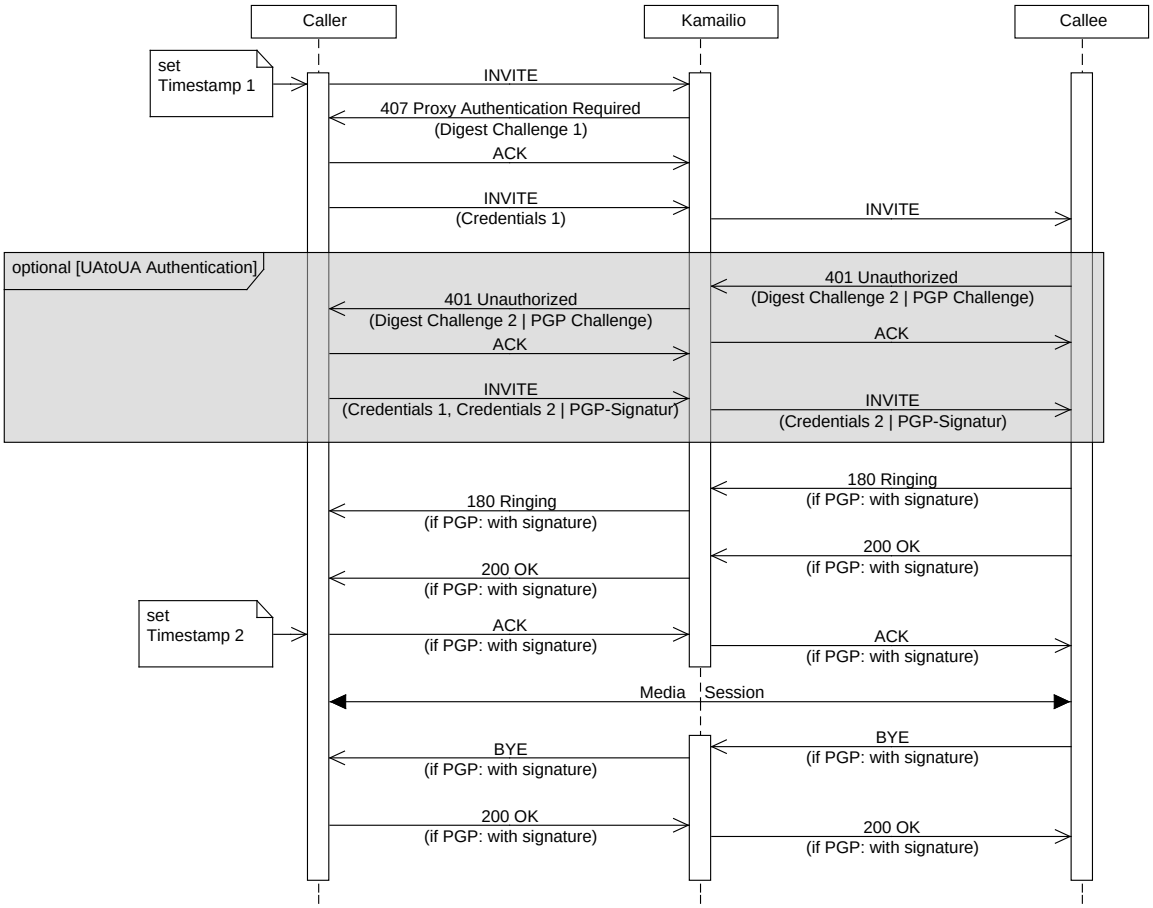


Figure 6. Measurement Scenarios

D. Integration in SIP

Appropriate SIP messages and header fields are necessary for the transmission of the authentication challenge and the corresponding signatures. SIP already defines the responses *401 Unauthorized* (sent by UAS) and *407 Proxy Authentication Required* (sent by proxy) to transport the authentication challenge for SIP Digest. Therefore, the messages contain a *WWW-Authenticate* header and a *Proxy-Authenticate* header. These messages and header fields can also be used to transport a PGP Authentication challenge. After receiving the challenge the UAC has to extend the initial request by a signature and has to send it again. RFC 3261 defines the request-header fields *Authorization* and *Proxy-Authorization* to transport the response of a received digest challenge. Again, these elements can also be used to transport a PGP signature. Hence, no special SIP extension is necessary. Moreover, the PGP Authentication does not affect the existing authentication mechanism in SIP. A message header, which already contains the information for a SIP Digest Authentication (e.g., between endpoint and proxy) can also carry the PGP Authentication (e.g., between the endpoints).

V. Security Analysis

Signatures are applied in communication infrastructures to provide authenticity, data integrity, and non-repudiation. With the awareness of the PGP trust concept and its weaknesses (cf. IV-C), our concept provides countermeasures against the following threats and harassments:

URI-Spoofing: By the lack of authenticity of signaling messages, it is possible to falsify the identity of communication partners to obtain sensitive information or to use personalized services. In our concept this is avoided by signing the proper header fields (*To* header, *From* header) and using the key-entity binding in PGP.

Call Hijacking: Without authentic localization information a call can be redirected toward an attacker's device. For example, the attacker can act as Man-In-The-Middle. As a countermeasure, our concept provides integrity for the identities and the localization information as well.

 ISBN: 978-1-61208-183-0

Registration Hijacking: Similar to Call Hijacking, an unauthenticated registration information allows redirection of calls as well. Moreover many Denial of Service attacks are caused by unauthenticated or unauthorized REGISTER requests [14]. Hence, the affected endpoints are not available anymore. To counter this attack our concept has to be applied between all components of a SIP infrastructure which act as UA (cf. IV-B).

Impersonation: Without a proper authentication an attacker can impersonate every component of a network. Similar to URI-Spoofing and Registration Hijacking this is avoided by providing authenticity of the communication partners and applying our concept between the different SIP entities.

Terminating Sessions: Within established sessions or during their establishment, requests can be sent which take effect on the dialog state. An attacker can inject falsified BYE or CANCEL requests and terminate the call or its establishment. For that reason it is crucial to consider the entire SIP session, not only the establishment of a call, as presented in our concept.

VI. Implementation

The presented concept was implemented at the endpoint side to analyze its behavior in practice and getting aware of the involved overhead. For the underlying SIP stack, the PJSIP - Open Source SIP Stack [12] was used. PJSIP is a complete SIP stack written in C. The PGP functionality is provided by GnuPG [18], which is an implementation of OpenPGP. To get access to GnuPG in PJSIP the library GnuPG Made Easy (GPGME) [9] was used. The following functionalities of the endpoints are implemented:

Callee: generation of the PGP challenge after receiving the initial *INVITE*, verification of the repeated *INVITE* and (if verification was successful) calculation of the signature and sending signed *200 OK* response

Caller: processing of the received PGP challenge, calculation of the signature and repeating the *INVITE*, verification of the signature in the received *200 OK* response and (if verification was successful) sending signed *ACK*

After the session initiation, all SIP messages were signed by the endpoints.

VII. Measurements

Since we wanted to investigate the overhead introduced by our approach, we compared the authentication with PGP to SIP Digest, and a call setup without any authentication.

The mechanisms were compared regarding their performance, not their security aspects. In our measurements, only the call setup between caller and callee was considered. The measured parameters were duration, memory consumption and CPU utilization.

A. Testbed and Scenarios

We used three nodes (each with Intel Core Duo E7500 CPU (2,93GHz), 2 x 2048MB Dual Channel DDR2 RAM, Gigbit Ethernet Interconnection) to setup a SIP Proxy (Kamailio v3.1.1) [11] and two SIP endpoints (PJSIP v1.8.5 with implemented PGP functionality). The underlying operating system was Debian 5.0.5 (Lenny) on each node.

We measured three scenarios: a) no authentication b) SIP Digest Authentication and c) PGP Authentication between the endpoints. The proxy was used with activated Digest Authentication in each scenario.

The message sequence is shown in Figure 6. For the measurement of the duration and the CPU utilization we only considered the call set up, which is labeled by two timestamps. The first timestamp is set when the initial *INVITE* is sent by the caller. When the caller sends the *ACK* after receiving the *200 OK* of the callee the call is successfully set up and the second timestamp is set. The memory consumption was measured for the whole SIP session. All measurements were done on the caller's side.

B. Results

For each scenario we performed 51 measurements and calculated the median for call set up duration and CPU utilization. The measurement of the memory consumption was done once by using valgrind [16]. The results are shown in Figures 7–9.

The overhead for authentication with PGP, i. e., the compute and memory demands at the caller's site, only slightly increase compared to no authentication or SIP Digest. The caused overhead is still acceptable. The CPU utilization rises by 10ms (see Figure 8), the need of memory increases by 2MB (see Figure 9). PGP Authentication increases the duration of the call setup by about 40ms compared to SIP Digest (see Figure 7). However, this delay is tolerable. The quality of the telephone service is not particularly affected.

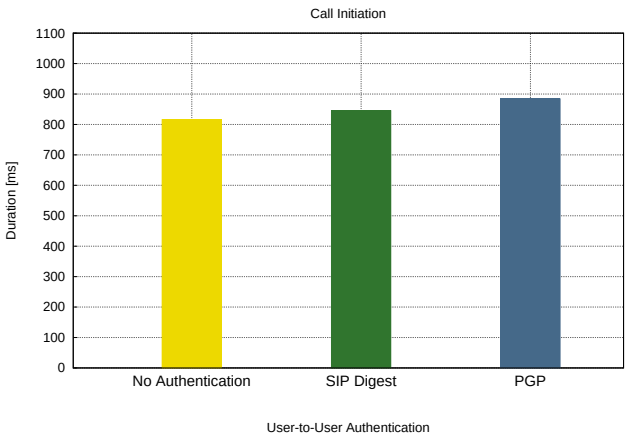


Figure 7. Median of Duration of Call Initiation

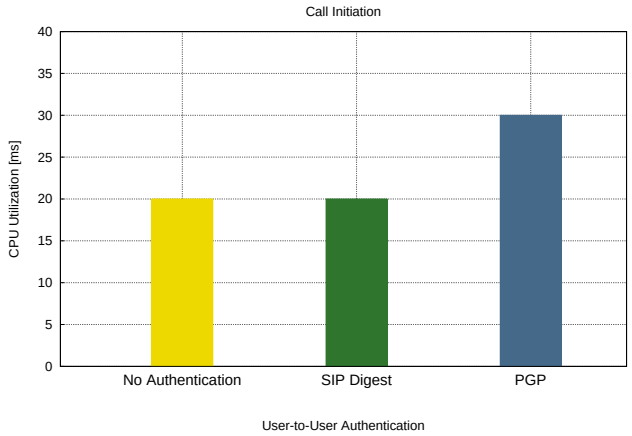


Figure 8. Median of CPU Utilization

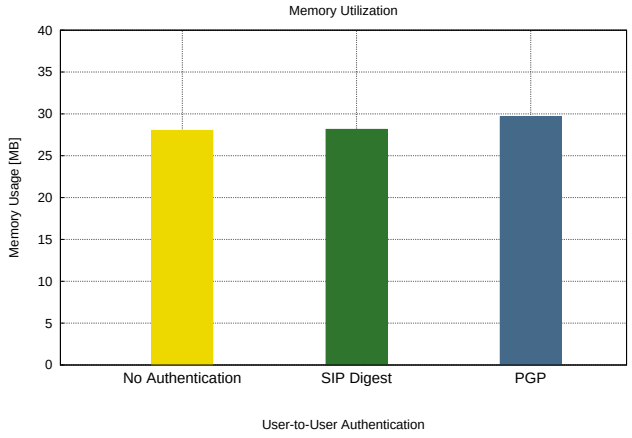


Figure 9. Memory Consumption

VIII. Conclusion and Future Work

We have shown how PGP signature can be used to secure SIP messages. We argued that an end-to-end and mutual authentication is necessary.

The measurements with our prototype have shown that the overhead is tolerable at the caller's side.

The PGP Authentication mechanism can be easily integrated in SIP infrastructures since necessary messages and header fields are already defined in RFC 3261.

The next step is to implement the mechanism also on the proxy side. The aim is to evaluate whether it is possible to use PGP Authentication with tolerable overhead also on these components of a SIP infrastructure.

References

[1] The H Security - Telecommunications regulator bars DigiNotar from issuing certificates. http://www.h-online.com/security/news/item/Telecommunications-regulator-bars-DigiNotar-from-issuing-certificates-1344786.html. Website accessed: September 30, 2011.

[2] Jon Callas, Lutz Donnerhacke, Hal Finney, David Shaw, and Rodney Thayer. OpenPGP Message Format. RFC 4880 (Proposed Standard), November 2007. Updated by RFC 5581.

[3] Tim Dierks and Eric Rescorla. The Transport Layer Security (TLS) Protocol Version 1.2. RFC 5246 (Proposed Standard), August 2008. Updated by RFCs 5746, 5878.

[4] Carl Ellison and Bruce Schneier. Ten Risks of PKI: What You're not Being Told about Public Key Infrastructure. In *Computer Security Journal*, 2000.

[5] Thomas Guillet, Ahmed Serhrouchni, and Mohamad Badra. Mutual Authentication for SIP: A semantic meaning for the SIP opaque values. In *Proceedings of New Technologies, Mobility and Security (NTMS)*, pages 1–6, November 2008.

[6] Mark Handley, Henning Schulzrinne, Eve Schooler, and Jonathan Rosenberg. SIP: Session Initiation Protocol. RFC 2543 (Proposed Standard), March 1999. Obsoleted by RFCs 3261, 3262, 3263, 3264, 3265.

[7] Weirong Jiang. A Lightweight Secure SIP Model for End-to-End Communication. In *Proceedings of the 10th International Symposium on Broadcasting Technology (ISBT)*, 2005.

[8] Alan B. Johnston. *SIP: Understanding the Session Initiation Protocol*. Artech House, 2^{nd} edition, 2004.

[9] Werner Koch and Wojciech Polak. Gnupg made easy [v1.2.0]. http://www.gnupg.org/, 2006.

[10] Lei Kong, Vijay Arvind Balasubramaniyan, and Mustaque Ahamad. A Lightweight Scheme for Securely and Reliably Locating SIP Users. In *Proceedings of 1st IEEE Workshop Voip Management and Security*, pages 9 – 17, April 2006.

[11] Ramona-Elena Modroiu, Bogdan Andrei Iancu, Daniel-Constantin Mierla, et al. Kamailio (openser) [v3.1.1]. http://www.kamailio.org/, December 02^{th}, 2010.

[12] Benny Prijono et al. Pjsip [v1.8.5]. http://www.pjsip.org/, October 20^{th}, 2010.

[13] Blake Ramsdell and Sean Turner. Secure/Multipurpose Internet Mail Extensions (S/MIME) Version 3.2 Message Specification. RFC 5751 (Proposed Standard), January 2010.

[14] Jonathan Rosenberg, Henning Schulzrinne, Gonzalo Camarillo, Alan Johnston, Jon Peterson, Robert Sparks, Mark Handley, and Eve Schooler. SIP: Session Initiation Protocol. RFC 3261 (Proposed Standard), June 2002. Updated by RFCs 3265, 3853, 4320, 4916, 5393, 5621, 5626, 5630, 5922.

[15] Henning Schulzrinne, Stephen Casner, Ron Frederick, and Van Jacobson. RTP: A Transport Protocol for Real-Time Applications. RFC 3550 (Standard), July 2003. Updated by RFCs 5506, 5761.

[16] Julian Seward, Nicholas Nethercote, Tom Hughes, Jeremy Fitzhardinge, et al. Valgrind [v3.6.0]. http://www.valgrind.org/, October 2010.

[17] Lars Strand and Wolfgang Leister. Improving SIP authentication. In *Proceedings of The Tenth International Conference on Networks (ICN)*, pages 164 – 169, January 2011.

[18] The GNU Privacy Guard Team. The GNU Privacy Guard [v1.4.9]. http://www.gnupg.org/, 2008.

[19] Alexander Ulrich, Ralph Holz, Peter Hauck, and Georg Carle. Investigating the OpenPGP Web of Trust. In *16th European Symposium on Research in Computer Security (ESORICS 2011)*, pages 489–507, September 2011.

 ISBN: 978-1-61208-183-0

Developing Trust and Reputation Taxonomy for a Dynamic Network Environment

Tanja Ažderska and Borka Jerman-Blažič
Jožef Stefan Institute
Ljubljana, Slovenia
e-mail: {atanja, borka}@e5.ijs.si

Abstract—**Trust and reputation are the pillars of many social phenomena that shape the Internet socio-economic scene. The few existing taxonomies provide only initial insights into the ways trust benefits can be felt, but they are neither complete nor elaborated in a systemic manner. In this paper, we propose a multidimensional framework for designing and assessing the completeness and consistency of reputation mechanisms. Our framework is based on systemic principles; it identifies reputation system components, the factors that influence the system-design, defines the interrelations between the former and the dependencies on the later. By considering the human-centric, dynamic and context-dependent trust-establishment, we detect five major factors that guide reputation systems' design. The presented framework is applied to BarterCast, a reputation mechanism that extends the current P2P network protocol – BitTorrent, and is deployed in the BitTorrent-client Tribler.**

Keywords–trust; taxonomy; reputation mechanisms; system theory; context

I. INTRODUCTION

Catering the variety entities and interactions between them, the Internet is an environment where the pervasive risk and inherent uncertainty pose a requirement for new tools to support decision making in such circumstances. Apart from the commercial expansion of the Internet, traditional networking among people relies on unwritten social protocols, like gossiping and rumors, to judge about one's trustworthiness and reliability. A global consensus on person's reputation has neither been required nor needed, yet the social model has been successfully supporting legitimate interactions by identifying untrustworthy individuals. The advent of social networking and computational semantics opens up a myriad of opportunities for merging the social and dynamic character of trust with the technical possibilities offered by Information and Communication Technologies. The growth of user-generated content, the vast offer of service providers, and the wealth of collaborative and market-based platforms, have introduced additional levels of complexity in the processes of information filtering and decision making. They require systemic approaches for treating trust and reputation (T&R). Hence, the success of online trust-based methods depends largely: a) on the research aimed at identifying where these methods offer the most benefit and b) on the quality of the frameworks where the principles of system design reside. Our work is a contribution in both of these directions. The framework defined here is guided by the principles of system theory and taxonomical categorization. To present the outlined topics and the results, the paper is organized as follows: the following section briefly examines related work in T&R, defines the notions of T&R and the progress towards their formalization. The succeeding sections outline the methodology used and introduce the proposed framework based on the principles of General Systems Taxonomy. Practical observations, supplemented with insights from other trust taxonomies and proposals, are elaborated through the framework, enabling the addition of a new level of granularity to the existing research map on T&R. The next section illustrates the application of the newly designed approach for the specific case of distributed environments, mapping the BarterCast reputation mechanism across the dimensions of the framework. The paper concludes with a review of the presented topics and a constructive discussion, outlining our future research plans.

II. THE NOTION OF TRUST AND REPUTATION IN A NETWORK ENVIRONMENT

Trust is a social manifestation we face on a daily basis. However, its definition is hard to grasp. One reason for this is its strong contextual dependence. However, another reason that is crucial and that refers to the practical side of system design is the non-linear nature of the social phenomena ascribed to trust, such as belief, regret, forgiveness, subjective judgment, etc. These comprise the affective (emotional, and thus the human) side of trust, and do not allow the system to be designed according to the elegant principles of mathematical linearity and probabilistic averaging. Therefore, incorporating trust into online scenarios analogous to those in the traditional social networks has not been very fruitful. The literature on T&R in social sciences is exhaustive [1–3]. The common attitude supports the aspect of relying on others' willingness to perform beneficial actions for one's welfare. Based on Gambetta's attitude on trust [4], we give the following initial definition:

Definition 1. Trust is the belief, i.e., the subjective probability that an entity will perform in a way likely to bring the expected benefit, or not to do unexpected harm.

Despite the interchangeable use of the concepts of T&R, reputation deserves its own and more specific definition that would stress how it differs from trust.

 ISBN: 978-1-61208-183-0

Definition 2. Reputation is the empirical memory about an entity's past behaviour, performance, or quality of service, in a specific context, i.e., domain of interest.

Hence, reputation is the amount of context-aware trust that an entity has created for itself, i.e., a quantitative representation of trustworthiness bounded by the domain of interest. Reputation results from calculation and assessments and is based on facts rather than mere opinion and belief (e.g., I trust you because of your good reputation), unlike trust, which is a more subjective form of evaluating someone's performance (e.g., I trust you despite your bad reputation).

In circumstances where one entity relies on another entity, trust choices include a certain level of risk. Josang defines two different types of trust – Reliability and Decision trust [5]. The former covers the aspect of trust as stated by Definition 1. The latter considers the risk brought about by the uncertainty of transactional outcomes and is used to extend the first definition, which now gains the following structure:

Definition 3. Trust is the extent to which one entity is willing to depend on others' decisions, accepting the unpredictable risk of a negative (undesired) outcome.

Much of the research on trust evaluation has its roots in Game Theory, where concepts like quality, cost and utility are more formally defined [6]. The most fundamental trust problems in game theory are captured by the Prisoner's Dilemma [7], a principle that demonstrates the trade-offs in people's decisions to maximize either their own profit or the overall outcome of the game. The Prisoner's Dilemma is also used in strategies for fostering contribution in some technical implementations online, such as BitTorrent's tit-for-tat policy [8]. Despite the early work on trust relations and conflict resolution in game theory, the notion of computational trust appears significantly later, when Marsh establishes the basis of formal trust in distributed artificial intelligence [9].

A work that relates quality and uncertainty within the framework of reputation is the Akerlof's study on the "market of lemons" [10]. Reputation mechanisms (henceforth denoted as RMs) are used to balance the information asymmetry, by helping buyers make better-informed decisions and incentivizing sellers to offer high-quality goods. Akerlof makes an instructive distinction between the *signaling* and the *sanctioning* role of RMs, which was only recently considered in computer science [11]. The computational formalization of T&R is mainly done by the use of a mathematical and formal logics apparatus. We restrain from presenting that body of work here, as this paper is part of the *identification* phase of a RM, rather than its *modeling* process.

III. TRUST TAXONOMIES AND THE NEW APPROACH

Several taxonomies of trust have been designed in the past decade [5], [12–14]. As a categorization of system entities, components and their interrelations, taxonomy is hardly a useful systemic approach if it only identifies the RM entities. Cohesive factor for all systems, which has not been tackled by any of the known taxonomies, is the identification of connections between the RM components. The framework presented in this paper not only specifies that, but it also provides analysis in several dimensions across the factors influencing RM's design. To entitle this work a systemic approach, we turn to the principles of General Systems Taxonomy and determine the position of RMs in the general systems space. Our taxonomy differs from the existing in the field in a few crucial aspects: 1) It follows a systemic approach of revealing the design issues in building RMs and relies on simple systemic principles; 2) It relates the RM subsystems in a way that allows understanding of their interrelations, but also of their connection to the environment where the overall system evolves; 3) It sets a common ground for the widespread, but scattered, research on computational T&R; 4) Most importantly, it determines the 'system' concept applicability of the defined taxonomy and detects the factors required for its completeness. The main content of this framework is outlined in the text that follows.

One of the most prominent works in General Systems Taxonomy is that of Nehemiah Jordan [15], according to which a system's taxonomy has three organizing principles: 1) Rate of change, 2) Purpose, and 3) Connectivity. Each principle defines two antitheses, resulting in the three pairs of properties shown in Table 1. Within this general framework, we also position the systemic properties of RMs, and use them later in developing the novel reputation taxonomy.

Dynamicity (D): Static systems are those that exhibit no change in a defined time-span. RMs are expected to provide long-term incentives and support decision-making in a dynamic manner. To do that, they consider the quality of experiences of the system entities and the history of transactions among them.

Environmental-orientation (E): The principle of purpose determines the direction of energy/information flow inside or outside the system. The two possibilities are a system-directed flow or environment-oriented. The former tends to maintain stable and constant conditions inside the system, whereas the latter modifies the system to obtain a desired state or bypass certain disturbances.

TABLE I. ORGANIZING PRINCIPLES OF JORDAN'S SYSTEMS TAXONOMY (the categories to which we assign RMs are bolded and italicized)

Rate-of-change	**Purpose**	**Connectivity**
Structural (static)	Purposive (system-directed)	Mechanistic (non-densely connected)
Functional (dynamic)	***Non-purposive (environment-directed)***	***Organismic (densely connected)***

Dense connectivity (C): The principle of system connectivity states two possibilities: systems are a)

mechanistic, i.e., not densely connected and the removal of parts or connections produces no change in the remaining components; or b) organismic, i.e., densely connected and the change of a single connection affects all the others. RMs depends heavily on the interactions among system entities. They are of inherently non-linear nature, implying that the outcome of each interaction has no predictable impact on the overall RM.

The significance of considering General Systems Taxonomy is in the clarification and simplification of the often-misused concept of a system. Our work establishes RMs as real systems, and by using sufficient generality and simplicity, categorizes them as dynamic (D), densely connected (C) and environment-oriented (E). In the next section we move to identification of the RM components, and determine their interrelations.

IV THE TAXONOMY FRAMEWORK

The new taxonomy proposed here covers more aspects of the issue and applies to the trust taxonomies and to the RM design: 1) It categorizes common and important concepts in the research on RMs, establishing a common systemic vocabulary; 2) It represents a novel approach to multi-dimensional mapping and assessment of the completeness and consistency of a RM; 3) It introduces additional granularity in the current taxonomic map of RMs, considering the notion of reputation and its application to the RM components; 4) It employs the D-C-E nature of RMs to detect additional factors that influence RMs design, providing better completeness of the taxonomy.

As a skeleton, we take Stanford's taxonomy [12], shown in Table II. The framework resulting from our work that was imposed on the skeleton allows a direct mapping of the models across the factors-dimension and subsystems-dimension in a consistent manner. This enables an immediate establishment of the interdependence between: a) the various RM subsystems; b) the subsystems and the RM as a whole; c) the RM and the general system where the RM is deployed; d) the RM and the environment where the overall system resides.

TABLE II. BREAKDOWN OF THE REPUTATION SYSTEM COMPONENTS (Marti *et al.*)

Reputation Systems		
Information Gathering	***Scoring and Ranking***	***Response***
Identity Scheme Information Sources Information Aggregation Stranger-Policy	Good vs. Bad Behavior Quantity vs. Quality Time-dependence Selection Threshold Peer Selection	Incentives Punishments

In order to specify the requirements and the implications of designing an efficient reputation mechanism, Marti et al. considered the following factors of impact: a) The limitations and opportunities imposed by the system architecture where the RM is deployed; b) The expected user behaviour; c) The goals of adversaries. As stated in Section III, RMs are of a D-C-E nature. Table III contains an assessment of the factors of impact on a D-C-E scale. It demonstrates which of these factors do not consider one or more system properties (D, C or E).

TABLE III. EVALUATING THE FACTORS OF IMPACT ON D-C-E SCALE (Y denotes "Yes" – does consider; N denotes "No" – does not consider)

Factor / Property	**User behavior**	**System Architecture**	**Goals of adversaries**
Dynamicity *(D)*	Y: through churn	***N:needed to capture environment evolution***	Y: accounted for in the adversarial strategies
Densely connected *(C)*	***N: very small number of users can have a large impact on the system***	***N: the reputation mechanism as a subsystem of the overall system has a huge impact***	***N: necessary to take into consideration for providing the resilience of the system***
Environment-oriented *(E)*	***N: so far only as system-oriented, neglecting the influence of the environment on user behavior***	Y: by considering the various properties of a centralized, distributed, hybrid	Y: few types of attacks (Sybil attack, collusion) resemble this nature of the reputation system

The content of Table III shows that the C-nature of the RMs is not considered at all. The interactions and relations between entities and the environment presented are not captured by any of the known trust taxonomies, and consequently, by none of the computational trust models.

Active Entity behaviour. As a first distinctive element from Stanford's taxonomy, we introduce the more general concept of reputation entity and recognize "users" as only one type of these entities. Entity refers to a party who participates in the process of reputation evaluation, either as an evaluator or as an evaluated side. We distinguish two types of reputation entities, active and passive. The former are enrolled actively in the reputation process: aggregating and disseminating information, acting upon certain triggers, and evaluating each other's and the trustworthiness of the passive entities. Examples are agents, users, peers in P2P networks, etc. In contrast, passive entities are those whose trustworthiness is evaluated by the active entities; they do not provide any feedback, and do not participate in the aggregation of reputation scores. Examples are items, comments, video/audio content, etc.

RMs must exhibit a high adaptive capability to address the issues outlined above. An important part of the solution is both the hard-technical and the soft-usability aspects of the system. The former may include availability and connectivity checking to form an overlay of reliable entities, while the latter will require bootstrapping techniques for the new-coming entities, and incentive policies for those who have already established some history of experiences.

Resilience and evolutionism. The circular, interlocking and time-dependent relationships among RM components

 ISBN: 978-1-61208-183-0

are also important in determining entities' behaviour. There often are properties of the overall solution that might not be found among the properties of its building components, leaving the behaviour of the whole system impossible to be explained in terms of the behaviour of its parts. In fact, this is a common property of complex systems that depend on social dynamics.

Context. Reputation information becomes significant only after it is put into a relevant context. Context is the set of circumstances or facts that surround a particular event or situation. Despite the various types of trust defined in the literature, only a few definitions consider its context-dependency. However, none of the known approaches considers the impact of context on the separate RM components [16]. Most of the current proposals employ it for content-filtering purposes. By including context information in the reputation evaluation, not only can the level of the entities' expertise be obtained, but also the domain of interest where this expertise is relevant.

Time. The time as well is an insufficiently considered factor that influences many of the design choices. Some relations between reputation and time have been studied extensively; however, many important time-properties have not received the expected attention. Each subsystem of the RM is influenced by decisions that should consider the permanency of the identifiers, the recentness of information, the time-stamp of feedback actions, the convergence of the reputation value, synchronization of time-driven actions, updates of the reputation values, etc. The time-issues in a RM depend on the given subsystem where they appear. Some of the ways to approach these issues include: introduction of a sliding window over which the reputation information gains certain importance; time-discounting of the various (meta) results obtained at a certain point in time or a combination of the discounting factors together with the entities' reputation in a certain context.

Privacy. The interest in information is accompanied by privacy requirements. Although privacy is a research field on its own, some design points of RMs directly face privacy challenges. RMs are expected to keep balance between the heterogeneity of users and their interest in information. As the main purpose of RMs is the embodiment of trust on the Internet, it would be useful to investigate where the offline forms of regulation-by-law fit in the online world and whether they can be incorporated to help the establishment of trust.

On the Internet, people tend to tolerate worse experiences, acknowledge lower competences, exhibit lower privacy requirements, accept greater risks and act under higher uncertainty. The fast convergence of the reputation effects degrades reputation as soon as the information propagates the network. By limiting this effect to the relevant context, RMs will exhibit better adaptability and flexibility to user demands. It is multidimensional as it is based on the factors identified to capture the RM's D-C-E nature and defines their relation to the RM subsystems.

V. THE EXAMPLE OF BART CAST

The reason we have chosen BarterCast (BC) [17] for taxonomical mapping is that it is fully distributed, but also a deployed RM in the BitTorrent content-sharing client Tribler [18]. Its design premise is that social phenomena (friendship, trust and sense of community) affect positively the system usability and performance. We briefly introduce BC, and then map it across the framework dimensions.

Information Gathering: For peers (client software), BC uses permanent IDs (*PermIDs*) based on a public key scheme, validated by a challenge-response mechanism to prevent spoofing. Users are referred to by *pseudonyms*. The social network creation is facilitated by the ability to import contacts from other networks (MSN, Gmail). Context information is stored in *MegaCaches* to support trust-based social groups. For content discovery, a *semantic overlay* of *taste buddies* (peers with similar taste) is maintained and discovered by a *gossiping protocol*. Exchanging data is done by 1) *exploitation*, with the buddies, or 2) *exploration*, with a random new peer. Only *direct experience* (for aggregated amount of service) is exchanged during the gossip. Peers maintain *private* (based on an entity's interactions with a single entity) and *shared history* (about interactions with all entities) and *subjectively* calculate the reputation. BC considers paths of two hops, due to the small-world effect in P2P file-sharing networks.

Scoring and Ranking: The network of interacting entities in BC is represented as a graph. As input statistics, both the quantity (upload in MB) and the quality (the positive contribution) of the service are considered in the scoring algorithm. The private and the shared history form the peer's local graph, which is used as an input for the maxflow algorithm. It computes the maximum flow over all possible paths, from a source node to a sink (target) node. The result is the highest reputation that a source node can give to a target node, and it is a scalar value in the [-1, 1] interval.

Response: BC introduces a few types of incentives. First, a *cooperative download* is used to improve the download performance of group members. Second, in addition to the BitTorrent's tit-for-tat (which gives peers only a short-term incentive to upload), BC incorporates long-term incentives by implementing a ranking policy, which allows interested peers an initial cooperation in the order of their reputation. Third, it cherishes the peers' sense of community, which on the long run acts as a social norm for contributive behavior. Finally, by introducing costly procedures for using system resources, BC discourages malice, providing an additional incentive for contributive peers. In order to select interacting partners, BC introduces a banning policy. The choice of whom to allow the use of resources is made according to the peers' reputation, where a reputation is required to be

 ISBN: 978-1-61208-183-0

above a negative threshold (to differ strangers from disreputable peers).

Stranger Policy: Strangers are tackled by the bootstrapping process in Tribler, in two ways. To obtain an initial list of neighbors, peers use a set of pre-known super-peers to bootstrap into the network. Then, there is also an overlay swarm with no central component that can also be used for initial bootstrapping, content discovery, and other information exchange.

Discussion. The way BC maps to the framework is presented in Table IV. The results suggest a space for substantial improvements. BC does not implement any type of integrity check of reputation entities and their relations across any of the defined factors. This can be achieved by introducing witnessing scheme, similar to that in [19]. Furthermore, coping with the dynamics is mainly handled on a network level through availability and connectivity check, considering only the node-churn in the network. Thus, many time-properties important for achieving consistency among the components are not taken into account. Although the validity of the reputation information is based on the 10 most recent transactions, this choice is made in a fixed manner rather than according to the system or interaction dynamics. One way to include the timeliness of reputation information in this RM is by introducing a time-discounting factor that will give different weights to the information according to its recentness. Another thing that BC lacks is a policy for penalizing malice. In an open, anonymous and dynamic environment, providing mechanisms that hold community members responsible for their actions is of crucial importance. Despite accounting for *taste similarity*, *taste* is much more subtle than preference. Results from Behavioral Economy show that users are often unaware of their taste, even for experiences from previously felt outcomes [20]. The possibility of importing contacts in Tribler from other social networks requires well-defined privacy policies, assurance for the system interoperability, and context-switching awareness. None of this is elaborated enough to justify the design choice for this kind of property. Although there is an *erase from profile* option, the download history for each peer is publicly visible for exploration and discovery. BC is based on the premise that, although non-resistant to cheating, real-world communities work well with millions of users. However, this does not speak about the impact these entities can have on the overall system welfare. For instance, only a small percentage of peers in a file-sharing community contribute the largest amount of resources in the network. False self-representation, as well as collusion, can have an impact on the cost that largely outweighs the benefit of designing and maintaining a RM. Finally, despite exploiting the small-world phenomenon for better gossiping in BC, this phenomenon is not an indication of any organizing principle of the nodes in the network. There is a certain structure a network should have in order for the small-world concept to be applied in the first place [21], [22]. In addition to applying re-organizing principles of the nodes' positions for satisfying the necessary structure, the BC reputation mechanism would benefit a great deal (with respect to both performance and accuracy of the result) from performing a full gossiping, instead of the current two-hop message exchange.

VI. CONCLUDING REMARKS AND FUTURE WORK

Building reputation is primarily a social process. Online environments can largely benefit from trustworthy choices. Handling numerous online experiences in a short time-span requires highly scalable solutions for trust establishment. In such a dynamic environment, having no RM to capture interaction trends is equal to being equipped for a world that no longer exists. The presented framework is a systemic approach to designing dynamic, densely connected and environment oriented RMs. As major factors that influence RM design we included *context, time, privacy, active entity behavior, resilience and evolutionism*, in addition to *system architecture*. The insights were incorporated into a multidimensional framework, together with the RM subsystems, to establish their interconnections and dependencies. The result is a more granular categorization of design choices/decisions. Finally, we mapped BC as a representative distributed and socially inspired RM onto our framework, revealing some weaknesses and proposing improvements of its design.

Future step in our work will be a system-modeling approach for resolving the design issues for a novel RM. According to the principles outlined in this work, the model will be premised on dynamicity, adaptability and evolutionism. We will employ System theory methods, allowing the use of sophisticated tools for evaluation and verification, something that has not been proposed so far by any of the approaches in the field. Moreover, it is a step towards the standardization of the design process of RMs. A multi-disciplinary approach is thus essential for limiting or extending the possibilities offered by ICT for preserving practicality, but adding innovation as well.

REFERENCES

[1] J. H. Fowler and N. A. Christakis, "Cooperative behavior cascades in human social networks," *Proceedings of the National Academy of Sciences*, vol. 107, no. 12, pp. 5334-5338, Mar. 2010.

[2] John Conlisk, "Why Bounded Rationality?," *Journal of Economic Literature*, vol. 34, no. 2, pp. 669-700, 1996.

[3] C. Castelfranchi and R. Falcone, "Trust is much more than subjective probability: Mental components and sources of trust," *32nd Hawaii International Conference on System Sciences - Mini-Track On Software Agents, Maui*, vol. 6, 2000.

[4] D. Gambetta, "Can We Trust Trust?," *TRUST: MAKING AND BREAKING COOPERATIVE RELATIONS*, p. 213--237, 1988.

 ISBN: 978-1-61208-183-0

[5] A. Josang, R. Ismail, and C. Boyd, "A survey of trust and reputation systems for online service provision," *Decision Support Systems*, vol. 43, no. 2, pp. 618-644, Mar. 2007.

[6] S. H. Chin, "On application of game theory for understanding trust in networks," in *2009 International Symposium on Collaborative Technologies and Systems*, Baltimore, MD, USA, 2009, pp. 106-110.

[7] D. Fudenberg and J. Tirole, *Game Theory*. The MIT Press, 1991.

[8] B. Cohen, "Incentives Build Robustness in BitTorrent," 2003.

[9] S. P. Marsh, "Formalising trust as a computational concept," 1994.

[10] G. A. Akerlof, "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism," *The Quarterly Journal of Economics*, vol. 84, no. 3, pp. 488-500, 1970.

[11] C. Dellarocas, "The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms," *Management Science*, vol. 49, no. 10, pp. 1407-1424, Oct. 2003.

[12] S. Marti and H. Garciamolina, "Taxonomy of trust: Categorizing P2P reputation systems☆," *Computer Networks*, vol. 50, no. 4, pp. 472-484, Mar. 2006.

[13] H. Alani, Y. Kalfoglou, and N. Shadbolt, "Trust strategies for the semantic web," *PROCEEDINGS OF THE TRUST, SECURITY AND REPUTATION WORKSHOP AT THE ISWC04*, vol. 7, p. 78--85, 2004.

[14] T. D. Huynh, "Trust and Reputation in Open Multi-Agent Systems," Jun-2006.

[15] N. Jordan, *Themes in Speculative Psychology*. Routledge, 2003.

[16] T. Heath, E. Motta, and M. Petre, "Computing Word-of-Mouth Trust Relationships in Social Networks from Semantic Web and Web2.0 Data Sources."

[17] M. Meulpolder, J. A. Pouwelse, D. H. J. Epema, and H. J. Sips, "Bartercast: A Practical Approach to Prevent Lazy Freeriding in P2P Networks."

[18] J. A. Pouwelse et al., "TRIBLER: a social-based peer-to-peer system," *Concurrency and Computation: Practice and Experience*, vol. 20, no. 2, pp. 127-138, Feb. 2008.

[19] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina, "The Eigentrust algorithm for reputation management in P2P networks," in *Proceedings of the twelfth international conference on World Wide Web - WWW '03*, Budapest, Hungary, 2003, p. 640.

[20] D. Ariely, G. Loewenstein, and D. Prelec, "Tom Sawyer and the construction of value," *Journal of Economic Behavior & Organization*, vol. 60, no. 1, pp. 1-10, May 2006.

[21] J. Kleinberg, "The Small-World Phenomenon: An Algorithmic Perspective," *IN PROCEEDINGS OF THE 32ND ACM SYMPOSIUM ON THEORY OF COMPUTING*, p. 163--170, 2000.

[22] O. Sandberg, "Searching in a Small World," p. 39--57, 2005.

TABLE IV. MAPPING BARTERCAST ONTO THE NEW FRAMEWORK

Subsystem \ Factor			Context	Time	Privacy	RE	AEB
Information gathering	ID Scheme		Non-linkable; Verifiable	permanent ID (PermID)	pseudonyms	N (machine-dependent ID)	challenge-response; combats free-riding; Sybil-vulnerable
	Info Sources		taste-buddies; subj. Reputations	10 most recent interactions	N	semantic overlay	considers 2 hops; employs small-world concept
	Info. Aggregation		MegaCaches for context-info; private and shared history	N	gossiping only about direct experience	exploitation & exploration	false feedback restricted by the information capacity of edges; collusion possible
	Integrity check		N	N	N	N	N
Scoring and ranking	Inputs		Quantity (Upload in MB); Only positive contribution;	N	N	N	History of transactions
	Comp. engine		Maximum-flow algorithm based on *arctan* function	N	Privacy as a metric	cooperative downloading protocol	No learning; Depends on system vulnerability
	Outputs		Single value in the interval [-1, 1]	N	N	optimistic un-choking	GUI for browsing peers
Response	Threshold		negative reputation threshold	Sliding window over 10 transactions	Preference similarity	N	Reputation-based peer selection
	Incentives	Reward	Improved service; Rank policy; tit-for-tat	N	N	Cooperation driven	relies on social altruism of taste-buddies; does not take risk into account
		Punish	N	N	*Erase from profile* option	N	N
Stranger policy			N	N	N	N	bootstrapping; connectivity & availability check

A Reference Model for Future Computer Networks

Hoda Hassan
Computer Science and Engineering Department,
American University in Cairo
Cairo, Egypt.
mhelali@aucegypt.edu

Abstract—**Future Internet design demands revolutionary approaches unfettered by legacy constraints and concepts. This paper presents a clean-slate Concern-Oriented Reference Model (CORM) for architecting future computer networks based on novel network design principles. CORM realizes the network as a software-dependent complex system. It defines the network design space in terms of function, structure and behavior, and perceives each of these design space elements within the context of network concerns, identified as Application, Communication, Resource and Federation. CORM adopts a bottom-up approach in network construction, focusing on the network building block, whose structure and behavior are inspired by evolutionary bacterium cell. Hence, CORM refutes the long endorsed concept of layering, and intrinsically accounts for emergent behavior, while ensuring network congruency. CORM's basic abstraction unit is validated using the Function-Behavior-Structure engineering framework. The paper concludes by presenting and evaluating an architecture derived from CORM.**

Keywords- Complex systems; Computer Network design; Computer Network Reference Model.

I. INTRODUCTION

Designing future computer networks dictates an eclectic vision capable of encompassing ideas and concepts developed in contemporary research unfettered by today's operational and technological constraints. However, unguided by a clear articulation of core design principles, the process of network design may be at stake of falling into similar pitfalls and limitations attributed to current network realizations. We opine that deficiencies apparent in current network realizations can be traced to the following underlying causes [1];

- The general trend towards network science and engineering lacks a systematic formalization of core principles that expresses essential network features required to guide the process of network design and protocol engineering;
- The prevalence of a top-down design approach for computer network architecture demonstrated as confining intelligence to network edges, and maintaining a dump core; and
- The absence of a general reference model, which embodies core network design principles, and acknowledges the multidimensionality in design entailed in architecting computer networks that reach beyond core networking requirements.

In this paper, we present a clean-slate Concern-Oriented Reference Model (CORM) for architecting future computer networks. CORM has sprouted as a generalization to our concepts, design principles and methodology presented in [2, 3]. Our initial endeavor, CellNet [3], was a bio-inspired network architecture, which was tailored to operate in accordance to the TCP/IP suite. However, CORM is a reference model for architecting any computer network. It expresses the most fundamental design principles for engineering computer networks at the highest level of abstraction. CORM stands as a guiding framework from which several network architectures can be derived according to specific functional, contextual, and operational requirements or constraints. CORM conceives computer networks as a distributed software-dependent complex system that needs to be designed along two main dimensions: a vertical dimension addressing structure and configuration of network building blocks; and a horizontal dimension addressing communication and interactions among the previously formulated building blocks. For each network dimension, CORM factors the design space into function, structure and behavior, applying to each the principle of separation of concerns (SoC) for further systematic decomposition. Perceiving the network as a complex system, CORM constructs the network recursively in a bottom-up approach (In this research work the term bottom-up refer to network composability as opposed to its more frequent use to refer to layer organization in the Internet layered architecture). CORM defines the network cell (NC) as the network building block. The NC's structure and behavior mimic the structure and behavior of evolutionary bacterium cell. The network is then synthesized from NCs according to a structural template that defines different structural boundaries.

Being a reference model for computer networks, CORM can be considered a definitional model; it expresses the required characteristics of a system at an appropriate level of abstraction [4]. CORM expresses the characteristics of adaptable complex systems, and network functionalities within its basic abstraction unit (CORM-NC), and enforces both to be synthesized into the network fabric by construction. Therefore, we validated CORM by validating the derivation of CORM-NC. In this respect, we used the Function-Behavior-Structure (FBS) framework [5] as our validation model. The FBS is applicable to any engineering discipline for reasoning about, and explaining the nature and process of design [13]. Furthermore, we present CAHN as an architecture for ad hoc networks derived from CORM, and evaluate CAHN's performance through simulation.

The paper is organized as follows; Section 2 overviews related work. Section 3 introduces CORM, and validates CORM's basic abstraction unit using the FBS framework. In Section 4, we derive CAHN, an architecture for ad hoc networks based on CORM, and evaluate CAHN's performance through simulation. Section 5 concludes the paper.

II. RELATED WORK

The Internet has been criticized for lack of security, difficulty in management, incognizant protocol operation, and

 ISBN: 978-1-61208-183-0

inadequate support for mobility [6], thus motivating a plethora of proposals; some attempting point solutions to specific problems, while others aimed architectural innovations. We identify two network dimensions along which most architectural proposals can be classified; a vertical dimension addressing structure and configuration of protocols, and a horizontal dimension addressing communication. We claim that most of the proposals focus on one dimension while diminishing the other. Below, we present proposals in [7], [8], and [9] as examples supporting our previous claim.

The SILO Project in [7] proposes an architecture based on fine-grained service elements that can be composed based on ontology of functions and interfaces. A SILO-enabled application can thus specify high-level functional requirements, and request service elements to be composed accordingly to meet these requirements. The Recursive Network Architecture (RNA) presented in [8], is based on recursive composition of a single configurable protocol structure. RNA avoids recapitulation of implementation, as well as encourages a cleaner cross-layer interaction. This is achieved by using, a single meta-protocol module, which facilitates the inter-protocol interactions at different layers. Content Centric Networking (CCN) presented in [9], creates a network architecture based on named data instead of named hosts by making the address in packets correspond to information or elements reachable on the Internet, rather than machines. CCN proposes a layered node model that resembles the structure of TCP/IP layering model, but differs in layers' responsibilities.

SILO and RNA have been presented as clean-slate architectural attempts towards Future Internet. However, layering, as a design paradigm, is still the prevailing model. An essential goal of both proposals is to gracefully embrace cross layering into the present network stack. Although considered clean-slate architectures, we argue that by adhering to layered stacks as the underlying model, both proposals might suffer form shortcomings attributed to the Internet model. First, both architectures do not give guidance to engineers as how to handle cross interests among composed protocols: The single control agent in SILO, as presented, is a monolithic unit representing a single point of failure for all protocols working under its control, as well as imposing scalability problem as service diversity, granularity, and operational parameters increase. As for RNA, we note that confining the logic for horizontal and vertical interlayer communication into a single entity, is a very challenging task that is error prone. Furthermore, it lacks explicit representation for interactions leaving it to be decided on at runtime. This allows for implicit assumptions to creep into protocol design and implementations. Second, both architectures have undermined monitoring and resource management failing to express both functions as first class architectural constructs. Finally, as presented, both architectures focus on the vertical dimension of the network without suggesting how the horizontal dimension will be incorporated in terms of naming and addressing. On the other hand, CCN focus mainly on naming and addressing and disregard the need for managing on-node interactions. Similar to SILO and RNA, CCN also adheres to a layered stack and fails to provide explicit specifications for inter-protocol interactions and cross-interest management.

III. CORM: A Concern-Oriented Reference Model for Computer Networks

For completeness, this section gives a synopsis of CORM's design principles and methodology presented in [2, 3]

A. CORM Design Principles and Methodology

CORM derivation process was initiated by identifying two core network-design principles that, we assert, are applicable to all computer networks regardless of their size, purpose, operational context, or capabilities. The first principle states that a computer network is a complex system, while the second principle states that a computer network is a distributed software system. From a complex system perspective, computer networks need be composed of *autonomous entities* capable of *emergent behavior* that can act coherently to perform the global system function, in spite of *intricate interactions* occurring at the micro and macro level [10]. On the other hand, as a distributed software system, computer networks need to be designed according to Software Engineering (SE) principles and concepts. Separation of Concerns (SoC) is a prominent SE principle that was extensively applied to the design of CORM for systematic decomposition of the network system. Guided by our principles, we formulated a Concern-Oriented Bottom-Up design methodology for deriving CORM. The Bottom-Up approach is motivated by our first design principle in general, and network composability of autonomous entities in specific, thus accentuating the importance of the entities composing the network system. These network-building entities need to imitate entities in a Complex Adaptive System (CAS is a complex system whose emergent behavior always lead to overall system stability, in contrast to unstable complex systems whose emergent behavior may result in system meltdown. In this paper the term complex system indicates CAS unless otherwise stated), by possessing adaptability, self-organization and evolvability as intrinsic features. The network will then be recursively synthesized from these network-building entities in a bottom-up approach substantiating the two main network-dimensions; a vertical dimension that addresses structure and configuration of network building entities, and a horizontal dimension that addresses communication and interactions among the previously formulated building entities. For the synthesized networks, the Concern-Oriented paradigm represents our vision in network functional decomposition realized at the micro (network-building entities), as well as at the macro (network horizontal and vertical dimensions) level.

As a direct consequence of our Concern-Oriented Bottom-Up design methodology, CORM does not differentiate between network core and network edge in terms of capabilities, thus contradicting the End-to-End (E2E) principle that has been central to the Internet design. It has been argued that the E2E principle has served the Internet well by keeping the core general enough to support a wide range of applications. However, we contend that, taken as an absolute rule, the E2E principle constrained core evolvability rather than fostered its capabilities rendering the Internet biased to those applications that can tolerate its oblivious nature, and forcing designers and protocol engineers to adopt point solutions to compensate for core deficiencies. Another consequence to our proposed bottom-up network composition is contradicting the prevailing misconception of abstracting a network in terms of an inter-network. Adopting a bottom-up approach to network

 ISBN: 978-1-61208-183-0

composition implies recursive construction of the inter-networks from networks, which are likewise recursively constructed form network components, which are constructed from one or more network building blocks.

B. CORM Components

A network reference model is an abstract representation of a network. It conveys a minimal set of unifying concepts, axioms, and relationships to be realized within a network [11]. For expressing a multi-dimensional system, such as a computer network, multiple abstract representations are required to capture the system from different perspectives. CORM abstracts a computer network in terms of function, structure and behavior, which are represented respectively as, the network-concerns conceptual framework (ACRF), the network structural template (NST), and the information flow model (IFM). Both the ACRF and the NST have been previously defined in relation to CellNet in [3]. However, in the following subsections, we will revisit their definition at the level of a reference model. The ACRF will be redefined in terms of the network requirement specification while the NST will be abstracted in terms of the basic network building block (NC).

1) ACRF:Conceptual Framework for Network Concerns

We postulate that the requirement specification of a computer network can be expressed as follows: *"The network is a communication vehicle that allows its users to communicate using the available communication media"*. Accordingly, we identify the network users, the communication media (physical and logical), and the communication logic as primary requirements, which the network design need to address and plan for. Applying the concept of SoC to the above requirement specification statement, we identify four main network concerns; Application Concern (ACn), Communication Concern (CCn), Resource Concern (RCn), and Federation Concern (FCn). The first three are core network concerns encompassing the network functional requirements, while the fourth is a crosscutting concern (non-functional requirement) representing the area of intersection or common interests among core concerns. Elaborating on each concern we have:

- The ACn encompasses the network usage semantics; the logic and motivation for building the network, where different network-based end-applications (network users) can be manifested.
- The CCn addresses the need for network route binding to provide an end-to-end communication path allowing network elements to get connected (communication logic)
- The RCn focuses on network resources, whether physical or logical, highlighting the need for resource management to efficiently address different trade-offs for creating and maintaining network resources (communication media).
- Finally, FCn orchestrates interactions, resolving conflicts and managing cross interests, where areas of overlap exist among the aforementioned core concerns.

These four network concerns are manifested as CORM conceptual framework for network concerns, referred to hereafter as ACRF. The ACRF represents the blueprint for the network functional design that need to be realized along both network dimensions; vertically on the network component and horizontally among network components.

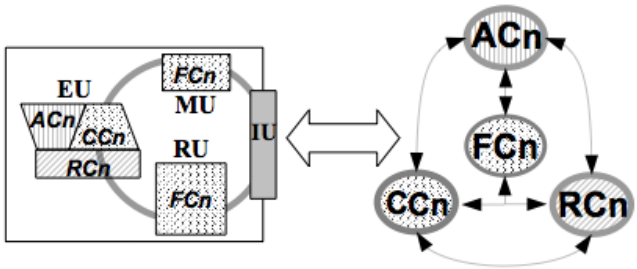


Figure 1. ACRF realization within an NC

Analyzing the Internet model (vertical dimension) and the current network realizations (horizontal dimension) with respect to the ACRF framework, we note that both RCn and FCn are absent. Vertically, the Internet-layered model accounts for ACn and CCn. However, the model did not apply the correct concern separation; a single concern was split along two layers. Moreover, the strict layered paradigm for functional decomposition curtailed all possibilities for considering FCn. As for RCn, it was assumed that resource-management functionalities, are either applications of specific type, and thus will be overlaid on top of the protocol stack, or are to be handled locally by the physical media. For the horizontal dimension, current network realizations account for both ACn and CCn, while the RCn and FCn are usually realized as point solutions. Servers and server farms represent ACn, while routers, switches, and DNS represent CCn. Both RCn and FCn, are implemented as add on functionalities conducted by the use of special protocols for network management and traffic engineering.

2) NST: Network Structural Template

The NST defines the structure of the network building-blocks, and the logic by which these blocks are grouped to compose the network. We classify network building-blocks into computational/decision capable e*ntities,* and a communication *substrate*. The former encompass the network-concern space (ACRF framework), while the latter is a passive interaction media for information exchange. Being the primary constituents of a software-based complex system, network-entities need to possess adaptability, self-organization and evolvability as intrinsic features thus mimicking bacterial cells in a bacterial colony; our adapted model of complex systems [12]. Hence, we define the Network Cell (NC) to be the primary network building block. The NC is a self-contained computational/decision entity capable of monitoring its state, adapting to perceived conditions, inferring decisions, recording its experience, and eventually evolving through self-learning and intelligent adaptations. One or more NCs make up a Network Component (Ncomp), which we define as the basic network entity capable of end-to-end communication. The ACRF is realized within the NC as illustrated in Fig. 1, thus forming the basic abstraction unit of CORM; the CORM-NC. For further details on the internal structure and units of the NC we refer the reader to [1, 2, 3].

Network Compositional Logic (NCL) defines the bottom-up network construction out of network-entities, and identifies the different interaction boundaries that can occur among network-entities (NC and/or Ncomp). NCL stems from our bottom-up definition of network and inter-network construction. NCL defines a computer network as two or more Ncomp connected by a communication substratum, where Ncomp interactions are

 ISBN: 978-1-61208-183-0

sustained, despite the heterogeneity of the hardware, middleware, and software of the connected Ncomps. As for a computer inter-network, NCL defines it as two or more computer networks connected by communication substrate, where interactions among Ncomps residing within each of the connected networks are sustained, despite the heterogeneity of the hardware, middleware, and software employed by the Ncomps composing the connected networks. Integrating NC, Ncomp, and NCL, we derive CORM NST, and define it using EBNF as follows:

CORM NST EBNF formal Definition:

Notations

- Trailing * →repeat 0 or more times
- Trailing + →repeat 1 or more times

Abbreviations

- MU = Monitoring Unit
- RU = Regulation Unit
- EU = Execution Unit
- IU = Interface Unit
- NC = Network Cell
- CCS = Cell Communication Substratum
- Ncomp = Network Component
- Net = Network
- NCS = Network Communication Substratum
- INet = Inter-network

Grammar Definitions

- NC = MU RU EU IU CCS
- Ncomp = NC (CCS NC)*
- Net = Ncomp (NCS Ncomp)+
- INet = Net (NCS Net)+ = Ncomp (NCS Net NCS)+ Ncomp

3) IFM: The Information Flow Model

The Information Flow model (IFM) represents the horizontal dimension of the network. IFM depicts the interactions occurring among network entities, giving rise to the emergent behavior required for network adaptation and evolution. The IFM captures the aspects of information exchange by defining two sub-models: Data Representation sub-model (DR) and Data Communication sub-model (DC). Data representation and communication in CORM exist at both the vertical and the horizontal network dimensions. Vertically, data representation and communication occurs within an NC, as well as between the different NCs making up a Ncomp. Horizontally, data representation and communication occurs between Ncomps in the same network, or across networks. DR will provide categorization for the different types of information flowing in the system, according to the ACRF framework. As such, DR is mainly concerned with the "meaning" of information flowing within the network system. DR need to handle complexity in terms of the amount of information required to depict the system-states at the macro and micro level, taking decisions on the details that need to be exposed and those that need to be suppressed. DC, on the other hand, is concerned with communication aspects including interface compatibilities, data formating across different communication boundaries and majorly routing functions, including addressing, naming and forwarding. Similar to DR, the DC will need to address characteristics of complex systems, such as the free-scale small-world layout, when devising the routing functions. Detailing DC and DR is the focus of our future work.

C. CORM Features

CORM refutes the long endorsed concept of layering, introducing the CORM-NC as a novel abstraction unit. To our knowledge, CORM is the first reference model that addresses the need for engineering for emergent behavior by accentuating monitoring, knowledge acquisition, and regulation as first class intrinsic features of the basic abstraction unit (BAU) –the CORM-NC. Furthermore, we argue that CORM maintains system integrity due to network construction congruency, where Ncomps, networks and inter-networks are defined recursively in terms of the BAU. In addition to the previously mentioned features, CORM facets acknowledge the multidimensionality of the networks, and accounts for concepts and notions proposed by contemporary designs and architectures including protocol composability out of fine-grained micro-protocols, dynamic protocol adaptation, protocol extensibility and flexibility, cross interest management and control, context awareness through monitoring, resource management as a standalone requirement, and inspired biological behavior and evolution. Table 1 highlights the differences between CORM and the more conventional layered network models (e.g., Internet, OSI, ATM, etc..).

D. CORM Validation

The FBS framework developed in [5], and illustrated in Fig. 2 is applicable to any engineering discipline, for reasoning about, and explaining the nature of the design process [13]. In this section, we aim to validate the derivation of CORM's BAU, the CORM-NC, using the FBS framework. The inception point for CORM-NC design is marked by our design principles. According to which, computer networks need to be designed as a software-dependent CAS that exhibit emergent behavior. CORM design principles formed our first set of requirements F_1 and expected behavior Be_1 as follows:

F_1 = CAS (autonomous entities, complexity)

Be_1 = Emergent Behavior (adaptation, self-organization, evolution)

Shifting to the structure that can deliver F_1 and Be_1, we attempted a *catalog lookup* by exploring natural complex systems, and studying their structure (S), and the individual behavior of their components (Bs). Our research led us to a recent study on primordial bacterial colonies [12]. This point marked our first functional reformulation. We formulated new requirements F_2 for designing a network cell that mimics the bacterium cell behavior Be_2. Accordingly, we synthesized the

TABLE I. CORM VS. LAYERED NETWORK MODELS

Features/ Model	CORM	Layered Models
BAU	NC	Layer
Operation of BAU	Independent: CORM-NC can exist and operate by itself	Dependent: a single layer can never exist or operate by itself
BAU responsibilities	(1)Execution of assigned network functions (2)Self-monitoring and regulation	Execution of assigned network function
BAU Relationships	Interdependent: NC realizes other NCs and cooperate to maximize the over all performance by adapting to context	Incognizant: A layer at one level uses services from the layer below and provides services to the layer above, while being incognizant to the presence of other layers that are at one level further
System Level awareness	Global awareness: CORM-NCs have a sense of global system goal	Unaware of the global system: Awareness is restricted to the layer boundary
Network Composition	Bottom-up recursive	Top-down incremental overlaying

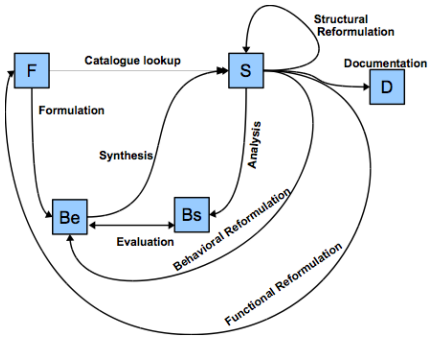


Figure 2. Gero's FBS (adapted from [5])

structure S_2 from Be_2 presenting the NC. However, F_2, Be_2, and S_2 needed further reformulation to detail network requirements. At this point, we defined the network requirement specification that led to the derivation of the ACRF framework for network concerns, yielding a new set of requirements F_3. F_3 was integrated with F_2, and super-imposed over Be_2, and S_2 to customize each towards the context of computer networks leading to the derivation of the CORM-NC.

CORM-NC delineates the BAU from which the network can be recursively built. However, at this point of our research, we still have not completely defined Bs for CORM-NC, since this will involve defining performance variables, and their range of values for the software code that will run within each unit of the NC structure. Nevertheless, Bs is accounted for by specifying the IFM as an essential part of CORM.

IV. Deriving and Evaluating a CORM-Based Architecture

The key difference between a reference model and an architecture is the level of concept abstraction that the model conveys, as well as the degree of requirement specifications that the model addresses. CORM expresses the most fundamental design principles for engineering computer networks at the highest level of abstraction. To derive an architecture from CORM further specifications regarding network operational context, performance requirements, and/or constraints need to be identified.

A. CAHN: A CORM-Based Architecture for Ad Hoc Networks

We define CAHN's requirement specifications as follows;

- Minimal architecture that provides core network functionalities: CAHN should be able to provide basic communication and transport services equivalent to that supported by the TCP/IP suite.
- Cross-interest management: CAHN should provide a systematic way for dealing with cross interests among the supported network functionalities.
- Modular: CAHN abstractions should separate functions into modules with clear defined interfaces.

Based on CORM's NST and ACRF, and guided by the above requirements, we define CAHN-Ncomp to be composed of four CORM-NCs, each instantiating the concerns defined by the ACRF. Accordingly, CAHN abstractions are the following concern-specialized CORM-NCs; Application Network Cell (ANC), Communication Network Cell (CNC), Resource Network Cell (RNC), and Federation Network Cell (FNC). CAHN networks will be composed of CAHN-Ncomps, each of which will be composed of ANC, CNC, RNC and FNC.

B. Engineering Protocols for CAHN

Protocol engineering in CAHN need to be classified according to the ACRF framework, and thus executed by the corresponding concern-specialized NC. Moreover, the task performed by each protocol (NC) will be internally classified according to the ACRF, as defined by the CORM-NC. To clarify this recursive assignment of the ACRF framework, we present an example for the routing function in CAHN.

According to the ACRF classification, the routing function is a CCn, which will be represented as a CNC in CAHN. However, routing as a function is a composite task that can be further divided into several subtasks such as, naming, addressing, forwarding, routing table creation and maintenance, etc. These identified subtasks will be recursively classified according to the ACRF. Following is an example of such classification:

- CNC-ACn: The application concern (ACn) of the CNC will be responsible for setting the routing protocol policies, which determines the quality of the routes to be discovered, and how the routes will be maintained. The CNC-ACn decisions will partially depend on the communication profile that is received from the ANC. This communication profile will indicate the destination and priority of the flow that is to be administered into the network, and the quality required for the end-to-end route
- CNC-CCn: Depending on the CNC-ACn requirements and policies, the CNC-CCn will decide on the appropriate routing protocol to be instantiated. The instantiation of a routing protocol depends on the micro-routing-protocols available on the CNC, from which a routing function can be devised. Alternatively, a default routing protocol can be adapted to the ACn requirements. CNC-CCn will also decide on link parameters, since route definition depends mainly on link characteristics. This introduces a cross interest between CNC and RNC, which will be handled by the FNC. Other communication tasks handled by the CNC-CCn include resolving routes, sending and receiving route requests and replies, communicating with neighbors, forwarding packets, etc.
- CNC-RCn will be responsible for estimating and managing the resources assigned to the CNC.
- CNC-FCn is responsible for monitoring and regulating the performance of the CNC . Parameter monitored by the CNC-FCn are specified once the CNC get specialized, and are subject to adjustments and/or amendments if required. Parameters monitored can either be specific, pertaining to the communication task assigned to the CNC, or general, relating to the over-all performance of the CNC. The CNC-FCn has a regulation cycle that will constantly check the performance of the communication related functions in specific, and the CNC operation in general, by comparing the values of the monitored parameters to thresholds values previously defined in a knowledge database stored in the FCN . If the monitored values fall below the indicated thresholds, the FCN will

 ISBN: 978-1-61208-183-0

interfere to regulate the operation of the CNC. Furthermore, the FCn can decide on any optimizations required to improve the performance, or it can interfere to resolve any cross interests that might rise among the core-concerns within the CNC. For example, the memory required by the routing table could exceed the space assigned to the CNC. In such a case, the FCn, after consulting its knowledge-base, could either instruct the CCn to alter its route-purging policy, or command the RCn to request more memory space.

C. CAHN Evaluation

CAHN is evaluated by simulating a CAHN-based network in the ns2 simulator [14]. Our simulation is based, in part, on the simulation in [15], in which a cross-layer power adaptation algorithm was devised for ad hoc networks. The algorithm in [15] integrated the operation at the Network, MAC and the physical layers to tune the transmission power of a node according to the number of its neighbors, in an attempt to minimize MAC contention, while maintaining network connectivity. However, such optimization had adverse effects on TCP traffic due to network oscillation between connectivity and dis-connectivity. This highlights the pitfalls of cross-layer adaptations that result in unintended consequences, when protocols at different layers operate with conflicting interests. We conjecture that CAHN-based networks can counteract such conflicting interests. Hence, we simulated CAHN-Ncomp on ns2 nodes by adjusting the ns2-code for the TCP, AODV and MAC to comply with the ACRF framework, as well as with CAHN-NCs. Thus any subsequent reference to these protocols will relate to their modified version. We define the performance parameters in CAHN simulation as; 1) the power level that results in minimum MAC contention, while sustaining next-hop transmission at RNCs and CNCs, respectively, 2) next hop neighbor at the CNCs, 3) and the TCP congestion window size at the source ANC (refer to [3] for details of adjusting congestion window size to path capacity). These parameters will be monitored and regulated by the FCns of the corresponding CAHN-NCs. Our simulation is divided into two phases. Phase 1is a learning-adaptation phase, where an adapted version of the power adaptation algorithm in [15] controls the transmission power. In this phase, the FNC populates its knowledge-base with information about the level of performance attained relative to the values assumed by the monitored parameters. In phase 2, the FNCs residing on CAHN-Ncomps, manage cross-interests among the performance parameters, and choose combined optimal-values that support the TCP flow. Hence, the FNCs prevent the oscillations reported in [15]. Figs. 3 and 4 plot the recorded TCP throughput at the sink nodes in ns2 simulations, in case of the cross-layer power algorithm as implemented in [15], versus CHAN.

V. CONCLUSION

This paper proposes CORM, a concern-oriented reference model for future computer networks. CORM is based on two design principles that realize the network as a software-dependent CAS. CORM refutes the long endorsed concept of layering, intrinsically accounts for emergent behavior, and ensures network congruency. We used the FBS engineering framework to validate CORM's BAU, the NC, then derived and evaluated an architecture based on CORM through simulation.

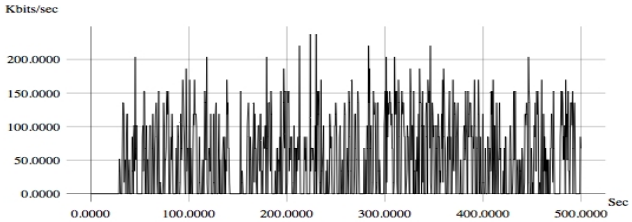


Figure 3. TCP-Sink throughput using cross-layer power adaptation[15]

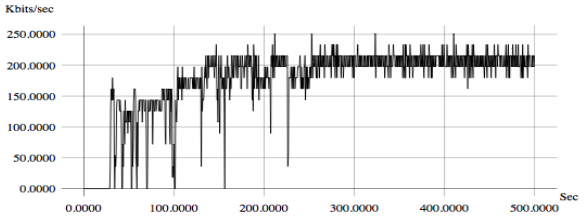


Figure 4. TCP-Sink throughput using CAHN

VI. REFERENCES

[1] H. Hassan, "A Reference Model and Architecture for Future Computer Networks," Ph.D. Dissertation, Virgina Polytechnic and State University, Blacksburg VA, May 26, 2010

[2] H. Hassan, R. Eltarras, and M. Eltoweissy, "Towards a Framework for Evolvable Network Design," Collaborate Com 2008, Orlando, FL, USA.

[3] H. Hassan, et al., "CellNet: A Bottom-Up Approach to Network Design," 3rd International Conference on New Technologies, Mobility and Security, Cairo, Egypt, December 2009.

[4] F. Polack et. al., "Complex Systems Models: Engineering Simulations," in ALife XI. MIT press, 2008.

[5] J.S. Gero "Design Prototypes: A Knowledge Representation Schema for Design," AI Magazine, vol. 11, pp. 26–36, Winter,1990.

[6] Clark, "NSF Future Internet Summit," Meeting Summary, Washington, DC October 12-15, 2009. Version 7.0 of Jan. 5, 2010

[7] D. Stevenson, et. al., "On the Suitability of Composable Services for the Assurable Future Internet," Military Communications Conference, Orlando, FL, USA, 2007.

[8] J. Touch, Y. WangV. Pingali, "A Recursive Network Architecture," http://www.isi.edu/touch/pubs/isi-tr-2006-626/isi-tr-2006-626.pdf. <retrieved, 12, 2011>

[9] V. Jacobson, et. al., "Networking Named Content," CoNEXT'09, December 1–4, 2009, Rome, Italy.

[10] Melanie Mitchell,"Complex systems: Network thinking," Artificial Intelligence, Volume 170, Issue 18, December 2006, pages 1194-1212.

[11] "Reference Model for Service Oriented Architecture 1.0," Committee Specification1. Aug.2006. http://docs.oasis-open.org/soa-rm/v1.0/soa-rm.pdf <retrieved, 12,2011 >

[12] E. Jacob & H.Levine, "Self-engineering Capabilities of Bacteria," Jr of Royal Society. Interface, http://star.tau.ac.il/~eshel/papers/Interface.pdf.<retrieved, 12, 2011 >

[13] P. Kruchten, "Casting Software Design in the Function-Behavior-Structure Framework," IEEE Software, vol.22, pp.52-58, 2005.

[14] "The Network Simulator – ns-2," http://isi.edu/nsnam/ns/<retrieved, 12, 2011 >

[15] V. Kawadia & P. R. Kumar, "A Cautionary Perspective on Cross Layer Design," IEEE Wireless Commun., vol. 12, no. 1, Feb. 2005, pp. 3–11.

 ISBN: 978-1-61208-183-0

Application Design over Named Data Networking with its Features in Mind

Sen Wang, Jianping Wu, Jun Bi
Network Research Center, Tsinghua University
Beijing, China
wangsen@netarchlab.tsinghua.edu.cn; {jianping, junbi}@cernet.edu.cn

***Abstract*—Designed around host-reachability, today's Internet architecture faces many limitations while serving data-oriented applications, which produce most traffic load to the Internet. Many clean-slate designs of the content/data oriented network have emerged to adapt to these needs. Named Data Networking (also known as CCN) is one of these designs to address these limitations from the fundamental level by building network architecture around named data. In this paper, we identify five key features crucial to application design over Named Data Networking and take the voice conference system as an example to show how this features impact the application design significantly in detail. We identify three major challenges facing current voice conference system and illustrate how NDN could help to solve these challenges. A NDN-based design of voice conference system is presented along with discussing its reliability and congestion control.**

Keywords-Named Data Networking; Application Design; Conference System.

I. INTRODUCTION

Internet was designed around a host-to-host model, which is much suitable for most applications at that time (e.g., telnet, ftp, etc.). But, today, most current Internet usage is data-centric [1]. The overwhelming use (>99% according to most measurements) of today's networks is for an entity to acquire or distribute named chunks of data (like web pages or email messages) [2]. Actually, users want to get data or service rather than communicate with the host which holds these data or service. With this insight, some clean-slate redesigns of Internet Architecture have emerged including CCN (Content Centric Networking) [3], DONA (A Data-Oriented Network Architecture) [4], etc. Therefore increasing attention has been attracted into this research area.

Named Data Networking (also known as CCN) [3] is a newly proposed Internet architecture which is designed around named data to address the limitations of today's Internet from the fundamental level. We expect that the success of NDN would largely depend on whether the new architecture can support various application needs more effectively and efficiently as it promises. So, designing applications over NDN is an extremely important issue to solve. In this paper, we identify five key features crucial to application design and take the conference system as an example to show how this features impact the application design significantly in detail.

The rest of this paper is organized as follows. In Section 2, five NDN Features are elaborated. Section 3 identifies three main challenges facing conference system and explains how NDN could help to solve these challenges with its embedded features. Section 4 takes the conference system as an example to show how features of NDN impact the application design significantly in detail. Finally, we conclude in Section 5.

II. NDN FEATURES FOR APPLICATIONS

As a promising, clean-slate network architecture, NDN is designed from a data-centric perspective. Differing from conventional connection-based TCP/IP architecture, NDN has its own features and its effects in design of applications which is summarized in this paper as follows:

First, NDN adopts the Publish/Subscribe communication paradigm to build a data-centric network architecture. The Publish/Subscribe paradigm is a vital ingredient for future services and applications. It allows asynchronous and decoupled many-to-many communication and typically supports data-centric information dissemination [12]. Sending Interests can be viewed as some kind of subscribe and the data delivery can be seen as a publishing process. The Publish/Subscribe paradigm decouples the producers and consumers of data in both time and space [13], which is the nature of most applications [12].

Second, NDN is receiver-controlled by nature. The original objective of the TCP/IP Internet architecture is to interconnect all existing networks and hosts uniformly and efficiently [5]. When a host connects to the Internet, it can communicate with arbitrary host connected to the Internet by its IP address. This enforces today's Internet a sender-control manner naturally. The Publish/Subscribe communication paradigm decoupled the producers and consumers of data. Producers don't need to hold references of consumers and know how many subscribers are participating in this interaction [13], and vice versa. In this paradigm, the conventional sender-controlled manner is not effective. We speculate that a receiver-controlled manner is more suited for NDN. As a clean-slate Internet architecture, this transition of NDN will turn Internet from push mode to pull and impact application design and implementation significantly.

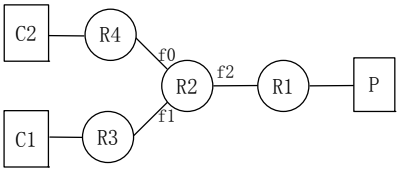
Third, NDN provides an auto-organized and asynchronous multicast distribution mode. In NDN, each chunk of data is named and can be transmitted and stored independently, which provides a substrate for the multicast distribution mode together with the Publish/Subscribe communication paradigm. Specifically, by compressing the interests with the same name and responding to interests with data cached in the intermediate routers, an auto-organized and asynchronous multicast distribution mode is provided in NDN network. As Figure 1 (a) shows, when the two consumers, say C1 and C2, send the Interests to the same datum published by P almost simultaneously, NDN router R2 will compress the two Interests and send just one Interest to R1. After the datum arrives at R2, R2 would find that the Interest requesting this datum has two corresponding interfaces f0 and f1in the PIT table (Pending Interest Table, a recording list of Interests which have been forwarded, while their corresponding Data have not been received yet) and send two copies of the datum through f0 and f1 respectively. The datum will be cached in the Content Store of R2. It should be noted that this

 ISBN: 978-1-61208-183-0

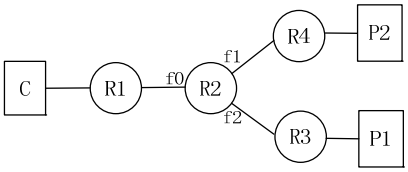
synchronization of the two interests is not necessary. Assume that C1 sends the Interest before C2 does. Before C1 gets the datum, the Interest sending by C2 can always join in the entry in the PIT table of R2, which is generated by the prior Interest. After the C1 gets the datum, the Interest sent by C2 can be satisfied by the cached copy in the Content Store of R2. It is found that this kind of multicast is auto-organized, and there is no need for any extra routing state or control traffic.

Besides this kind of one-to-many distribution mode where many consumers are interested in the same data, NDN also provides some kind of many-to-one distribution mode where many producers publish different data with the same name prefix, and a consumer sends a series of interests with the name prefix to get all these data matching the name prefix. We refer to this kind of one-to-many distribution mode as Enumeration Process. As Figure 1 (b) shows, the two producers, say P2 and P1, publish two data with the two names, ccnx://thu.edu.cn/course-A/homework/sam and ccnx://thu.edu.cn/course-A/homework/alice, and have the same prefix ccnx://thu.edu.cn/course-A/homework/. A consumer, namely C, who wants all the homework of course A, sends an Interest with the name ccnx://thu.edu.cn/course-A/homework/. When the Interest reach NDN router R2, R2 looks up the entry for this prefix in its FIB table (Forwarding Information Base. It is much similar to the FIB of current IP router) and forwards it to R4 and R3 from f1 and f2 respectively. Both P1 and P2 will receive the Interest and respond with its datum respectively. The two data will arrive at R2, and only one of them will be send to the R1 because one interest can just get one datum. Assume that the datum tagged with the name ccnx://thu.edu.cn/course-A/homework/sam is received by C, then C would send another Interest with the same name ccnx://thu.edu.cn/course-A/homework/ but with an attribute Exclude set with the parameter sam which means data with the name constructed by suffixing the interest's name with sam are not viewed as matching this Interest. So, this Interest will get the datum with name ccnx://thu.edu.cn/course-A/homework/alice, which has been cached in R2. Repeating this process, C can get all data tagged with the name prefix ccnx://thu.edu.cn/course-A/homework/. The process will not finish until an Interest gets no datum in an expiration time.

Forth, NDN offers infrastructural support for applications to be designed in a server-less manner. In NDN, named data are the first-class residents and Interests are routed directly according to their names. So for NDN, there is no need to map the wanted data names to their locations. Taking traditional VoIP software based on SIP as an example, the major reason for the existence of a central server is to provide with some kind of name resolution which resolves the human-readable name of a user to current IP address of the host, from which the user register to the server. This kind of complexity of structure and configuration results from a mismatch between the user's goal and the network's means of achieving it [15].



(a) One-to-many



(b) Many-to-one

Figure 1. Simple scenarios for one-to-many and many-to-one distribution mode

Fifth, NDN transmits each piece of data with a signature which is generated by the data's publisher by signing the readable name and its corresponding datum with its public key. The consumer can validate the integrity of the datum received and the association of the datum and its name. Applications can use this signature and some key distribution mechanism based on NDN itself as a foundation to satisfy their own secure demands.

III. SOLVE THE MAJOR CHALLENGES OF CONFEFENCE SYSTEM

In this section, we take the conference system as an example and identify three major challenges facing current conference systems and illustrate how NDN could help to solve these challenges.

IP multicast model is viewed as a scalable and efficient pattern for multi-party communication [6]. But for lack of extensive deployment of IP multicast, designs of conference system based on IP multicast are not accepted widely. Many researchers turned to design conference systems based on centralized server [7]. These designs transfer the scalable problems of endpoint to the server and make the situation even worse for the server must deal with media flows of all the endpoints. In [8][9], it is proposed to construct an application-level multicast overlay over IP to delivery data for conference system. In spite of these solutions, the scalability of conference system is still an open issue. As the VoIP market is growing rapidly, for those who don't want to transfer from traditional phone system to VoIP system, the main concern is the quality of actual VoIP calls [10]. Kushman et al. [10] shows that the qualities of current VoIP systems are unacceptable due to network outages. The main cause was identified as the poor performance of BGP update. Another main concern about conference system as well as many other applications is secure issue. How to keep the privacy and integrity of the calls and allow only granted users to access conference resources is still an open problem in the context of poor secure infrastructure of the Internet. In short, the three major challenges of designing a conference system are i) Scalability, ii) Quality of calls, iii) Security.

With its data-oriented nature, NDN brings enormous potentials and challenges to application design. We argue that NDN provides a substrate for resolve the aforementioned three challenges. First, NDN names data directly instead of naming host and involves Publish/Subscribe paradigm of communication, which make it possible to automatically embrace some kind of multicast providing a scalable and efficient pattern for multi-party communication [6]. Second, NDN does not have routing loops for its data-name-based design [11]. Interest can be forwarded along multiple paths. This feature allows rapid recovery from network outage, as [10] suggests that multi-path routing is a promising direction to

 ISBN: 978-1-61208-183-0

deal with unintelligible quality of VoIP calls caused by BGP update. Third, naming data makes it possible to secure data itself instead of securing the transmission channel. Today's connection-based network architecture does not provide essential infrastructure for securing data, which is the main concern of most applications. As an add-on function, many solutions were proposed to provide various-kind and various-stage security of communication channels. NDN realizes the transition from channel-oriented security to data-oriented security. The task of securing the data can be accomplished by end-to-end cryptographic signatures and encryption (when data secrecy is needed), leaving open only the task of key management among the data sending/receiving parties, but not any channel or boxes in the middle of the data delivery paths [11].

IV. AN EXAMPLE OF COFERENCE SYSTEM DESIGN

In this section, we take the voice conference system as an example to show how this features impact the application design significantly in detail. We identify three main issues of a conference system to be resolved, i) how could a participant know the names of active conferences without centralized server; ii) how could a participant get the name list of other users in a conference; iii) how does a participant get the audio data of other participants in the same conference. We refer to these issues as conference discovery, speaker discovery and voice data distribution respectively. In the following two subsections, some discussions about these three parts are presented, and more details can be found in [14]. After these three main parts, we would discuss some extended features including reliability and congestion control for a conference system and show the real potential of NDN for application design.

A. Conference and Speaker Discovery

Without the existence of central server, a participant needs to communicate with all other conference creators or participants in the same conference for conference discovery or speaker discovery process respectively. These two use cases match the NDN enumeration process aforementioned in Section 2 perfectly. The Interest with the names used for the two processes would be routed by either broadcast or multicast. This kind of multicast in NDN can be achieved by some kind of mechanism where the names like a multicast IP address in that the publishing process resembles the group joining process of IP multicast and the forwarding of Interest resembles data transmission of IP multicast. But it should be noted that this process is used for getting data from multiple parties, and IP multicast is used for sending data to all group members. In contrast, to fetch voice data of other participants, the location-dependent names of participants are used because there is no need for broadcasting Interest. So there would be no additional state imposed to the routing system. Actually the voice data will be efficiently delivered to multiple receivers as Section 2 shows. For conference and speaker discovery and voice data distribution, using separate namespaces makes the system more scalable.

B. Voice Data Distribution

As Section 2 shows, a participant's voice data can go through an automatically-formed spanning tree and arrive at each other participant more efficiently than unicast. This property makes the NDN conference system more scalable than traditional unicast-based conference system. Besides, the producer and consumer of voice data are decoupled in both time and space through Publish/Subscribe communication mode. In terms of time, the producer just publishes its voice data independently and does not need to generate responding datum for the arrival of an Interest designedly. In terms of space, the producer does not know how many and who are receiving its voice data. It can be observed that the Publish/Subscribe mode makes the design simple and efficient. The transmission of real-time stream can be decoupled and appears to be of Publish/Subscribe mode by nature. On the other hand, delivery of voice data is controlled by the consumer in that the consumer controls which chunk of voice data it want to get and how fast these data would be transmitted.

C. Reliability

In this subsection, we discuss the reliability of data distribution of NDN conference system here. Considering the extended function of whiteboard, we could borrow the ideas from literature [16], which designs a reliable framework for IP multicast. IP multicast can be viewed as a special case of Publish/Subscribe communication mode. Joining a group is to express interest in certain subject and delivering data is to publish messages to the interested. The difference is of the granularity in that the NDN makes use of the late-binding technique, but for IP multicast, a receiver keeps a session relationship after joining a group. For IP unicast, the sender has control of data transmission in terms of flow control, reliability, etc. When it comes to reliability of IP multicast, it seems not work well. Floyd et al. [16] shows a transition from sender-based to receiver-based control in the context of reliable multicast due to the fact that the sender cannot keep controlling the transmission any more for so many and delay-diverse receivers. For NDN, a receiver-based reliability mechanism is much more natural. Each receiver is responsible for its reliability of data delivery and keeps independence on correct reception of data.

Besides, as [16] suggests, the "naming in application data units (ADUs)" model works far better for multicast than IP address-based one. NDN architecture provides applications with unique and persistent names, which would eliminate the delay and inefficiency imposed by separate protocol namespace [18]. Furthermore, the performance of retransmission could be improved by data cached in NDN router or other participants who have received the data already. As [18] argues, to design a performance-optimal and efficient transport protocol, some application information (e.g., application data units) should be involved in the protocol design. The concept of networking named data could be viewed as an application of this viewpoint into the network layer, which provides significant efficiency and flexibility for the design of the upper layers.

In summary, with receiver-control mode and application-specified name which are embedded in NDN architecture, reliability can be naturally achieved with some mechanism similar with SRM [16].

D. Congestion Control

The Internet's heterogeneity and scale make multipoint communication design a difficult problem [17]. If a participant generates only one kind of quality of audio data (e.g., with a certain encoding rate), other participants will

 ISBN: 978-1-61208-183-0

have a uniform transmission rate of audio data of the participant. This means some low-capacity regions of network suffer congestion and some high-capacity regions are underutilized. To solve this problem in the context of IP multicast, McCanne et al. [17] proposes a receiver-driven layered solution. This solution can be transplanted into NDN circumstance naturally. The NDN is receiver-driven by nature, and its application-specified name is well-suited for a layered solution. We can give different qualities of audio data different names e.g.

Ccnx://thu.edu.cn/bob/audio/high_quality/seq<20>
Ccnx://thu.edu.cn/bob/audio/low_quality/seq<20>

A participant can try to get the audio data of higher quality periodically. If congestion is detected, it would give up this trial and stays on its original quality-level. This process is of lower cost than that in the context of IP multicast in that joining and leaving an IP multicast group is costly, but for NDN, it is costless for NDN's late-binding property.

V. CONCLUSION

It could be found that the many-to-many data distribution mode of NDN allows multi-communication applications, like the voice conference system, to be designed more naturally and efficiently. By sending Interests with different names of voice data, a conference participant can migrate smoothly from one quality level of voice data of other participants to another according to its bandwidth. Furthermore, the content caching mechanism makes the reliability of multicast transmission mode more simply and efficiently. Our future works include studying the reliability and congestion control of NDN for the voice conference system in more detail and extending the implementation presented in [14] with these functions. We will also attempt to address some limitations of NDN in some special application scenarios as our future work. For example, in the scenario of emergent report such as earthquake alarm, data are generated unpredictably. Therefore, either a long-lived Interest is needed, which would occupy the PIT entry for an extremely long time. Or, applications need to send interests periodically. Both solutions aforementioned seem not to be as efficient as current sender-based IP approaches.

REFERENCES

[1] S. Shenker, "We Dream of GENI: Exploring Radical Network Designs," presentation, CRA Computing Comm unity Consortium, 2007;

[2] V. Jacobson, "If a clean slate is the solution what was the problem", Stanford" Clean Slate" Seminar, 2006.

[3] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard, "Networking named content," ACM CoNEXT '09, 2009, pp 1-12.

[4] T. Koponen, M. Chawla, B. G. Chun, A. Ermolinskiy, K. H. Kim, S. Shenker, and I. Stoica. "A Data-Oriented (and Beyond) Network Architecture," ACM SIGCOMM Computer Communication Review, 2007, vol. 37, no. 4, pp. 181-192.

[5] J. Wang, E. Osterweil, C. Peng, R. Wakikawa, L. Zhang, C. Li and P. Cheng, "Implementing Instant Messaging Using Named Data," Proceedings of the Sixth Asian Internet Engineering Conference, 2010, pp. 40-47.

[6] D. Pendarakis, S. Shi, D. Verma, M. Waldvogel, "ALMI: An Application Level Multicast Infrastructure," Proceedings of the 3rd conference on USENIX Symposium on Internet Technologies and Systems, 2001, pp. 5-5.

[7] J. Rosenberg, "A Framework for Conferencing with the Session Initiation Protocol", RFC 4353, 2006.

[8] C. Luo, J. Li, and S. Li, "DigiMetro-an application-level multicast system for multi-party video conferencing," GLOBECOM'04, 2004, vol. 2, pp. 982-987.

[9] X. Wu, K.K. Dhara, and V. Krishnaswamy, "Enhancing Application-Layer Multicast for P2P Conference," Proc. of IEEE Consumer Communications and Networking Conference, 2007, pp. 986-990.

[10] N. Kushman, S. Kandula, and D. Katabi, "Can you hear me now?! it must be BGP," ACM SIGCOMM Computer Communication Review, 2007, vol. 37, no. 2, pp. 75-84.

[11] M. Meisel, V. Pappas, and L. Zhang, "Ad hoc networking via named data," Proceedings of the fifth ACM international workshop on Mobility in the evolving internet architecture, 2010, pp. 3-8.

[12] M. Särelä, T. Rinta-aho, and S. Tarkoma, "RTFM: Publish/subscribe internetworking architecture," ICT Mobile Summit, 2008.

[13] P.T. Eugster, P.A. Felber, R. Guerraoui, and A.M. Kermarrec, "The many faces of publish/subscribe," ACM Computing Surveys, 2003, vol. 35, no. 2, pp. 114-131.

[14] Zhenkai Zhu, Sen Wang, Xu Yang, Van Jacobson and Lixia Zhang, "ACT: Audio Conference Tool Over Named Data Networks," ACM SIGCOMM Workshop on Information-Centric Networking (ICN 2011), 2011, vol 11.

[15] V. Jacobson, D.K. Smetters, N.H. Briggs, M.F. Plass, P. Stewart, J.D. Thornton, and R. L. Braynard, "VoCCN: voice-over content-centric networks," Proceedings of the 2009 workshop on Re-architecting the internet, 2009, pp. 1-6.

[16] S. Floyd, V. Jacobson, S. McCanne, C.G. Liu, and L. Zhang, "A reliable multicast framework for light-weight sessions and application level framing," ACM SIGCOMM Computer Communication Review. 1995, vol. 25, no. 4, pp. 342-356.

[17] S. McCanne, V. Jacobson, and M. Vetterli, "Receiver-driven layered multicast," ACM SIGCOMM Computer Communication Review, 1996, vol. 26, no. 4, pp. 117-130.

[18] D.D. Clark and D.L. Tennenhouse, "Architectural considerations for a new generation of protocols," ACM SIGCOMM Computer Communication Review, 1990, vol. 20, no. 4, pp. 200-208.

 ISBN: 978-1-61208-183-0

Feedback, Transport Layer Protocols and Buffer Sizing

Shankar Raman*, Shashank Jain* and Gaurav Raina†
India-UK Advanced Technology Centre of Excellence in Next Generation Networks
**Department of Computer Science and Engineering, †Department of Electrical Engineering*
Indian Institute of Technology Madras, Chennai - 600 036, India
Email: mjsraman@cse.iitm.ac.in, shashank@cse.iitm.ac.in, gaurav@ee.iitm.ac.in

***Abstract*—A key aspect of network performance is coupled with the design of transport layer protocols, the choice of feedback from queues, and by the buffer sizing requirements at routers. In this paper, we consider some transport protocols which use different feedback mechanisms to manage their flow and congestion control. We study the performance of these protocols under the influence of different buffer sizes. The transport protocols considered include CUBIC TCP, Compound TCP and an illustrative protocol that could utilize Explicit Congestion Notification (ECN) marks. CUBIC TCP, which is the current default implementation in Linux, uses *packet loss* as the primary feedback signal. Compound TCP, which is the current default implementation in the Windows platform, uses both *packet loss and queuing delay*. In the aforementioned transport protocols, using NS-2 simulations and some analysis, we exhibit that irrespective of the feedback signal used, buffer sizes play a very important role in network performance. In particular, we highlight that even minor variations in buffer size can readily lead to the emergence of limit cycles. These limit cycles tend to destabilize the queue dynamics, induce deterministic oscillations in the packet losses and can degrade link utilization. Using a combination of currently deployed protocols and an illustrative protocol, our work serves to exhibit the importance for a combined study of transport protocols, different feedback mechanisms and sizing router buffers.**

Keywords-Feedback; Transport protocols; Buffer sizing.

I. INTRODUCTION

Transport protocols play an integral part in delivering end-to-end quality of service. However, the design of transport protocols is affected by the choice of feedback mechanisms from the queue. Router buffers, which traditionally have been used to smooth statistical fluctuations in the demand for transmission capacity, also play a key role in providing end-to-end performance. In this paper, we highlight the inter-related nature of transport protocols, feedback and buffers.

A. Buffer sizing

Buffers in routers are a key architectural component of the Internet, and have played an important role in store-and-forward communication networks. Despite their importance, they can also have a detrimental effect by introducing queuing delay and jitter. In the Internet, buffers are currently sized using a rule of thumb which says that each link needs a buffer of size $B = C * \overline{RTT}$, where C is the data rate, and $\overline{RTT}$ is the average round-trip time of the flows passing across the link which is currently taken to be 250 ms [19]. For example, a 10 Gbps router line card needs approximately 10 Gbps * 250 ms = 2.5 Gbits of buffers, which is enough to hold roughly 200k packets. This rule of thumb is clearly not scalable with the growth of transmission capacity. Additionally, such large buffers also have a significant influence on the energy consumption of routers [16].

B. Transport protocols and small buffers

A body of work is emerging that takes a rather radical approach to the issue of buffer sizing: it suggests that it *might* be possible to have buffer sizes of the order of tens of packets. This small buffer sizing rule does not depend on C or $\overline{RTT}$. For work on the development of scaling regimes for queuing delay see [5], for work on TCP see [11], [12], and for a more recent overview see [16]. A key conclusion of [12] is that small buffers have a stabilizing effect on the end-to-end dynamics of Additive Increase Multiplicative Decrease (AIMD) TCP flow control. In essence, in a large bandwidth-delay product environment with small drop-tail buffers, anything larger than a few dozen packets may lead to synchronization effects. Synchronization, in this context, is synonymous with (stable) limit cycles; for definitions and an exposition of the requisite theory, see [10]. The aforementioned analysis was, however, limited to the standard AIMD TCP.

C. Feedback, transport protocols and small buffers

One way to classify transport protocols is via the feedback signals that they use to manage flow and congestion control. The feedback signals that the end-systems may use are queuing delay, packet loss, explicit congestion notification (ECN) marks, or rates. ECN marks are intended to be used in conjunction with transport protocols, and there are numerous proposals for queue management strategies on how to mark packets; for example RED [2].

Recent work has begun to focus on the aspect of sizing router buffers under the influence of different forms of feedback and queue management strategies. For example, the study of rate based feedback with different notions of fairness among the flows was analyzed in [6], [17]. For a study of some queue management schemes, with small

 ISBN: 978-1-61208-183-0

buffers, see [9] and for the impact of delay based transport layer protocols (like FAST TCP [18]), see [13]. Additionally, some recent work has also been carried out on a mixture of real time traffic (open-loop) and TCP traffic (closed-loop) with respect to the issue of sizing router buffers [15].

Today, there is no consensus on the desired feedback mechanism, the transport protocol variant for a given feedback mechanism, or on the optimal rule for next-generation router buffer size. Our work exhibits a relationship between feedback, transport protocols, router buffer sizing and network performance. Using a combination of simulations and some theory, we show how the choice of router buffer size may affect performance irrespective of the feedback used; incorrect buffer sizes may induce the onset of limit cycles.

The rest of the paper is organized as follows. In Section II, we briefly describe CUBIC TCP, Compound TCP and also consider a model for an ECN based transport protocol. In Section III, we conduct simulations with CUBIC and Compound TCP in the Network Simulator (NS-2) [20]. In Section IV, we analyze the ECN based protocol with different resource design functions. In Section V, we present our conclusions and some discussions.

II. Some Transport Protocols

In this section, we describe the variants of transport protocols that have been implemented in Linux and the Windows platforms. CUBIC TCP [3] is currently deployed in Linux and uses loss as the feedback signal. Compound TCP [14] is the current default implementation in the Windows platform, and uses both delay and loss for congestion control. In addition to these TCPs, we also consider a theoretical model of a transport protocol that may use ECN marks.

A. CUBIC TCP

In CUBIC TCP, upon detecting the loss of a packet, the congestion window is reduced by a *multiplicative factor* β, where β is a constant multiplication decrease factor. The window size prior to the reduction is set to W_{max} and the current window is increased using the following cubic window growth function

$$W(t) = C_s(t-K)^3 + W_{max}, \tag{1}$$

where C_s is a parameter called a *scaling factor*, t is the elapsed time since the last window reduction, and K is the time period the above function takes to increase from W to W_{max} when no loss is detected. The functional form of K is given by

$$K = \sqrt[3]{W_{max}\beta/C_s}. \tag{2}$$

If standard TCPs like TCP Reno increase their window size by α per RTT, then the window size of CUBIC in terms of elapsed time is given by

$$W_{tcp(t)} = W_{max}(1-\beta) + \frac{3\beta}{2-\beta}\frac{t}{RTT}, \tag{3}$$

where

$$\alpha = 3\beta/(2-\beta)\cdot \tag{4}$$

Depending on the value of the current window size *(cwnd)*, CUBIC operates in the following three different regimes:

$$cwnd = \begin{cases} W_{tcp(t)} & cwnd < W_{tcp(t)} \\ cwnd + \frac{W(t+RTT)-cwnd}{cwnd} & cwnd < W_{max} \\ \text{probe for new } W_{max} & cwnd > W_{max}. \end{cases} \tag{5}$$

The increased growth rate helps to achieve scalability, whereas the fairness and stability is maintained by forcing an almost linear growth when the window size is far from W_{max}. For further details on the protocol design see [3].

B. Compound TCP

Compound TCP is a loss-based congestion control algorithm with a scalable delay-based component [14]. This additional delay-based component, derived from TCP Vegas, serves for better efficiency, RTT fairness and TCP friendliness. The delay-based component is effective only in the congestion avoidance phase where the sender side congestion window is determined by

$$win = min(cwnd + dwnd, awnd), \tag{6}$$

where $cwnd$ is the normal loss-based component, $dwnd$ (delay window) controls the delay-based component and $awnd$ is the advertised window from the receiver. Compound TCP also maintains the number of backlogged packets in the queue, *Diff*, for every connection.

$$\textit{Diff} = (Expected - Actual) * BaseRTT, \tag{7}$$

where $Expected = WindowSize/BaseRTT$ and $Actual = WindowSize/RTT$. The delay-based component gracefully reduces its window if *diff* $> \gamma$ (the threshold value), i.e. we need at least γ packets in the system to detect an early congestion. The changes in the window size for Compound TCP can be summarized as

$$dwnd(t+1) = \begin{cases} dwnd(t) + (\alpha.win(t)^k - 1) & \textit{diff} < \gamma \\ dwnd(t) - \zeta.\textit{diff} & \textit{diff} \geq \gamma \\ win(t).(1-\beta) - cwnd/2 & \text{loss} \end{cases} \tag{8}$$

where α, β and k are tunable parameters. ζ is a parameter which determines how rapidly the window size should be reduced when early congestion is detected.

 ISBN: 978-1-61208-183-0

C. An ECN-based transport protocol

We outline an illustrative transport protocol, mentioned in [5], and analyse it with different resource design choices. Consider a network with a set J of *resources*. Let a route r be a non-empty subset of J, and write R for the set of possible routes and suppose that route r carries a flow of rate x_r, for each $r \in R$. Consider the following equations

$$\frac{dx_r(t)}{dt} = k_r\Big[w_r - x_r(t)\sum_{j\in r}\mu_j(t)\Big] \tag{9}$$

for $r \in R$, where k_r is the gain factor and

$$\mu_j(t) = p_j\left(\sum_{s:j\in s} x_s(t)\right) \tag{10}$$

for $j \in J$, and where the weights w_r determine the share of the scare resources obtained by the different flows. We can interpret equations (9) and (10) as follows. Suppose that resource j *marks* a proportion $p_j(y)$ of packets with a feedback signal when the total flow through resource j is y. Thus equation (9) corresponds to a rate control algorithm for user r that comprises two components: a steady *increase* at a rate proportional to w_r, and a steady *decrease* at rate proportional to the stream of congestion indication signals received. We now consider two functional forms for the resource. Suppose that

$$p_j(y) = (y/C_j)^{B_j} \tag{11}$$

This form arises if the resource j were to be modelled as a $M/M/1$ queue, with a service rate of C_j packets per unit time, at which a packet is marked with a congestion indication signal if it arrives at the queue to find at least B_j packets already present. This functional form has also been proposed to represent small buffer drop-tail networks while modeling long-lived TCP flows; see [11], [12]. Another simple functional form for the resource could be

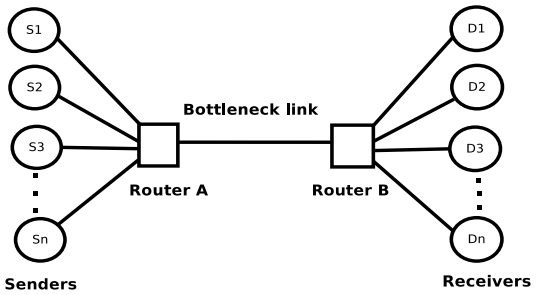
$$p_j(y) = [y - C_j]^+ / y, \tag{12}$$

where $p(.)$ is the proportion of packets overflowing a large buffer. This functional form has been devised to represent drop-tail networks; see [11], [12] and references therein.

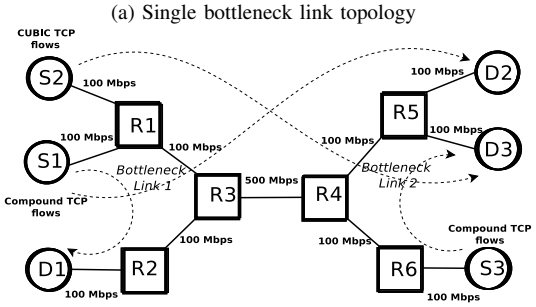
III. Simulations for CUBIC and Compound TCP

Given that both CUBIC and Compound TCPs are implemented today, it is appropriate to perform simulations when both these protocols are present in the network. We highlight some representative simulations over a single and multi-bottleneck topology where we vary buffer sizes, round-trip times and also the traffic mix.

In a previous evaluation, the slow convergence rate of CUBIC TCP was noted [7]. A comparison of CUBIC with Compound TCP [1], [8] has revealed that CUBIC TCP has a propensity to be very aggressive, which readily translates into unfairness towards competing Compound TCP flows. However, neither CUBIC nor Compound TCP have undergone evaluation with respect to the issue of buffer sizing prior to their implementation in Linux and Windows platforms.



(a) Single bottleneck link topology



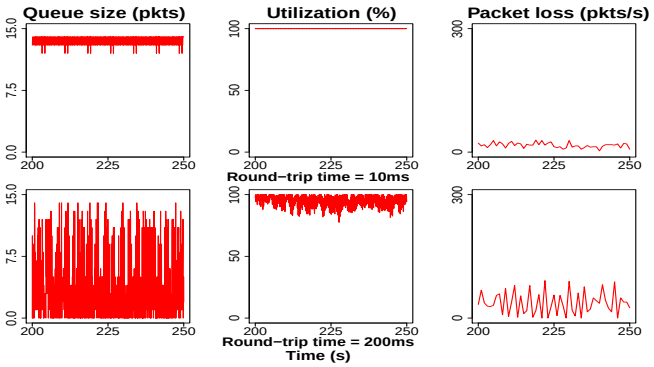
(b) Multi-bottleneck link topology

Figure 1: Simulation set-up for CUBIC and Compound TCPs

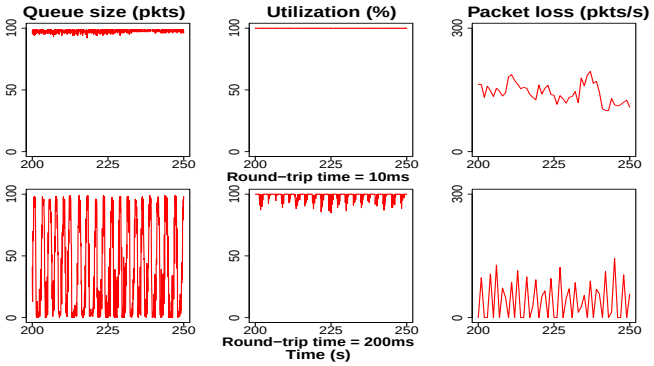
The following parameters are used for simulations: buffer size = 15, 100, $C * \overline{RTT}/\sqrt{N}$ and $C * \overline{RTT}$ packets, round-trip time (RTT) = 10 ms and 200 ms, bottleneck link capacity (C) = 100 Mbps, number of flows (N) = 60, and packet size = 1500 bytes. The currently deployed industry recommendation for $\overline{RTT}$ is 250 ms. Given the proliferation of the Windows platform, we choose a scenario where 80% of the flows are Compound TCP. This ratio between Compound and CUBIC TCP is just representative and a fuller set of experiments are left for further study. The topologies we used are depicted in Figure 1. In the figures, R's, S's and D's refer to the routers, the sources and the destination end-points, respectively. In the multi-bottleneck link topology the dotted lines represent the flows between the source and destination end-points.

The decision to choose small buffers in the range of 15 to 100 packets comes from the previous analysis of AIMD TCP Reno [11], [12]. These papers exhibit that small buffers have a stabilising effect on the end-to-end dynamics of TCP Reno. They also exhibit that even minor variations in buffer size can readily lead to the emergence of stable limit cycles. As new protocols are often designed to out-perform the standard TCP Reno, it is natural to begin an investigation of other transport protocols in similar buffer sizing regimes.

 ISBN: 978-1-61208-183-0

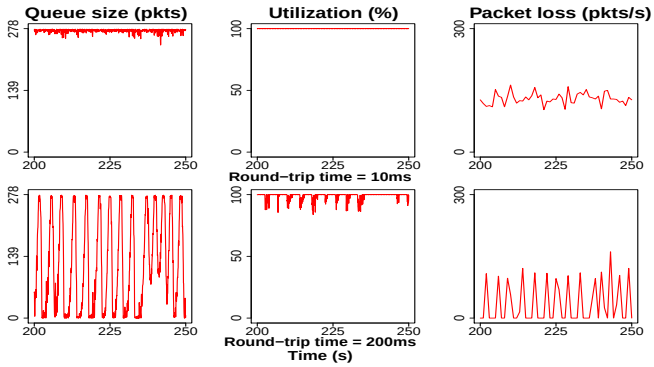


(a) Buffer size = 15 packets

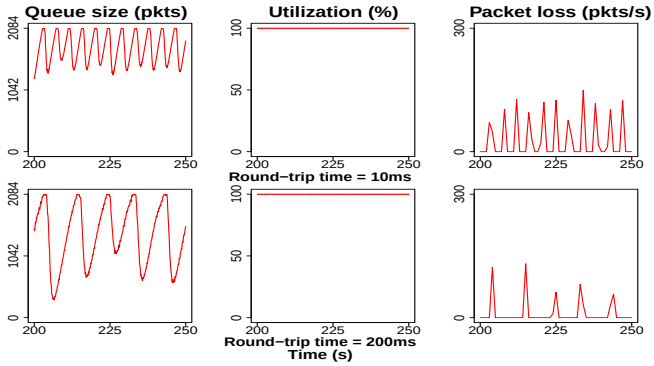


(b) Buffer size = 100 packets

Figure 2: Single bottleneck link with a capacity of 100 Mbps, 60 long-lived flows (80% Compound, 20% CUBIC) with round-trip times of 15, 100 ms.



(a) Buffer size = $\frac{C*\overline{RTT}}{\sqrt{N}}$ packets



(b) Buffer size = $C * \overline{RTT}$ packets

Figure 3: Single bottleneck link with a capacity of 100 Mbps, 60 long-lived flows (80% Compound, 20% CUBIC) with round-trip times of 15, 100 ms.

1) Single bottleneck, long-lived flows: In small buffers, see Figure 2, with 15 packet buffers the queue does not exhibit non-linear instabilities and there is a minor loss in utilization with larger RTTs. With 100 buffers, and with larger RTTs, the emergence of deterministic non-linear oscillations is clearly visible. With larger buffers, see Figure 3, we again witness non-linear oscillations which can also start to hurt utilization.

2) Single bottleneck, long-lived and short-lived flows: We observed the emergence of non-linear oscillations even with minor variations in a small buffer regime with long-lived flows. It is natural to investigate the impact a traffic mix of long-lived flows with HTTP flows in such a regime. Even with this traffic mix, see Figure 4, the non-linear instabilities prevail and there is an impact on utilization.

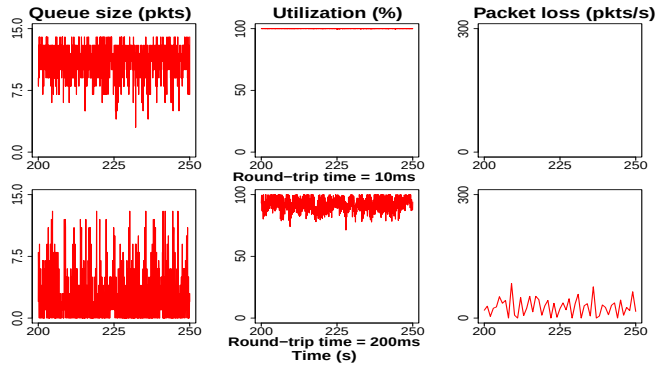
3) Multi-bottleneck link, Long-lived flows: It is natural to investigate the presence of multiple bottlenecks on the qualitative nature of the results observed with a single bottleneck topology. Due to space limitations, we only show results for the case of long-lived flows in a small buffer regime. Even with multiple bottlenecks, with cross traffic, the impact of small buffers on long-lived flow remains the same. From Figure 5, which shows the parameters of interest for Bottleneck Link 2, we again observe that 100 packet buffers induce non-linear oscillations whereas 15 packets have a stabilizing effect on the mix of CUBIC and Compound TCP flows. Further, flows with longer RTTs lead to a reduction in utilization.

So far we focused on the impact of buffer size on the stability of the queue size, but the issue of TCP fairness is also important. The work in [1], [8] showed, using current design rules for buffers, that CUBIC TCP can be unfair to other CUBIC flows as also to other TCP variants. We now briefly comment on the issue of fairness between competing TCP flows in a small buffer regime; see Figure 6. Observe that CUBIC TCP still is unfair to other CUBIC flows, and also to Compound flows. This was despite CUBIC TCP flows being a small proportion of the overall flows.

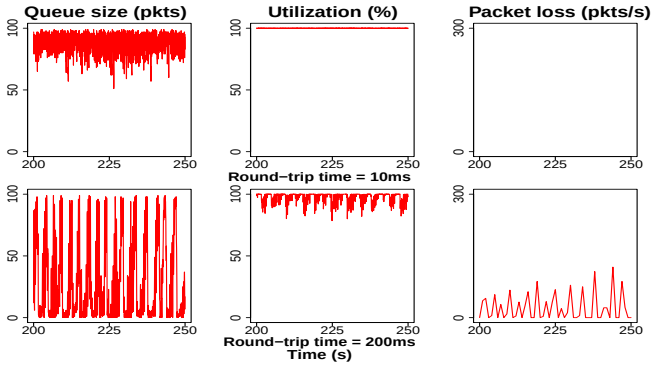
IV. Analysis of ECN-based transport protocol

In this section, we provide some analysis of the ECN-based transport protocol model that was outlined in Section II-C.

 ISBN: 978-1-61208-183-0

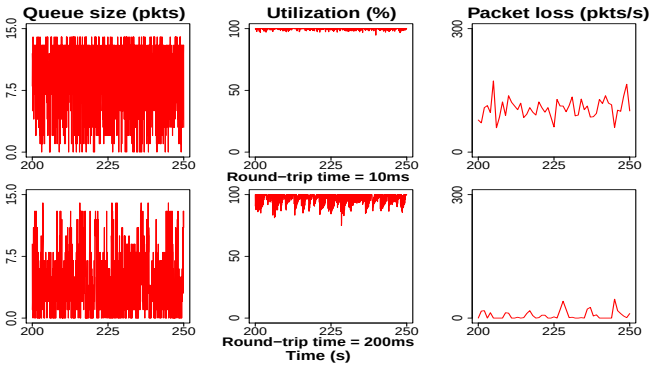


(a) Buffer size = 15 packets

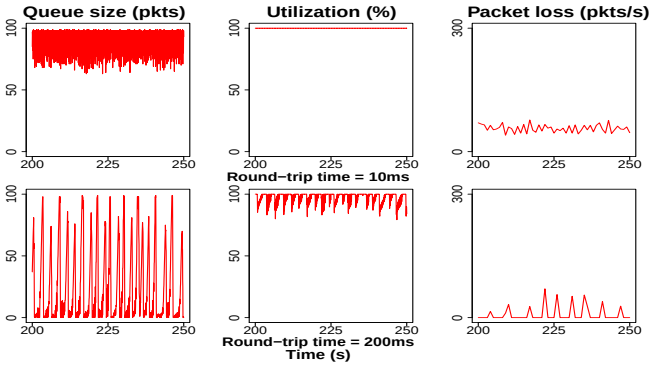


(b) Buffer size = 100 packets

Figure 4: Single bottleneck link with a capacity of 100 Mbps, 60 long-lived flows (80% Compound, 20% CUBIC), 20% short-lived HTTP flows with round-trip times of 15, 100 ms.



(a) Buffer size = 15 packets



(b) Buffer size = 100 packets

Figure 5: Multi-bottleneck link with a capacity of 100 Mbps, 60 long-lived flows (80% Compound, 20% CUBIC), with round-trip times of 15, 100 ms.

Consider a collection of flows all using a single resource, and that all the flows share the same gain parameter κ. Let $x(t) = \sum_r x_r(t)$ be the total flow through the link, and further let $w = \sum_r w_r$ represent the total weight. Additionally, we assume that a congestion indication signal generated at the link is returned to the source after a fixed and common RTT τ. Summing equation (9), and taking the time delay into account, we have

$$\frac{dx(t)}{dt} = \kappa\big(w - x(t-\tau)p\left(x(t-\tau)\right)\big)\cdot \qquad (13)$$

Let x be the equilibrium point of equation (13), let $x(t) = x + u(t)$, and write p and p' for the values of the function $p(.)$ and $p'(.)$ at x. Then, linearising, we get

$$\frac{du(t)}{dt} = \kappa\tau(p + xp')u(t-\tau)\cdot \qquad (14)$$

Using results from [10] we now state the following conditions for stability, and the onset of a Hopf type bifurcation. With the function (11), a necessary and sufficient condition for system (13) to be locally stable is $\kappa\tau p(1+B) < \pi/2$; further, the system undergoes a Hopf type bifurcation at $\kappa\tau p(1+B) = \pi/2$ producing an oscillatory solution with

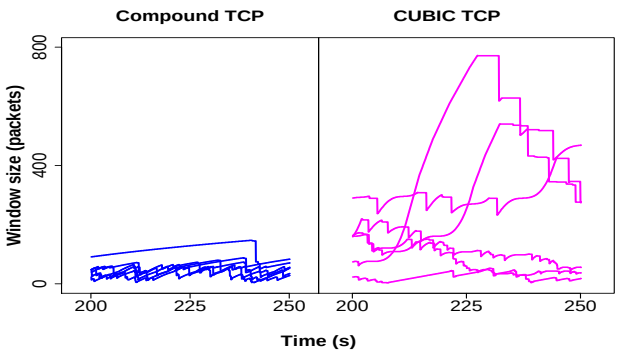


Figure 6: Sample window sizes for Compound and CUBIC. Single bottleneck, buffer 15 packets, and RTT of 200 ms.

period 4τ. This simple example shows us that the larger the value of B, the greater the chance of the transport protocol undergoing a Hopf type bifurcation to induce limit cycles. Such deterministic periodic oscillations were also observed with the protocols we simulated, in a small buffer regime, as we varied the buffer from about 10 to 100 packets. With the

 ISBN: 978-1-61208-183-0

functional form (12), a necessary and sufficient condition for system (13) to be locally stable is $\kappa\tau < \pi/2$; further, the system undergoes a Hopf type bifurcation at $\kappa\tau = \pi/2$ producing an oscillatory solution with period 4τ.

Let us explore another simple functional form for the resource. Let us suppose that the workload arriving at the resource over a time period δ is Gaussian with mean $x\tau$ and variance $x\delta\sigma^2$. Further, suppose that an incoming packet is marked, with an Explicit Congestion Notification bit, if when it arrives the workload that is already present in the queue is larger than the threshold level B. From the stationary distribution for a reflected Brownian motion [4]

$$p(x) = \exp\left(\frac{-2B(C-x)}{x\sigma^2}\right). \quad (15)$$

With the aforementioned resource design function, the condition for the first Hopf bifurcation becomes

$$\kappa\tau(1 + 2BC/(x\sigma^2))p(x) = \pi/2, \quad (16)$$

with period 4τ. Noting that the left-hand side of the above relation is increasing in $w(= xp(x))$, thus for any $w < C$, a condition for local stability is

$$\kappa\tau(1 + 2B/(\sigma^2)) < \pi/2. \quad (17)$$

These conditions clearly serve to highlight the destabilising impact of the threshold B. The threshold may be motivated in terms of buffer size, or in terms of thresholds for marking packets in active queue management schemes. So even if we had a largish buffer size, these models suggest that threshold for marking, or dropping packets may have a destabilising effect on queue dynamics. Now let us explore the design considerations and trade-off that arises for stability. An increase in the factor B causes p' to increase, causes an increased sensitivity in the resource's load. To counter the potentially destabilising effect of this increased sensitivity, there will have to be a reduction in the factor $\kappa\tau$ which represents the sensitivity of the response of the end-systems to the congestion indication signals. Thus, with ECN based transport protocols the form of the resource design again plays an important role for performance.

V. Conclusion and Future Work

Today, the Internet has CUBIC TCP, Compound TCP and large buffers. With growth in communications capacity, router design with large buffers will not be scalable. Using simulations, for CUBIC and Compound, and the analysis of an illustrative ECN-based protocol, we reveal the rather subtle influence that small buffers could have on performance. A key phenomena which arises with even minor variations in buffer size is the emergence of limit cycles. These periodic cycles exhibit the loss of control theoretic stability, they induce periodic oscillations in the queue size and in the losses, and can also reduce link utilization.

Utilization is an important metric, but the network should not strive for a 100% utilization at the cost of large queue sizes which contribute to extra queuing delay. A small reduction in link utilization could well be acceptable if next-generation routers could be made faster or cheaper.

Queuing delay is a key concern for real-time services and is an added justification for having small buffered routers as an architectural consideration for a future Internet. *Packet loss* is important, but only within reason. In fact, loss is the primary feedback signal that is used in the Internet today and TCP has mechanisms to cope with loss. Packet loss can be handled; say, by forward error correction for real time traffic and by appropriate retransmission algorithms for other traffic. On the other hand it is rather difficult to compensate for queuing delay. *Non-linear oscillations* are observable in large bandwidth-delay product environments, when buffer sizes are not dimensioned appropriately. Deterministic, and periodic, queue size fluctuations will lead to bursty losses, they will induce jitter and can hurt link utilization. One really cannot predict their influence on quality of service for end-users. For example, they may prompt time-outs for web transfers, and may also defeat the purpose of forward error correction. We recommend the dimensioning of router buffer sizes to avoid such non-linear oscillations.

Our work shows that to develop a comprehensive understanding of next-generation network performance, we will have to investigate jointly the design of transport layer protocols, feedback from queues, and router buffer sizing.

Acknowledgments

The authors would like to acknowledge the UK EPSRC Digital Economy Programme and the Government of India Department of Science and Technology (DST) for funding given to the IU-ATC. The authors would like to thank Professor V. Kamakoti for extending the facilities of the Reconfigurable Intelligent Systems Engineering (RISE) Lab at IIT-Madras.

References

[1] I. Abdeljaouad, H. Rachidi, S. Fernandes, and A. Karmouch, "Performance analysis of modern TCP variants: A comparison of Cubic, Compound and New Reno", Symposium on Communications, May 2010, pp. 80–83, doi: 10.1109/BSC.2010.5472999.

[2] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance", IEEE/ACM Transactions on Networking, vol. 1, no. 4, August 1993, pp. 397–413, doi: 10.1109/90.251892.

[3] S. Ha, I. Rhee, and L. Xu, "CUBIC: a new TCP-friendly high-speed TCP variant", ACM SIGOPS Operating Systems Review, vol. 42, no. 5, July 2008, pp. 64–74, doi: 10.1145/1400097.1400105.

[4] J. M. Harrison, "Brownian motion and stochastic flow systems", John Wiley & Sons, New York, 1985, ISBN: 0471819395.

 ISBN: 978-1-61208-183-0

[5] F. P. Kelly, “Mathematical modelling of the Internet”, Mathematics Unlimited - 2001 and Beyond, Springer-Verlag, 2001, pp. 685–702, ISBN: 9783540669135.

[6] F. P. Kelly, G. Raina, and T. Voice, “Stability and fairness of explicit congestion control with small buffers”, Computer Communication Review, vol. 38, no. 3, July 2008, pp. 51–62, doi: 10.1145/1384609.1384615.

[7] D. Leith, R.N. Shorten, and G. McCullagh, “Experimental evaluation of Cubic-TCP”, Sixth International Workshop on Protocols for Fast Long-Distance Networks, March 2008, pp. 1–9, doi: 10.1.1.64.9364.

[8] K. Munir, M. Welzl, and D. Damjanovic, “Linux beats windows!–or the worrying evolution of TCP in common operating systems”, Sixth International Workshop on Protocols for Fast Long-Distance Networks, 2007, pp. 43–48, doi: 10.1.1.69.905.

[9] G. Patil, S. McClean, and G. Raina, “Drop tail and RED queue management with small buffers: stability and Hopf bifurcation”, ICTACT Journal on Communication Technology, vol. 2, no. 2, June 2011, pp. 339–344, ISSN: 2229-6948(ONLINE).

[10] G. Raina, “Local bifurcation analysis of some dual congestion control algorithms”, IEEE Transactions on Automatic Control, vol. 50, no. 8, August 2005, pp. 1135–1146, doi: 10.1109/TAC.2005.852566.

[11] G. Raina, D. Towsley, and D. Wischik, “Part II: Control theory for buffer sizing”, ACM SIGCOMM Computer Communication Review, vol. 35, no. 3, July 2005, pp. 79–82, doi: 10.1145/1070873.1070885.

[12] G. Raina and D. Wischik, “Buffer sizes for large multiplexers: TCP queuing theory and instability analysis”, Proceedings of Euro-NGI Conference on Next Generation Internet, April 2005, pp. 173–180, doi: 10.1109/NGI.2005.1431663.

[13] P. Raja and G. Raina, “Delay based feedback, transport protocols and small buffers”, ACM/IEEE International Workshop on Quality of Service (IWQoS), San Jose, June 2011, pp. 1–3, doi: 10.1109/IWQOS.2011.5931330.

[14] K. T. J. Song, Q. Zhang, and M. Sridharan, “Compound TCP: A scalable and TCP-friendly congestion control for high-speed networks”, 4th International Workshop on Protocols for Fast Long-Distance Networks, 2006, pp. 1–8, doi: 10.1.1.130.1595.

[15] A. Vishwanath, V. Sivaraman, and G. N. Rouskas, “Anomalous loss performance for mixed real-time and TCP traffic in routers with very small buffers”, IEEE/ACM Transactions on Networking, vol. 19, no. 4, August 2011, pp. 933–946, doi: 10.1109/TNET.2010.2091721.

[16] A. Vishwanath, V. Sivaraman, and M. Thottan, “Perspectives on router buffer sizing: recent results and open problems”, ACM SIGCOMM Computer Communication Review, vol. 39, no. 2, April 2009, pp. 34–39, doi: 10.1145/1517480.1517487.

[17] T. Voice and G. Raina, “Stability analysis of a max-min fair rate control protocol (rcp) in a small buffer regime”, IEEE Transaction on Automatic Control, vol. 54, no. 8, August 2009, pp. 1908–1913, doi: 10.1109/TAC.2009.2022115.

[18] D. X. Wei, C. Jin, S. H. Low, and S. Hegde, “FAST TCP: motivation, architecture, algorithms, performance”, IEEE/ACM Transactions on Networking, vol. 14, no. 6, December 2006, pp. 1246–1259, doi: 10.1109/TNET.2006.886335.

[19] D. Wischik and N. McKeown, “Part I: Buffer sizes for core routers”, ACM SIGCOMM Computer Communication Review, vol. 35, no. 3, July 2005, pp. 75–78, doi: 10.1145/1070873.1070884.

[20] NS2, [Online]. Available: *http://www.isi.edu/nsnam/ns/*, [Accessed: 20th December 2011].

 ISBN: 978-1-61208-183-0

A Generic Service Model for QoS Management

Service Management in Next Generation Network

Tatiana Aubonnet
Département Informatique
Conservatoire National des Arts et Métiers, CEDRIC
Paris, France
tatiana.aubonnet@cnam.fr

Noemie Simoni
Department INFRES
TELECOM ParisTech, LTCI, UMR 5141 CNRS
Paris, France
simoni@telecom-paristech.fr

***Abstract*—We introduce a generic service model aiming at ensuring Quality of Service management. This modelling approach has been proposed by the Next Generation Network and Service Management project. A fairly high integration level of the tools has been reached using object-oriented paradigm and Model Driven Interoperability approach. We provide an example of Quality of Service management through a Service Level Specification of the Virtual Private Network service. We consider important points to reflect the complexity introduced by the Service Level Agreement and Quality of Service management: reaction model, cooperation model and co-ordination model.**

Keywords-QoS management; SLS; proactive SLA; VPN

I. INTRODUCTION

Today, new services have to be rapidly deployed upon various types of networks. Moreover, the provisioning and the assurance of a wide range of services depends on the orchestration of heterogeneous, widely distributed software components, which can be owned by different service providers and operate over diverse networks. In such a scenario, designing and providing complex, value-added services, ensuring their nominal quality levels with traditional, service deployment, provisioning, monitoring and management becomes increasingly difficult and costly.

To answer this problem, one possible key of the service management is to build new services based on SLS (*Service Level Specification*) templates. According to the object-oriented concept (OMG standards) and MDI (*Model Driven Interoperability*) approach [2], we propose the object model description of the SLS template for QoS (*Quality of Service)* management. The main advantages of the object-oriented approach are the modelling, the overall behaviour of the system and the flexibility which permit modularity, portability, re-usability and easily extensible object classes.

Our contribution to this problem is to take into account three management responsibilities: user, application and network. Additionally, we use the same modelling to which we add the following models: the *co-ordination model* which addresses the dynamic management process by identifying the different steps which should be taken in a running (changing) context; the *interaction behaviour (reaction) model* further which specifies the autonomous degree of the distributed components (delegated agents: passive, active, interactive, proactive). By assigning the proactive behaviour type to the delegated agent according to different cases, more proactive SLA (*Service Level Agreement*) can be obtained for the QoS management.

We present in this paper the feedback of our experience on the QoS dynamic management. This paper is organized as follows. The SLS context for QoS management is described in Section 2. Section 3 presents characteristics of existing SLS templates. Section 4 is devoted to our propositions for generic service model, the specification of a VPN service is developed as an example. We give an example of VPN architectural model in Section 5. Our propositions for a proactive SLA are presented in Section 6. Finally, in Section 7, we exhibit the advantages of our approach in Next Generation Network.

II. SLS TEMPLATES USAGE

A Service Level Agreement (SLA) is a formal negotiated agreement between two parties. It is designed to create a common understanding about services, priorities, responsibilities, etc. [25]. A Service Offer or a Commercial Offer may by a set of elementary services.

A Service Level Specification (SLS) is the technical part of a SLA. More formally it has been defined in [3] as a protocol independent representation of a set of technical parameters and their associated semantics that describe the transport service that a (packet) flow is to receive over the transport domain, between ingress and egress interfaces.

The TMF (*TeleManagement Forum*) SLA Description corresponds to the SLS [14], the SLA template corresponds to the SLS template (Figure 1). A SLS template is associated to an elementary service. The SLS is a totally instantiated SLS template that can be used to provision, activate and monitor the corresponding elementary service. The question will remain in the latter case on how to manage the services consistency at the provisioning stage (synchronisation, rollback in case one of the services is unavailable, subscription sequence, etc.) and for assurance (correlation of elementary services alarms on the service offer, execution of proactive and reactive maintenance activities).

 ISBN: 978-1-61208-183-0

A structure of information blocks that can be seen as concrete classes in the oriented-object paradigm.

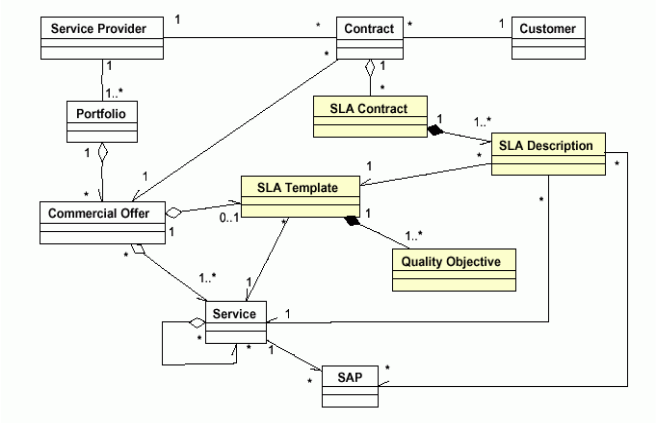


Figure 1. TMF SLA Model.

Moreover, the set of QoS parameters that the service should follow and that should be monitored is instantiated.

III. CHARACTERISTICS OF SLS TEMPLATES

In this section, we present a panorama concerning the SLS templates in different projects: TEQUILA consortium (Traffic Engineering for Quality of Service in the Internet, at Large Scale), Eurescom P1008 project "Inter-operator interface for ensuring end-to-end IP QoS" and Eurescom P1103 project "Inter-Operator IP QoS Framework - ToIP and UMTS Case Studies" [8, 9, 10, 11]. In these projects the template is given as an example and SLS negotiation is described. The basic information to be included in SLSs, and lists a set of basic parameters, which will actually compose the elementary contents of an SLS. The common characteristics of existing SLS templates are:

- Scope – topological region (ingress, egress) interfaces.
- Flow description– SLS is per flow (diffserv info, source info, destination info, app info).
- Traffic description– test if in- or out-of-profile [peak rate, MTU, bkt depth].
- Excess treatment – how to handle out-of-profile traffic [dropped, shaped, remarked].
- Performance Parameters – service guarantees the network offers customer [delay, jitter, pkt loss and throughput].
- Service Schedule – start time and end time – i.e., when the service is available.
- Reliability – downtime and time to repair.
- Others Parameters – route, reporting guarantees, security etc.

We analyzed these characteristics and the documents of these projects and in the following section we propose a *generic service model*, i.e., a template generic compatible with the different next generation services.

IV. A GENERIC SERVICE MODEL

In this section, we consign our propositions for a generic service model. We will present our propositions for:

- Information model (Section A).
- The different levels of visibility for the SLS description (Section B).
- QoS model (Section C, D).

The information model proposed in Section 4.1 is used to analyze the NGNSM SLS template, in order to obtain a generic NGNSM SLS.

A. *Information Model*

The Information Model (IM) is an approach to the management of systems and networks that applies the basic structuring and conceptualisation techniques of the object-oriented paradigm. The approach uses a uniform modelling formalism that-together with the basic repertoire of object-oriented constructs, supports the cooperative development of an object-oriented schema across multiple organizations. Ideally, information used to perform tasks is organized or structured to allow disparate groups of people to use it. This can be accomplished by developing a model or representation of the details required by people working within a particular domain. Such an approach can be referred to as an information model. An information model requires a set of legal statement types or syntax to capture the representation, and a collection of actual expressions necessary to manage common aspects of the domain of QoS management.

This section describes a generic QoS information model object-oriented. This model includes expressions for common elements that must be clearly presented to management applications. The purpose of the Information Model is to give the structure to the management information and to model management aspects of the related resources [16]. The information model deals with managed objects which provide abstract views of the physical and logical resources for the purposes of management. It provides guidelines for describing the logical structure of the managed objects and other pertinent management information about such objects.

A generic information model is essential to the generation of uniform fault, configuration, performance, security, and accounting management which can be applied to the heterogeneous and distributed environment. On the basis of analysing and comparing with the existing work which has been done in [15, 16] the ENST has proposed an information model.

The information model is presented as a set of structured classes of objects in *different levels of visibility*. The class Network Element (NE), which represents each network objects and is the root of this logical structure, is described as an element consisting of:

- *Network Elements* (v) which are in the same level (v) with the considered NE.
- *Network Elements (v-1)*, which are in the lower level(v-1) and provide *a service to the level* (v).

 ISBN: 978-1-61208-183-0

- *Architecture Element*, whose behaviour is expressed with the help of static and dynamic properties.
- A *service class*, which is used to express the service it offers (role *server*) and the operations called by the element (role *client*).

Precisely, with the managed *object model* [23] it is possible to set a managed object as a delegated autonomous agent, as each managed object is provided, in addition to its basic *Operational Service Interface*, with a *Management Interface*. This interface can be very well used for today's management where simply there are notification transits and sometimes tunings of QoS parameters. But it can also, and especially, be used for our distributed perception of QoS management. It makes it possible, indeed, to set management rules and to endow managed objects with some delegated intelligence in order to have some distributed LDPs (*Local Management Decision Points*), thus avoiding depending solely on a unique MDP (*Management Decision Point*). These different behaviours correspond to different object status that appears in the management class description in the information model. QoS information model is dependent on QoS instrumentation, which could by done during the design phase.

B. *Different Levels of the SLS Description*

The information model described in the preceding section is used to analyze the NGNSM SLS template, in order to obtain a generic NGNSM SLS. Initially we analyzed the report of Alcatel "SLS template and principles" and in the present section, we will present our propositions for modelling. Our first proposition contains one of our important rules: the different levels of visibility. Considering all this we would recommend, it would be necessary to have the QoS constraints for each level and the QoS metrology associated to handle the QoS contracted.

We proposed the following levels for the description of information:

- Generic Service Level (for the SLA).
- Traffic flow level.
- Network connectivity level.

The QoS per flow is proposed for the traffic flow level (Figure 2).

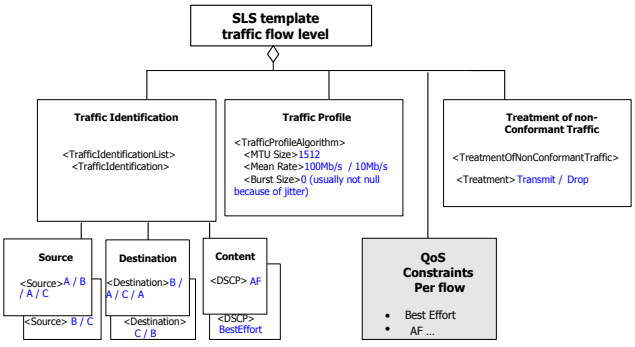


Figure 2. IP/VPN information model : traffic flow level

The information model in these different levels, which is object-oriented, provides an abstraction of the resources and flexibility of the development. The link between the levels of visibility is made thanks to the architectural model (see Section 5) and to the aggregation of levels V-i.

C. *QoS Model : Principles*

Always on the basis of the document [20] we analyze the SLS template for Generic Service Level. Our "grid of reading" is always done through our conceptual tools [1, 12, 13, 16, 17, 21].

QoS model provides the basic support for organizing management activities. In the Figure 3 we introduce some of the propositions for QoS model:

- As much as it is necessary to validate its generic information model by checking that all the applications will find information which they need, as much as it is necessary that the information model is independent of the applications. This is why, it is necessary to have a generic terminology and to choice a good vocabulary (e.g., Provisioning, Monitoring, etc.). Nevertheless, if we want to mention some application, we propose the "application QoS scope" (Figure 3).
- We think that the "Provisioning QoS" which contains information describing the QoS of service used for the provisioning process, represents *contract QoS information* it must be monitored. So "provisioning QoS" and "Monitoring QoS" contain the parameters which we indicate by QoS "Design_value".
- Commitment: Parameter that should be monitored with all the needed configuration in term of assurance is the parameter which we indicate by QoS "thresholds_value" [19].

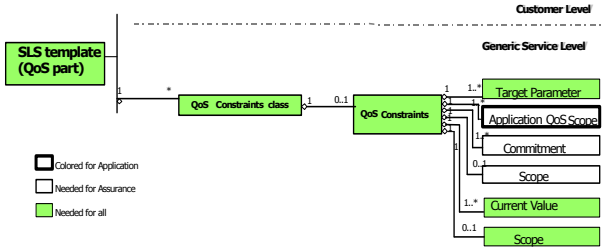


Figure 3. Propositions for SLS Template (QoS part) in Generic Service Level

D. *QoS Model : Parameters*

According to the description given in [5], the following set of eight QoS criteria are needed for a comprehensive QoS appraisal: Availability, Fidelity, Capability, Speed, Reliability, Flexibility, Usability and Security. Among the eight listed above, only four are essential to describe the behaviour of the service: speed, fidelity/accuracy, capacity and availability [7]; all of which will be taken into consideration.

 ISBN: 978-1-61208-183-0

Service Components / Medium	QoS criteria and parameters					QoS class Y.1541	CoS UMTS	QoS class G1010	PHB	QoS criteria depending classes INTRADIFF
	Delay	*Delay variation*	*Fidelity (Information loss)*	*Capacity*	*Availability*					
Interactive games /data	< 200 ms	U	Zero	DBW	UAT	Class 2	Interactive	EI Interactive	AF1.1	CoS 6.4
Telecontrol / Data	< 250 ms	U	Zero	DBW	UAI	Class 2	Interactive	EI Interactive	AF1.1	CoS 6.4
Telnet / Data	< 200 ms	U	Zero	DBW	UAI	Class 2	Interactive	EI Interactive	AF2	CoS 6.4
Video TeleConfer. service (VTC) / Video	<150 ms	U	Error tolerant PLR < 1%	DBW	UAT	Class 0	Interactive	ET Interactive	AF1.2	CoS 6.2
	400 ms with echo control	U	Error tolerant PLR < 1%	DBW 16-384 kbit/s	UAT	Class 1	Interactive	ET Interactive	AF2	CoS 6.3
Audio-conference /Audio	<150 ms	< 1 ms	Error tolerant PLR < 3%	DBW 4-64kbit/s	UAT	Class 0	Convers.	ET Interactive	EF	CoS 6.0
	400 ms with echo control	< 1 ms	Error tolerant PLR < 3%	DBW 4-64kbit/s	UAT	Class 1	Convers.	ET Interactive	EF	CoS 6.1
Telephone service / Audio	<150 ms	< 1 ms	Error tolerant PLR < 3%	DBW	UAT	Class 0	Convers.	ET Interactive	EF	CoS 6.0
	400 ms with echo control	< 1 ms	Error tolerant PLR < 3%	DBW	UAT	Class 1	Convers.	ET Interactive	EF	CoS 6.1
Voice messaging record / Audio and playback / Audio	< 2 s for record	< 1 ms	Error tolerant PLR < 3%	DBW 4-32kbit/s	UAT	Class 1	Interactive	ET Responsive	AF3.2	CoS 5.2
	< 1 s for playback								AF3.2	CoS 5.0
Electronic mail SMTP/POP server access / Data	< 2 s (< 4 s/page acceptable)	U	Zero	VBW	UAT	Class 4		EI Responsive	AF3.1	CoS 5.2
Web Browsing / Data	<2 s/page (< 4 s/page acceptable)	U	Zero	VBW	UAT	Class 3		EI Responsive	AF3.1	CoS 5.2

TABLE I. QOS PARAMETERS FOR EACH SERVICE COMPONENT

Table 1 gives an example of QoS requirements for the above services and a quantitative comparison between the proposed model and other SLS templates models. The difference consists in using only one model. The QoS agent is included in each component and it manages QoS according to the four criteria defined for the current value: Delay / Delay variation, Fidelity (Information loss) Capacity, Availability. Each one these of these criteria should be expressed in quantifiable and measurable parameters (see five columns of the QoS criteria and parameters).

A state-of-the-art effort has been performed in order to situate this model with respect to other generic models of the international community (ITU-T M3100 [15], ETSI GOM [6], TINA-C NRIM [24]) and to propose our SLS template model which is in this context instantiated to the VPN service.

V. ARCHITECTURAL MODEL: VPN APPLICATION

The following section is the connection between the levels of visibilities. Indeed, it is necessary to be able to have the traceability between the levels. Our answer is given by the architectural model which translates the aggregation and the co-operation of the whole of the network components.

In this section, we well examine our proposition of architectural model for end to end QoS, by using DiffServ/IP/VPN case study [4, 22]. After having introduced our QoS model in the previous section, it would be interesting to explore its capabilities and contributions through the case study on DiffServ/ IP / VPN. We consider for this purpose a distributed system consisting of a carrier's network built from multiple DiffServ [12] domains. The network is intended to provide customers with differentiated services.

In IP, VPN service relies on the VR (Virtual Routing) functionality that may include tunneling (encapsulation) and securing (IPSec). In addition to this functionality, the DiffServ VPN region handles, through the IDC (Inter Domain Connection) function, differentiation interoperation among the DiffServ domains regarding traffic aggregation (TA), traffic conditioning (TC) and aggregate forwarding (PHB). A domain PHB shares to the bearer IP network elements IPF (IP Forwarding) and IPR (IP Routing). Thus, the manageable distributed components of the system are VRs, IDCs, TAs, TCs, PHBs, IPFs and IPRs. This leads to the network abstract model depicted in Figure 4.

 ISBN: 978-1-61208-183-0

In accordance with the <Node, Link, Network> abstract model [23], we see that the DiffServ VPN is a network resource composed of nodes (VR, IDC) and links (VPN link) of the same visibility level. It relies on DiffServ domains, which are networks of lower visibility level. Each DiffServ domain is, at its turn, composed of nodes and links of the same visibility level and relies on a network of lower visibility level. We have assigned DiffServ domain components TA, TC and PHB to different visibility levels to be able to make accurate decisions according the services performed by each one of them.

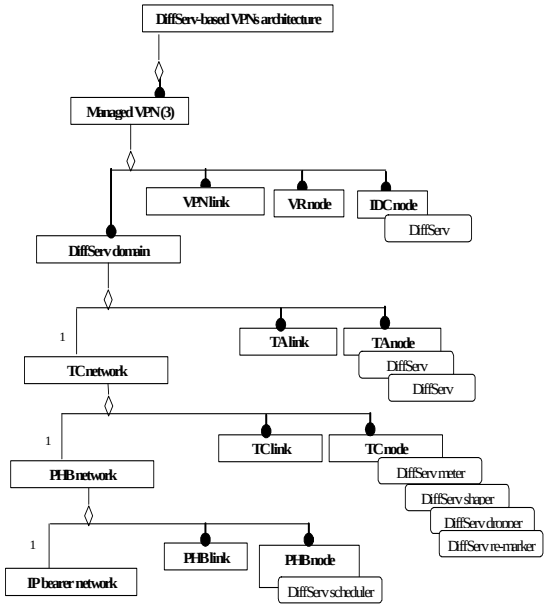


Figure 4. VPN architectural model

Each DiffServ domain (and its sub-networks) relies on the IP bearer network. Visibility level IP and visibility levels below than IP are not developed here. Application of the managed object model leads to the following managed objects: VR nodes, IDC nodes, TA nodes, TC nodes, PHB nodes, IPF nodes and IPR nodes. Note that since manageable objects belong to different visibility levels, some of the managed nodes may share the same physical equipment.

VI. A PROACTIVE SLA

In this section, the request is: how to define a proactive SLA between a service provider and a user. To answer this request, we present:

- Modelling supports with our view point for SLA/QoS modelling (Section A).
- Different models to confer the proactively capacity (Section B).

The first answer is the reaction capacity during the operating phase, i.e., to meet the dynamics requirements. Then the possibility *to anticipate* and to decide in an autonomous way, the nodes will confer to them the proactively capacity.

A. A View Point for SLA/QoS Modelling

Until now, we considered the SLA as the user QoS request with respect to his applications. However the nature of these applications induces their own constraints. This is why, we think that to take into account the user "desired", it is necessary to pass at the higher visibility level, above service visibility level.

We can say that the service visibility level is constrained by application dependent or ISP dependent where as the user visibility level (SLA) could be:

- Agenda dependent.
- Localization dependent,
- Terminal dependent, etc.

In other words, on the service level we have the decisions which rise of the responsibility for the applications. On the network level we have the tactical decisions which concern the responsibility for the network and thus for the operator. Whereas, for the user level we have the strategic and organizational type decisions, concerned with the enterprise and its users responsibilities.

Therefore, we propose to consider the SLA/user level as a "service" and to keep the same model and the same modelling support for other levels services.

B. Our propositions for proactive SLA

To fulfil the requirements indicated in the precedent sections, a powerful and flexible approach should be adopted. In accordance with these objectives, in the present section, we propose the different sub-models for dynamic SLA/QoS :

- *Reaction model* which applies in every node and which classifies four types of object behaviours reflecting different autonomous levels.
- *Co-operation model* which identifies the roles engaging in the distributed management activities, as well as the relations among them.
- *Co-ordination model*, which is proposed to provide a means to support dynamic management to guarantee an end-to-end QoS.

Reaction model (behaviour model)

Management tasks are performed via the interactions among the objects. The interactions between the objects are performed by sending messages from one to another. The different behaviours exhibited by the objects during their interactions. Specifically, four types of interaction behaviours are identified: passive, active, interactive, pro-active.

A passive object encapsulates some resource and a set of routines and operations that can be performed on the resource. It provides services which are used by one or more active objects. A passive object can only be involved in the manager-agent relation, and plays the role of agent. All the manageable objects should be at least passive.

An active object performs some function and may also encapsulate some resource and the operations for accessing it, but it may invoke operations on other objects. It can be

 ISBN: 978-1-61208-183-0

used to describe the object behaviour exhibited in the manager-agent relation. An example of this type of behaviour occurs when an agent is requested to perform an m-action by a manager using the protocol CMIP (Common Management Information Protocol) defined in [18]. An interactive object describes the interaction behaviour of an active object in which the object has needed to obtain interactively complementary information in order to continue the on-going process. It can be used in the negotiation between the manager of managers and a manager, or in the peer-to-peer relation such as the relation between two managers who are not in the same domain. It also can be applied in the manager-agent relation.

A proactive object describes the interaction behaviour of an active object in which the object, who is highly-autonomous, does not simply act in response to their environment stimulus (changes), they are able to exhibit goal-directed behaviour by taking the initiative thereby reacting to indicators rather than reacting to severe problems as perceived by the user. This is the case where a managed object can automatically detect problems and find the pre-determined solutions when an event occurs without the manager's intervention, allowing for network self-healing. The proactive agent can be used to maintain the QoS dynamically.

A fundamental approach to achieve the proactive management is to characterize carefully different problem conditions in the network and to address appropriately their resolution for recovering from complicated situations or situations that require higher levels of reasoning and the correlation of multiple, seemingly, disparate problem conditions.

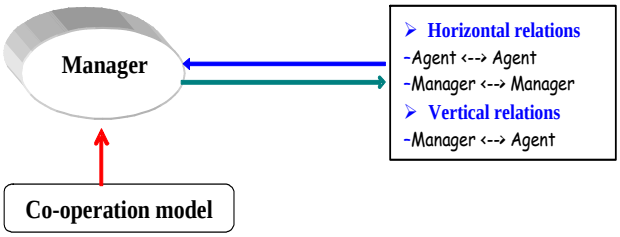
Under different conditions (e.g., in different contexts or when receiving different stimuli), the object behaviour can change among status of passive, active, interactive, and pro-active. These four types of interaction behaviours outline how the network components support the management policies in order to maintain the contracted QoS, especially to contribute to the dynamic QoS management.

Cooperation model (relation model)

Managing distributed systems introduce more complexity. Management responsibilities are structured and partitioned to the sub-systems. Each sub-system is responsible for only a local portion of the overall area. In order to reflect these above characteristics, the roles and their relation model are needed to identify the roles of an entity involved in the management activities and the relations between these entities, which is shown in Figure 5.

According to their different responsibilities taken in the management, four types of roles. These objects can be:

- Manager is used to refer to any entity, human or automated, that can perform management activities such as control, co-ordination and monitoring.
- Manager of Managers (MoM), similar to manager, but in a higher-level in comparison with other managers;
- Agent refers to any entity that provides the access and performs the operations requested by the manager on the managed objects. It reflects to the manager a view of these objects and sends notifications reflecting the behaviour of these objects
- Managed objects (Mos) provide abstract representations of the managed resources. Managed objects may be organized into sets called management domains as a result of organizational requirements. These domains achieve a partitioning of the management environment based on functional areas or according to geographical, technical or organizational criteria.



<Agent <--> Agent, Manager <--> Manger, Manager <--> Agent>

Figure 5. A cooperation model

This supports the distribution and delegation of management functionality and also supports co-operation between different components of the management infrastructure. It is, therefore, not only possible to delegate functionality from managing systems (managers) to managed systems (agents) but also between managed systems. Another point is that the roles participating in management activities are subject to dynamic changes: roles like manager and agent are temporary and bound to the tasks to be done. The change between these two roles results from the relations with other entities during the management activities.

Co-ordination model (organization model)

The co-operative management process can be represented by the co-ordination model which contains seven phases:

- Request
- Translate
- Hop-by-hop negotiate
- Accord
- Reject
- Supervise
- Re-negotiate

The objective of the management is to guarantee the end-to-end QoS required by the user. This is achieved by managing the co-operative management process among the individual object QoS. The QoS is requested and then translated into comprehensible parameters (QoS parameters). All the negotiation results should be reported to the corresponding responsibility level according to management policies.

 ISBN: 978-1-61208-183-0

VII. CONCLUSION

In today's deregulated and competitive market, telecom operators and ISPs need to be able to define their individual management goals and to adjust their individual decisions to meet their specific needs while respecting the general management agreement that governs their global cooperation.

Some ideas from them, especially the idea of the integration of the information model in different levels of abstraction to re-use the information and to provide a template generic compatible with the different NGN services.

To enable implementation of consistent end-to-end QoS for such environments, the QoS model applied in this article provides capabilities to structure and to partition management in a large distributed system as well as to adapt flexibly to changing requirements:

- Through the managed object model and the network abstract model, it makes it possible to organize the management system into distributed domains and to support dynamic QoS management by distributing management tasks and decisions among the system domains.
- Through the architectural model it provides the traceability.

In this article, we considered two important points to reflect the complexity introduced by the SLA/QoS management:

- We proposed to take into account the SLS template based object-oriented paradigm.
- We used the same modelling to which we added the following models: *interaction behaviour (reaction) model, cooperation model and co-ordination model.*

By assigning the proactive behaviour type to the delegated agent according to different cases, the proactive SLA can be obtained for the QoS dynamic management.

ACKNOWLEDGMENT

We would like to thank partners Alcatel: Gilles Désoblin, Olivier Martinot and Véronique Daurensan for their fruitful participation in NGNSM (*Next Generation Network and Service Management*) project, providing a SLS reference model and support of the tool suite.

We would also like to thank Michel Cotten and Kirill Polishchouk for their help and advice in finalizing this article.

REFERENCES

[1] T. Aubonnet and N. Simoni, “PILOTE: A Service Creation Environment in Next Generation Network,” Proceedings IEEE Intelligent Network Workshop, IN’2001, pp. 36-40, Mai 6-9, 2001, Boston, USA.

[2] J. Bézivin, R.M. Soley, and A. Vallecillo, “Model-Driven Interoperability: MDI 2010,” Proceedings of The First International Workshop on Model-Driven Interoperability, MDI 2010, pp. 147-149, October 5, 2010, Oslo.

[3] E. Bouillet, D. Mitra, and K. Ramakrishnan, “The Structure and Management of Service Level Agreement in Networks,” IEEE Communications Magazine 41(7), pp. 102–109, 2003.

[4] M. Conti, A. Hasani, and B. Crispo “Virtual private social networks,” Proceedings of The First ACM conference on Data and application security and privacy, CODASPY 2011, February 21-23, 2011, San Antonio, USA.

[5] ETSI EG 202 009-1, “User Group; Quality of Telecom Services; Part 1 : Methodology for identification of parameters relevant to the Users,” 2009.

[6] ETSI NA 43316, “Telecommunications Management Network (TMN) Generic Managed Object Class Library for the Network Level View,” Mars 1995.

[7] ETSI TR 000 029-1 V0.1.8, “User Group; End-to-end QoS management at the Network Interfaces; Part 1: User’s E2E QoS – Analysis of the NGN,” 2009.

[8] EURESCOM IST project TEQUILA “Traffic Engineering for Quality of Service in the Internet, at Large Scale,” April 2000.

[9] EURESCOM, Project P1008 “Specification of Inter-domain Quality of Service Management Interfaces,” May 2001.

[10] EURESCOM, Project P1103 “Inter-Operator IP QoS Framework - ToIP and UMTS Case Studies,” Jan. 2002.

[11] EURESCOM, Project P806-d1 “A Common Framework for QoS/Network Performance in a multi-Provider Environment, Deliverable 1: The EQoS Framework,” Janvary 2002.

[12] IETF, Network Working Group, “Quality of Service (QoS) Mechanism Selection in the Session Description Protocol,” RFC5432, March 2009.

[13] ITU-T Recommendation E.800, “Terms and definitions related to quality of service and network performance including dependability,” August 1994.

[14] ITU-T Recommendation GB921 “Enhanced Telecom Operations Map (eTOM): The business Process Framework. TeleManagement FORUM,” Jun 2002.

[15] ITU-T Recommendation M.3100, “Generic Network Model,” July 1995.

[16] ITU-T Recommendation X. 723, “Structure of management: generic management information,” 1993.

[17] ITU-T, Recommendation X.641, “Information technology - Quality of service: Framework,” December 1997.

[18] ITU-T Recommendation X.711, “Information technology - Open Systems Interconnection - Common Management Information Protocol: Specification,” 1997.

[19] S. Kessal, N. Simoni, "A Deployment of Service Elements Based on QoS," services, Proceedings IEEE World Congress on Services, SERVICES 2011, pp. 93-94, July 4-9, 2011, Washington.

[20] E. Marilly, O. Martinot O, S. Betge-Brezetz, and G. Delegue, “Requirements for service level agreement management,” Proceedings IEEE Workshop on IP operation and Management, IPOM 2002, pp. 57-62, 2002, Dallas.

[21] N. Simoni, “Dés réseaux intelligents à la nouvelle génération de services,” ISBN: 978-2-7462-1218-3, Edition Hermes, 2007.

[22] N. Simoni, X. Xiong, and C. Yin, "Virtual Community for the Dynamic Management of NGN Mobility," Proceedings of The Fifth International Conference on Autonomic and Autonomous Systems, ICAS 2009, pp. 82-87, April 20-25, 2009, Valencia.

[23] N. Simoni, “Gestion de Réseau et de service: Similitude des concepts, spécificité des solutions,” ISBN: 2225829802, Edition Masson, 1998.

[24] Telecommunications Information Networking Architecture TINA-C, “Network Resource Information Model,” Version 3.0, 1997.

[25] TeleManagement Forum, GB917, “Service Level Agreement Management Handbook," Volume 2 Concepts and Principles, Release 3.0, May 2010.

 ISBN: 978-1-61208-183-0

Towards Opportunistic Data Dissemination in Mobile Phone Sensor Networks

Viet-Duc Le, Hans Scholten and Paul Havinga
Pervasive Systems
University of Twente
Enschede, the Netherlands
{*levietduc, hans.scholten, P.J.M.Havinga*}*@utwente.nl*

***Abstract*—Recently, there has been a growing interest within the research community in developing opportunistic routing protocols. Many schemes have been proposed; however, they differ greatly in assumptions and in type of network for which they are evaluated. As a result, researchers have an ambiguous understanding of how these schemes compare against each other in their specific applications. To investigate the performance of existing opportunistic routing algorithms in realistic scenarios, we propose a heterogeneous architecture including fixed infrastructure, mobile infrastructure, and mobile nodes. The proposed architecture focuses on how to utilize the available, low cost short-range radios of mobile phones for data gathering and dissemination. We also propose a new realistic mobility model and metrics. Existing opportunistic routing protocols are simulated and evaluated with the proposed heterogeneous architecture, mobility models, and transmission interfaces. Results show that some protocols suffer long time-to-live (TTL), while others suffer short TTL. We show that heterogeneous sensor network architectures need heterogeneous routing algorithms, such as a combination of Epidemic and Spray and Wait.**

Keywords-data dissemination; opportunistic network routing; heterogeneous architecture; mobility model; delivery speed.

I. INTRODUCTION

Mobile phones are getting attention as means to collect data. Data gathering can take place in the background, where the mobile phone user once gave permission to do so (e.g., location tracking), or involves continued active user participation (e.g., friend applications, Foursquare, Crowd sourcing, etc.). Collecting data is particularly meaningful when performed by many phones simultaneously. In such a way, the measurements have significantly enhanced reliability and accuracy. Thus, monitoring safety in public spaces becomes a "natural" application for mobile phone sensor networking.

Wireless Sensor Networks (WSNs) have been taken into consideration to replace the existing Wired Sensor Networks, since WSNs provide a wide range of context-awareness for real-time applications at low costs. A variety of sensor types with dense deployment forms a connected wireless mesh network via low power, short-range radios, collaborating to acquire and transmit the target data to sink nodes [1]. But still, the cost of deploying all kinds of such required sensors is considerably high in terms of time and money.

The next step in sensor networks is to enhance, or even replace, wireless sensor networks with mobile phones. Thanks to developments in sensor technology, smart phones, such as the iPhone or Android-based phones, are equipped with a large number of sensors, including GPS, accelerometers, gyroscope, proximity sensors and cameras. But, even regular phones have sensors, although we might not realize they have them: microphones, light sensors, and onboard radios. Not all mobile phones can access 3G mobile internet, especially when a disaster happens, for example, an earthquake or tsunami. But, still mobile phones have the means to participate in the sensor network. Through WiFi or Bluetooth radio, mobile phones can collaborate with nearby ones or the existing infrastructure-based sensor network in the sensing network. As requiring a connected path from source to sink, traditional routing algorithms may perform poorly in scenarios where the communication path is disrupted due to damaged infrastructure or overload in the infrastructure. Opportunistic routing algorithms in Mobile Sensor Networks (MSN) have been proposed in a number of recent studies to evaluate the performance of routing algorithms on sensor data gathering [2][3][4][5]. However, these algorithms use either basic scenarios or simple simulation architectures that are still quite far from real-world applications.

This paper investigates the performance of existing opportunistic routing algorithms in a heterogeneous architecture. We consider heterogeneous means of communication, especially WiFi and Bluetooth. The proposed architecture includes most of real-world components such as Roadside Units (RSUs), buses, cars and pedestrians. To achieve a realistic setting, the architecture is mapped on a real city, the city of Enschede, Netherlands. In addition, a new mobility model will be introduced based on available Shorted Path Map Based model in The ONE simulator [6]. By means of simulations, the proposed architecture and mobility model are used for the comparison of opportunistic routing protocols.

The paper has the following structure: related work is discussed in Section 2. Section 3 presents the architecture, a new mobility model and evaluation metrics. The simulations and an analysis of simulation results are the subject of Section 4. Based on the results, Section 5 gives possible directions for current and future research.

 ISBN: 978-1-61208-183-0

II. Related Work

In this paper, we focus on performance of message delivery in opportunistic networks that are essentially comprised of the existing wireless sensor networks (intelligent lampposts) and the mobile sensor networks (flocks of mobile phones). The network can be characterized as intermittently connected and sparse mobility. Traditional wireless ad-hoc networks routing protocols require end-to-end connectivity for a data packet delivered. In other words, if the destination is not available on the connected path, the packet delivery will fail and no further effort is taken to secure future transmission of the data. Consequently, routing protocols must be adapted for these new networks. Numerous opportunistic routing algorithms have been proposed in the last few years with different mechanisms, which are generally categorized based on either the type of network (*without infrastructure* and *with infrastructure*) or the pre-known information of the networks (*Stochastic* and *Context-based*) [7]. These categorizations slightly overlap as depicted in Figure 1. If networks are sparse and most nodes possess unpredictable movement, the stochastic protocols are more appropriate. In our opinion, the context-based protocols are more suitable for our considered networks, since the global knowledge of fixed infrastructure and mobile infrastructure can be used to improve the routing performance.

Stochastic routing protocols deliver messages by simply disseminating them all over the network. Being passed from node to node, messages will be gradually delivered at the destination. Epidemic Routing [8] diffuses messages similar to the way virus/bacteria spread in biology. When encountering others, a node will replicate and broadcast the messages to them. These nodes that just received the messages will move to other places and continuously replicate and transmit messages to other nodes whenever they are in range of communication. Though increasing the possibility of message delivery, the method results in flooding the network, and rapidly exhausts available resources. Direct Delivery (DD) [9] only delivers the holding messages directly to the destination; therefore, DD saves huge amounts of resources but decreases significantly the delivery ratio. Spray and Wait (SnW) [10] is a tradeoff between multi-copy scheme (Epidemic) and single-copy scheme (Direct Delivery) by finding an optimal number of copies of messages and dividing the message delivery process into 2 phases (*spray phase* and *wait phase*). First Contact (FC) [11] is a variant of single-copy scheme, which sends messages to the first encountered node or a random node if there are more than one.

Probabilistic Routing Protocol using History of Encounters and Transitivity (PRoPHET) [12] is a well-known Context-based routing protocol. PRoPHET estimates the delivery predictability for each known destination at each node before passing a message. The estimation relies on the history of encounters between nodes.

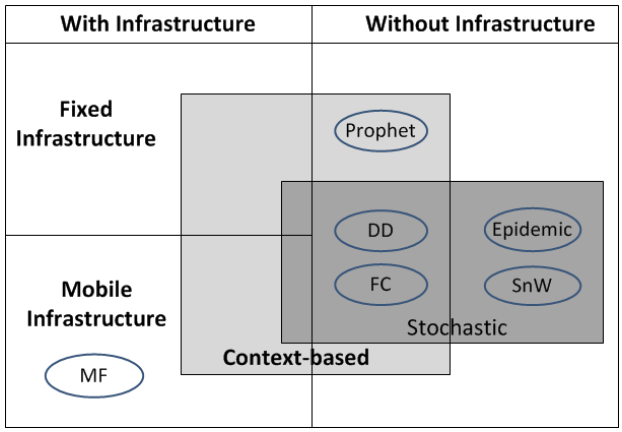


Figure 1. Categorizations of routing protocols in opportunistic networks.

A lot of attention has been given on how to apply above opportunistic routing algorithms in data dissemination for public safety applications. DTF-MSN [2] shows a scheme to gather information in the Delay/Fault-Tolerant Mobile Sensor Network based on an improvement of Direct Delivery and Epidemic. The proposal consists of two key components: queue management and data transmission. Queue management decides the importance of messages, and data transmission decides the node with high delivery probability to send messages to. However, the scenario used to evaluate the proposal has only one mobility model, where both source and sink are mobiles nodes, and is far from realistic for the public safety application domain. Camara et al. [3] present a good mechanism for the distribution of public safety warning messages, but the mechanism is limited to vehicle-to-vehicle and infrastructure-to-vehicle. The work uses only the basic Epidemic routing and there is no comparison with other routing protocols. A variant of Message Ferry (MF) [13] looks ahead at route information of ferries and then schedules messages to be exchanged based on the route information and the priority of messages. However, MF algorithm uses a simple architecture with only few fixed nodes (gateways) and mobile nodes (ferries). This algorithm is entirely constrained by the route and time schedule of ferries. Without the route information, the proposed routing algorithm will perform poorly. Recently, Keranen et al. [14] evaluate opportunistic networks with various mobility models and routing algorithms by using the ONE. Nevertheless, the used architecture does not include fixed infrastructure and the results only show the simulation speed.

This paper uses partially the ONE simulator [6] for simulations. The ONE includes several opportunistic routing algorithms and mobility models. The simulator also allows researchers to import their own maps and to configure the simulator with their own settings by many parameters, such

 ISBN: 978-1-61208-183-0

as speed of mobility, message size, buffer size, and etc.

III. Proposed Opportunistic Mobile Sensor Network (OppMSN)

Most traditional public safety applications are based on fixed and mobile wireless sensor networks and consider nodes to be connected. However, the very recent innovation of mobile phones with different kinds of onboard sensors and available low power consumption radios has brought on a new interest of using mobile phone as the main part of sensor networks. The network becomes an opportunistic network mainly comprised of the existing wireless sensor network and the mobile sensor network. Our proposal focuses on opportunistic mobile sensor networks for public safety applications.

A. Architecture

The considered opportunistic network is separated into several regions based on communities as shown in Figure 2. In order to link these regions, each of them has base stations equipped with long-range interfaces such as satellite, GSM, Internet. Each region consists of the following components: a fixed infrastructure, a mobile infrastructure (e.g., data mules) and mobile nodes.

- *Fixed infrastructure:* Road side units (RSUs) are deployed along main roads of the region. RSUs will be physically integrated in or fixed to the existing infrastructure, like lamppposts, GSM base station, or walls. RSUs form an ad-hoc wireless network, acting as a backbone, connecting mobile nodes with central servers or data sinks. The fixed infrastructure can also be used to disseminate information from central servers to the regions. The distance between RSUs is approximately 50 meters, using WiFi to build the network. There are two types of wireless interfaces for the RSUs, short-range Bluetooth and WiFi 802.11. Messages are transferred among RSUs through WiFi. The Wifi interface is also used to connect to buses, trams, cars, and smart phones. Bluetooth is designed for communication between RSUs and regular phones.
- *Mobile infrastructure:* Equipped with WiFi 802.11, busses and trams with known routes and known stops are considered as the mobile infrastructure in OppMSN applications. Since busses and trams move relatively fast, Bluetooth characterized by short-range (< 10 m) and low speed (< 2 Mbit/s) is not an appropriate option for busses and trams.
- *Mobile nodes:* The last component of the heterogeneous architecture consists of cars and mobile phones (used by pedestrians). There is no information of possible paths towards the sink because mobile phones and cars move unpredictably. Mobile phones are classified into either smart phones or regular phones. Smart phones typically have both WiFi and Bluetooth interfaces, while regular phones have only Bluetooth. For the same reason buses and trams use WiFi only, cars are equipped with WiFi.

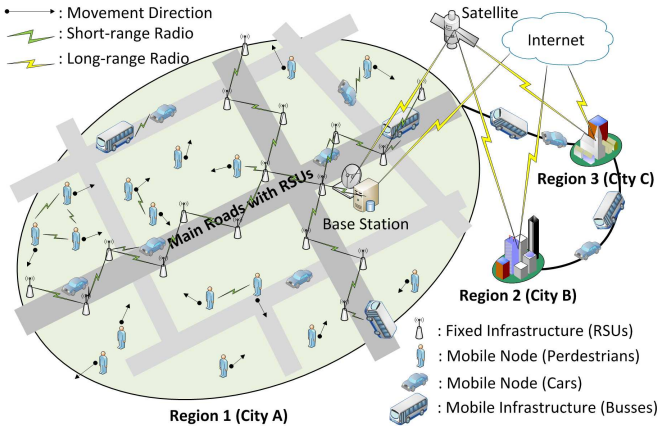


Figure 2. Architecture for Opportunistic Mobile Sensor Networks.

B. Architecture Performance Requirements

Depending on the physical characteristic, each of proposed components has a different degree of performance requirements such as reliability, throughput, latency, and electric power consumption. Fixed infrastructure has unlimited electric power supply, strong and stable signal strength, and large data storages. Therefore, latency and throughput are the most considerable performance requirements, and reliability and power consumption can be ignored. A message should be transferred as fast as possible via the ad-hoc connected network based on fixed infrastructure. Since the RSU network is not a sort of mesh networks, the bottleneck phenomenon probably decreases throughput and increases latency.

Mobile infrastructure, such as busses and trams, has no constraint on power supply, signal strength, and storage capacity. Thus, mobile infrastructure also has no problem with reliability and power consumption. As busses or trams play the role of messengers shuttling between sources and sinks in the network, latency depends significantly on velocity and distance. In addition, mobile infrastructure may become a bottleneck point because many passengers try to connect to a bus or a tram. As a result, the throughput of mobile infrastructure needs to improve as well.

Since mobile phones suffer limited power supply and intermittent connectivity, power consumption and throughput are the most critical performance requirements. Reliability is another considerable performance requirement because mobile nodes are sparse and dynamic. That some people are not willing to turn on the wireless interfaces all the time also makes the network less reliable. Moreover, velocity and unpredictable movement patterns of mobile nodes deter obtaining low latency and high throughput.

 ISBN: 978-1-61208-183-0

C. Network Operations

When a node sends a message to the data sink (base station) by using an opportunistic protocol, the message is transferred towards the base station by the store-carry-forward paradigm. The message is stored in phones or vehicles, and then forwarded to other nodes during opportunistic contacts. The node receiving the message is either the base station, a car, a phone, or a lamppost (RSU). The nodes, except the base station, continuously forward the message when in communication range of other nodes. Eventually, nodes may carry the message to the base station. If reaching a lamppost, the message usually takes the paths based on connected RSUs to go to the base station.

RSUs with a large storage capacity also act as a relay node in the network. Messages are stored at lampposts for a period of time until they expire due to a limited time-to-live (TTL). In some cases, reliability of event detection is enhanced by aggregating data provided by lampposts. A mobile node perhaps receives messages from the fixed infrastructure and then forwards them to other nodes. As a result, a message containing event information will not only be transferred to the base station, but also disseminated to nodes in network.

Busses and trams are not only message ferries as described in [13], but also gateways for passengers. Because the contact durations of mobile phones carried by passengers on a bus are quite long, messages may be fully exchanged among the passengers. Furthermore, these messages are stored at the bus and then disseminated to next passengers or delivered to the base station at the last bus stop.

When moving from one region to another, a mobile node will act as a gateway, transferring messages between regions. The transfer will be slow compared to using the fixed infrastructure. As the anticipated application domain is safety in public spaces, (emergency) messages should reach their destination as fast as possible.

D. Mobility Model

To increase the realism of the mobility model, five basic movement models are applied for different groups of nodes in our architecture. This approach represents the heterogeneous nature of reality, with road side units, cars, busses and pedestrians.

We assume that a portion of mobile nodes represents pedestrians wandering around without any specific purpose. The existing Map Based Movement (MBM) provided by the ONE is likely the most suited. MBM is Random waypoint movement with map-based constraints, in which a mobile node moves from one map node to another by randomly selecting a neighboring map node. This movement is repeated a randomly chosen number of times.

Naturally, people do not just wander around. They want to go somewhere for a purpose, using the shortest or fastest path possible. The choice between walking or taking the car or bus is often decided by the Euclidean distance to the destination. These destinations are very diverse [15], ranging from points of interest in the public domain (e.g., restaurants, parks, offices) to the more private ones (e.g. friends, home, family). The density of mobile nodes will differ accordingly. We propose a new movement model called Random Shortest Path Map Based Movement (RSPMBM) to model the behavior of human-like mobility. A node selects an arbitrary destination within a predefined range and then moves along the shortest path. Euclidean distance ranges are configurable in a setting file, for example, the distance ranges can be set $[50, 500]$ and $[500, 5000]$ meters for pedestrians and cars, respectively.

The new Road Side Unit Placement model defines where RSUs are placed on a map, along side roads with a certain distance between RSUs. The RSUs are stationary and form a wireless ad-hoc network or wireless sensor network.

For people who always take the bus, the Bus Traveler Movement and Bus Movement models are used for bus travelers and busses respectively. These movement models are provided by the ONE simulator.

E. Evaluation Metrics

To evaluate the proposed architecture and the proposed mobility model, we use the inter-contact time, first defined by Chaintreau et al. [16]. Inter-contact time is the time interval between two successive contacts of a pair of nodes, from the end of one contact to the next contact with the same node. Inter-contact time characterizes the frequency of opportunities for nodes to send packets to other nodes. The distribution of inter-contact time has an impact on the performances of different routing algorithms. It also shows that the inter-contact times are power-law distributed with the power-law exponent less than one.

Four metrics are used to evaluate the aforementioned performance requirements of different routing algorithms. Two of them are metrics implemented in the ONE: delivery ratio, and latency. Hop-count metric is no longer an informative metric to assess the delivery cost in time and distance in OppWSNs as it is used in connected ad-hoc WSNs. Instead, we define Delivery Speed and Delivery Cost for a more accurate evaluation.

- *Latency:* The time between the moment that a message is sent at the source and the time it is delivered at the destination.
- *Delivery ratio:* The number of successfully delivered messages divided by the total number of unique sent messages.
- *Delivery speed:* The speed of a message traveling from origin to destination. It is defined by Euclidean distance divided by latency.
- *Delivery cost:* The total number of messages including replicates divided by the number of successfully delivered messages.

 ISBN: 978-1-61208-183-0

IV. Simulation and Evaluation

In order to evaluate our proposed architecture and mobility model, a realistic simulation environment is set up, using a real city map. The results of running selected routing protocols are analyzed and compared to gain a better understanding on performances of existing routing protocols. From that, we may attain implications for future work.

A. Environment Setup

The simulation uses the center of the city of Enschede as a realistic setting. In the center of the map, there is the central bus station. The map shown in Figure 3 takes up approximately 4000 by 4000 meters and is exported from Openstreetmap.org. To this map several layers, as submaps, are added for lampposts, roads for cars, paths for pedestrians and routes for busses. Lampposts are positioned at the outer and inner ringroads, and four main roads radiating from the center. Cars are restricted to roads, but pedestrians may roam everywhere. Busses follow routes from the real city bus system. Roads in the ONE simulation have zero width. To overcome this limitation, roads are defined by two parallel routes as the lanes of a real road. In this way, communication with vehicles or pedestrians at both sides of the road is more realistic.

The simulation is carried out with 336 intelligent lampposts manually fixed on main roads, 50 cars, and 600 pedestrians moving around inside the city. The initial position of cars and pedestrians is randomly distributed. There are quite many bus lines in the city, but only four are chosen because others have routes overlapping the lamppost lines. Since the lampposts can transfer messages to the base station much faster than busses do, busses that run along lamppost lines have small contributions to the message delivery. Each bus line has two busses shuttling their routes. Since our basic goal is to investigate the contribution of pedestrians in disseminating data, only a small portion of cars, 50 over 650 mobile nodes, are simulated in the simulation. We also assume that the speed of pedestrians remains almost constant, $0.5 - 1.5$ m/s. Therefore, the mobility speed has a minor effect on performance results.

Since our proposed architecture also aims to reduce the use of mobile services, we only consider available short-range interfaces, particularly Bluetooth and WiFi. All mobile phones have Bluetooth Version 2.0 at 2 Mbit/s net data rate with 10 m radio range, while smart phones have only WiFi interface at net data rate of 10 Mbit/s with 60 m radio range. We assume that fifty percent of pedestrians own smart phones and the rest uses normal phone. Lampposts have both interfaces. The remaining nodes, cars and busses, use WiFi only, because they move at speeds that make Bluetooth communication unrealistic.

From the 600 pedestrians, 500 move with a purpose, while 100 are just strolling. Because cars likely possess predetermined routes, RSPMBM would be most suited.

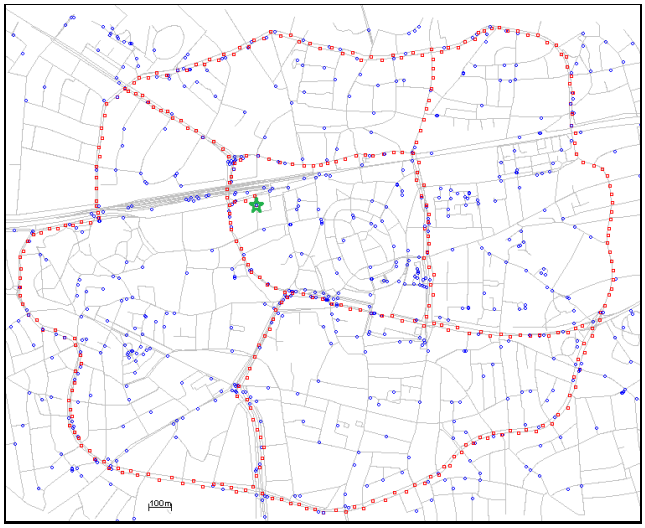


Figure 3. Inner-city of Enschede.

Busses follow fixed routes with predefined stops, and are modeled with the Bus Movement mobility model. Finally, pedestrians in busses are modeled with the Bus Traveler Movement model.

Data dissemination in the above heterogeneous scenario is simulated with a number of opportunistic routing protocols: Epidemic [8], Direct Delivery (DD) [9], FirstContact (FC) [11], and PRoPHET [12], and Spray and Wait (SnW) [10] with the number of copies (n) to be 6. This setting value is default in the ONE simulator. Since Message Ferry (MF) [13] is only useful for busses to transfer messages among base stations of cities, in our simulation with a single city, busses are just considered as a vehicle to transport passengers and do not implement MF.

Messages are generated every $25 - 35$ seconds by random cars and pedestrians. Lampposts do not generate messages, but act as a communication backbone. Messages may contain pictures, video and soundbites and are 500 KB to 1 MB in size. The buffer of normal mobile phones is set to 5 MB, and smart phones, cars, lampposts, and busses have 50 MB buffers.

B. Architecture and Model Evaluation

Figure 4 plots the complementary cumulative distribution (CCDF) of the inter-contact times. The graphs show that the inter-contact time distribution of RSPMBM has a power-law distribution with the exponent approximate 0.3 and similar to the real iMote trace [17]. This power-law distribution does contradict the exponential decay implied by previous mobility models that have been used to design routing algorithms (see [16]). Because the exponent and shape of the distribution may vary between environments, we did not configure parameters to produce the exact same exponent and shape as the iMote trace. Note the match between the iMote

 ISBN: 978-1-61208-183-0

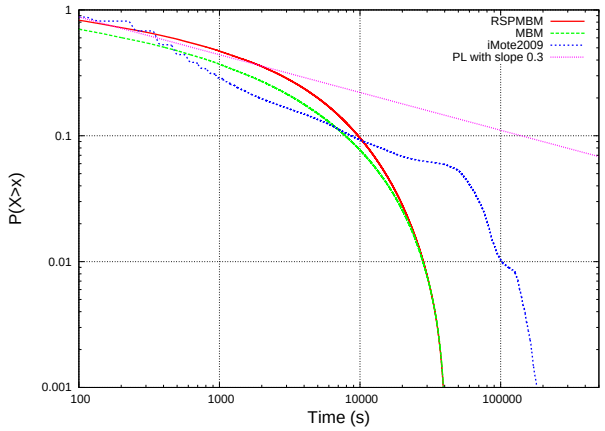


Figure 4. Inter-contact times for RSPMBM compared to the iMote trace.

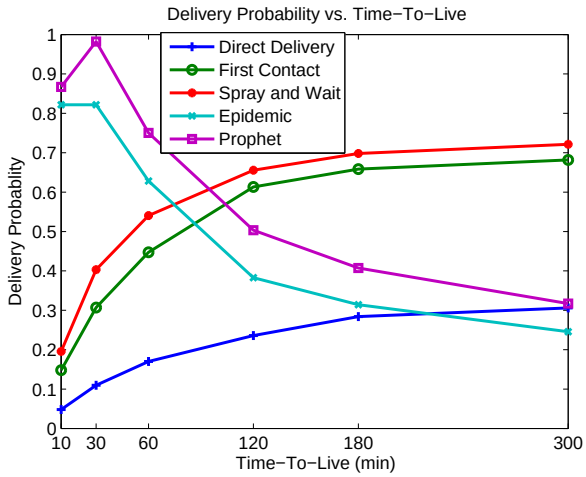


Figure 5. Message delivery probability.

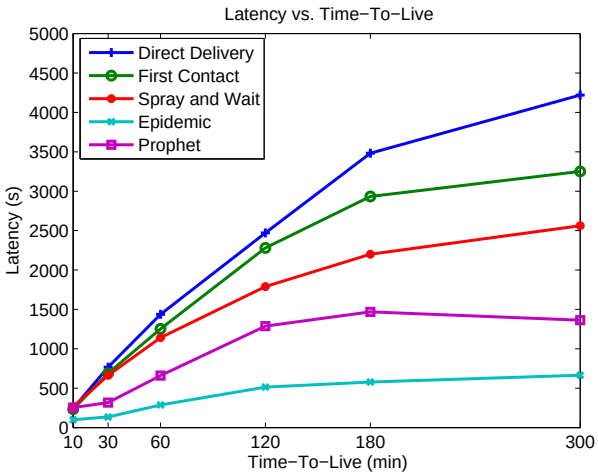


Figure 6. Average latency of message delivery.

trace and RSPMBM in the first two thirds of the graph. The difference in the last part of the graph is due to the longer trace (in time) of the iMote, leading to more contacts with low distribution probabilities. RSPMBM has shorter contact times due to the lamppost communication backbone. In other words, nodes in our simulation environment meet more frequently that those in the iMote experiment.

Figure 4 also shows the inter-contact time distribution for MBM used in the Enschede City Scenario (ECS) for comparison. Surprisingly, both RSPMBM and MBM produce similar tails of distribution (exponent coefficients). However, the inter-contact time distribution of RSPMBM has higher probability than that of MBM. This is expected, inter-contact times usually get shorter with increasing reality [6].

C. DTN Routing Algorithm Evaluation

Time-to-live (TTL) is an important variable for data dissemination, and strongly influences data delivery probability, latency, delivery speed, and delivery cost in opportunistic networks. In safety applications, emergency messages should be delivered with high probability, low latency, and high speed. Setting a high value for TTL is useless, i.e., a message that keeps a high TTL, probably would have a high latency, low speed, and less importance. Though TTL has a huge impact on the performance of routing protocols, it is hardly studied in existing literature. In the remainder of this section, we will investigate the influence of TTL on delivery ratio, latency, speed, and costs of messages.

Figure 5 shows the delivery probability of each routing algorithm as TTL in the scenarios increases from 10 to 300 minutes. In the graph two very different trends in delivery probability can be observed. DD, FC and SnW have increasing delivery probability with increasing TTL, with a highest gain in the lower TTL values. This is as one would expect. The longer the TTL of a message, the more opportunities for message transferring. Counter-intuitive is the decreasing delivery probability with increasing TTL for Epidemic and PRoPHET. This is explained as follows. Epidemic and PRoPHET are multi-copy, thus the number of relayed messages increases exponentially when TTL is long. Eventually, with a limited buffer and limited contact duration, the delivery probabilities of Epidemic and PRoPHET will dramatically suffer. This explanation is reconfirmed in Figure 8, which depicts the delivery cost for each routing protocol.

Figure 6 plots the average latency of message delivery as TTL increases. From the graph, one can see that increasing TTL results in increasing delays in message delivery. This is as expected. Since flooding the network with messages, Epidemic scores best. Although Epidemic has the lowest delivery probability at high TTL values, when a message reaches its destination, the message will have low latency. Direct Delivery scores lowest with high latency. DD delivers messages directly to the destination. So it may take some

 ISBN: 978-1-61208-183-0

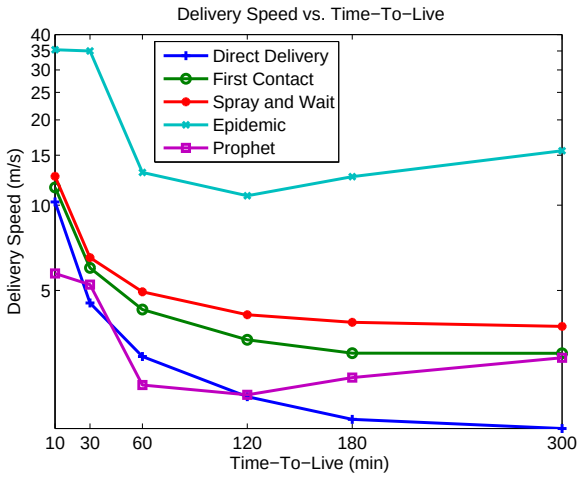


Figure 7. Average speed of message delivery.

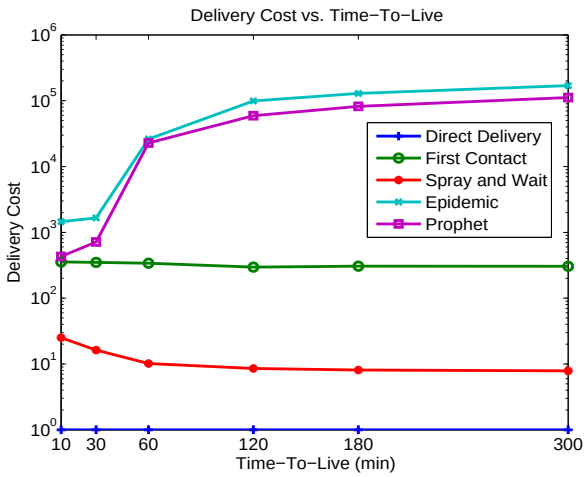


Figure 8. Delivery cost.

time for this opportunity to happen.

The speed of message delivery is depicted in Figure 7. The speed decreases sharply in the first part of the graph for all protocols and then remains almost constant. For 10-min TTL, only messages near the base station or lampposts can reach the destination. Other messages would be dropped before arriving at the base station. Increasing TTL causes more messages farther away from the base station to be delivered. This explains why the average delivery speed declines sharply. However, when TTL is greater than 60 minutes, most messages have sufficient lifetime. Therefore increasing TTL further does not affect the delivery speed.

The delivery speed of Epidemic and PRoPHET goes up slightly when TTL is greater than 120 minutes. Due to overhead, there are fewer messages that could be delivered. Hence, the average delivery speed rises slightly again.

Epidemic has the highest delivery speed since it floods messages over the network. DD has the lowest delivery speed on account of sending messages only when mobile nodes encounter the base station.

As PRoPHET has the second lowest latency in Figure 6, one would expect it to have the second highest delivery speed. On the contrary, the graph in Figure 7 shows that PRoPHET has the lowest delivery speed. The reason lies in the fact that PRoPHET transfers messages based on the frequency of node encountering, called delivery predictability. Owing to the lamppost connected network, most nodes have almost the same delivery predictability. Consequently, messages are wastefully transferred around before reaching the destination. In such way, even the average delay of a message is low, but the Euclidean distance from its source to the base station is short too. That is why the delivery speed of PRoPHET is low even though its latency is not high. This behavior also proves that delay of message delivery is not sufficient enough to evaluate quality of message delivery.

Because the majority of nodes have limited power supply, the delivery costs of opportunistic routing algorithms must be taken into account. The delivery cost represents the ratio between the number of total transmissions needed over that of successfully delivered messages. Figure 8 shows that Epidemic and PRoPHET have the highest delivery cost because they maximize the opportunities of message delivery by replicating copies of messages as much as possible. DD and SnW have the least overhead, as DD has only one single copy of a message and SnW has 6 copies of messages at maximum. Clearly DD has the lowest delivery cost of all routing algorithms. The delivery costs for Epidemic and PRoPHET increase sharply with increasing TTL, but stabilize after a while. The reason is that only a limited number of messages can be transferred during the limited contact duration.

V. Conclusion and Future Work

In this paper, we have proposed a heterogeneous architecture comprising fixed infrastructure, mobile infrastructure, and mobile nodes. In addition, we proposed the a realistic mobility model and metrics. Several well-known opportunistic routing protocols were tested with this architecture. Our observation shows that none of the evaluated protocols performs well with a heterogenous scenario, such as the one described in this paper. Since a single simple routing algorithm does not suffice to improve the overall message delivery performance, a contribution of several algorithms should be considered:

- Road Side Units (RSU), as used in the lamppost backbone network, should not only carry received information to a central server, but also disseminate information to nearby passing nodes. This communication shortcut leaves the base station out of the loop and contributes a better delivery speed and delivery cost. The Epidemic

 ISBN: 978-1-61208-183-0

routing protocol with a flooding control mechanism is best suitable for the RSU network.

- Busses, which act as data mules or message ferries, have a mobility pattern based on fixed routes and time schedules. The Message Ferry routing protocol is most appropriate.
- Pedestrians and cars are best served by stochastic and context-based schemes. However, exchanging messages between nodes that use different routing protocols is a challenge. For examples, nodes running PRoPHET fail to update the delivery predictability of nodes running Epidemic due to the unavailability of delivery predictability in Epidemic router.

We also plan to take message priority into consideration. Because designing an optimal routing protocol with a delivery probability of 100% under all conditions is difficult, prioritizing messages becomes a necessity. Message prioritization perhaps relies on the importance of information, creation time, or source location. Priorities must be defined by a specific application, for instance, public safety applications define the priority based on the source location, creation time, and seriousness of detected events. One last point of concern is the security and privacy of information. A leading principle should be that the creator owns the data and decides how the data can be used by others. However, one may argue that in situations of emergency this principle may be overruled by authorities. This issue will be addressed in future research. Following this research, a testbed is planned to implement and evaluate the proposed heterogeneous DTN architecture.

ACKNOWLEDGMENT

This work is supported in part by the SenSafety project in the Dutch Commit program, in part by SI4MS (NWO/STW Grant, dossier 655.010.209), and by EIT ICT LABS, Digital Cities of the Future.

REFERENCES

[1] I. Akyildiz, W. Su, and Y. Sankarasubramaniam, "A survey on sensor networks," *IEEE Comm. Magazine*, vol. 40, pp. 102–114, 2002.

[2] Y. Wang and H. Wu, "Delay/fault-tolerant mobile sensor network (dft-msn): A new paradigm for pervasive information gathering," *IEEE Trans. Mobile Computing*, vol. 6, pp. 1021 – 1034, 2007.

[3] D. Camara, C. Bonnet, and F. Filali, "Propagation of public safety warning message: A delay tolerant approach," in *Proc. IEEE Communications Society WCNC*, 2010.

[4] A. T. Erman, A. Dilo, and P. Havinga, "A fault-tolerant data dissemination based on honeycomb architecture for mobile multi-sink wireless sensor networks," in *Proc. of the Sixth International Conference on Intelligent Sensors, Sensor Networks and Information Processing, ISSNIP 2010*, 2010, pp. 97–102.

[5] R. Schwartz, R. Barbosa, N. Meratnia, G. Heijenk, and J. Scholten, "A directional data dissemination protocol for vehicular environments," *Computer Communications*, vol. 34, pp. 2057–2071, 2011.

[6] A. Keranen, J. Ott, and T. Karkkainen, "The one simulator for dtn protocol evaluation," in *Proc. of the 2nd International Conference on Simulation Tools and Techniques(SIMUTools)*, 2009.

[7] Newcom++, "State of the art of research on opportunistic networks, and definition of a common framework for reference models and performance metrics," Downloaded from http://www.newcom-project.eu/public-deliverables/research/ (Febr. 2012).

[8] A. Vahdat and D. Becker, "Epidemic routing for partially connected ad hoc networks," Department of Computer Science, Duke Univeristy, Durham, NC, Tech. Rep., 2000.

[9] T. Spyropoulos, K. Psounis, and C. Raghavendra, "Single-copy routing in intermittently connected mobile networks," in *Proc. Sensor and Ad Hoc Communications and Networks (SECON)*, 2004, pp. 235–244.

[10] T. Spyropoulos, K. Psounis, and C. Raghavendra, "Spray and wait: An efficient routing scheme for intermittently connected mobile networks," in *Proc. of ACM SIGCOMM Workshop on Delay-Tolerant Networking (WDTN)*, 2005.

[11] S. Jain, K. Fall, and R. Patra, "Routing in a delay tolerant network," in *Proc. of ACM SIGCOMM on Wireless and Delay-Tolerant Networks*, 2004.

[12] A. Lindgren and A. Droia, "Probabilistic routing protocol for intermittently connected networks," *Internet Draft draft-lindgren-dtnrg-prophet-02, Work in Progress*, 2006.

[13] Y. Xian, C. Huang, and J. Cobb, "Look-ahead routing and message scheduling in delay-tolerant networks," in *Proc. IEEE Conference on Local Computer Networks (LCN)*, 2010.

[14] A. Lindgren, T. Karkkainen, and J. Ott, "Simulating mobility and dtns with the one," *Journal of Communications*, 2010.

[15] F. Ekman, A. Keranen, J. Karvo, and J. Ott, "Working day movement model," in *Proc. of the 1st ACM SIGMOBILE workshop on Mobility models (MobilityModels)*, 2008.

[16] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on the design of opportunistic forwarding algorithms," in *Proc. IEEE Infocom*, 2006.

[17] J. Scott, R. Gass, J. Crowcroft, P. Hui, C. Diot, and A. Chaintreau, "CRAWDAD trace cambridge/haggle/imote/infocom2006 (v. 2009-05-29)," Downloaded from http://crawdad.cs.dartmouth.edu, Febr. 2012.

 ISBN: 978-1-61208-183-0

Towards A Theoretically Bounded Path Key Establishment Mechanism in Wireless Sensor Networks

Aishwarya Mishra
School of Information Technology
Illinois State University
Normal IL 61790 USA
amishra@ilstu.edu

Tibor Gyires
School of Information Technology
Illinois State University
Normal IL 61790 USA
tbgyires@ilstu.edu

Yongning Tang
School of Information Technology
Illinois State University
Normal IL 61790 USA
ytang@ilstu.edu

***Abstract*—Random Key Pre-distribution Scheme (RKPS) guarantees any pair of neighboring nodes in a Wireless sensor network (WSN) can build a secure connection either directly or through a path key establishment mechanism (PKEM). For any pair of neighboring sensor nodes without a direct secure connection due to unfound common key, a node can resort to PKEM to flood a keyrequest in the connected graph to reach the neighboring node and build a secure connection thereafter. One remaining challenge in PKEM is to find an optimal transmission radius for flooding. Commonly used empirically or probabilistically bounded flooding mechanisms may cause high power consumption on sensor nodes and also easily be exploited to launch power exhaustion Denial of Service (DoS) attacks to sabotage a WSN. In this paper, we tackle this challenge by first theoretically analyzing the upper bound of diameter in Erdős-Rényi (ER) random graph theory, and then verifying the performance of theoretical bounded PKEM using simulations. The performance evaluation shows both the correctness and effectiveness of our proposed theoretically bounded path key establishment mechanism.**

Keywords- sensor networks, random key predistribution, graph diameter, random graph, theoretical bound, path key establishment mechanism.

I. INTRODUCTION

With the growing prevalence of wireless sensor networks (WSNs), security becomes extremely important for WSN-based applications [1], [5], [17], [19], [20], [28], especially when they are deployed in hostile environment. One of proposed solutions for resource constraint WSNs is Random Key Pre-distribution Scheme (RKPS) [1], which guarantees any pair of neighboring nodes in a WSN would be able to build a secure connection either directly or through a path key establishment mechanism (PKEM).

The first Random Key Predistribution Scheme was proposed in [1] and emerged as a promising solution for WSNs. RKPS is fully distributed and allocates shared secret keys to each sensor node in such a manner that the adversarial compromise of a fraction of the nodes does not impact the security of the complete network. This scheme relies on allocating on each sensor before deployment a small random subset of keys (keyrings) from a large universal set of random keys (keypool), such that each keyring overlaps with any other keyrings with a small probability.

RKPS pre-distributes the keys in such a way that each sensor node in a deployed WSN can directly build secure wireless connections with at least a fraction of its neighboring nodes, where common keys can be found in their pre-distributed key pools. By properly selecting the RKPS parameters (e.g., keyring size), a connected graph among all sensor nodes in the WSN can be constructed, in which a network path composed of one or multiple wireless connections can be found for any two nodes according to Erdős-Rényi (ER) random graph theory. For any pair of neighboring sensor nodes without a direct secure connection due to unfound common key, a node can resort to PKEM to flood a keyrequest in the connected graph to reach the neighboring node and build a secure connection thereafter. It is worth noting that flooding is the required messaging mechanism at the initial phase of trust establishment among sensor nodes. More effective routing mechanisms [2] can be applied later among trusted sensor nodes.

One remaining challenge in PKEM is to find an optimal transmission radius for flooding. Commonly used empirically or probabilistically bounded flooding mechanisms may cause high power consumption on sensor nodes and also easily be exploited to launch power exhaustion Denial of Service (DoS) attacks [29] to sabotage a WSN.

In this paper, we tackle this challenge by first theoretically analyzing the upper bound of diameter in Erdős-Rényi (ER) random graph theory, and then verifying the performance of theoretical bounded PKEM using simulations. The performance evaluation shows both the correctness and effectiveness of our proposed theoretically bounded PKEM.

The rest of the paper is organized as the following. Section II discusses the related work. Section III provides the background of PKEM and derives the theoretical bound of flooding radius in PKEM. Section IV and Section V present our simulation design and results. Finally, Section VI concludes the paper.

 ISBN: 978-1-61208-183-0

II. Related Work

Research on RKPS was first introduced in [1]. A variety of schemes have been proposed [16], [17], [19]–[21], [28] built upon the basic RKPS by combining with other key predistribution schemes for improving sensor network security [18], [26]. These schemes have been reviewed in [5], which also covered an extensive survey on the state-of-the-art in sensor network security. Since our problem is tangential to the RKPS that is assumed in our work, we focus on reviewing PKEM related research work in the past [2], [7]–[9].

Further results pertinent to our work are found in [11], which discussed the application of graph theory to RKPS in the context of sensor networks, and produces validating results for specific ranges of its parameters. The work in [1] presented empirical observations that the length of any keypath does not exceed an estimated constant number for their simulation cases with $1000 \sim 10000$ nodes, but did not provide formal mathematical guidance which could characterize how PKEM will behave for much larger node populations. Another contribution in [1] was the explicit statement of the assumptions related to the minimum degree of the underlying connectivity graph, which had been assumed to be higher than the maximum number of neighboring nodes supported by modern wireless MAC layer protocols.

The first attempt of using a TTL limited path key establishment appears in [11], which aimed at limiting the overhead of the RKPS scheme. However, they were mainly interested in observing the average lengths of various keypaths by repeating the same experiments as in [1]. The result from [11] also noted that most of the actual keypath lengths were much smaller than the observed maximum length. However, they did not characterize the asymptotic behavior of the PKEM and how the length of a keypath was affected by the node population, and deployment density.

The work in [12] followed the same directions as [6] and proposed a theoretic graph framework for parametric decision making for RKPS, optimal keyring size, and network transmission energy consumed in PKEM. Some simulation guidance can be found in [12] showing the approach to the construction of a high performance simulation, which is also adopted in our simulation design. However, impractical full-visibility was assumed in [12]. Moreover, only the average length of keypaths was investigated other than the nature of the longest keypaths.

In contrast to the previous research discussed above, our work focuses on finding the maximum required length of keypath under the practical sensor network model with limited visibility and the consideration of expected node populations, node connectivity, and the power resources of a sensor node.

III. Theoretical Bound Analysis

For a given network with node population n, RKPS applies Erdős-Rényi random graph theory to choose the sizes of keyring k and keypool K, such that the secure network formed resembles a connected Erdős-Rényi random graph. In this paper, an Erdős-Rényi random graph is represented by $G_{(n,p)}$, where n is the number of vertices and p represents the probability that a vertex is connected to any others within the graph. A graph where all the nodes are connected into a single giant component is denoted as a connected graph.

A. Trust Graph

Secure connectivity between neighboring nodes in a sensor network can be represented by a trust graph [3], in which each sensor node is represented by a vertex and a secure connection between any two nodes is represented by an edge. Similarly, the underlying wireless connectivity in the sensor network can also be represented in the form of a connectivity graph, where each sensor represented by a vertex is connected to all other sensors within its transmission range. It is worth noting that the trust graph is contained within the connectivity graph and by definition is a sub-graph of the connectivity graph. Figure 1 shows a trust graph example built on the top of a deployed WSN.

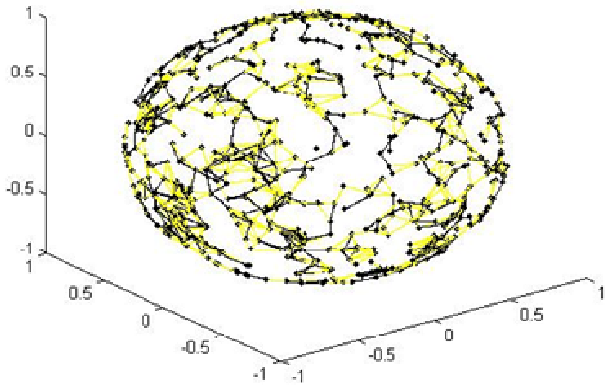


Figure 1: An example of trust graph: lighter edges represent wireless connectivity and darker edges represent secure connectivity.

B. Generalized RKPS Model

Random subsets of keys (keyrings) are chosen from a large pool of keys (keypool), such that any two keyrings may share at least a common key with certain probability. After being deployed, each sensor attempts to establish trust with its neighbors by discovering common key(s) through keyrequests. For any given node u, the small size of keyrings only allows a fraction of u's neighbors directly authenticate the received keyrequest from u. For any u's neighbor node, for instance, v that are unable to directly authenticate u's keyrequest, a path key establishment mechanism (PKEM) can coordinate the trust establishment between u and v.

 ISBN: 978-1-61208-183-0

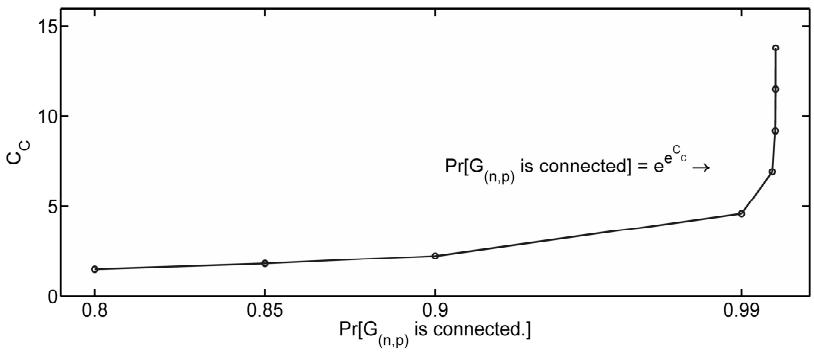


Figure 2: The impact of C.

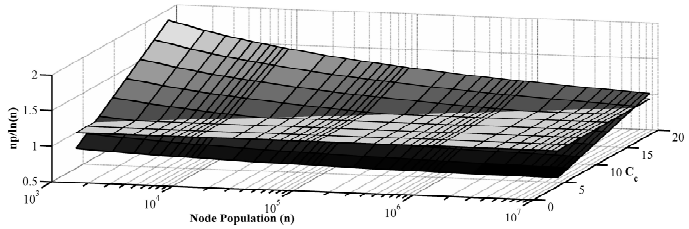


Figure 3: Plot of $np/\log(n)$ showing the value of C for various ranges of n and C.

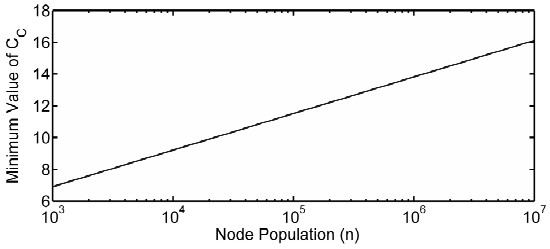


Figure 4: Plot showing $C = \log(n)$, values of C where $np/\log(n) = 2$.

With the support of PKEM, v forwards the indirectly authenticated keyrequest from u to its trusted neighbors which either authenticate or forward it to their trusted neighbors until a transitively trusted node authenticates the targeted neighboring node u. RKPS chooses the keyring and keypool sizes such that the secure network formed by the deployed sensors can be modeled as a connected ER graph, and the keyrequest would potentially be forwarded to all nodes connected securely to each other.

A repeatedly forwarded keyrequest describes a path through the network, where each node within the path trusts the next node in the path, termed as a keypath [1]. As a consequence of PKEM, multiple keypaths emanate from the node requesting PKEM authentication of a particular keyrequest, and a large number of the connected sensor nodes within the network consume power in computation and communication to authenticate a single keyrequest. Initial research on RKPS [1] investigated the varying length of keypath to propose an empirical mechanism to limit the length of keypath using Time-To-Live (TTL) parameter on the process of keyrequest. However, the recommended TTL depended upon empirical observations, which may not be applicable to different sizes of WSNs using different deployment schemes.

The deployment model of a sensor network is generally assumed to be uniformly random and the neighboring nodes of any particular sensor after deployment cannot be predicted.

For a random graph $G_{(n,p)}$ [4], [13], [25], we have:

$$if \quad p = \frac{\ln(n)}{n} + \frac{C}{n} \tag{1}$$

$$then \lim_{n\to\inf} P(G_{(n,p)} \; is \; connected) = e^{e^{-C}} \tag{2}$$

where C is a constant and should be chosen such that the chance of having a connected graph $P(G_{(n,p)} \; is \; connected)$ is close to 1.

Prior research [1] on RKPS has recommended choosing the value of C between 8 and 16, as shown in Figure 2. which can yield the desired value of p, and further derive the keyring size (k) for a given keypool size (K).

It is worth noting that the ER graph theory assumes that any node within a given graph can be connected to any others, i.e., every node can see any others within the network (full visibility model). However, in sensor networks a sensor node is only connected to a small subset of n_a ($n_a \ll n$) randomly deployed nodes, which are within its communication range (limited visibility model). In order to overcome this practical limitation, the work in [1] proposed adjusting p to the effective probability (p_a).

By introducing the concept of effective probability, a node can connect to any of its neighboring nodes, such that the average degree d of the nodes in the graph remains constant as shown below:

$$d = (n_a - 1)p_a = np \tag{3}$$

 ISBN: 978-1-61208-183-0

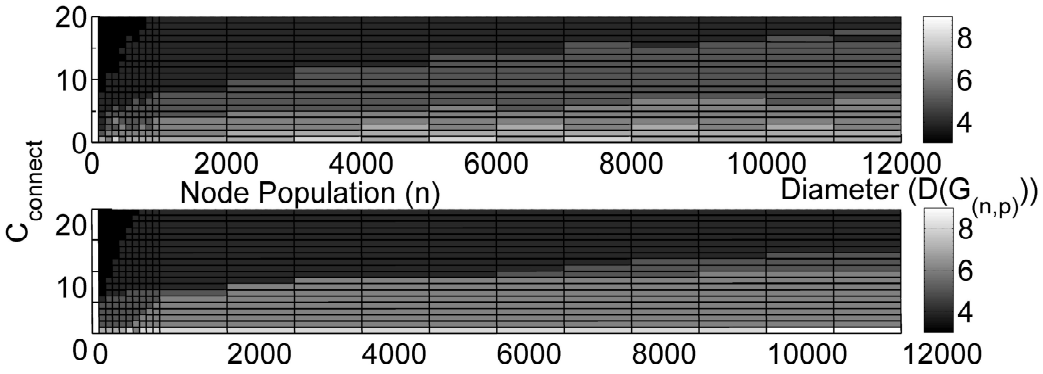


Figure 5: The comparison of practical and theoretical graph diameters considering the impact of C.

With this calculated value of p_a, the work in [1] derived k according to the following equation:

$$p_a = 1 - \frac{(K-k)!^2}{K!(K-2k)!} \tag{4}$$

The results identifying the upper bound on a random graph diameter for the parameter ranges assumed in the discussion above have also been proposed in Theorem 4 in [13], where $p \geq c\log(n)/n$.

C. Diameter of a Sparse Random Graph

Theorem 4 in [13] states that given Eq. 5, the diameter of the graph is concentrated.

$$\frac{np}{\log n} = c \geq 2 \tag{5}$$

$$diam(G_{(n,p)}) \leq \lceil \frac{\log n}{\log np} \rceil \tag{6}$$

This formula gives the theoretical upper bound on the diameter of a sparse random graph. Please note that we are assuming $c \geq 2$ because the value of the constant C typically chosen sufficiently high.

We utilized the Matgraph [14] library in MATLAB to verify by simulating the above theoretical results on several instances of random graph for various values of n when $c = 1$. Figure 3 confirms the theoretical results above. It is worth noting that the diameter values remain relatively stable for large increments of n, which should allow the future extension of a sensor network, even with the current limited diameter. We also notice that the observed diameter value is far below the one predicted by the theoretic analysis, which would make it robust against transmission failures in the shortest path.

As discussed earlier, most empirical studies of RKPS have assumed a value of C in the range of 8 to 16. Figure 3 and Figure 4 plot the value of C in Eq. 5 and Eq. 6 showing C can be safely assumed higher than 2 for lower ranges of n and higher ranges of C in Eq. 1. These values are coincident with the range assumed in prior research on RKPS schemes.

The value of C in Eq. 1 has significant impact upon whether c in Eq. 5 is in a range where the diameter of the random graph remains $O(\log(n)/\log(np))$. Figure 5 implies that lower values of C in Eq.1 will not allow the diameter of the graph to remain small.

IV. Simulation Design

Our simulations are designed to verify the characteristics of trust graph using RKPS scheme with various ranges of n, p and C to validate whether the obtained trust graph from simulations follows the theoretical results. We generated random topologies for sensor networks by varying the number of nodes from 1000 to 5000, and calculate the corresponding keyring sizes from a keypool of 100000. While pursuing the construction of our simulations, we also identified an important implicit assumption that the minimum degree of the underlying connectivity graph of a sensor network should be higher than the maximum expected degree of the trust graph as the results from [15].

In order to investigate the effective diameter of a trust graph in a WSN, we created a sensor network simulator along the directions discussed in [22]. Our simulator model derives the keying size based on [1], and allows for variations in the sensor network deployment densities through node range variation.

Most of the simulation studies in recent studies [1], [6], [15] have used a unit square as the deployment area with varying transmission ranges to simulate different node densities. More recently, the work in [22] identified the boundary effect that occurs at the borders of any sensor network, where the boundary sensors do not enjoy the average neighborhood connectivity available to nodes away from the boundary. To eliminate this effect, the work in [22] proposed the deployment of the sensor network on a spherical surface to eliminate the boundary effect and produce a sensor network model, which can be used to test the hypothesis assuming homogeneous node connectivity. Boundary effect can significantly influence the degree distribution of a trust graph in simulations but its impact in practical deployments

 ISBN: 978-1-61208-183-0

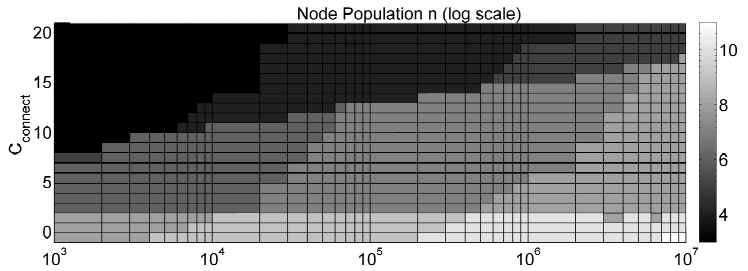


Figure 6: Log range plot of diameter for various ranges of n and C.

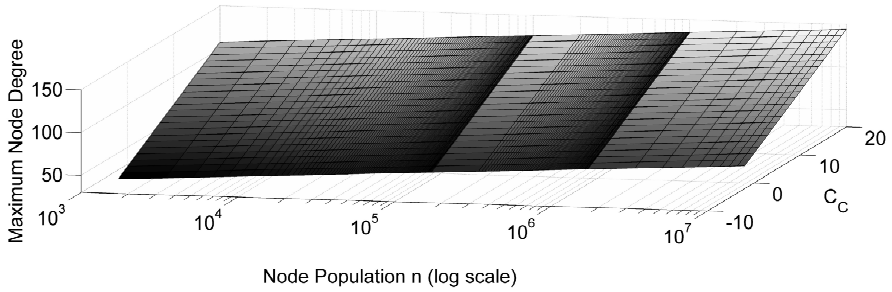


Figure 7: Asymptotic plot of required maximum node degree for large sensor networks.

is considerably less and further mitigated if the nodes on the boundary resort to dynamic range extension as suggested by [23]. Following the directions from the work in [22], we model our node deployment using Ziggurat method due to Marsaglia [24] to generate uniformly distributed points on a spherical surface. We calculate the node distances using the greatest circle arc length. But we also assume that the node range is a disk shaped area on the surface of the sphere equivalent to the one formed on a plane, which allows us to model practical planner deployment, while eliminating the boundary effect.

V. Simulation Result

Figure 6 and Figure 7 plot the log range theoretical predictions from the theoretical analysis results, shown by Eq. 1 to Eq. 5. Several observations and conclusions can be drawn on the basis of these simulations.

The diameter of a deployed sensor network increases very slowly with the increase of network size, and remains constant for large ranges of node populations. This observation shows the promise in the extensibility and graceful degradation of a sensor network deployment, even if the TTL value is controlled as a constant. On the other hand, this shows that controlling the TTL would only provide limited control over the number of nodes visited by a keyrequest and the consequent power consumption of PKEM. The number of nodes which may receive a PKEM request rises rapidly with each increment of TTL in a large network.

Further, Figure 7 also shows that node degrees may rise as high as 140, which is prohibitively high for current sensor node platforms. We note that several methods have been proposed to mitigate this problem including range extension. Another method to allow higher node degrees could be to allow neighboring sensors to transparently repeat a keyrequest broadcast so as to allow a larger number of nodes to respond to authentication. Recent research in the power consumption of available sensor node platforms shows that each wireless transmission can cause very high power consumption.

VI. Discussion and Conclusion

This paper formally studies the communication overhead in path key establishment mechanism (PKEM) and the possible improvement through state-of-the-art research combining sensor network deployment schemes and communication mechanisms with the theoretical results from ER random graph study. PKEM is a variant of flooding broadcasting and specifically an instance of probabilistic broadcasting. While we have focused on PKEM specifically, our results are also extendable to the sensor node revocation protocol for RKPS, which also relies on broadcasting.

We have presented and tested an analytical model which provides simplified guidance on the TTL configuration of PKEM for large sensor network deployments. We have shown that certain assumptions regarding the modeling of the trust graph are necessary to preserve its properties as embodied in an ER random graph model. Lastly, we studied the predictions of our analytical model for large scale deployment and identified their impact on the feasibility of large scale sensor networks. Our simulations have demonstrated that the theory on random graph approximates the

 ISBN: 978-1-61208-183-0

practical observations and can prove to be highly effective especially in the design of large scale sensor networks.

Our work also shows that the secure connectivity and diameter of the corresponding trust graph is intimately related to its deployment density and node connectivity. A graph with poor connectivity would significantly weaken the trust graph and may result in undesirable partitioning of the corresponding sensor network.

Through this work, we hope to trigger a discussion of the problem existed in keyrequest broadcasting methods. In order to securely limit the overhead of randomized broadcasting, generally reducing transmission complexity may be more suitable for wireless sensor networks, especially when they are deployed in large scale. This paper serves to provide a skeleton of theoretical assumptions, which may facilitate the application of ER graph theoretic results to the problem of broadcasting at large.

REFERENCES

[1] L. Eschenauer and V. D. Gligor, "A key-management scheme for distributed sensor networks," in Proceedings of the 9th ACM conference on Computer and communications security, Washington, DC, USA, 2002.

[2] C. Karlof, D. Wagner, Secure routing in sensor networks: attacks and countermeasures, First IEEE International Workshop on Sensor Network Protocols and Applications (2003).

[3] P. Roberto Di, V. M. Luigi, M. Alessandro, P. Alessandro, and R. Jaikumar, "Redoubtable Sensor Networks," ACM Trans. Inf. Syst. Secur., vol. 11, pp. 1-22, 2008.

[4] P. Erdos and A. Renyi, "On the evolution of random graph.," Institute of Mathematics Hungarian Academy Of Science, 1959.

[5] Y. Xiao, V. K. Rayi, B. Sun, X. Du, F. Hu, and M. Galloway, "A survey of key management schemes in wireless sensor networks," Comput. Commun., vol. 30, pp. 2314-2341, 2007.

[6] J. Hwang and Y. Kim, "Revisiting random key pre-distribution schemes for wireless sensor networks," in Proceedings of the 2nd ACM workshop on Security of ad hoc and sensor networks, Washington DC, USA, 2004.

[7] S. Zhu, S. Setia, S. Jajodia, LEAP: efficient security mechanisms for large-scale distributed sensor networks, in: Proceedings of The 10th ACM Conference on Computer and Communications Security (CCS 03), Washington D.C., October, 2003.

[8] W. Du, J. Deng, Y.S. Han, P.K. Varshney, A pairwise key predistribution scheme for wireless sensor networks, Proceedings of the 10th ACM Conference on Computer and Communications (SecurityCCS 03) (2003) 4251.

[9] D. Liu, P. Ning, Establishing pairwise keys in distributed sensor networks, Proceedings of the 10th ACM Conference on Computer and Communications Security (CCS 03) (2003) 5261.

[10] H. Chan, A. Perrig, D. Song, Random key predistribution schemes for sensor networks, in: Proceedings of the 2003 IEEE Symposium on Security and Privacy, May 1114, pp. 197 213.

[11] T. M. Vu, R. Safavi-Naini, and C. Williamson, "On applicability of random graphs for modeling random key predistribution for wireless sensor networks," in Proceedings of the 12th international conference on Stabilization, safety, and security of distributed systems, NewYork, NY, USA, 2010.

[12] V. Tuan Manh, W. Carey, and S.-N. Reihaneh, "Simulation modeling of secure wireless sensor networks," in Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools, Pisa, Italy, 2009.

[13] F. Chung and L. Lu, "The Diameter of Sparse Random Graphs," Advances in Applied Mathematics, vol. 26, pp. 257-279, 2001.

[14] Beno, t. Otjacques, F. Feltz, G. Halin, and J.-C. Bignon, "Mat'Graph: transformation matricielle de graphe pour visualiser des changes lectroniques," in Proceedings of the 17th international conference on Francophone sur l'Interaction Homme-Machine, Toulouse, France, 2005.

[15] R. Durrett, Random Graph Dynamics. New York, NY: Cambridge University Press 2006.

[16] A.S. Wander, N. Gura, H. Eberle et al., Energy analysis of public-key cryptography for wireless sensor networks, in: Proceedings of the Third IEEE International Conference on Pervasive Computing and Communications (PERCOM), 2005.

[17] D. Malan, M. Welsh, M.D. Smith, A public-key infrastructure for key distribution in TinyOS based on elliptic curve cryptography, in: Proceedings of 1st IEEE International Conference Communications and Networks (SECON), Santa Clara, CA, October 2004.

[18] P. Ning, R. Li, D. Liu, establishing pairwise keys in distributed sensor networks, ACM Transactions on Information and System Security 8(1) (2005) 4177.

[19] M. Eltoweissy, M. Moharrum, R. Mukkamala, Dynamic key management in sensor networks, IEEE Communications Magazine 44 (4) (2006) 122130.

[20] X. Du, Y. Xiao, M. Guizani, H.H. Chen, An Effective Key Management Scheme for Heterogeneous Sensor Networks, Ad Hoc Networks, Elsevier, vol. 5, issue 1, January 2007, pp. 2434.

[21] J. Lee, D.R. Stinson, Deterministic key pre-distribution schemes for distributed sensor networks, To appear in Lecture Notes in Computer Science (SAC 2004 Proceedings) (2004).

[22] T. M. Vu, C. Williamson, and R. Safavi-Naini, "Simulation modeling of secure wireless sensor networks," in Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools, Pisa, Italy, 2009.

[23] H. Joengmin and K. Yongdae, "Revisiting random key predistribution schemes for wireless sensor networks," in Proceedings of the 2nd ACM workshop on Security of ad hoc and sensor networks, Washington DC, USA, 2004.

[24] G. Marsaglia and W. Tsang, The Ziggurat method for generating random variables, 2000.

[25] B. Bollobs, Random Graphs: Academic Press, London, 1985.

[26] M.F. Younis, K. Ghumman, M. Eltoweissy, Location-aware combinatorial key management scheme for clustered sensor networks, IEEE Transactions on Parallel and Distributed Systems 17 (8) (2006) 865882.

[27] Carlo Blundo, Alfredo De Santis, Amir Herzberg, Shay Kutten, Ugo Vaccaro, and Moti Yung, "Perfectly-Secure Key Distribution for Dynamic Conferences," in Proceedings of the 12th Annual International Cryptology Conference on Advances in Cryptology, 1993.

[28] P. Traynor, H. Choi, G. Cao, S. Zhu, T. Porta, Establishing pair-wise keys in heterogeneous sensor networks, in: Proceedings of IEEE INFOCOM 06.

[29] A.D. Wood, J.A. Stankovic, Denial of service in sensor networks, Computer 35 (10) (2002) 5462.

 ISBN: 978-1-61208-183-0

Selective Link Cost Alteration in Reservation-Based Multi-Hop Wireless Mesh Networks

Christian Köbel*, Walter Baluja García* and Joachim Habermann†

*Department of Telecommunications, Polytechnic University ISPJAE José Antonio Echeverría
Calle 114, No. 11901, Marianao, Havana, Cuba
{Kobel, Walter}@Tesla.CUJAE.edu.cu
†Department for Information Technology, Electrical Engineering & Mechatronics
TH Mittelhessen University of Applied Sciences
Wilhelm-Leuschner-Str. 13, D-61169 Friedberg Germany
Joachim.Habermann@iem.thm.de

***Abstract*—Link state protocol-driven wireless mesh networks are known for their flexible and highly scalable structure, due to a high degree of individuality, in terms of routing table generation in each connected node. QoS-focused research allows these nodes now to make accurate next-hop decisions based on QoS-sensitive routing metrics. This development enables significantly improved QoS-performance on the network layer. To further adapt link cost calculation to QoS-demands posed by digital media services on higher layers, we propose the Selective Link Cost Alteration (SLCA) scheme, which includes resource reservation demands into the process of neighbor link evaluation, in a distributed fashion. SLCA's goal is to avoid that best effort packets competing with a protection-worthy QoS stream, therefore routing conditions are altered in order to keep the best available path free for QoS-related packets.**

***Keywords*-Wireless Mesh Network; WirelessLAN; Multi-Hop; Multi-Path; Reservation-Based; Policy-Based; OLSR.**

I. INTRODUCTION

Wireless Mesh Networks (WMN) based on IEEE 802.11 W-LAN links have long passed the border to offer simple connectivity between wireless nodes and gateways, since a lot of valuable research has been conducted on this field. Due to these improvements, mostly achieved by well supported mesh routing protocols, research on WMNs now merely concentrates on bringing efficient Quality-of-Service (QoS) approaches to WMNs. To create digital media-ready WMNs is the next important step in WMN research, to finally satisfy also consumer demands on modern wireless networks. Integrated QoS enables to handle high definition content and real time communication, without forcing real-time traffic to compete with best effort traffic on a shared medium. Besides classic approaches to encounter QoS problems, such as bandwidth shaping of disturbing traffic or packet prioritization in the MAC layer, WMNs offer a third highly effective feature: rerouting. In the best case, several routes are available to a destination. This allows to chose a different route for QoS streams in order to avoid bottlenecks or saturated links, or even bypass disturbing traffic to the second-best path, in order to keep the best path "clean" for QoS-related packets. In practice, classic destinations are merely gateways to external host networks, such as the Internet. Intra-mesh client traffic is less common.

The most common approach to improve QoS support in mesh clouds or backbones is by implementing more sophisticated routing metrics. Often cross-layer architectures, which include channel usage and congestion information are used. Since the presented approach is based on, respectively extends the popular Optimized Link State Routing Protocol (OLSR) [1], one of the first QoS-favored routing metrics to consider would be the Expected Transmission Time (ETT). It's a combination of the standard Expected Transmission Count (ETX) metric [1] and the actual bandwidth of a link. Bandwidth, a crucial factor for QoS, is determined via a packet-pair technique [2] within ETT. But making a node aware of such complex link state parameters sometimes comes with a conceptual misbehavior of the load distribution: when more than one next hop is available towards a destination, the entire load (QoS- and potentially disturbing best-effort traffic) tranceived by one of the two end-to-end communication partners, will first be routed towards the best link until it gets saturated and available bandwidth decreases. With ETT then link cost increases due to the usage of the link, which results in a load shift to the next best next-hop. After a while the first link recovers, the load shifts again and the first link is used anew. This results in an oscillating load shift between two next-hops. Such issues make load distribution unpredictable in a larger network scale.

We propose a novel *Selective Link Cost Alteration* (SLCA) scheme with n metric alterations resulting in n routing table variations, which exploits the initially described rerouting ability by distributing reservation messages along the network to reserve concrete bandwidth resources between a source and a destination. Packets not belonging to a reservation are urged to take the next best path. SLCA also solves the described 'load oscillation issue right from

 ISBN: 978-1-61208-183-0

the start, since several paths are used simultaneously, if available.

This paper is structured as followed: Section 2 deals with other research work related to SLCA. Section 3 summarizes the system model. Both Section 4 and 5 then deal with the impact of SLCA. Section 4 depicts SLCA basic algorithm in a concrete example, whose measurement results are summarized in Section 5. Section 6 concludes the work; Section 7 outlines future SLCA improvements.

II. Related Work

Exploiting path diversity with parallel transmissions from a source to a destination is a widely discussed research topic. It has led to a fair amount of multi-path routing protocols in the past few years. One of the more recent ones, developed by Hu et al. [3], is named Multi-Gateway Multi-Path Routing Protocol (MGMP) and extends the Hybrid Wireless Mesh (single path) Protocol (HWMP), which was included in the IEEE 802.11s draft. Their research reveals that using multiple paths to a destination clearly improves QoS parameters such as throughput, delay and packet loss. Similar to standard OLSR and OLSR with SLCA extension, Ghahremanloo [4] compares the standard Ad-hoc On Demand Vector (AODV) protocol with its multi-path version Ad-Hoc On Demand Multi-Path Distance Vector (AOMDV). He comes to the conclusion that AOMDV outperforms single-path AODV in terms of total throughput and end-to-end packet delay. Path diversity was a strong motivation for the SLCA development, but rather in a way that the best path shall be blocked for a priority QoS stream so that alternative paths remain for other traffic. This led to the inclusion of distributed reservations. An initial motivation to investigate routing metric manipulation and link cost alteration was to merge the receiver-initiated Resource Reservation Protocol (RSVP) [5] with the layer 3 routing engine: instead of pure resource reservation through bandwidth reduction of non-reserved traffic along a single path, the impact of using rerouting capabilities for such traffic in every node is investigated here. Concerning RSVP QoS levels, SLCA applies *rate-sensitive* reservations. Köhnen et al. extended the RSVP concept in a broad manner with their QoSiLAN approach [6]. QoSiLAN aims on providing access technology/layer 2 independent and self-organized QoS resource reservations, in originally unmanaged heterogeneous single path LAN networks. QoSiLAN is server-based and relies on collaborative bandwidth shaping of all hosts involved in a packet stream. A host generates an end-to-end reservation message and unicasts it to a control server (QoSiLAN manager). This server proactively monitors the network topology and advises all involved hosts of this stream/path to lower the bandwidth of their ongoing processes, in order to keep free the to-be-reserved resource, on a per link basis. The QoSiLAN system includes monitoring and classification for outgoing traffic, using a variation of the Statistical Protocol IDentification (SPID) algorithm [7] and optionally deep packet inspection features. Such a capability describes a mandatory component for SLCA-ready OLSR-based wireless mesh networks, since actual bandwidth demands of services have to be identified by the originator node. Still, this feature is not included; SLCA furthermore offers a QoS framework for networks, which support QoS-traffic detection and classification. SLCA now extends the QoSiLAN scheme, in a way that bandwidth used for non-reserved traffic does not have to be lowered by involved nodes, as it is necessary in single-path Ethernet or infrastructure networks, where mostly only one path is available. Furthermore SLCA reroutes those packets, if several paths are available. Also, a distributed solution is preferred to a central one, since it better suits the ad-hoc character of a mesh network. Routing topology consistency across the entire mesh network is a crucial deployment factor for SLCA. It relies on all mesh nodes to individually calculate and maintain the same topology, all with the same link cost values. A condition, which is favored by increasing the OLSR Topology Control (TC) message interval, or by using the OLSR fisheye algorithm [1]. Couto et al. deal with the problem of routing table inconsistency due to high OLSR signaling packet loss rates, which might have a severe impact in large-scale mesh networks. For instance, different views on the real topology might lead to routing loops. To further increase common routing table stability, they propose to include control packet loss rates in the development of new routing metrics. Furthermore, the level of inconsistency in routing tables may be increased by adding receive-acknowledgments for topology control updates to new routing protocols.

III. Routing Table Management

A. Concept

The proposed system follows a distributed policy-based, or more precisely said, a reservation-based routing approach, implemented in the mesh protocol. A reservation message basically contains the regarding source and destination sockets, the to-be-reserved bandwidth plus flow identification and may be initiated by any single service running on any node in the mesh cloud and is valid for an entire path. The residual bandwidth always remains available for other traffic. Every node individually increases selected link costs on the best path between the source and the destination (their addresses are included in a reservation r_n), according to the demanded bandwidth value. n is the reservation index. This virtually increases the overall cost of the to-be-reserved path. Routing decisions for packets, which match r_n will not be affected by this alteration and therefore will favor the best route/next hop. Routing decisions for all other packets will see the reserved route as virtually burdened. Thus, a rerouting of potentially disturbing traffic is facilitated, no matter if such traffic occurs or not; a contribution to the proactive character of OLSR. Each new reservation n takes up the routing table

 ISBN: 978-1-61208-183-0

valid during the previous reservation r_{n-1} and adds new routing entries for the affected source and destination nodes, according to the demanded resources in r_n.

If the route with the best conditions (when considered in a load-free state) between a source and a destination is now already loaded with other traffic before r_n is distributed, rerouting this non-reserved traffic to alternative routes is facilitated, instead of forcing the QoS-traffic of r_n to take less loaded paths. If the overall path capacity, partially used by a running stream of an active r_n suddenly decreases due to changing link conditions, it is probable that best effort traffic using the same route is shifted to other routes, before the *entire* load (including r_n traffic) would eventually be rerouted. Such load shifting aspects fundamentally differ from those of existing mesh routing concepts.

It is important to notice that SLCA does not describe a full multi-path system, since packets of a single reservation will never be scheduled over multiple paths simultaneously. It is possible and intended though, that best effort traffic to a common destination might take a different route. Due to stream differentiation, load balancing and packet reordering between multiple paths, as common issues in multi-path systems, are of no importance here, since a single flow always takes a single path. SLCA combines concrete and strict resource reservation methods, described in protocols like RSVP or MPLS, with the dynamical mesh routing character, represented by individual next-hop decisions made by every node. The design reflects a "soft" reservation method, since the reservation is not strictly forced and traffic is rerouted only if applicable. As an example, it wouldn't make sense to reroute disturbing traffic over a route with 4 hops, if the destination is only 1 hop away. On the contrary, SLCA offers strong advantages if several routes are available, which have more or less similar routing conditions. SLCA especially improves the QoS level if the bandwidth of a reserved stream is not constant during transmission, but the maximum peak bandwidth shall still be available constantly. Details on the SLCA algorithm are described in Section IV.

B. Requirements

There are certain general requirements for the used mesh network. At first, r_n has to be known to every node. Therefore, network-wide OLSR topology control messages are chosen for distribution. Every node, respectively its services, must be able to determine bandwidth demand for a stream included in r_n. The ability to predict QoS demands of a service or application, define all necessary parameters required for a valid reservation and generate and trigger a concrete reservation message is therefore considered as given by an external entity, module or program. Our modified OLSR daemon has to run on every node, otherwise topology inconsistency is likely to occur and routing may become unstable, due to differing link cost calculation bases. If SLCA is not active on all nodes in the network, the overall routing behavior might become unpredictable. SLCA also requires every node to have the same view on the topology, since nodes must be able to frequently calculate the best path between the source and the destination of r_n. SLCA is only effective on multi-hop routes, for example evoked through long inter-node distances or obstacles. If the source and the destination of r_n plus potential disturbers roam all within the same coverage area on a shared medium, the impact of reservation is low.

C. OLSR extensions

Here we briefly mention required additions to OLSR's core functionalities. OLSR does not naturally generate routing tables with multiple entries for the same target. This ability was therefore added to its routing engine. The routing core is now also able to process the IP source address of a passing packet. This aspect, combined with the ability to poll the port number from the packet's transport header, is mandatory to finally determine the source and destination sockets of a packet. Finally, a new Wireshark dissector was written to interpret modified TC signaling messages (Wireshark OLSR message type 202).

D. Routing Metric

The SLCA scheme is compatible with any routing metric supported by OLSR, as long as its link cost calculation scheme includes bandwidth as one link quality parameter. Examples for such metrics are WCETT [9], MIC [9] or ETT [2]. For our testbed, we extended ETT (see Eq. 1) in 2 steps:

1) By replacing the bandwidth B in the original ETT Equation with the *residual bandwidth* B_{residual} (see Eq. 2). This extension is already more effective than the regular ETT, as it includes the capacity of a link and therefore offers a more accurate picture of the link condition [10]. In our implementation, the necessary maximum link capacity is obtained by a simple statistical bandwidth analysis of the link: the peak bandwidth is measured and stored over a variable time window. After this period, peak bandwidth is supposed to be the highest achievable bandwidth $B_{\max}$ of the link. Each time a higher peak is reached, previous $B_{\max}$ is replaced with the fresh value.
2) By subtracting the reserved bandwidth from the residual bandwidth (see Eq. 3).

$$ETT_{\text{original}} = ETX * \frac{S}{B} \tag{1}$$

$$ETT_{\text{residual}} = ETX * \frac{S}{(B_{\max} - B)} = ETX * \frac{S}{(B_{\text{residual}})} \tag{2}$$

$$ETT_{\text{SLCA}} = ETX * \frac{S}{(B_{\text{residual}} - B_{\text{reserved}})} \tag{3}$$

where S is the packet size in Bytes, ETX is the Expected Transmission Count metric and B is the actual bandwidth value obtained by link probing.

 ISBN: 978-1-61208-183-0

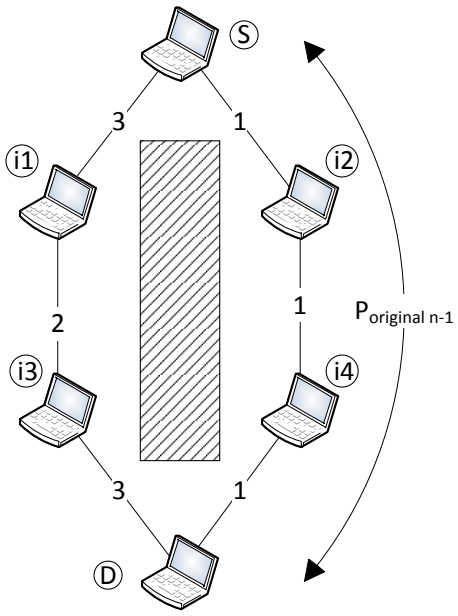


Figure 1. Test setup, no present reservations

Note, that the common advantages of a link cost routing protocol still apply. If the actual link quality changes, whether or not a reservation is active or traffic is present, it will have a proper effect on proactive routing decisions. While using OLSR here, SLCA may be applied to any other link state protocol, as long as routing tables are created in a proactive fashion.

IV. Validation of Concept

Here, the general steps to put a new, network-wide distributed reservation to practice, are explained through a simplified example. Observing only 6 nodes allows to follow the changing routing table entries in detail. Note that this scenario was also used for performance evaluation in Section V. The distance between nodes is 80m, to favor a multi-hop scenario. A solid obstacle, placed in the middle, also serves to shape the desired topology. For the sake of simplicity, $n = 1$ here. Figure 1 shows the topology present in the network layer; The source (S) will distribute a reservation of 5 Mbit/s for bidirectional traffic with the destination node (D). In this concrete setup, the maximum available bandwidth per link is set to 7 Mbit/s. The corresponding dimensionless link cost values are shown in Figure 2, with the resulting routing table for S. The following steps are performed:

1) S is about to start a media stream to D and generates a reservation message m_n, which specifies: $\text{socket}_{\text{source}}$, $\text{socket}_{\text{destination}}$, demanded bandwidth [bits/s] by the stream/application/service, flow ID, transport protocol information (TCP/UDP) and DSCP/DiffServ class priority (yet unused)

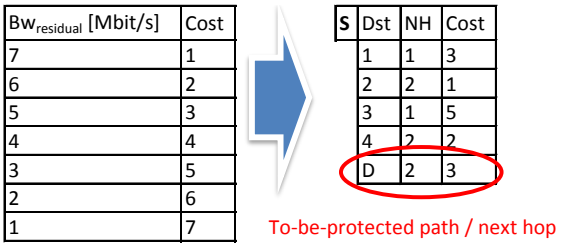


Figure 2. Initial routing table for S with raw link cost map

2) m is included in OLSR TC messages and flooded using the OLSR Multi-Point-Relay (MPR) mechanism
3) S and the intermediate nodes 1-4 individually determine the best route available from S to D. The best path naturally is $P_{\text{original}-\text{n}-1}$ due to the lowest overall cost of 3, which shall be protected.
4) All nodes change and extend their original routing tables $(n-1)$ by recalculating $i1_n$ - $i4_n$ and by adding additional entries for D_n and S_n, as shown in Figure 3. Note, that intermediate nodes have two entries for S and D, in case they forward packets (mis)matching the reservation criteria. S and D also have routing entries for each other, for packets between them, which do not belong to any reserved stream between the two-end-to-end points (e.g., best effort traffic).
5) All nodes adapt the new costs for every single link of $P_{\text{original}-\text{n}-1}$: the bandwidth component within the routing metric and it's output is manipulated according to the reserved bandwidth. Link costs on the reserved route are virtually increased accordingly. The updated costs are used by all nodes to recalculate paths to all destinations. A new topology is created as described by Figure 4. Despite virtual worsening, all paths are still load free. $P_{\text{alternative}-\text{n}}$ shall now be preferred by disturbing traffic.
6) Packets, which match *all* reservation's criteria are allowed to use unaltered routing entries D_{n-1} and S_{n-1}, all other packets are routed according to S_n, 1_n, 2_n, 3_n, 4_n and D_n. In this way S, D and all forwarding nodes between them use the best route for the reserved stream, while all other traffic is urged to not use the virtually more expensive route between S and D, if possible. This leads to less congestion on $P_{\text{original}-\text{n}-1}$, which increases QoS level on this path.
7) The media stream is running on reserved route. Even if the load situation in the network changes dynamically, selective alteration is always applied to changing link costs.
8) Reservation is either actively relinquished by S when the media stream ends, or becomes invalid automatically, due to a validity time counter t added to OLSR (*soft state* - S has to refresh its reservation constantly)

 ISBN: 978-1-61208-183-0

S	Dst	NH	Cost
	1_n	1	3
	2_n	2	6
	3_n	1	5
	4_n	2	12
	D_{n-1}	2	3
	D_n	**1**	8

1	Dst	NH	Cost
	S_{n-1}	S	3
	S_n	S	3
	2_n	S	9
	3_n	3	2
	4_n	**3**	11
	D_{n-1}	3	5
	D_n	3	5

2	Dst	NH	Cost
	S_{n-1}	S	1
	S_n	S	6
	1_n	S	9
	3_n	**S**	11
	4_n	4	6
	D_{n-1}	4	2
	D_n	4	12

3	Dst	NH	Cost
	S_{n-1}	1	5
	S_n	1	5
	1_n	1	2
	2_n	**1**	11
	4_n	D	9
	D_{n-1}	D	3
	D_n	D	3

4	Dst	NH	Cost
	S_{n-1}	2	2
	S_n	2	12
	1_n	D	11
	2_n	2	6
	3_n	D	9
	D_{n-1}	D	1
	D_n	D	6

D	Dst	NH	Cost
	S_{n-1}	4	3
	S_n	**3**	8
	1_n	3	5
	2_n	4	12
	3_n	3	3
	4_n	4	6

n = 1
(n - 1)... previous/original table
to-be-protected
improvement through re-routing

Figure 3. Simulation Results: Routing Table Development

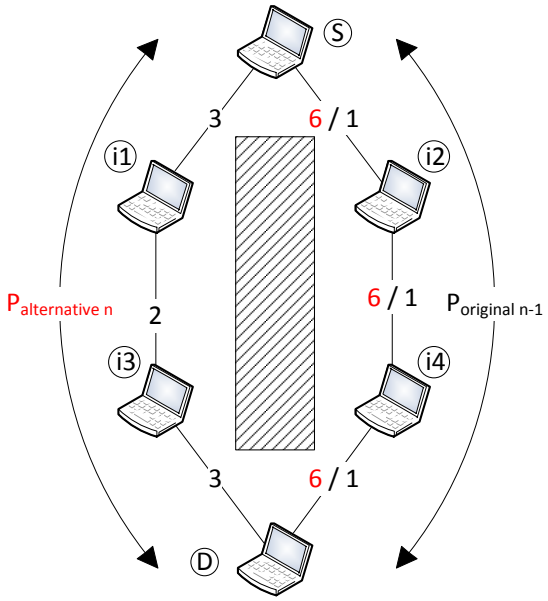


Figure 4. Altered topology with 1 active reservation

Figure 3 reveals that 5 of 6 routing tables show improvements, marked in red, in a way that the next hop decision was altered in favor of the relief of $P_{\text{original}-n-1}$.

V. Measurements

The QoS performance in terms of bandwidth for the measurement setup depicted in Figure 1 is evaluated. As

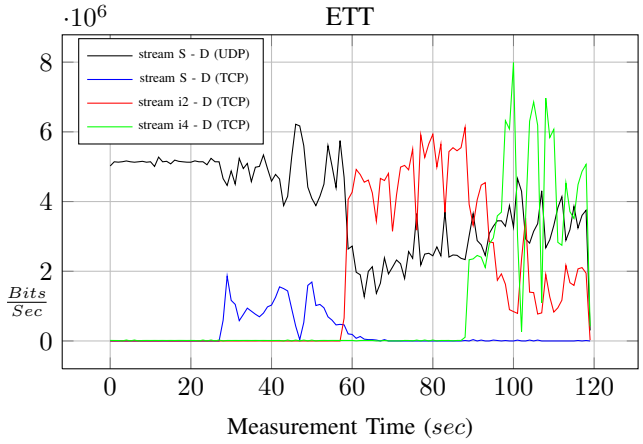


Figure 5. Results: BW comparison of 4 streams with regular ETT metric

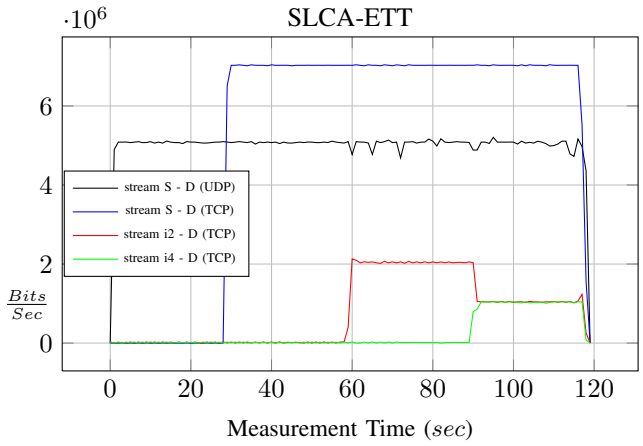


Figure 6. Results: BW comparison of 4 streams with SLCA-ETT metric

a multi-hop simulator, Omnet++ [11] with INETMANET framework [12] is deployed. The IEEE 802.11g mode is used on all links. The overall measurement time is 120 seconds. At first, node S initiates a 5 Mbit/s UDP stream to node D. After 30s, a TCP stream is initiated by S on the same path. The intermediate nodes 2 and 4 start a TCP stream to D after 60, respectively 90 seconds. At first, the regular ETT metric is used; results are shown in Figure 5. Secondly, SLCA is applied. Therefore, S broadcasts a reservation for its UDP stream at 5 Mbits/s, according to Section IV. Now, the next-hops in each node are chosen on a per packet-basis, depending on the to-be-routed packet (whether it matches criteria described in r_1 or not). The results in Figure 6 reveal that the reserved UDP stream suffers from less disturbances by best-effort TCP traffic and offers a more stable performance, in contrast to the unmanaged OLSR scenario in Fig. 5. Also, the two streams initiated by S (TCP and UDP) are separated on both available routes, which exploits the present path diversity.

 ISBN: 978-1-61208-183-0

VI. Conclusions

The Selective Link Cost Alteration is an experimental resource reservation scheme for WMNs, which facilitates rerouting of potentially disturbing traffic on reservation-protected multi-hop routes. SLCA provides a flexible framework for QoS-related media services in mesh networks. Its concept is adaptable to other QoS metrics as well, since the to-be-altered routing parameter, and its *actual* and *required* value, might be replaced by the packet delay, packet error rate or any other QoS parameter instead. A combination of several parameters is also feasible, in order to enable more precise QoS definitions.

VII. Open Issues and Outlook

As a typical problematic condition for both, centralized and distributed reservation systems, its resources are always finite. More than the available resources can't be assigned or managed. Now, research investigates into balancing a certain threshold, in a way that reservations are denied, or not even triggered, if single link capacities are physically limited. Similar to RSVP, a reservation cannot be guaranteed. Contrary to RSVP, the initiator of a reservation is not informed about its feasibility by intermediate nodes. It is intended to solve this issue by adding an unicast signaling message, to confirm a reservation to its originator. Although, such confirmations are not included yet, research has shown that alternating the link costs by considering QoS needs clearly works in favor of intended reservations.

Future performance evaluations of SLCA will also deal with a realistic maximum number of reservations per node. Although the routing table management is scalable and theoretically does not limit the amount of active r_n, too many additional routing entries for each new reservation might result in an unmanageable routing table, or fully occupied computation hardware. n is therefore always finite, which concludes in a threshold for $n_{\max}$. This threshold defines the state, from where new r_n (apart from pending ones) will have a contrary effect on QoS performance and is yet to be determined by further investigations. Using SLCA in larger mesh clouds without scaling $n_{\max}$ might therefore cause unpredictable network performance problems.

Also SLCA allows for some general, conceptual amplifications. A prioritization of active reservations is desired. In the latest version, the earlier a reservation is registered by OLSR, the more unspoiled bandwidth resources it has available for link cost recalculation; all following reservations then only manage residual resources. This behavior might be replaced by a fairness scheme, which prioritizes reservations based on certain characteristics, even if they have arrived later than reservations for less relevant packet flows. Reservation usage feedback remains also subject to refinement: if a reservation is successfully applied on a multi-hop route, it should be used by the following related traffic as well. Future SLCA versions must register if reserved packets are actually passing; an action, which requires further cross-layer elements, like statistical packet analysis, in a node. If not, expected packets, which haven't arrived or have arrived only in unsatisfying quantities, have to be announced and the reservation might be canceled ahead of schedule.

As our work is closely related to multi-interface, multi-channel mesh networks [13], it is intended to include channel diversity as another resource in SLCA, to further improve capacity utilization in wireless mesh networks.

References

[1] A. Tonnesen, A. Hafslund, and Ø. Kure, "The UniK- OLSR plugin library", in *OLSR Interop Workshop*, San Diego, August 6-7 2004.

[2] P. Esposito, M. Campista, I. Moraes, L. Costa, O. Duarte, and M. Rubinstein, "Implementing the expected transmission time metric for olsr wireless mesh networks", in *1st IFIP Wireless Days WD 08*, 2008.

[3] Yun Hu, Weiqing He, Shoubao Yang, and Yuan Zhou, "Multi-gateway multi-path routing protocol for 802.11s WMN", in *2010 IEEE 6th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, 2010, pp. 308-315.

[4] P. Ghahremanloo, "Multi-path routing challenging single-path routing in Wireless Mesh Networks: Network modeling of AODV and AOMDV", in *Wireless Advanced (WiAd) 2011*, 2011, pp. 176-179.

[5] Network Working Group, "Resource ReSerVation Protocol (RSVP)", RFC 2205, September 1997. Online available: http://tools.ietf.org/html/rfc2205

[6] F. Adamsky, C. Köhnen, C. Überall, V. Rakocevic, M. Rajarajan, and R. Jäger, "A novel concept for hybrid quality improvements in consumer networks", in *2011 IEEE International Conference on Consumer Electronics - Berlin (ICCE-Berlin)*, pp.39-43, 6-8 Sept. 2011

[7] C. Köhnen, C. Überall, F. Adamsky, V. Rakocevic, M. Rajarajan, and R. Jäger, "Enhancements to Statistical Protocol IDentification (SPID) for Self-Organised QoS in LANs", in *2010 Proceedings of 19th International Conference on Computer Communications and Networks (ICCCN)*, pp.1-6, 2-5 Aug. 2010

[8] R. S. Couto, M. E. Campista, L. H. M. Costa, and O. C. M. Duarte, "An experimental analysis of routing inconsistency in indoor wireless mesh networks", in *2011 IEEE Symposium on Computers and Communications (ISCC)*, 2011, pp. 609-614.

[9] S. S. Ahmeda, and E. A. Esseid, "Review of routing metrics and protocols for wireless mesh network", in *2010 Second Pacific-Asia Conference on Circuits,Communications and System (PACCS)*, 2010, Vol. 1, pp. 27-30.

[10] Hao Zhou, Chuanhe Huang, Yong Cheng, and Gang Wang, "A New Multi-metric QoS Routing Protocol in Wireless Mesh Network", in *International Conference on Networks Security, Wireless Communications and Trusted Computing*, NSWCTC '09, Vol. 1, pp.459-467, 25-26 April 2009

[11] "OMNeT++ Network Simulator Framework", http://www.omnetpp.org

[12] "INETMANET Framework for OMNEST/OMNeT++", https://github.com/inetmanet

[13] C. Köbel, W. Baluja, and J. Habermann, "A Survey on Multi-Channel Based, Digital Media-Driven 802.11 Wireless Mesh Networks", in *ICWMC 2011: The 7th International Conference on Wireless and Mobile Communications*, 2011, pp. 169-175.

 ISBN: 978-1-61208-183-0

Reliable Technology for Wireless Mesh Networks with Low System Requirements

Vladimir Sulc
MICRORISC s.r.o.
Jicin, Czech Republic
sulc@microrisc.com

Radek Kuchta | Radimir Vrba
Faculty of Electrical Engineering and Communication
Brno University of Technology
Brno, Czech Republic
kuchtar | vrbar @feec.vutbr.cz

***Abstract* – In this paper IQMESH, new networking technology for wireless mesh networks, its basic principles and related routing algorithms are presented. The presented technology was developed especially for applications in the field of buildings automation and telemetry. However, other applications such as smart grids or street lighting can also benefit from straightforward implementations and low system resources requirements. An implementation for IQRF communication platform, described at the end, shows actual system resources requirements in a specific scenario limited by 240 hops and 65 thousand devices.**

Keywords-wireless; mesh; networking; routing; algorithm; IQRF; IQMESH.

I. INTRODUCTION

IQMESH is a networking technology developed for Wireless Mesh Networks (WMN) utilizing packet transmissions. In such networks, messages are sent in smaller parts called packets. A packet has information about a recipient and, in general, the mesh network. This is transmitted from a sender to its recipient through nodes connected in the mesh network. A strategy for sending packets from one node to another is commonly known as routing, and its goal is to deliver packets efficiently and reliably.

A mesh network in which every node has a direct link to another node is a fully connected mesh network. In real WMN only partially connected mesh networks are used, which means that there is no universal direct link between devices. As an illustrative example of such mesh network topology and related routing is to compare it to vehicles traveling between cities. The whole path from the origin to the destination consists of numerous individual roads connecting cities. Searching for the best connection between two cities is similar to mesh networks finding the shortest and most efficient path, from the origin to the destination, between two selected nodes.

A link can be established between any two nodes in a mesh network. In a network consisting of n nodes and one coordinator, the number of possible links will always be lower or equal to N_{MAX} calculated as (1).

$$N_{MAX} = \frac{n\,(n-1)}{2} \quad (1)$$

Devices in WMN communicate with each other wirelessly, so communication between two devices is usually limited by the communication distance, or so called range limitation of these two devices. Positions of devices in general WMN are not known and the range limitation can depend on many conditions, therefore the routing is usually a great algorithmic challenge to find the best path the packet should travel along. More nodes result in more possible links and consequently to more combinations of links.

Many different routing algorithms are used practically. A flooding, routing based on tables and random routing are just a few basic examples of such algorithms. Unfortunately, due to the many specifics of WMN and limitations of the target, application is not possible to easily utilize such algorithms.

Flooding in a general mesh network is based on propagation of the packet over the whole network and is to be considered as the most reliable for WMN. Real implementations of WMN in industrial, scientific and medical ISM radio bands are limited physically by the connection speed, the so called bit rate, resulting in big delays and low network responsiveness.

Routing algorithms based on the sharing and distribution of routing tables or vectors are usually considered to be the most efficient. High memory demands and big overload in the case of distribution routing tables usually limit usage of such algorithms for larger WMNs where resources of nodes controllers are not limited by the economy of the project.

Possible packet collisions in connection with lower bit rates limits real implementations of random routing in WMN and practically disqualifies telemetry and control applications where the highest reliability should be achieved.

WMNs are nowadays considered and already used as a communication platform for many different applications in the field of telemetry and automation. Automatic meter reading AMR, street lighting control or smart grids are just few examples of such applications utilizing networks with hundreds or thousands of devices. Therefore both the cost of communication devices and high reliability of the routing should be priorities. Technology described in this paper provides both reliable and effective packet delivery solutions with minimal demands on system resources.

In this paper IQMESH, networking technology for wireless mesh networks, is presented. Basic principles of the technology is followed by explanation of discovery and routing. Reliability and efficiency aspects will be discussed further on. Specific implementation of IQMESH routing technology will be described in the end.

II. RELATED WORK

Efficient and reliable packet delivery in large wireless networks consisting of hundreds or even thousands of

 ISBN: 978-1-61208-183-0

devices and supporting up to several hundred hops is a big algorithmic challenge. Considering actual speed, output power and spectrum limitations, as well as economic factors, a flooding mechanism seems to be the most viable for most target applications.

Therefore, flooding is commonly used in wireless ad-hoc networks. There are many techniques of flooding differing in control algorithms, efficiency, reliability and overhead.

The simplest flooding technique is based on re-transmitting only new, not yet registered packets. In this scheme every packet should be identified and is re-transmitted only once. This mechanism guaranties that a packet is delivered to all nodes at minimal costs. In the real environment of a WMN, collisions would affect functionality and result in high traffic [1]. Reducing flooding traffic is the goal of many approaches to make the flooding mechanism more reliable and efficient [1]. Proposing a probable flooding scheme, e.g., distance-based, location-based or cluster-based flooding.

Schemes in category 1-hop neighborhood are based on knowledge of the closest neighbors reachable directly in one hop. Different approaches [2-5] are based on 1-hop neighborhood knowledge. Cai et al. [2] propose adding the list of its 1-hop neighborhood to the packet, and recommends to the receiver not to forward the message if its complete 1-hop neighborhood is already included in the received list.

Schemes in the category 2+ hops neighborhood are based on storage of the neighborhood, which is limited by the number of hops from each node. In this category every node knows the network topology up to n-hops. In [6] Qayyum et al. propose a heuristic algorithm to compute multiple relays. Ko et al. in [7] propose improving broadcast operations for ad-hoc networks using 2-hop connected dominating sets.

Spohn et al. in [8] argue by simulations that protocols focused on making an optimal broadcast tree with the implicit assumption that all nodes should be reachable from the source may no longer be true because flooding protocols in wireless sensor networks are used to deliver data packets towards a single, or only subset, of destination nodes and proposes a new flooding protocol for utilizing directional information to achieve efficiency in data delivery.

Time Synchronized Mesh Protocol, described in [9] is based on TDMA and requires sharing time information and precise time synchronization of all nodes. Overall power consumption would benefit from the time synchronization, but on the other hand interference, collisions and environment influences would impact delivery reliability.

III. IQMESH

IQMESH is the networking technology developed for WMN with a coordinator and utilization of packet transmissions. Reliability is achieved through a flooding mechanism, collisions are avoided by TDMA and its routing efficiency is based on the virtual routing structure VRS, created by the coordinator during discovery. The following paragraphs provide a step by step analysis of particular parts of the IQMESH technology.

A. *Basic principles*

WMN is a general network of devices connected wirelessly. Every device in the network has some unique information (address) enabling addressing inside of the network - MAC, node ID, index, address, etc. Packets sent in such network include the address of the recipient. The principle of IQMESH technology is to extend this addressing space and define a new virtual routing structure in the network. A coordinator will dedicate to every device, found during discovery process, a unique Virtual Routing Number VRN, which will be used for future routing. Figure 1a shows an illustrative example of a standard network, where its nodes can be addressed by their address N1 – N5, after discovery, additional routing information is added as shown in figure 1b. Only VRN are used for the routing, while devices' addresses are used solely for addressing. Flooding and other routing algorithms can benefit from systematic indexing of nodes by VRN, e.g., if the VRN reflects distance by the number of routing hops from coordinator to the node.

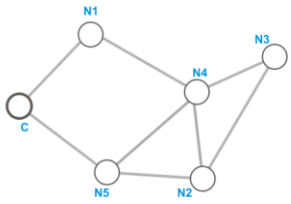


Fig. 1a Network example

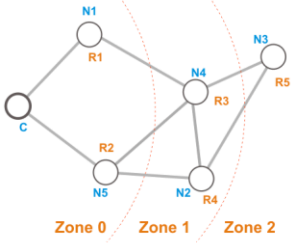


Fig. 1b Network after discovery

B. *Discovery*

During discovery, the coordinator seeks out nodes connected to the network and dedicates them to a unique VRN, reflecting their distance from the coordinator. For example, an incremental indexing can be used. The coordinator starts to search its neighbors. All devices responding to its "Answer Me" type message will receive their VRN. Based on received response it is assumed that the link between coordinator and responding nodes should be symmetrical, enabling future routing both to and from the coordinator. All nodes directly responding to the coordinator have a direct link to the coordinator, so they should belong to Zone 0, being directly accessible without routing. Then, the coordinator incrementally asks all nodes from Zone 0 to discover their 1-hop neighborhood and then dedicates a VRN to all newly found nodes which have not been found in the previous step, and thus not belonging to the Zone 0. Each node can also store some additional information in this step, e.g., respective zone number, parent's VRN, parent's network address or VRN of the first node in respective zone. This information would later be used for routing optimization. Processing all answers from nodes belonging to Zone 0, the coordinator will know all nodes belonging to Zone 1, which are nodes accessible to the coordinator by one routing hop. The same procedure will be then invoked recursively for all nodes belonging to zones Zone 1 and higher until all nodes are found or until there are some further zones available. At

 ISBN: 978-1-61208-183-0

the end of discovery every found, and thus discovered, node has a unique VRN reflecting its distance from the coordinator. In typical applications such as smart buildings, telemetry systems and street lighting the discovery is made just once during the installation phase.

C. Routing

The goal of packet routing in target applications is to reliably and efficiently deliver data over the network. In IQMESH based networks, the flooding mechanism is primarily used. VRS created during discovery process is directionally flooded. The network would be flooded from the coordinator to the node for all control purposes or from the node to the coordinator for data collection. A special order of VRS together with TDMA enables a directional, efficient and collision free flooding mechanism based on VRN. Every node routing packet in its dedicated time slot will also add to the packet its own VRN_X enabling other nodes to know and consequently synchronize to their respective time slot. The coordinator uses VRNC equal to 0. The network routing mechanism is illustrated in Fig. 2.

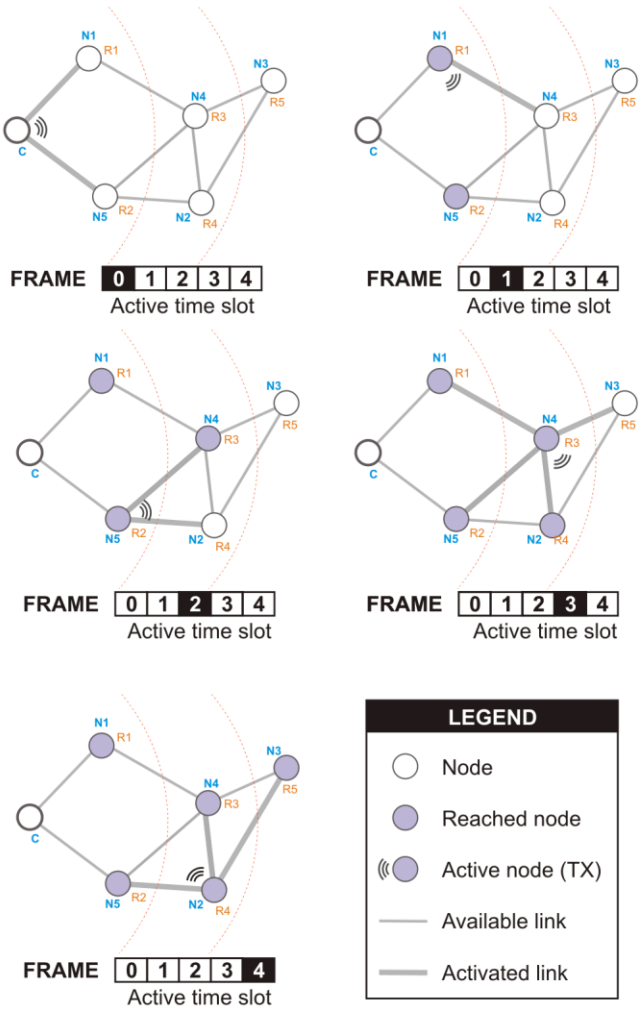


Fig. 2 Routing mechanism

D. Time synchronization

In contrast to many other techniques, e.g., [9], there is no need to share precise time information over the network. Routed packets keep track of number of hops made. This number corresponds with the respective time slot based on VRN for every device. This mechanism, together with adding information of length of time slot, enables efficient re-synchronization of all devices based on packet reception. In addition, dynamic timing based on packet length can be supported.

E. Resources

Generally, the coordinator needs to hold the information of the discovered nodes, whereas each node should keep VRN information only. This means no special HW resources are required to implement IQMESH technology. Storing some additional information recognized during discovery, as described in future paragraphs, can dramatically increase the efficiency of the routing. Specific IQRF implementation showing real system requirements will be mentioned further on.

F. Optimizations

Additional system resources dedicated to the coordinator or to each node can increase future routing efficiency. Storing the address or VRN of a parent node by each node could be mentioned as a good example. In such case, every node can reach the coordinator quite efficiently by tree topology via parent nodes. Also, the storing of minimal VRN in the respective zone, or equivalent saving of the maximal VRN of the previous zone by each routed node can be used to increase routing efficiency. Software techniques used during discovery or following discovery process can also increase the final efficiency of the routing, e.g., the coordinator can exclude all leaf nodes without child links from VRS during discovery.

G. Reliability

The described IQMESH technology ensures reliable and efficient directional flooding. Due to the expected redundant links in many real WMNs, it dramatically increases reliability. A temporarily lost link will obviously not cause failure of packet delivery. Routing mechanisms making use of underlying TDMA avoid conflicts as every routing node has just one dedicated time slot corresponding to its VRN. Many tests with environment noise simulations have confirmed reliability increase. Usually only noise generated during time slots dedicated for devices without redundant links can cause failure in packet delivery. Noise generated during the first time slot usually affects delivery, as no redundant links have yet to be created, which is a first time slot issue FTSI. Overall performance in standard office building environments was measured and a fail rate based on several weeks of experimentation resulted in 1 not delivered packet from 17 250 transmissions. Two additional slots for the coordinator were used to fix the FTSI.

 ISBN: 978-1-61208-183-0

H. *Efficiency - routing from coordinator*

As redundant paths resulting from the principle of VRS flooding are expected and packet collisions are eliminated by TDMA, it might not be obvious to use reception acknowledgment during flooding. However, this assures fair routing efficiency without any impact to reliability. Based on the TDMA, flooding routing realized via VRS for nodes with $VRN_X < VRN_A$, where VRN_A is VRN of addressed node, and assumption that every node is addressed in the same frequency, the average frame will consist of time slots, where n is the number of nodes in a WMN:

$$T_{AVG} = \frac{\sum_{k=1}^{n} VRN_k}{n} = \frac{n}{2} \quad (2)$$

Generally, blind flooding efficiency in similar cases would be calculated as a number of links to ensure 100% reliability of packet delivery. Comparing (1) and (2), we can see dramatic efficiency increase.

For any addressed node within the zone Z_x, only nodes belonging to previous zone Z_{x-1} should make the routing without any strong impact to the reliability. The resulting efficiency based on this presumption will be higher, but always dependent on the topology of the specific WMN. The following formula reflects expected system efficiency for such a scenario:

$$T_{AVG} \leq \frac{n}{2} \quad (3)$$

Efficiency of this routing scenario, skipping redundant routes by nodes in the same zone Z_x for specific node X, can be expected as (4), where VRN_{Zx-1} is VRN of the last node related to the zone Z_{x-1}. Based on this principle, all nodes related to the zone Z_0 can be addressed directly without routing.

$$T_{AVG} = VRN_{Z_{x-1}} \quad (4)$$

I. *Efficiency - routing to coordinator*

As the matter of fact, there is information about parent nodes recognized and stored during discovery of every node. This information means that there is a tree topology available for routing packets from nodes to the network coordinator. In such a scenario, every routing node in its time slot sends a packet exclusively to its parent. Therefore in using TDMA, the number of time slots is equal to the number of hops and, for each frame, corresponds to the zone number Z_X for the node originating communication. Assuming Z_{MAX} as a maximum zone number in the network while indexed from 0, it would be generally proven:

$$T_{AVG} \leq Z_{MAX} \quad (5)$$

Reliability increase is mostly preferred in typical WMN applications. Oriented flooding with redundant backup paths can be used for such applications. In this case, each node originating communication to the coordinator will use its own VRN_X number as a limit of hops. For such routing, similar efficiency like (2) is expected.

Avoiding redundancy of routing by nodes from the same zone, routing efficiency for a specific node X can be expressed similarly like in formula (4), where is a VRN of the last node related to the zone Z_{X-1}.

IV. IQMESH IMPLEMENTATION IN IQRF PLATFORM

IQRF is the communication platform and related technologies [10]. The name IQRF stands for an Intelligent Radio Frequency. Basic specifications and early designs were made in 2004, when, in Malaga, Spain, the first integrated modules were introduced. IQRF is the platform integrating a variety of components for building LR-WPAN in an easy way, simplifying and shortening the design phase of a wireless communication system. Specific implementation of IQMESH routing technology will be described in following paragraphs.

A. *Network abstraction and limitations*

IQRF platform addresses mainly LR-WPAN applications, such as building automation systems and telemetry systems. In such applications many devices can be connected in a fixed infrastructure, usually created during the installation process, and provides permanent links to other devices in this infrastructure. There are commonly other devices connected in the network without permanent links to the other devices in the network, e.g., because of the mobility of such devices or because of power supply limitations. Based on these criteria, IQRF abstracted network topology is described in Fig.3.

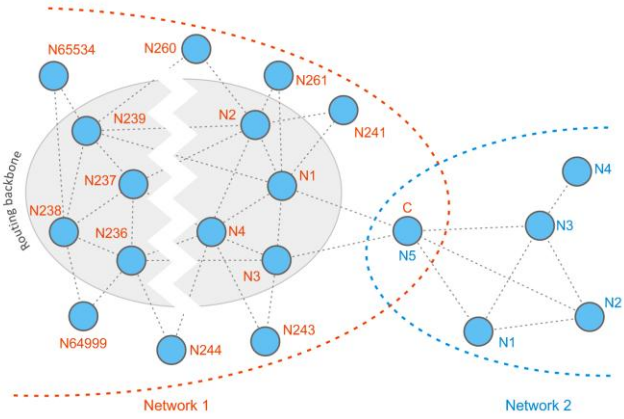


Fig. 3 IQRF abstracted network topology

The routing backbone usually consists of devices providing permanent links and bonded to the network during the installation process. Typical examples for such devices would be equipment like gateways, actuators or sensors mounted in specific locations. Such devices can be used as a routing backbone. IQMESH defines Virtual Routing Structure (VRS) to be used for directionally oriented flooding. VRS would be created from such infrastructure during the installation process (or later on). IQRF OS

available on every transceiver module provides functionality for automation of such processes through function *discovery(x)* instigated by the coordinator and via servicing of system packets by node devices.

As visible from Figure 3 and description available in [10], several limitations were applied to increase efficiency of the network. Routing backbone consists of up to 239 devices, allowing very efficient one byte addressing, broadcast and also group messages. VRS is created from routing backbone devices very simply by calling *discovery(x)* function, where parameter *x* limits the number of zones, which is related to a depth of the network. Other devices will receive logical network address 0xFE from the coordinator during bonding. To extend the addressing scheme for addressing 65k devices, additional two byte user addresses are used. Every IQRF device can work simultaneously in two wireless networks, further extending the network or chaining of several networks allowing the possibility to build up larger systems.

IQRF specific implementation of IQRF OS 3.0x on TR-5xx transceiver modules supports several non-routing and routing schemes. A specific routing scheme is chosen in application based on requested efficiency and purpose by setting system variable *RTDEF*. Routing based on network logical address, tree routing to the coordinator and routing based on VRS are three basic routing schemes supported by current implementation of IQRF OS. Addressing in the network is realized via the logical network address obtained during bonding which is one byte long or via a two byte user address dedicated by the user by calling function *setUserAddress(x)*.

B. Dynamic time slots

To avoid conflicts within the network during routing, TDMA is used. One frame can include up to 240 time slots, allowing for up to 240 hops in the network. In IQRF, time slots are measured and set up in ticks, every tick is 10 ms long. As data load in the packet can consist from 1 byte up to 128 bytes, and 19.2 kb/s is the typical bit rate used for transmissions, 1.2 kb/s up to 115 kb/s are supported, the length of the frame would be too long if a fixed time slot is used.

Support of dynamic time slots based on the data load in the packet and requested purpose dramatically increased routing efficiency. Time slot is defined by setting the variable *RTSLOT* to the number of ticks convenient for a specific purpose. Polling request of the coordinator, e.g., can include just a specification of one or more nodes which should send data to the coordinator. In such a case, a minimum time slot 1 tick long can be used to propagate this request over the network, assuring delivery in 2.4 s in the worst case to any device. Time slot 5 ticks long together with simultaneous choosing of tree routing schemes can be used for a 128 byte long answer to assure maximum time efficiency.

Routing description, such as examples demonstrating routing and right parameters setup for specific purposes are available at User's and Reference guides, download-able from [10].

C. System resources

IQRF OS, including complete support both for the coordinator and nodes, is ported to TR-52Bx modules based on a PIC16F886 microcontroller. System resources used for routing and related services:

Program memory:	< 2k instructions
Data memory:	< 40 bytes (node mode)
	< 300 bytes (during discovery)
EEPROM:	< 40 bytes (node mode)
	< 2k bytes
Packet overload:	+ 6 bytes

D. Development, testing, results

The standard testing environment during development was based on a set of 200 node devices, each device including transceiver module TR-52BA inserted into a DK-EVAL-03 evaluation board and a GW-USB-04 In Circuit Wireless Programmer enabling bulk programming of all devices by one click and from one coordinator device consisting of transceiver module TR-52BA and CK-USB-02 universal programmer / debugger.

In-building applications mainly for lighting and dimming control realized in networks consisting of hundreds and thousands of devices confirmed reliability and the ability to work in real time. On the other hand, due to such a local environment, just a few hops were needed to cover the whole building.

The real challenge was street lighting control, covering large parts of towns, with networks composed of up to 200 devices with different networks using different channels to avoid spectrum concurrency. Fig. 4 shows one of such implementations realized in the suburb of Nitra, Slovakia, EU. Several kilometers range were covered by devices based on transceivers supporting only 3.2 mW of output power with small PCB antenna.

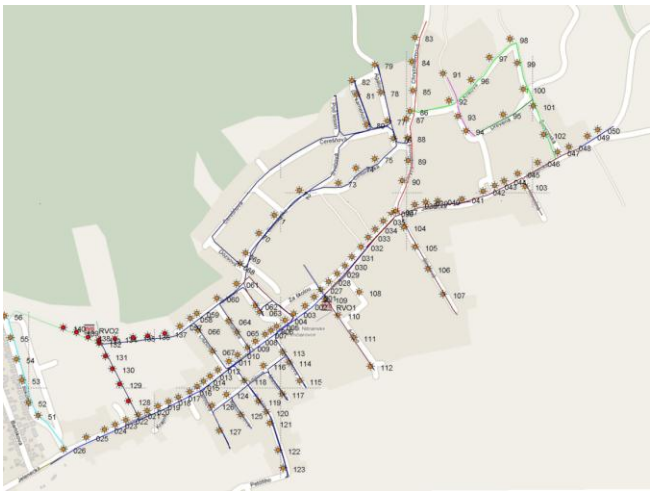


Fig. 4 Real implementation of street lighting application

 ISBN: 978-1-61208-183-0

V. CONCLUSION

IQMESH networking technology for wireless mesh networks, its basic principles and related routing algorithms were presented. Specific implementation in IQRF communication platform was described. Like with any other technology or algorithm, IQMESH applications can benefit from technological advantages and would be affected by its limitations. The flooding scheme would be an excellent option for telemetry systems, e.g., AMR applications for water meters providing data just a few times a day or for street lighting applications. On the other hand, many redundant links and consequent time delays can generate difficulties in real time applications and missing support for node to node communication would create more programming work on the application layer.

VI. FUTURE WORK

Spreading over frequency spectrum instead of TDMA, a combination of both methods, achieving higher efficiency of the routing, increasing reliability in noisy environments and the usage of described technology are just few topics for future research. Future advanced data aggregation algorithms can benefit from VRS and routing schemes.

ACKNOWLEDGEMENT

This research has been supported and co-financed by the Ministry of Industry and Trade of the Czech Republic under contracts FR-TI1/058 "Project Smart House - Open Platform" and FR-TI3/27, project "Open Platform for Modern Cities".

REFERENCES

[1] Sze-Yao Ni, Yu-Chee Tseng, Yuh-Shyan Chen, and Jang-Ping Sheu. The Broadcast Storm Problem in a Mobile Ad Hoc Network. ACM MOBICOM, pp. 51-162, Aug' 1999.

[2] Ying Cai, Kien A. Hua, and Aaron Phillips, “Leveraging 1-hop Neighborhood Knowledge for Efficient Flooding in Wireless Ad Hoc Networks”, 24th IEEE International Performance Computing and Communications Conference (IPCCC), April 7-9, 2005, Phoenix, Arizona.

[3] Xinxin Liu, Xiaohua Jia, Hai Liu, and Li Feng, “A Location Aided Flooding Protocol for Wireless Ad Hoc Networks”, MSN 2007, LNCS 4864, pp.302-313, 2007.

[4] Hyojun Lim and Chongkwon Kim, “Multicast Tree Construction and Flooding in Wireless Ad Hoc Networks,” In Proc. of the ACM Int’l Workshop on Modeling, Analysis and Simulation of Wireless and Mobile System (MSWIM), pp 61-68, Aug. 2000.

[5] Hai Liu, Pengjun Wan, Xiaohua Jia, Xinxin Liu and Frances Yao, “Efficient Flooding Scheme Based on 1-hop Information in Ad Hoc Networks”, in proceedings IEEE infocom, Communications Society subject, 2006.

[6] Amir Qayyum, Laurent Viennot, and Anis Laouiti, “Multipoint Relaying for Flooding Broadcast Messages in Mobile Wireless Networks,” In Proceeding of the 35th Hawaii International Conference on System Sciences, 2002.

[7] Marco Aurelio Spohn, Jose Joaquin Garcia-Luna-Aceves, “Improving Broadcast Operations in Ad Hoc Networks Using Two-Hop Connected Dominating Sets,” In Proceedings of IEEE Global Telecommunications Conference Workshops, 2004.

[8] Young-Bae Ko, Jong-Mu Choi and Jai-Hoon Kim, " A New Directional Flooding Protocol for Wireless Sensor Networks Information Networking. Networking Technologies for Broadband and Mobile Networks Lecture Notes in Computer Science, 2004, Volume 3090/2004, 93-102, DOI: 10.1007/978-3-540-25978-7_10

[9] Kristofer S. J. Pister and Lance Doherty, "TSMP: Time Synchronized Mesh Protocol," In Proceedings of the IASTED International Symposium, Orlando, 2008.

[10] MICRORISC. IQRF - wireless technology. <retrieved: 12, 2011> Available: www.iqrf.org

 ISBN: 978-1-61208-183-0

Forwarding and Routing Stateless Multi-Hop Protocol for Wireless Sensor Networks

Rivo S. A. Randriatsiferana*, Richard Lorion*, Frederic Alicalapa* and Fanilo Harivelo†
*University of La Reunion, LE2P Lab, Saint Denis, Reunion, France,
Email: rivo.randriatsiferana, richard.lorion, frederic.alicalapa@univ-reunion.fr
†University of La Reunion, LIM Lab, Saint Denis, Reunion, France, Email: fanilo.harivelo@univ-reunion.fr

***Abstract*—Energy conservation can be achieved by the use of clustering routing protocols in Wireless Sensor Networks (WSN). They have a great impact on the system lifetime. In these WSNs, the concepts of communication between Cluster Head (CH) nodes and Base Station (BS), data aggregation and round can be approached in a different way. This paper studies a cluster-based multi-hop routing protocol. The proposed approach introduces a progressive data aggregation, and a sequential data request during a round. Simulation results show that the new Forwarding and Routing Stateless Multi-Hop (FRSM) protocol extends network lifetime (improved by 50%) and lowers energy consumption. The comparison with an existing cluster-based protocol shows better performances.**

***Index Terms*—clustering, geographic routing, greedy forwarding, perimeter forwarding, network lifetime.**

I. Introduction

In general, a WSN consists of a large number of small and cheap sensor nodes that have limited energy, processing power and memory storage capacity. They usually monitor areas, collect data and report them to the Base Station. Due to the achievement in low-power digital circuits and wireless communication facilities, many applications of the WSN are developed and are already been used in building monitoring, military object and object tracking. They can also be used in hostile area where it is difficult to replace embedded batteries [1]. Hence, energy is the fundamental resource constraints. Its conservation represents one of the two issues addressed by this work to prolong the network lifetime. On the other hand, node's role and functions relative to network communication or the network topology cause some nodes to die quicker than the others. It is then essential to balance load among nodes. That constitutes the second purpose of this work.

The overall energy consumption can be reduced by allowing only a portion of the nodes, which are called Cluster Heads (CHs), to communicate with the base station (BS). Thus the data sent by each node is then first collected by Cluster Heads and compressed. After that, the aggregated data is transmitted to the BS. Low Energy Adaptive Clustering Hierarchy (LEACH) [2], is one of the first clustering protocols that was proposed for reducing power consumption. LEACH forms clusters by using a distributed algorithm, each node has a certain probability of becoming a CH per round, and the task of being a Cluster Head is rotated between nodes. A non-CH node determines its cluster by choosing the CH that can be reached with the least communication energy consumption. At the data transmission step, each Cluster Head sends an aggregated packet to the base station by a single hop communication. A well-known evolution of LEACH is Hybrid Energy-Efficient Distributed (HEED) [3]. In HEED, the initial probability for each node to become a tentative Cluster Head depends on its remaining energy, and the final CHs are selected according to the intra-cluster communication cost. HEED also consider one-hop communication between CH nodes and base station. Although clustering can reduce energy consumption, it has some problems. The main problem is that energy consumption is concentrated on the Cluster Heads, which have to transmit over long distance.

In the following, the BS is within the transmission range of every CH. Thus, each CH node can forward the data to the base station directly. However, it consumes much more energy in this way, and that does not necessarily balance the energy consumption among the network. So, a cluster routing method with equalized energy expenditure must be found. As mentioned in [4], short hops are generally more energy-efficient than one-hop with a few long hops. So, we propose a multi-hop routing between CHs in order to minimize energy consumption during transmission. Thus, GPSR (Greedy Perimeter Stateless Routing) protocol [5] [6] and his energy aware evolution GEAR (Geographical and Energy Aware Routing) [7] are two approaches aiming to improve the extensibility of the network in the presence of a large number of nodes. Their main advantages lie in the fact that the propagation of information on topology is necessary only for one hop. GPSR and GEAR make greedy forwarding decisions and perimeter forwarding using information about a router's immediate neighbors in the network topology. Consequently by using the same path to BS, the GPSR protocol leads to a premature failure of the nodes constituting the preferred way. Other approaches were proposed to improve the performance of clustering. EEUC [8] tackles the hot spot issue ; the Cluster Heads closer to the base station are burdened with heavy relay traffic and tend to die early. EEUC partitions the nodes into clusters of unequal size, and clusters closer to the base station have smaller sizes than those farther away from the base station. EECS [9] extends LEACH by realizing a localized election of Cluster Heads and a near uniform distribution of them. In cluster formation phase, a non-Cluster Head node chooses its Cluster Head by considering not only saving its own energy but also balancing the load of Cluster Heads. A

 ISBN: 978-1-61208-183-0

new weighted function is then introduced.

In this paper, we propose and evaluate a Forwarding and Routing Stateless Multi-Hop (FRSM) Protocol for WSNs mitigating the considered problems. FRSM combines a clustering scheme with a multi-hop routing while addressing the mentioned premature failure of the nodes by a rotation of roles. It consists of two phases: one relates to cluster management in sensor area, and the other handles the data transmission between clusters and the BS. In order to improve the efficiency, we consider an aggregation of data of each cluster during the transmission from a source CH to the BS. Hence, this work introduces a progressive data aggregation and a sequential data request during a round. In the present primary study, we choose not take into account energy remaining in sensors. Therefore we only evaluate the coupling method between clustering, multi-hop routing and aggregation method. This early evaluation aims to quantify the contribution of this coupling. For that reason, comparisons will be made with a single-hop protocol, namely, LEACH protocol. Simulation model is choosen to make those comparisons feasible. Future work will address comparisons with other multi-hop protocols.

The paper is organized as follows: Section II describes our FRSM protocol and its algorithm. In Section III, simulation context and results will be presented and discussed. Finally, in Section IV, conclusions will be given.

II. Forwarding and Routing Stateless Multi-hop Protocol

Forwarding and Routing Stateless Multi-Hop Protocol (FRSM) is proposed in order to increase the lifetime of the network, and to ensure the balance of energy consumption. The idea is to adopt GPSR protocol with Cluster Head elections by LEACH protocol. Data gathering is performed with a new algorithm: furthest CH data request, inter-CH data aggregation.

A. Assumptions and modeling of the system

We assume a wireless sensor network model with the following hypothesis: the system lifetime is the main objective and latency is not a major criteria. We consider a multi-hop homogeneous WSN where all nodes are alike. Each node can reach the BS if needed by controlling the transceiver power. Nodes location is recorded in the BS memory (using additional GPS function). The nodes have uniform initial energy allocation and the nodes are stationary. This assumption about node mobility is typical for sensor networks. The sensing field dimension is $100m * 100m$ and we consider that the BS is located at $(X_{BS} = 50, Y_{BS} = 0)$ in a two-dimensional XY plane.

The network is organized into clusters under the control of Cluster Heads. The BS collects the overall data using a sequential data request towards CHs. Each node senses the environment at a fixed rate and sends to the Cluster Heads. During the CHs selection process, each node n computes a random number between 0 and 1. If this number is lower than a threshold $T(n)$, the nodes becomes a Cluster Head. $T(n)$ is given by the following equation for the current round number (r) :

$$T(n) = \begin{cases} \frac{p}{1-p*(r \bmod \frac{1}{p})} & \text{if } n \in G \\ 0 & \text{otherwise} \end{cases}$$

where p is the average number of Cluster Heads during a round (10% will be considered in the simulation section). G is the group of nodes that have not been Cluster Head in the last $1/p$ rounds, otherwise $T(n)$ equals zero.

B. FRSM algorithm

The solution relies on the following algorithm :

1) $T(n)$ is calculated at each node to elect CH_i.
2) Cluster creation: all nodes organize themselves into clusters, under the control of the closest CH using the k-means algorithm [10].
3) Flooding phase to inform the BS and each CHs of other CH's position.
4) Physical data and CH identity aggregation is done in each cluster.
5) Furthest CH (defined as CH_{F1}), which has not been yet interrogated, is activated by the BS.
6) CH_{F1} sends its packets to a forwarding CH (which is closest to the BS, defined as CH_2) using GPSR routing algorithm.
7) Data aggregation and identity aggregation are computed at this second CH_2 (Data from CH_{F1} and from CH_2, and CH_2 and CH_{F1} identities).
8) From CH_2, data are sent to a third CH using GPSR routing algorithm (CH_3), which will again perform data and identities aggregation.
9) After x hops, the BS is reached. The data are then stored. Furthest CH (CH_{F2}), which has not been interrogated in the present round, is contacted by the BS.
10) Data are routed through inter-CH, which has never sent data during the current round. If needed RF transceiver is tuned to reach the BS.
11) This data collect process is repeated until each cluster have sent their data.
12) A new round begins with $T(n)$ calculation at each node (return to step 1).

Route or next hop selection is done on geographical routing basis as introduced in GPSR [5]: greedy forwarding, is used wherever possible, and perimeter forwarding, is used in the regions greedy forwarding cannot be. In greeding forwarding, a forwarding node make a locally optimal, greedy choice in choosing a packet's next hop. Upon receiving a greedy-mode packet for forwarding, a node searches its neighbor table for the neighbor geographically closest to the packet's destination. If this neighbor is closer to the destination, the node forwards the packet to that neighbor. When no neighbor is closer, the node marks the packet into perimeter mode. GPSR forwards perimeter-mode packets using a simple planar graph traversal. When a packet enters perimeter mode at node x bound for node D, GPSR forwards it on progressively closer faces of the planar graph, each of which is crossed by the line xD.

 ISBN: 978-1-61208-183-0

III. Simulation and Discussion

To evaluate the performance, our algorithm has been simulated in Matlab and the results were compared with LEACH. It is noted that the comparison with GPSR is illogical because GPSR does not consider clusters and CHs. Before presenting the simulation results, the radio model and some important parameters should be described.

A. Radio model

We use a simple model for the radio hardware energy dissipation introduced by [11] [12], where the transmitter and receiver dissipate energy to run the radio electronics (see Fig. 1). It was initially used to calculate the power consumption of LEACH protocol.

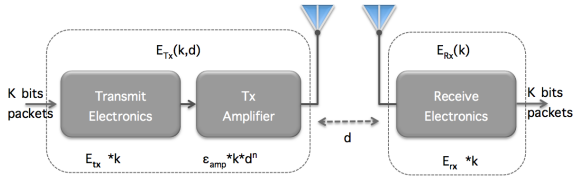


Fig. 1. Radio energy dissipation model

Power control can be used to invert this loss by appropriately setting the power amplifier. If the distance is less than a threshold, the free space (fs) model is used; otherwise, the multipath (mp) model is used. Energy consumed during transmission (E_{Tx}) is calculated with :

$$E_{Tx}(k,d) = E_{Tx-elec}(k) + E_{Tx-elec}(k,d) \quad (1)$$

while developing:

$$E_{Tx}(k,d) = \begin{cases} E_{tx} \cdot k + \epsilon_{fs} \cdot k \cdot d^2, & if \quad d \leq d_o \\ E_{tx} \cdot k + \epsilon_{mp} \cdot k \cdot d^4, & if \quad d > d_o \end{cases} \quad (2)$$

where E_{tx} is the transmit energy consumption by the radio transceiver for one transmission, k is the size of the message in bits, E_{fs} and E_{mp} represent the energy consumed by the radio amplifier, depending on the transmission distance d, and d_0 equals $\sqrt{\frac{E_{fs}}{E_{mp}}}$.

The energy consumed during message reception is calculated by:

$$E_{Rx}(k) = E_{Rx-elec}(k) = E_{rx} \cdot k \quad (3)$$

where E_{rx} is the receiving energy consumption by the radio transceiver and k is the size of the message in bits. Energy consumption during reception is only calculated when a message is received, i.e., the radio transceiver only expends power during message reception. Power consumption for the calculation operations is much weaker than the communication energy. In addition, data aggregation also costs some energy and the energy consumption for aggregating a certain data signal is represented as E_{da}. E_{da} is calculated by applying the following [13]:

$$E_{da} = 5\,\text{nJ/bit} \quad (4)$$

TABLE I
Parameters used in our simulation

Radio model	Description	Value
E_o	Initial energy	100 mJ
E_{tx}	Transmitting energy	0.208 $mJ/packet$
E_{rx}	Receiving energy	0.121 $mJ/packet$
E_{da}	Consume Energy data aggregation	5 nJ/bit
ϵ_{fs}	Transmit amplifier free-space	10 $pJ/bit/m^2$
ϵ_{mp}	Transmit amplifier for two-way	0.0013 $pJ/bit/m^4$
Other parameters	**Description**	**Value**
n	Number of nodes	100
p	CHs selection probability	10%
k	Packet size	320 bits
BS	Base station located	$(50, 0)$

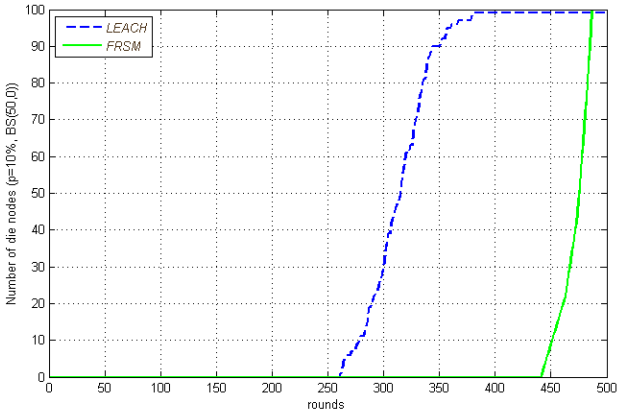


Fig. 2. Comparison of the number of dead nodes over time.

The energy consumption of each aggregation node for aggregating the data from itself and m neighbor nodes is represented as:

$$E_{DA} = (m+1) \cdot k \cdot E_{da} \quad (5)$$

The parameters used for the simulations of the implemented protocols are shown in Table I. Those parameters are taken from Chipcon RFIC datasheet [14].

B. Simulation results- Base station located at (50,0)

Fig. 2 gives the comparison between FRSM protocol and LEACH protocol in term of the lifetime per rounds. It is clearly shown that our FRSM algorithm outperforms LEACH in the number of alive nodes. If the lifetime metric is defined as number of rounds for which the first node died, FRSM can reach 442 rounds, whereas LEACH only reaches 261 rounds. For half of the nodes being alive, FRSM can reach 475 rounds, but LEACH only reaches 315 rounds. The lifetime metrics are improved by 50% for FRSM. Besides, the noticeable nearly linearity of the FRSM curve starting at 442th round proves the efficiency of load balancing.

Considering the two protocols, the remaining energy of nodes over the number of rounds has been presented (see Fig.

 ISBN: 978-1-61208-183-0

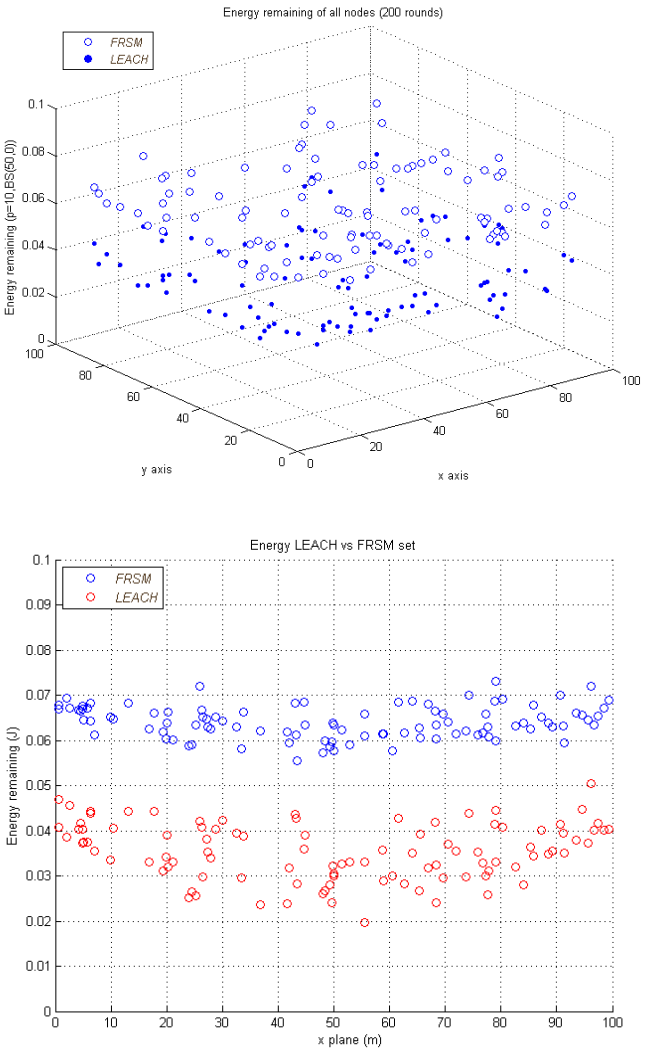


Fig. 3. Energy remaining after 200 rounds in the XY plane and X axis.

3) in the XY plane and in X axis. We show that the FRSM has much more desirable energy expenditure than the LEACH protocol. As the number of rounds is equal to 200, the average remaining energy of all the nodes in the networks for FRSM and LEACH are respectively 0.064 J and 0.035 J. So, the FRSM exhibits a 80% reduction in energy consumption over LEACH.

IV. Conclusion and Future Works

The Forwarding and Routing Stateless Multi-hop protocol for wireless sensor networks has been introduced in this paper. We adopt the cluster-based algorithm to make sure the well-balanced energy in the network. Thus, shorten communication distance among Cluster Heads and progressive aggregation data during transmission, were considered to reduce the global communication energy consumption. Based on specific network assumptions, simulation results show that this method obtains satisfactory performance on prolonging the network lifetime (increased by 50%). To go further, other points are under study: simulation of a fairly comparison with HEED, measurements on a more accurate energy consumption model and the use of energy aware information's for the routing process.

Acknowledgment

We acknowledge the financial support of structural funds of the structural funds of the European Community, and the Regional Council of Reunion Island (Region Reunion) for providing research grants (N/REF.: 201029758/DIREM/AS/lrs).

References

[1] M. G. Rashed, M. H. Kabir, and S. E. Ullah, "An energy efficient protocol for cluster based heterogeneous wireless sensor network," *International Journal of Distributed and Parallel Systems*, vol. 2, no. 2, pp. 54–60, 2011.

[2] W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," *IEEE Transactions on wireless communication*, vol. 1, no. 4, pp. 660–670, October 2002.

[3] O. Younis and S. Fahmy, "Heed: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks," *IEEE Transactions on Mobile Computing*, vol. 3, no. 4, pp. 366–379, 2004.

[4] Y. Bolian, S. Hongchi, and S. Yi, "A two-level topology control strategy for energy efficiency in wireless sensor networks," in *11th International Conference on Parallel and Distributed Systems.*, vol. 2, Washington, DC, USA. IEEE Computer, November 2005, pp. 358–362.

[5] B. N. Karp and H. T. Kung, "Gpsr: Greedy perimeter stateless routing for wireless networks," in *the Sixth Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom 2000), Boston, MA*, August 2000, pp. 243–254.

[6] B. N. Karp, "Geographic routing for wireless networks," Ph.D. Dissertation, Harvard University, Cambridge, MA, October 2000.

[7] Y. Yu, R. Govindan, and D. Estrin, "Geographical and energy aware routing: a recursive data dissemination protocol for wireless sensor networks," UCLA Computer Science Department, Tech. Rep., 2001.

[8] C. Li, M. Ye, G. Chen, and J. Wu, "An Energy-Efficient Unequal Clustering Mechanism for Wireless Sensor Networks," in *IEEE Mobile Adhoc and Sensor Systems Conference*, 2005, p. 8.

[9] M. Ye, C. Li, G. Chen, and J. Wu, "EECS: An Energy Ecient Clustering Scheme in Wireless Sensor Networ," in *IEEE Performance, Computing, and Communications Conference*, 2005, pp. 535–540.

[10] S. P. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, pp. 129–137, 1982.

[11] T. Voight and H. Ritter, "Solar-aware clustering in wireless sensor networks," in *Ninth IEEE Symposium on Computers and Communications*, vol. 1, 2004, pp. 238 – 243.

[12] L. Dehni, Y. Bennani, and F. Krief, "Lea2c: Low energy adaptive connectionist clustering for wireless sensor networks," in *MATA'05*, 2005, pp. 405–415.

[13] G. J. Pottie and W. J. Kaiser, "Wireless integrated network sensors," *Commun. ACM*, vol. 43, pp. 51–58, May 2000.

[14] (2006) Cc2420 data sheet. [Online]. Available: http://www.ti.com/product/cc2420⟨retrieved:12,2011⟩

 ISBN: 978-1-61208-183-0

Resource Management for Advanced, Heterogeneous Sensor-Actor-Networks

Matthias Vodel, Mirko Lippmann, and Wolfram Hardt
Dept. of Computer Science
Chemnitz University of Technology
Chemnitz, GERMANY
Email: { vodel | limir | hardt }@cs.tu-chemnitz.de

***Abstract*—Sensor-Actor-Networks (SANET) consist of several heterogeneous subsystems, which provide specific capabilities for measuring or manipulating its environment. During the runtime, the communication infrastructure as well as communication tasks and the available communication resources are changing dynamically. Furthermore, advanced application scenarios in this domain have strict requirements regarding to the minimal system uptime, QoS features or backup strategies. In this context, one of the most challenging objectives for researchers all over the world is the efficient integration and handling of heterogeneous, distributed SANET components to ensure a reliable and stable system operation. In respect of this issue, we present a novel cross-layer resource management approach for advanced SANET. We are now able to reallocate communication resources for each subsystem on-demand during the runtime. For this purpose, we developed a real-time radio standard integration concept and respective routing strategies with adaptive multi-standard, multi-interface metrics. A respective real-world demonstrator was designed and implemented. Based on this platform, we start a proof of concept evaluation and analyse the operational behaviour of a given SANET testbed configuration. The proposed measurements clarify the necessity as well as the feasibility of an intelligent, integrated resource management unit for advanced SANET architectures.**

***Keywords*-Energy Efficiency, Resource Management, Channel Reallocation, Dynamic Optimisation, Embedded Systems, Wireless Sensor Networks (WSN), Sensor-Actor-Networks (SANET)**

I. INTRODUCTION

Sensor-Actor-Networks (SANET) as well as Wireless Sensor Networks (WSN) represent distributed embedded systems which are able to sense its environment for specific events or behaviour. Based on wireless communication interfaces, the subsystems (*nodes*) are able to exchange information. Actuator nodes are additional entities of SANET, which allows the system to manipulate the environment based on a predefined set of rules. *Figure 1* illustrates a given SANET architecture, its different abstraction layers and the respective operational tasks on each layer.

To operate autonomously, each subsystem has limited energy resources. Here, the efficient management of these resources is essential. To maximise the system runtime, developers have to find a trade-off between working performance and power consumption of the hardware system architecture. In this respect, the trade-off starts with the used sensor components (accuracy, sample rate, size), resource limitations regarding to the μController (memory, number of I/O pins, speed) and ends with the wireless communication interfaces (data rate, transmission range, latency, interference liability). Besides these hardware aspects, the concrete application scenario implies further operational restrictions. In this context, scenario-specific communication protocols in the several abstraction layers are critical to optimise the system efficiency [1], [2].

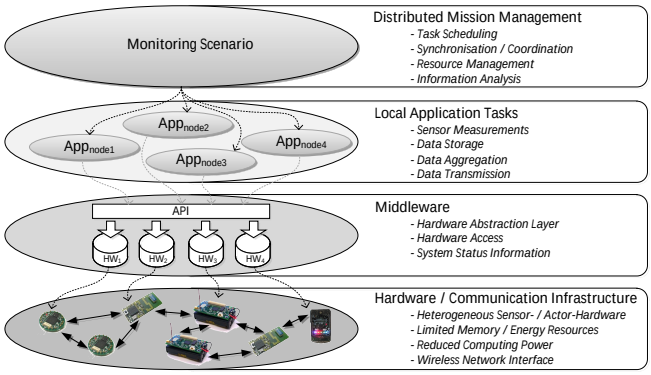


Figure 1. A given SANET architecture including four different abstraction layers. The first one includes the different hardware components. A middleware provides uniform access to the available subsystems for all applications. Single operational tasks are executed on capable nodes. A global mission management is responsible for calculating and coordinating all the tasks in order to fulfil the given mission objectives.

This paper proposes a resource management approach for coordinating the overall communication behaviour within SANET applications dynamically. It integrates different features for distributed, embedded system to ensure a reliable and robust communication infrastructure. It is structured as follows: After this introduction, section *II* provides an overview about related technologies in the domain of energy-efficient and robust SANET. This includes cooperative routing strategies, cross-layer approaches and further communication techniques. The proposed resource management concept and its basic components are introduced in section *III*. This section also provides respective examples and application scenarios. Section *IV* describes the chosen testbed configuration based of a capable hardware prototype platform. The results are discussed in section *V*. Finally, the

 ISBN: 978-1-61208-183-0

paper concludes with a summary and an outlook for future work in this research area.

II. Related Work

A primary objective of basic network optimisation techniques are stable and robust end-to-end communication *channels* through a given heterogeneous multi-hop topology. Channels represent the essential logical resource on top of the physical network interfaces. In order to balance the network load through these channels, a lot of research was done in the domain of multi-path routing [3], [4]. The idea is to split a data stream into multiple, potentially prioritised sub-streams and transmit these parts over different route path to the sink. Here, several problems have to be solved. On the one side, we have to find stable communication paths in dynamic, heterogeneous network infrastructure for a lossless date transmission. On the other hand, requirements for worst-case latencies and minimum transmission data rates have to be fulfilled. Most of the related multi-path concepts operate on a homogeneous network topology and uses unidimensional routing metrics. Regarding to our proposed work, these metrics have to extended for the multi-interface, multi-standard domain (e.g., *EBCR - Energy Balanced Cooperative Routing* [5], [6]).

Other routing approaches use multi-dimensional metrics for optimising the route paths. [7] and [8] describe concepts for gathering network information as well as additional system information from different abstraction layers. Such *cross-layer (X-layer)* approaches, like in [9], have a much better knowledge about the current network situation than traditional, uni-dimensional routing algorithms on the network layer.

In a further step, advanced research projects are looking for approaches to balance the network communication over multiple interfaces with different communication standards [1]. The main idea is to use the advantages of multiple radio standards. At the same time, we bypass the disadvantages of using one single technology, which result from their specific application fields. Accordingly, the developed radio standard integration concept provides a heterogeneous network infrastructure and an efficient real-time protocol conversion approach [10], [11].

Further technology integration approaches, like *Cognitive Radios (CR)* as well as *Software Defined Radios (SDR)* represent other concepts for optimising the communication in mobile application scenarios. CR is operating on the hardware-near layer to minimise radio interferences and to adapt the communication channel dynamically [12], [13]. SDR stands for a modular framework, which implements the whole protocol stack of a given communication standard in software. Accordingly, SDR provides an outstanding flexibility and allows a real-time conversion between different radio standard [14]. Unfortunately, due to the required hardware resources, SDR is not applicable for the embedded mobile domain like WSN or SANET [15]. Another promising research project represents *Ambient Networks* [16], [17], which are focusing on a platform-spanning communication infrastructure based on *Ambient Services* - an additional abstraction layer on top of the user application.

III. Dynamic Resource Manangement

Regarding to our proposed concept and with focus on dynamic scenarios, one challenging problem deals with the varying communication resources and changing environmental conditions during the runtime. Dependent on the application scenario, different capacities for the data transmission are required. An advanced resource management for multiple physical interfaces has to consider several additional parameters, which includes the local system status and distributed network information.

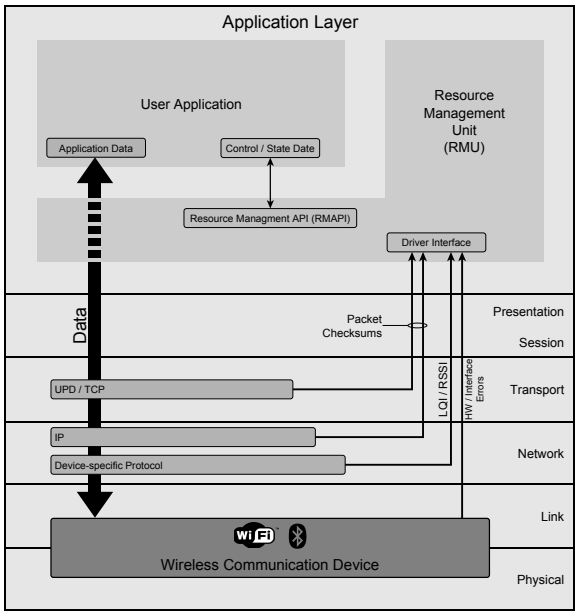


Figure 2. Resource Management Unit (RMU) and its integration into the protocol stack.

Figure 2 represents typical communication architecture and the integration of our proposed *Resource Management Unit (RMU)*. In contrast to related X-layer routing and communication approaches, the RMU uses standardised information, which are provided by the hardware components, the respective drivers or the embedded operating system. Specific modifications or adaptations in the hardware architecture or the protocol stack are not required.

In order to establish a logical communication channel from a source application to another remote application over a multi-hop network infrastructure, communication resources have to be allocated. Therefore, the usual way is to request a new communication socket from the operating system based on the respective transport protocol. In our

 ISBN: 978-1-61208-183-0

proposed concept, instead of opening a communication socket directly, each application handles its requests over the RMU. The RMU operates as a software component in the application layer and provides a dedicated *Resource Management API (RMAPI)*. Based on a set of services, the RMU is able to monitor the network communication and to allocate communication channels for the user applications.

Furthermore, the RMU allows a proactive channel analysis for high prioritised data streams. For this purpose, two or more nodes exchange special *RMU tracker packets*. Even if it takes more energy and computing time, this technique is essential for data critical application scenarios, in which a simultaneous and continuous channel monitoring is not capable. In such critical cases, the RMU is responsible for backup channels and the respective reallocation.

The RMU is able to manage multiple interfaces and radio standards simultaneously. In order to use these advantages within the system architecture, a real-time on-demand switching technique is required. Therefore, an efficient radio module integration concept has to operate directly on top of the hardware devices as a kind of embedded middle-ware. For this purpose, the *EAN (Embedded Ambient Networking)* concept was developed and allows a dynamic conversion between different radio standards [1], [10], [11]. Currently, several international cooperation projects research for an embedded high-performance platform based on this EAN approach. In combination with the RMU, an adaptive and flexible communication architecture will be created.

A. Channel Modeling & Reallocation Schemes

As already mentioned the resource management metric includes local system information and distributed network information. In this context, rules and calculations are very similar to related cross-layer routing metrics. In contrast to multi-dimensional routing metrics on the network layer, the channel management operates parallel to user application on the ISO/OSI layer 7. Accordingly, the RMU coordinates all communication requests between user applications and network interface. For providing a optimised, scenario-specific reallocation scheme, a multi-dimensional set of parameters is required for estimating the current situation. These parameters are categorised as follows:

1) Latency:

- hop count (flat network hierarchy)
- number of protocol conversion (use less different interfaces as possible)

2) Data throughput:

- minimum or average data rate
- stream splitting / multi-path capabilities

3) Energy consumption:

- interface power consumption (standby, rx/tx)
- trade-off transmission range and route path length

4) Security:

- channel stability / robustness (based on channel monitoring techniques)
- channel prioritisation

5) Capacity utilization:

- interface load
- protocol overhead

B. Example Scenario I - Balancing & Optimisation

The decision making processes of the RMU represents a challenging problem. *Figure 3* illustrates a typical scenario. In order to optimise the network communication, $Node_{new}$ can be integrated in different ways. It is possible to split the data streams into two subchannels between $Node_{new}$ and $Node_2$ over the radio standards RS_1 and RS_2. In this case, both interfaces are used to balance the net load or to realise a multipath data prioritisation. Otherwise, one interface has to be preferred. Hereby, the remaining communication capacities in $Node_2$ (20% left) and $Node_3$ (50% left) have to be considered.

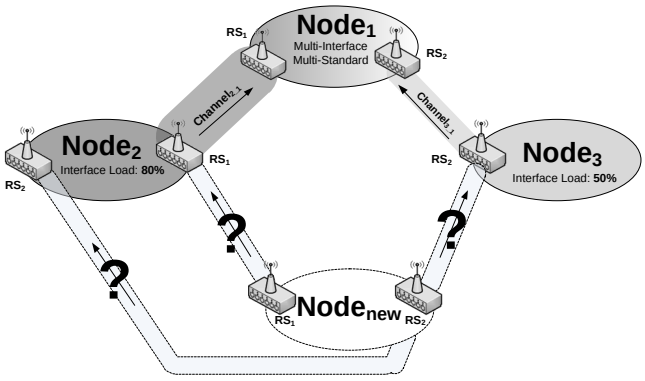


Figure 3. Channel allocation and reallocation scenario I. $Node_{new}$ has to be integrated into the existing node topology. Usually, each node has wireless network interfaces with different radio standards (RS_1 and RS_2). $Node_1$ and $Node_2$ integrate two interfaces. In $Node_3$, only one interface is available. The established channels between the nodes bind communication resources. The resource management has to decide about the channel balancing in a cooperative process.

C. Example Scenario II - Fault Response

Concerning the decision process, the RMU calculates the remaining interface capacities with theoretical parameters of the respective communication standard specifications. In real-world multi-hop application scenarios, environmental disturbances and unexpected effects also have an huge influence on the communication behaviour. Especially in dynamic scenarios, obstacles represent critical limitations for a stable, continuous data transmission. *Figure 4* visualises such a situation.

 ISBN: 978-1-61208-183-0

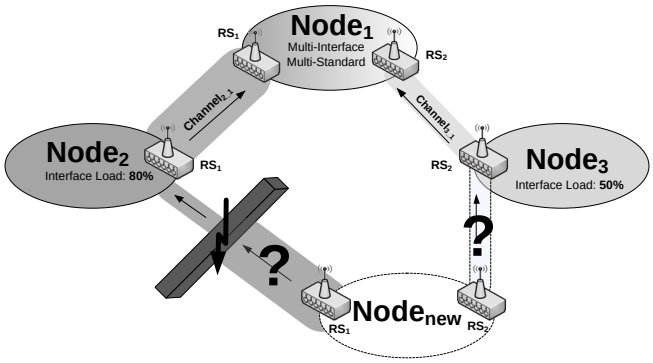


Figure 4. Channel allocation and reallocation scenario II. Typically, radio standard 1 (RS_1) provides more transmission capacities than RS_2 (backup channel). Due to obstacles, the usable channel capacity is not equal to the maximal capacities of the radio specification. The generated data stream in $Node_{new}$ requires situation-specific channel resources. In order to decide about the channel allocation, the RMU has to estimate the remaining capacities.

In consequence, the RMU monitors active channels for detecting bottlenecks in the communication. Accordingly, based on the given metric, critical data stream can be reallocated. Furthermore, such a proactive channel analysis allows a re-prioritisation of all active channels in order to optimise the network communication.

IV. Testbed Configuration

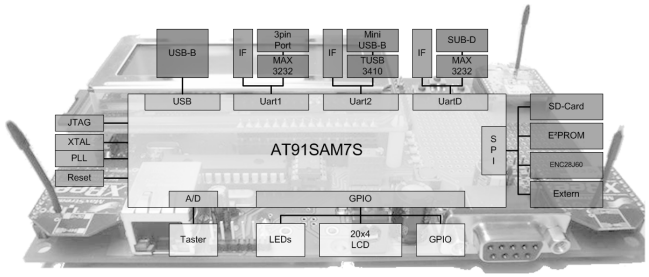


Figure 5. Prototype evaluation platform. The system architecture integrates different COTS wireless network interfaces into one integrated network node.

Based on the proposed concepts for a RMU, the radio standard integration and respective simulation results in [18], [11], we decided to design a prototype platform which implements the features in a multi-interface, multi-standard communication environment [10]. The platform interconnects up to four different network adapters with different communication standards. The wireless interfaces are connected via modular communication slots, which are compatible to COTS (*commercial off the shelf*) hardware components. *Figure 5* illustrates the system structure with the central ARM7 microcontroller. The platform is designed as an evaluation board on a proof of concept level. With respect to this application domain, the ARM7 provides a lot of computing performance for many possible test scenarios. Further developments will shrink the design to an ultra-low-power sensor board with a MSP430 microcontroller [19], [20]. Alternatively, an Artix-7 FPGA implementation is also possible.

This prototype platform allows us to test and analyse essential features of the proposed channel management concept, including the radio standard integration, the Ad Hoc communication standard switching as well as the dynamic resource reallocation. In this paper we present essential results regarding to the real-time protocol conversion and the respective channel reallocation capabilities.

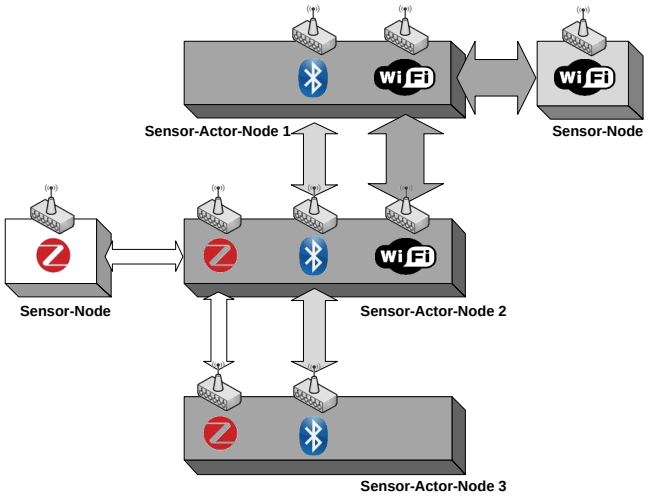


Figure 6. Demonstrator testbed environment. Each node provides several wireless network interfaces and capabilities for prioritised communication channels. The key challenge represents a cooperative management of different communication technologies.

Figure 6 shows the realised multi-hop network topology as a heterogeneous sensor-actor scenario. The given network infrastructure integrates three communication standards, based on IEEE 802.11, 802.15.1 and 802.15.4. During the test scenarios, each node generate sensor and control data with different priorities and different data volumes (acceleration, temperature, noise level, visual/audio data). Each communication standard is represented by a dedicated IP subnet. Accordingly, each node has the knowledge about technology-specific channel properties. Furthermore, additional information about the actual channel load and channel quality are available within the RMU. All the data streams have to be transmitted simultaneously. In consequence, the RMU allocates and reallocate various end-to-end channels. In case of a switching communication standard, the data payload will be converted dynamically in real-time. The conversion processes includes a header analysis, the packet reassignment and, if required, a re-segmentation of the payload.

 ISBN: 978-1-61208-183-0

V. RESULTS - PROTOCOL CONVERSION

The described testbed configuration represents an advanced multi-interface, multi-standard SANET. Based on this topology, we evaluate the channel reallocation capabilities, represented by the overall transmission times as well as the node internal protocol conversion times.

In a first scenario, we measure the latency for the protocol conversion during a bidirectional communication. As already mentioned, the conversion process is done on a hardware-near middle-ware between the ISO/OSI layer 2 and 3 (*EAN*). During the test cycles in *figure 7*, the data rate was increased step-by-step.

The illustrated diagrams visualise average values of 1000 continuous transmission cycles. As we can see, system architecture as well as the protocol conversion operate stable and efficient with delay times under 3ms. Anyway, each conversion process increases the communication overhead for a given channel. The overhead ratio is dependent on the data payload and the packet size. The key question is, how critical such a conversion process in relation to the overall multi-hop transmission is. If we take a closer look on our first scenario, the data forwarding latency increases minimally. *Figure 8* illustrates the results for a ZigBee to Bluetooth conversion with a normal packet size.

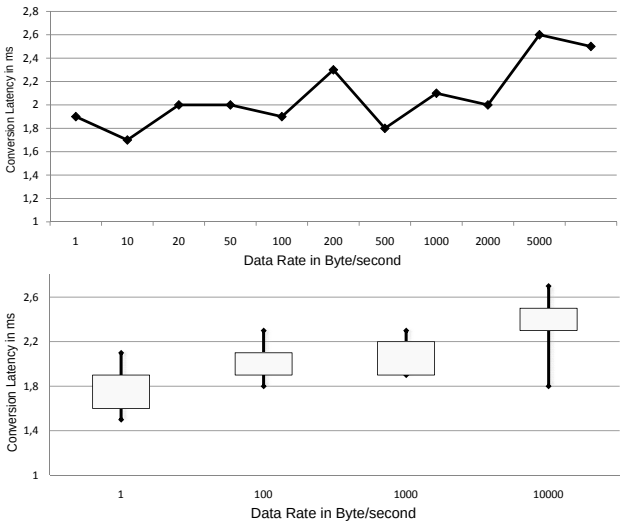


Figure 7. Top: Continuous protocol conversion measurements from Wifi (IEEE 802.11g) to ZigBee (IEEE 802.15.4). The transmission data rate starts with 1 Byte/second and ends with 10000 Bytes/second.
Bottom: Long term protocol conversion measurements from Bluetooth (IEEE 802.15.1) to Wifi (IEEE 802.11g) with different transmission data rates.

In contrast, the oscilloscope screenshot in *figure 9* represents a detailed waveform diagram of another conversion scenario from ZigBee to Wifi. This scenario uses very small data packets with minimal data payload. The environmental properties are similar to the first test scenario. The global addressing protocol is IP. As expected, the influence of the conversion process on the overall transmission delay is higher.

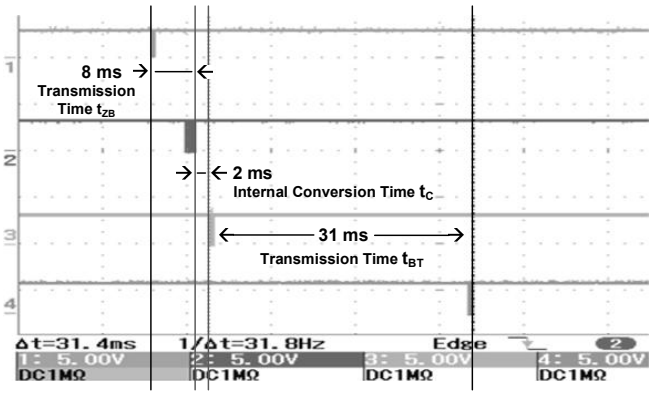


Figure 8. Bidirectional multi-hop communication measurement including a protocol conversion process from Bluetooth (t_{BT}) to ZigBee (t_{ZB}) and vice versa. The results were measured with an oscilloscope directly at the connectors without overhead from the operating system, especially by scheduling-based inaccuracies.

These results clarify the importance of an intelligent channel management, which is able to analyse the actual situation and considers both application-specific parameters and network behaviour.

Anyway, all test results provides a normal transmission behaviour without errors or disturbances. The dynamic channel reallocation between several multi-hop communication paths works stable. Hence, the proposed multi-interface channel management is feasible and very efficient for advanced application scenarios in the WSN and SANET domain.

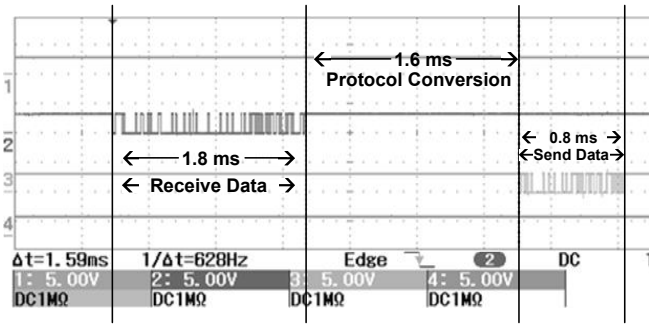


Figure 9. Detailed measurement of an Ad Hoc protocol conversion process from ZigBee (IEEE 802.15.4) to Wifi (IEEE 802.11g) standard on layer 3 (IP).

VI. CONCLUSION

In this paper, we propose a novel channel management concept for advanced WSN and SANET scenarios in heterogeneous communication environments. It focuses on dynamic channel switching techniques in order to realise an intelligent load balancing over the available wireless

 ISBN: 978-1-61208-183-0

capacities. We clarify the importance of such concepts for critical application scenarios to ensure guaranteed resources. In combination with an innovative radio standard integration concept, we are able to optimise the communication behaviour significantly.

The presented test scenarios were done on a research prototype platform. The results demonstrate the feasibility of the proposed concepts. The realised network topology integrates several COTS sensor entities as well as multi-interface, multi-standard sensor-actor-nodes. All the measured timings for the protocol conversion need less the 3ms. Accordingly, the protocol overhead within a multi-hop communication increases minimal. At the same time, we create a reliable network infrastructure and improve the connectivity significantly.

Further research work combines the proposed resource management approach with *wake-up-receiver* technologies (*WuRx* [19], [20]) to evaluate innovative communication concepts for WSN/SANET applications. Another point of research deals with the integration of advanced transport protocols for WSN and SANET scenarios [21], [22].

REFERENCES

[1] M. Vodel. *Radio Standard Spanning Communication in Mobile Ad Hoc Networks*. PhD thesis, Chemnitz University of Technology, Germany, 2010.

[2] M. Vodel, M. Lippmann, M. Caspar, and W. Hardt. Distributed High-Level Scheduling Concept for Synchronised, Wireless Sensor and Actuator Networks. *Journal of Communication and Computer*, pages 27–35, 2010.

[3] S. Mueller, R.P. Tsang, and D. Ghosal. *Multipath Routing in Mobile Ad Hoc Networks: Issues and Challenges*. Springer Verlag, 2004.

[4] D. Ganesan, R. Govindan, S. Shenker, and D. Estrin. Highly-Resilient, Energy-Efficient Multipath Routing in Wireless Sensor Networks. *SIGMOBILE Mobile Computing and Communications*, 5(4):11–25, 2001.

[5] M. Vodel, M. Caspar, and W. Hardt. EBCR - A Routing Approach for Radio Standard Spanning Mobile Ad Hoc Networks. In *Proceedings of the 4th Confernence on Computing and Inforamtion Technology*, pages 57–58, May 2008.

[6] M. Vodel, M. Caspar, and W. Hardt. Energy-Balanced Cooperative Routing Approach for Radio Standard Spanning Mobile Ad Hoc Networks. In *Proceedings of the 6th International Information and Telecommunication Technologies Symposium*, pages 42–47. IEEE, December 2007.

[7] G. Yang, M. Xiao, H. Chen, and Y. Yao. A Novel Cross-Layer Routing Scheme of Ad Hoc Networks with Multi-Rate Mechanism. In *Proceedings of the International Conference on Wireless Communications, Networking and Mobile Computing*, volume 2, pages 701–704. IEEE Computer Society, September 2005.

[8] L. Iannone, R. Khalili, K. Salamatian, and S. Fdida. Cross-Layer Routing in Wireless Mesh Networks. In *1st International Symposium on Wireless Communication Systems*, pages 319–323. IEEE Computer Society, September 2004.

[9] Y. Zhang, J. Luo, and H. Hu. *Wireless Mesh Networking: Architectures, Protocols and Standards*. 2006.

[10] M. Vodel, M. Caspar, and W. Hardt. A Capable, Lightweight Communication Concept by Combining Ambient Network Approaches with Cognitive Radio Aspects. In *Proceedings of the 17th IEEE International Conference on Telecommunications (ICT)*, 2010.

[11] M. Vodel, M. Caspar, and W. Hardt. Embedded Ambient Networking - A New, Lightweight Communication Concept. In *Proceedings of the 9th IEEE International Conference on Communications (ICC)*, 2010.

[12] W.Y. Lee and I.F. Akyildiz. Optimal Spectrum Sensing Framework for Cognitive Radio Networks. *Transaction on Wireless Communications*, 2008.

[13] K.R. Chowdhury and I.F. Akyildiz. Cognitive Wireless Mesh Networks for Dynamic Spectrum Access. *Journal of Selected Areas in Communications*, 2008.

[14] A.S. Harrington, C.-G. Hong, and A.L. Piazza. *Software Defined Radio - The Revolution of Wireless Communication*. International Engineering Consortium, 2004.

[15] F.K. Jondral. Software-Defined Radio-Basics and Evolution to Cognitive Radio. *EURASIP Journal on Wireless Communications and Networking*, 3:275–283, April 2005.

[16] P. Magnusson, F. Berggren, I. Karla, R. Litjens, F. Meago, H. Tang, and R. Veronesi. Multi-Radio Resource Management for Communication Networks beyond 3G. In *Proceedings of the 62nd Semiannual Vehicular Technology Conference*, pages 1653–1657. IEEE Computer Society, September 2005.

[17] B. Ahlgren, L. Eggert, B. Ohlman, and A. Schieder. Ambient Networks: Bridging Heterogeneous Network Domains. In *Proceedings of the 16th International Symposium on Personal Indoor and Mobile Radio Communications*, page no pp. given. IEEE Computer Society, September 2005.

[18] M. Vodel, M. Sauppe, M. Caspar, and W. Hardt. A Large Scalable, Distributed Simulation Framework for Ambient Networks. *Journal of Communications*, pages 11–19, 2009.

[19] nanett Germany. Nano System Integration Network of Excellence, http://nanett.org, 2011.

[20] M. Vodel, M. Caspar, and W. Hardt. Wake-Up-Receiver Concepts - Capabilities and Limitations. *Journal of Networks*, 2011.

[21] C.-Y. Wan, A.-T. Campbell, and L. Krishnamurthy. PSFQ: a reliable transport protocol for wireless sensor networks. In *Proceedings of the ACM International Workshop on Wireless Sensor Networks and Applications*, pages 1–11. ACM, 2002.

[22] K. Sundaresan, V. Anantharaman, H.-Y. Hsieh, and R. Sivakumar. ATP: A Reliable Transport Protocol for Ad Hoc Networks. *IEEE Transactions on Mobile Computing*, 4:588–603, 2005.

 ISBN: 978-1-61208-183-0

Improving Fairness in Wireless Mesh Networks

Jorge L S Peixoto
Serv. Federal Process. Dados (SERPRO)
Fortaleza - CE - Brazil
jorge.peixoto@serpro.gov.br

Marcial P Fernandez
Univer. Estadual Ceará (UECE)
Fortaleza - CE - Brazil
marcial@larces.uece.br

Luis F de Moraes
Univer. Federal Rio de Janeiro (UFRJ)
Rio de Janeiro - RJ - Brazil
moraes@ravel.ufrj.br

***Abstract*—Wireless Mesh Networks (WMNs) are networks that aim to establish hybrid wireless communications in areas with little or no telecommunication infrastructure available. The main motivation of these networks in Brazil and other developing countries, particularly in remote areas, is providing Internet access in places without commercial infrastructure and where the telecom operators have no economic interest due to its low demand. The idea is that a user in the network can forward the packets of distant users to the wired network connection point. Currently, wireless mesh networks use the IEEE 802.11 (WLAN) protocol due its availability and low-cost. However, IEEE 802.11 protocol favors the near users at expenses of a very low performance for distant users from the gateway node, connected to the wired network. The goal of this paper is to propose a mechanism using the IEEE 802.11e and QoS extension to provide a fairness network resource distribution for all participating WMN nodes, independent of their distance. The prototype was tested in simulation and the results demonstrated the proposal effectiveness in resource allocation in a wireless mesh network.**

***Keywords*-Wireless Mesh Networks; Fairness; IEEE 802.11e.**

I. INTRODUCTION

The use of wireless networks as an alternative to wired networks has fostered a large number of studies whose focus is improving the behavior of autonomous devices. Due to the lack of an infrastructure network, the communications' management is the responsibility of the nodes themselves.

Wireless Mesh Networks (WMNs) are cooperatives and self-configurable networks, that interconnect a set of fixed nodes that can route packets to each other through multi-hop [1]. They have the advantage to be a low-cost, easy to deploy and a highly fault tolerant network.

The main application of Wireless Mesh Networks is to provide access in areas with fair telecommunications infrastructure where Internet access is given only by Plain Old Telephone Service (POTS). A mesh infrastructure allows to bring Internet access to low-income places where telecom operators don't have interest to offer broadband services. The mesh networks thus become a viable alternative to promote digital inclusion in under developed areas.

Wireless Mesh Networks can be classified into three classes based on nodes' functionality: flat, hierarchical and hybrid[1]. The *flat* WMN are composed of routers (nodes) with gateway functions that provides additional functions to support the mesh network routing. The *hierarchical* WMN are created by users but with functions of routing packets in the mesh. The *hybrid* network are made of mesh routers and clients that provides mesh routing (or not) and wired network gateway. The hybrid network provides connections with different types of networks. This work will be based on this approach.

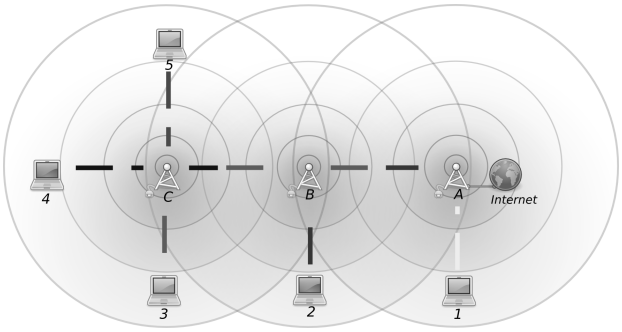


Figure 1. Unfairness forward in Mesh Networks.

In multi-hop wireless networks based on IEEE 802.11 [2], the performance is low, and in some cases, resources starvation, for client nodes far from the gateway node. This occurs due to some characteristics in such networks, such as: hidden and exposed terminal problem; IEEE 802.11 Binary Exponential Backoff [3]; the fact that DCF provides to stations equal access opportunity to the shared environment; and, data streams from distant stations (in number of hops) to the gateway, has more dispute on medium access that increases the packet loss and collision, reducing the throughput.

Figure 1 shows these problems. Suppose A, B and C are part of the wireless back-haul. Only A is connected to the wired network (Internet). Nodes A, B and C are configured with static routes and do not generate traffic, just forward the data from client nodes 1, 2, 3, 4 and 5. Now, suppose 1, 2, 3, 4 and 5 start (each one) a data flow to the Internet. The flow that starts at the node 4, before arriving at A, must pass through C and then B. The flow from node 2, before arriving at A, passes through B. Node 1 communicates directly with A, and then, to the Internet.

Through simulation experiments, we observed that the gradual increase in data rate from client nodes (1, 2, 3, 4

 ISBN: 978-1-61208-183-0

and 5), cause the increase of node 1 flow at the expense of the reduction of flow from other client nodes (2, 3, 4 and 5). Node A divide equally the opportunity for node 1 and B that consolidate the traffic from other nodes. In the Figure 1, while the client nodes forward only one stream each, node A forwards 5 streams (from 1, 2, 3, 4 and 5), B forwards 4 streams (from 2, 3, 4 and 5) and C forwards 3 streams (from a 3, 4 and 5).

This paper proposes a mechanism to share the *medium access time* among wireless back-haul routers and clients. The resource sharing should be proportional to the number of users connected to each wireless router. And also, it should maximize the network resource use in order to avoid resource waste.

The rest of the paper is structured as follows. In Section II, we present some related works and in section III we present the proposed mechanism. Section IV shows the validation methodology and Section V shows the results. Finally, Section VI shows the conclusions and some suggestions to future works.

II. Related Works

The medium access fairness in IEEE 802.11 wireless networks has been discussed in many papers. However, little work has been done about fairness on mesh networks. Some related works are presented below.

Bensaou [4] and Wang [5] proposed a new algorithm for quantitative back-off instead the DCF Binary Exponential Back-off. In their proposal, each station continuously estimates its own throughput and the throughput from other stations, which competes to the medium access. Then, each one calculates a fairness index used to adjust the contention window. The simulation shows that this algorithm achieves a better fairness than the IEEE 802.11 original algorithm.

Xu et al. [3] shows that, although the IEEE 802.11 MAC protocol support *Ad hoc* networks, it was not designed for it, i.e., the connectivity is basically multi-hop. It presents several problems, such as medium access unfairness and TCP instability, and proposes some solutions. Finally, they show that IEEE 802.11 MAC protocol does not work well in multi-hop networks.

Wang [6] proposed a mechanism to guarantee applications end-to-end delay in multi-hop wireless LANs. Due to node mobility and the distributed medium access, these networks suffer from severe delay and jitter variations. Thus, the perceived fairness on delay is essential to provide all nodes the same delays guarantees in a multi-hop WLAN network. They propose a new framework to guarantee delays using three modules: one is responsible for provisioning the delay information in the class of service, another for the adaptive selection among the available class of services and the last is responsible for monitoring the average delay of each network node and select the MAC priority.

Gambiroza et al. [7] proposed the IFA algorithm (Inter-transit access points Fairness Algorithm) to improve fairness in multi-hop networks. In IFA, each node calculates the amount of time it can use to transmit its data, improving overall fairness. The evaluation requires exchange of information control messages about the state of each link. The nodes send to their neighbors the amount of network resources they need to forward incoming traffic. After an exchange of control information, each node runs the algorithm to calculate its maximum rate allowed.

III. A prioritization mechanism to improve fairness in mesh networks

The mechanism proposed in this paper aims to share *medium access time* in a fair manner among nodes in a wireless back-haul. The mechanism functionality is to define a certain limited amount of shared resources for each wireless router from WMN back-haul. Then, it needs to know the number of client nodes connected in this router and its child nodes. The algorithm calculates the amount of necessary resource and it allocates for each connection.

It is important to know that the algorithm should run only in router nodes, but not in the client nodes. Then, the user does not need to update any software or hardware, they should use the normal WLAN protocol.

When we progressively increase the client nodes transmission rate, the rate of nodes close to the gateway router continues to grow at the expense of reduction on rate of distant nodes. By simulation experiments, we notice that when the network reaches saturation, the client nodes farther from the gateway router suffer starvation. The mechanism proposed aims to minimize this problem.

Another important factor is associated with WMN topology. Most of WMN has only one *path* between two nodes, as data preferentially flows from an external network to client nodes, then more flows are forwarded to the nodes near the gateway router. This gives a disproportion in traffic flows to the router nodes (Figure1), then, it should be given more medium access time to router nodes to minimize this disparity.

The prioritization mechanism allocates the resource *media access time* to the nodes of the WMN back-haul using a parameter that limits the maximum transmission time for each node. This parameter is defined as the Transmission Opportunity (TXOP) by IEEE 802.11e amendment. The resource allocation calculation is done as follows: each client node will receive a fixed amount of resource and the routers nodes receive an amount of resource proportional to the number of upward client nodes connected.

Figure 2 shows a WMN example and how the resource allocation works. Suppose, initially, that the TXOP value is configured to 1 for each client node (1, 2, 3, 4 and 5). Router C is configured with TXOP equal to 3, because there are three clients connected (5, 4 and 3). Router B receives a

 ISBN: 978-1-61208-183-0

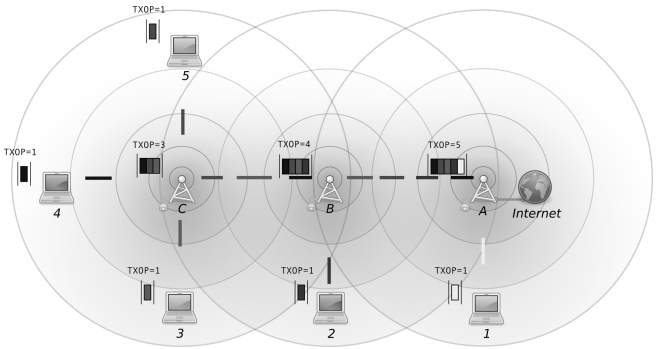


Figure 2. Resource allocation.

TXOP equal to 4, because there are four clients downward (5, 4, 3 and 2). Finally, A receives a TXOP equal to 5 because there are five clients downward (5, 4, 3, 2 and 1).

During the mechanism evaluation, it was noticed that the resources allocated to routers were underused. This was because all packets from the queue of the router were transmitted before the end of the time allocated, thus, on every transmission opportunity time a considerable amount of resource was wasted. To minimize this, it was added some intelligence in mechanism to dynamically adjust the router IEEE 802.11e AIFS parameter according the queue length. When it is near zero, the AIFS is increased and then: decreases the medium access probability for the node; increase the numbers of packets in the queue, and finally; maximizes the resource use because the allocated resource ends before the queue length reaches zero. When the queue length exceeds the threshold, the AIFS value is restored.

IV. Proposal Validation

This section describes the validation methodology and the implementation of proposed mechanism in Network Simulator version 2 (ns-2) platform. Only DCF functions are implemented in the IEEE 802.11 MAC layer of ns-2 core. We chose the TKN (Telecommunication Networks Group) [8] model, that supports an EDCA based on the latest version of IEEE 802.11e draft, which presents an improved binary exponential back-off algorithm that was adopted in the final version. The platform used was the Network Simulator (ns-2) version 2.28 patched with TKN EDCA from Vivek [9]. The operation system was GNU/Linux Ubuntu Hardy 8.04 and compiler was GNU C/C++ Compiler version 3.33 or 2.95.

A. Scenario

To validate the proposal, we performed simulations in three different scenarios. The first scenario, showed in Figure 2, is a **Typical Scenario** of a WMN where the mechanism was evaluated. To validate the proposed mechanism we evaluated in two other scenarios: the **Line Scenario** topology, showed in Figure 3, which represents the worst case for

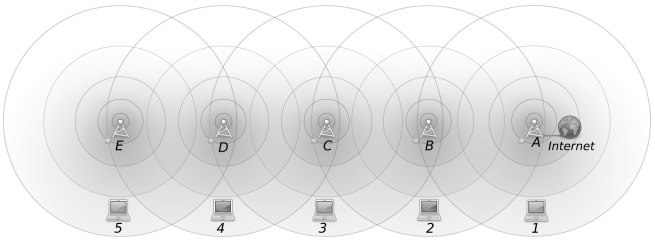


Figure 3. Line WMN scenario: nodes 1, 2, 3, 4 and 5 are clients, nodes B, C, D and E are routers and node A is the gateway.

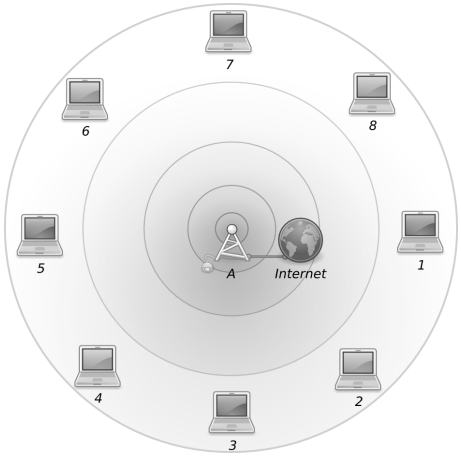


Figure 4. Star WMN scenario: nodes 1, 2, 3, 4, 5, 6, 7 and 8 are clients and node A is the gateway.

WMN (largest number of hops) and the **Star Scenario** topology, showed in Figure 4, which represents the best case for WMN, i.e., all stations are equidistant from the gateway node (one hop).

In Figure 2, nodes 1, 2, 3, 4 and 5 generate traffic Constant Bit Rate (CBR) using the UDP (User Datagram Protocol) transport protocol. These flows had 500 bytes packets to node A. TCP (Transmission Control Protocol) was not used because its congestion control mechanism would affect the measurement results [10]. Nodes A, B and C are fixed router nodes configured with static routes [9] that make the wireless back-haul. The node A is 100 meters from B and B is 100 meters from C. Nodes 5, 4 and 3 are within 100 meters of C, node 2 is within 100 meters of B and node 1 is within 100 meters of A.

The simulations were performed with and without the mechanism enabled. Each experiment consisted of 10 simulations of 240 seconds and the average with a 95% confidence interval was calculated. Each experiment used a random generated seed. In each round, the traffic rate generated by nodes varies from 0 to the maximum rate that cause the network congestion.

In the first test, we used the IEEE 802.11b DCF at 11 Mbps data and 1 Mbps for RTS/CTS/ACK, while in the second test we used the IEEE 802.11b EDCA at the same

 ISBN: 978-1-61208-183-0

rates. All the nodes have omni-directional antennas with the same transmission power, providing 100 m as transmission radius and 150 meters as carrier detection radius. In both tests, the nodes were configured with $\mathrm{CW}_{min} = 31$ and $\mathrm{CW}_{max} = 1023$. In the second test, the $AIFS = 1$ was used. The amount of reserved resource (TXOPLimit) for nodes C is $3t$ and for node B is $4t$, where $t = 0.915ms$, the time in milliseconds to transmit a 500 bytes frame and receiving a frame acknowledgment (ACK). The minimum queue length threshold at nodes C and B is 2 times the number of child nodes, i.e., the threshold of $C = 6$ and $B = 8$.

The number of nodes connected to each router in the prototype were assigned manually since it was not a goal of our work to perform node discovery. However, this function could be implemented using resource discovery protocols, such as Link Layer Discovery Protocol (LLDP).

V. RESULTS

The proposed mechanism was evaluated using the following performance metrics:

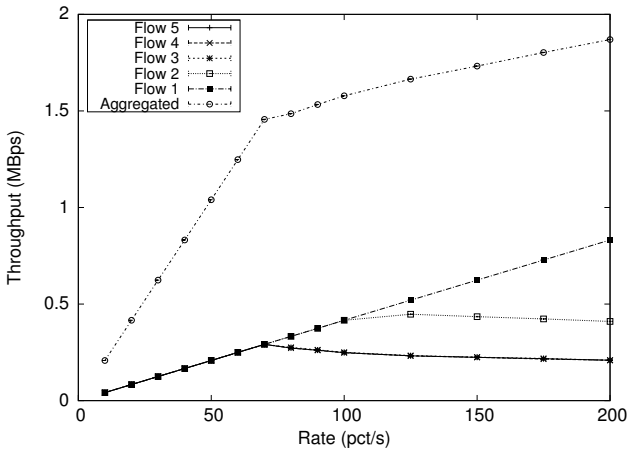
- Throughput in MBps: measuring the number of bytes per second of a specific flow received by the destination application;
- Number of transmission opportunities (TXOP): number of transmission opportunities obtained by each node per second;
- Packet Drops rate in pkts/s due full queue: the number of packets dropped per second for each node due full queue.

We plotted two graphs for each metric: one with standard IEEE 802.11 DCF (no mechanism) and another with proposed prioritization mechanism enabled. In all graphs, the x-axis represents the traffic load generated by the client node at the application layer.

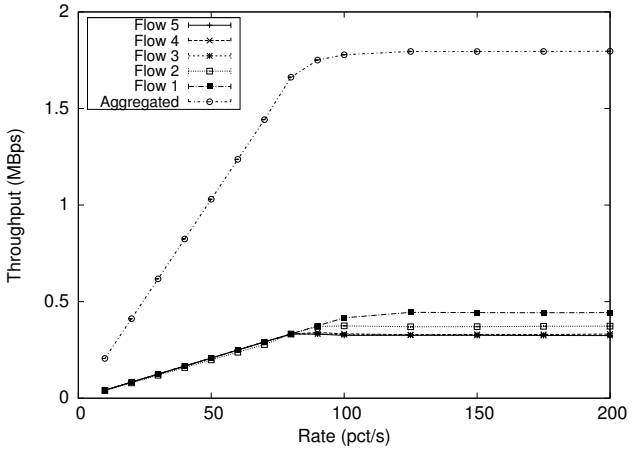
A. Results for Typical Scenario

Figure 5(a) shows what was written in the last paragraph of Section I, about increasing the flow rate in node 1 against other flows. The throughput reaches the maximum rate at 400 pkts/s, when the network becomes saturated. But we can note signs of saturation at 70 pkts/s, when the flow coming from the farthest nodes begins to decrease. The flow from node 2, one hop less than nodes 3, 4 and 5, begins to decrease the throughput at 125 pkts/s.

Figure 5(b) shows the behavior of throughput with the proposed mechanism enabled. We can see a better fairness among throughput from different nodes. When the system is saturated, the flow from node 1 no longer monopolizes the bandwidth. But even with the mechanism enabled, there is a little difference among the flows coming from nodes with different number of hops: node 1 was a little better than node 2 and better than flows from nodes 3, 4 and 5. We can also see a reduction on global throughput because



(a) Without mechanism.



(b) Proposed mechanism enabled.

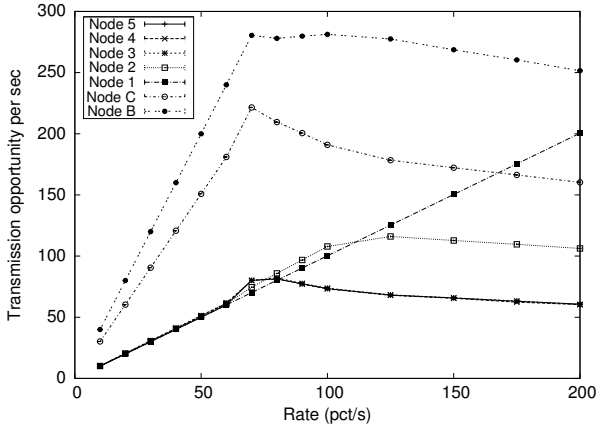
Figure 5. Typical Scenario: Throughput in MBps per flow

the mechanism gives more bandwidth for distant nodes that have a longer delay to gateway node reducing the overall throughput.

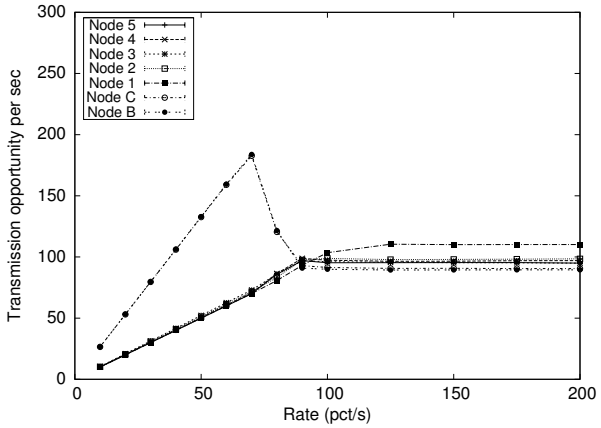
In Figure 6, the y-axis shows the number of transmission opportunities per second by node. The client nodes can only send one packet at every opportunity, so we can also consider that y-axis represents the number of packets sent per second (throughput in pkts/s). However, router nodes can send multiple packets at each transmission opportunity period. For example, on the simulation scenario showed in Figure 2, the node C was assigned a value of 2.745 ms, this means that every transmission opportunity given to C, the channel will be reserved for C for a maximum time of 2.745 ms (time enough to forward up to 3 packets of 500 bytes). With the mechanism off, all nodes send one frame every transmission opportunity period.

Figure 6(a) also demonstrates that the network becomes saturated above 70 pkts/s. After that, the graph shows the node 1 bandwidth monopolization.

 ISBN: 978-1-61208-183-0

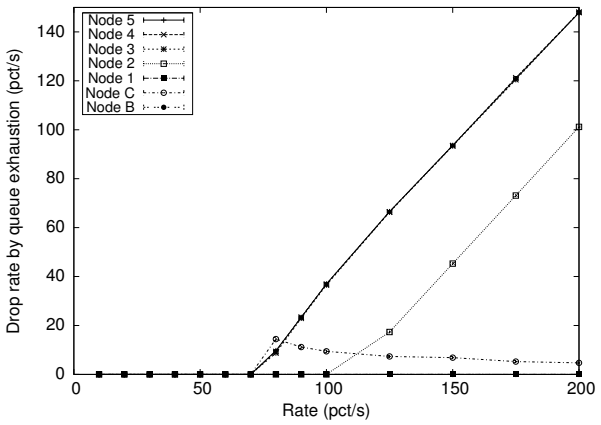


(a) Without mechanism.

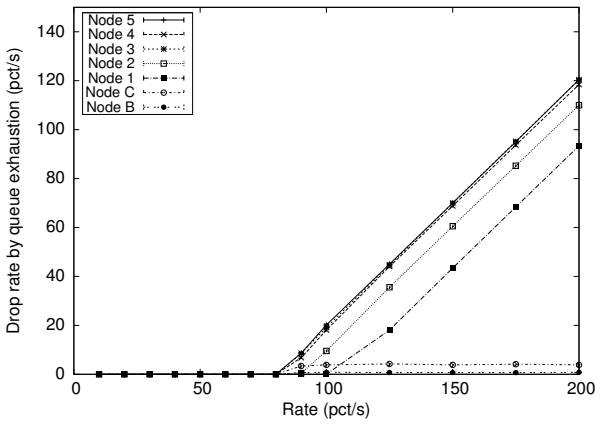


(b) Proposed mechanism enabled.

Figure 6. Typical Scenario: Transmission Opportunity per node



(a) Without mechanism.



(b) Proposed mechanism enabled.

Figure 7. Typical Scenario: Drops rate due full queue.

In Figure 6(b) we can see the router resource use yield. Yield is defined as the ratio of the number of packets transmitted by the number of transmission opportunities. When the system has low demand (until 70 pkts/s), the yield is low because there are no sufficient packets to fill the queue. When the system is near saturation (at 70 pkts/s), the yield increases, reaching 41% in node C. The maximum yield for the router C occurs when it always transmit 3 frames every transmission opportunity, reaching yield of 93% at 90 pkts/s rate.

After analyzing the results, we could notice that the reasons for dropping packets are full queue (Figure 7) and collision. Client nodes almost drop packets only due full queue. Since the routers only forward packets, not generate them, there is less queue use and hence less full queue drops. We can see in Figure 7 that drop rate is similar with and without the mechanism, but it is clear the improvement on drop fairness, now more distributed among nodes.

B. Results for Line Scenario (worst case)

The mechanism evaluation in a line scenario, worst case for WMN, shows that it was very efficient because in this topology the fairness is clearly evident.

In Figure 8(a), the maximum network throughput occurs at 375 pkt/s rate. But at 50 pkt/s rate there was a reduction on flow 5 throughput. Figure 8(b) shows the better fairness among flows after the mechanism was enabled. When the mechanism is enabled, we see an equalization of TXOP among client nodes and routers.

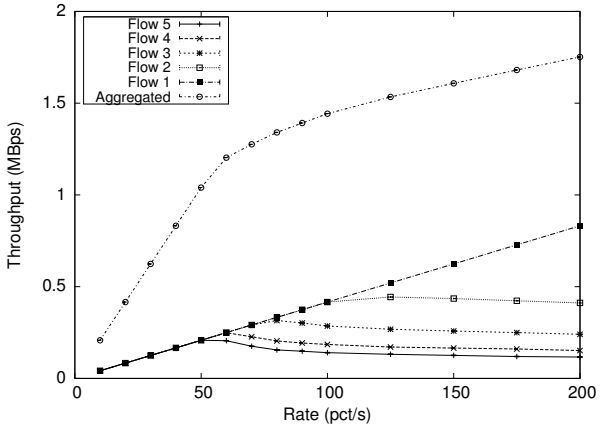
C. Results for Star Scenario (best case)

The third scenario evaluated a star topology, the best case of WMN. In this simulation we note that there is no improvement when the proposed mechanism is used, because all client nodes are equidistant (1 hop) from gateway node.

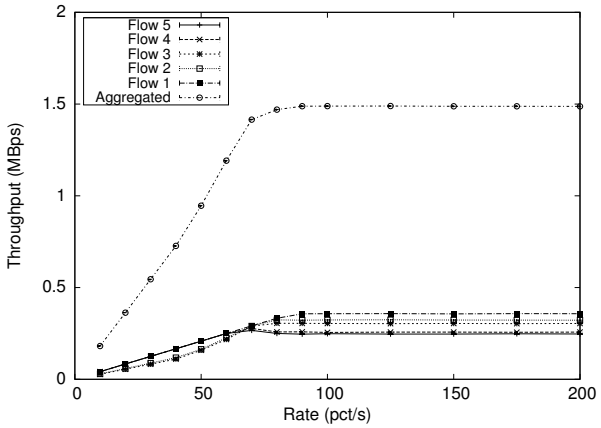
VI. Conclusion and Future Work

The WMN networks based on IEEE 802.11 are being used to provide Internet access in under-developed areas. Aiming

 ISBN: 978-1-61208-183-0



(a) Without mechanism.



(b) Proposed mechanism enabled.

Figure 8. Line Scenario: Throughput per flow in MBps

to extend the coverage of such networks in an economically viable manner, network designs have been studying platforms characterized by low-cost equipment, ease of configuration, installation and maintenance. When progressively increasing the transmission rate in client nodes, the throughput of the nodes near gateway router continues to increase against the throughput reduction from distant nodes. When the network reaches the bandwidth saturation, we observed that the farther client nodes from the gateway router suffer starvation.

This paper presented a new mechanism to improve the fairness at WMN back-haul. The mechanism changes the prioritization of medium access time of routers nodes to reach fair share of network resource. The basic operation is giving out to back-haul router nodes an amount of medium access time based on the number of client nodes connected.

The simulation experiments showed that the mechanism is effective, since it allocated the resource proportion to the numbers of flows routed by router node, minimizing the problem of unbalanced resource distribution among the WMN nodes. The efficiency was assessed by comparisons of the WMN performance with prioritization mechanism disabled and enabled.

Based on work presented in this paper we can suggest some future work: Dynamic allocation resource: as showed in Section III, the prioritization mechanism allocates an amount of resource based on the number of descendants clients. If the number of client nodes vary dynamically, it is necessary to adjust the resource allocation dynamically. Protocol to exchange information among routers: this protocol would be necessary to implement dynamic allocation resource mechanism to update status information (number of client nodes).

REFERENCES

[1] Ivan F. Akyildiz and Xudong Wang, "A Survey on Wireless Mesh Networks," *IEEE Communication Magazine*, vol. 43, no. 9, pp. S23–S30, 2005.

[2] S. Kim, S. Lee, and S. Choi, "The Impact of IEEE 802.11 MAC Strategies on Multi-Hop Wireless Mesh Networks," in *2nd IEEE Workshop on Wireless Mesh Networks (WiMesh 2006)*, 2006, pp. 38–47.

[3] Shugong Xu and Tarek Saadawi, "Does IEEE 802.11 MAC Protocol Work Well in Multi-hop Wireless Ad Hoc Networks?" *IEEE Communications Magazine*, vol. 39, no. 6, pp. 130–137, 2001.

[4] B. Bensaou, Y. Wang, and C. C. Ko, "Fair Medium Access in 802.11 based Wireless Ad-Hoc Networks," in *The ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc00)*, 2000, pp. 99–106.

[5] Y. Wang and B. Bensaou, "Achieving Fairness in IEEE 802.11 DFWMAC with Variable Packet Lengths," in *Global Telecommunications Conference (GLOBECOM01)*, vol. 6. IEEE, 2001, pp. 3588–3593.

[6] K.-C. Wang and P. Ramanathan, "End-to-End Delay Assurances in Multihop Wireless Local Area Networks," in *Global Telecommunications Conference (GLOBECOM03)*, vol. 5. IEEE, 2003, pp. 2962–2966.

[7] V. Gambiroza, B. Sadeghi, and E. W. Knightly, "End-to-end performance and fairness in multihop wireless backhaul networks," in *10th ACM Annual International Conference on Mobile Computing and Networking (MobiCom04)*. New York, NY, USA: ACM Press, 2004, pp. 287–301.

[8] Sven Wiethölter and Christian Hoene, "IEEE 802.11e EDCA and CFB Simulation Model for NS-2," http://www.tkn.tu-berlin.de/research/802.11e_ns2/, 2007, Last accessed Jul 2010.

[9] Vivek, "NS Static Routing," http://decision.csl.uiuc.edu/∼vivek/software/ns-manual/, Last accessed Jan 2010.

[10] K. Tang and M. Gerla, "Fair Sharing of MAC under TCP in Wireless Ad Hoc Networks," in *Proc. IEEE Multiaccess, Mobility and Teletraffic for Wireless Communications (MMT99)*, vol. 4. IEEE, 1999, pp. 231–240.

 ISBN: 978-1-61208-183-0

An Integrated Location Method using Reference Landmarks for Dead Reckoning System

Mingmei Li, Kazuyuki Tasaka and Kiyohito Yoshihara
KDDI R&D Laboratories Inc.
2-1-15 Ohara Fujimino-shi, 356-8502
Saitama, Japan
Email: {mi-li, ka-tasaka, yosshy}@kddilabs.jp

Abstract— **Dead reckoning system is a promising solution for pedestrian location, which determines the user's location by using multiple built-in sensors in the wireless devices or wearable sensors (3-axis, magnetic, gyro) on the pedestrian. However, most location estimation methods in dead reckoning systems meet the problem of accumulated location errors due to magnetic field disturbances. Some existing methods solved the problem by forcing the user to take the phone in a fixed gesture. However, they are not practical for the users in daily life and also accumulate errors with time flying. To solve the above problem, in this paper, we propose a new method to integrate wireless time-of-arrival landmark signals into dead reckoning system. The proposed method utilizes the hearable Time Difference of Arrival signals from the LMs, which are placed in the known locations to help correct the accumulated location estimation errors in dead reckoning systems. We evaluate the location estimation error with the proposed method when different number of reference LMs is placed and compare with location estimation in a dead reckoning system. Simulation results show that location estimation errors drop when different number of reference LMs are used.**

Keywords- dead reckoning system; integrated location method; mobile phone built-in sensors (3-axis, magnetic, gyro); TDOA landmarks.

I. INTRODUCTION

Recently, mobile devices are widely spreading. The location information of mobile devices is expected to be used in many new services, such as friend finding, shopping guide, etc. Many mobile phones have GPS receivers, but some services should be provided in the situations where the function of GPS receivers is not available, such as indoor areas, underground areas and complicated urban districts with a lot of buildings.

Many studies have investigated indoor location technologies, and some services based on time difference of arrival (TDOA) [1-4], wifi access points [5] are available today. However, providing many sensors with localization hardware (e.g., GPS) is expensive in terms of cost and energy consumption [6-7]. A more reasonable solution to the localization problem is to allow mobile phones to have their step information at all times, and allow users to infer information from these sensors [8-11]. Recently, mobile phones with built-in sensors (e.g., 3-axis, magnetic, gyro sensors, etc) have been widely spreading. These sensors help providing a lot of user's information that can be used in a location system: arm swing detection, step count estimation, direction estimation and step length estimation. Swing detection, step count estimation, direction estimation and step length estimation. Therefore, relative location systems- DR systems offer a promising solution [12-15]. Dead reckoning (DR) systems use sensors (e.g., accelerometer and gyroscope) to determine the user's current location without external infrastructures, they derive the characteristics of human such as the number of steps, step length, and direction.

However, most location estimation methods in DR systems meet the problem of accumulated location errors due to magnetic field disturbances. Existing works solved the problem by using the following methods: required the user to input their step size before the location estimation [12-13]; forced the user to mount the mobile phone in a fix gesture on his/her body [14]. Obviously, the first one is not accurate once the users step into crowded environment, user's step size is changed into smaller ones; and the second one brings inconvenience to the users. Taking these into considerations, in this paper, we propose a new method to correct the accumulated estimation errors. The proposed method integrates wireless TDOA LMs signals (LMs are placed in known places) into location estimation in a dead reckoning system. Once the user walks near the transmission range of a TDOA LM, the TDOA signal information are received, and the accumulated errors are compensated with the received ranges.

The paper is organized as follows. Section II briefly describes dead reckoning system. Section III presents the proposed integration location system. Section IV illustrates the details of the proposed integrated method. Section V compares simulation results of the proposed method and the dead reckoning location method. Finally, Section VI concludes the paper.

II. DEAD RECKONING SYSTEM

A. System Overview

In a pedestrian navigation system, it is necessary to locate the position of the user in any environment. Dead reckoning location system is such a system that can estimate the user's location based on a previously determined location, without external infrastructures. For this reason, a self-contained

 ISBN: 978-1-61208-183-0

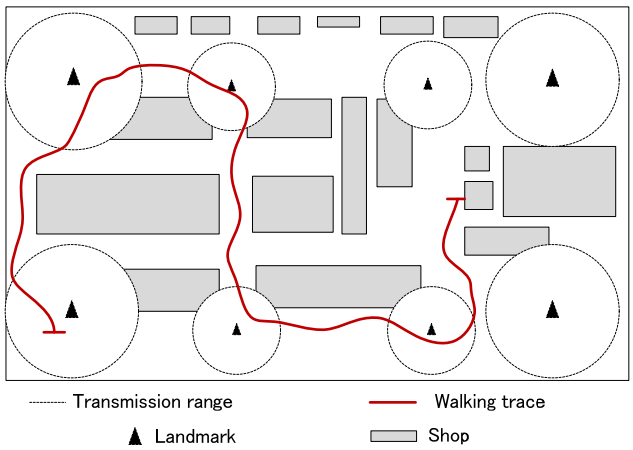


Figure 1. Integrated location system

navigation system based on a dead reckoning principle is ofinterest. To locate the position of the user, distance and heading from a known origin have to be measured at an acceptable level of accuracy [14-15]. In a pedestrian navigation system, an electronic pedometer can be used to count the number of steps, which can be combined with the step size for obtaining the distance traveled. In addition, a terrestrial magnetic compass can be used as a heading sensor.

B. Problem in Dead Reckoning System

The dead reckoning system is dependent upon the accuracy of measured distance and heading and the accuracy of the known origin. Recently, mobile phones with built-in sensors (e.g., 3-axis, magnetic, gyro sensors, etc) have been widely spreading. These sensors help providing a lot of user's information that can be used in a dead reckoning location system. The current relative location of the user is calculated as movement in an estimated direction, based on the step length estimated from the last position at each estimated step. However, most studies on dead reckoning system accumulated estimation errors, which greatly reduced location accuracy.

Apart from sensor measurement noise, the main factors that have effect on the estimation accuracy are step size error and direction bias error. The step size error is the difference between the actual step size and the predetermined step size entered by the user. Although the exact step size is not necessary for the distance calculation, the average step size over a short period has to be measured. The reason for this is that the step size of the user may vary according to the environment; for example, the step size of the user is shorter when the user is walking in a crowded area. Hence, the predetermined step size cannot be used effectively for the distance measurement. The direction bias error is a result of several causes such as magnetic declination and body offset [2, 8]. In a clean environment, the total bias can be changed slowly over a long period and may need to be re-calibrated occasionally. Therefore, how to alleviate the accumulated location errors is the challenging issue. We solve this problem by introducing the reference TDOA LMs in the system to alleviate the accumulated estimation errors. The details are described in the next sections.

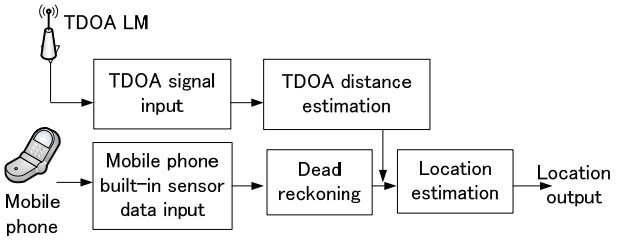


Figure 2. System architecture

III. INTEGRATED LOCATION SYSTEM

A. Requirements and Scope

Our method is to develop an integrated location method using some reference LMs and mobile phone built-in sensors. The requirements and scope are as follows,

- Commercial mobile phones with built-in sensors (3-axis, magnetic, gyro sensors).
- Reference LMs located at known places, e.g., the intersections or corners of the road.
- Location estimation should be compensated when the user received TDOA signals from LMs, accuracy guaranteed even in a magnetically noisy environment.

B. System Architecuture

The integrated location system is shown in Figure 1. It includes some reference LMs (RF/acoustic signals deployed at some known locations) which are deployed at the corners or intersections of shops. The transmission range of each LM is different based on the space of the corners and intersections of shops. The user walks in the room, and the walking trace is shown in red line. The system architecture is shown in Figure 2, which constitutes mobile phones with built-in (3-axis accelerometer / magnetic / gyro) sensors and TDOA LMs. Each LM broadcasts its ID and location periodically. The number of steps is counted by detecting the peaks of acceleration in the same way as a typical pedometer and the orientation of a mobile phone with the sensors is found in the same way as an electronic compass. The step size and the compass bias were calculated from step size average and compass bias average of the last 100 seconds. In location correction part, the mobile phone corrects its location, by using information from TDOA signals [14-15].

C. Reference TDOA LMs

The reference LMs use TDOA ranging method, which are based on the time difference of arrival between radio and acoustic signals. TDOA ranging utilizes the fact that two signals propagate at different speeds: an instantaneously at short distances for radio waves, and approximately 340m/s for sound. The TDOA LMs operate as follow; the sender broadcasts a radio message followed by an acoustic signal (chirp) with a known frequency signature. The mobile phone receives the radio message by starting to listen the chip using

 ISBN: 978-1-61208-183-0

Table 1. Definitions

Definitions:
r_{LM}- maximum transmission range of a LM;
r_{TDOA-} received TDOA signal distance from a LM;

Intersection- arcs of a circle of a radiu r_{TDOA} centered at LM by roads which locates within the circle.

(x_k, y_k)- LM_k location;
(x_{est}, y_{est})- mobile phone location estimated by DR method;
d_{est}- the distance between (x_{est}, y_{est}) and (x_k, y_k);

(x_{mk}, y_{mk})- centroid location of intersection m within LM_k;
$(x_{cross,ij}, y_{cross,ij})$- location of the road's cross-point i and j;

integrated RF reader [15]. Once the mobile phone detects the radio message, it estimates the distance by computing the difference in arrival time of the radio and acoustic signals. As an example, the MCS410CA Cricket mote can be used as the TDOA reference LM [1]. The Cricket Mote includes all of the standard MICA2 hardware and an ultrasound transmitter and receiver. This device uses the combination of RF and ultrasound technologies to establish differential time of arrival and hence linear range estimates [16]. The Cricket mote works at a frequency in the 433MHz band, but the frequency can be fine tuned within several megahertz either at compile time or runtime [17].

For the mobile phone users, Wireless Dynamics has announced a device called the iCarte that will add both RFID and NFC capabilities to the iPhone, which can be used for the users to receive RF signals [18].

IV. INTEGRATED LOCATION METHOD

As discussed in Section II, there are several sources of errors in the dead reckoning system, mainly as step error and direction error. Most existing methods solved the problem by using the following methods: required the user to input their step size before the location estimation; forced the user to mount the mobile phone in a special position on his/her body. Obviously, these methods accumulate estimation errors with time flying. The first one is not very accurate once the users step into crowded environment, they may change their step size into smaller ones; and the second one brings inconvenience to the users. Taking these into considerations, we propose an integrated location method utilizing the received TDOA range from reference locations to correct the accumulated estimation errors.

There are two parts in the proposed integrated location method. The first part is that, when TDOA LM signal is not available, the location is estimated using the method in a DR system [15]. The second part is that, when the TDOA LM signal is available, the received TDOA signal will be incorporated in the estimation process. We will describe the detail steps of the whole method, and two parts in the following subsections. Table 1 shows the definitions that are used in the paper.

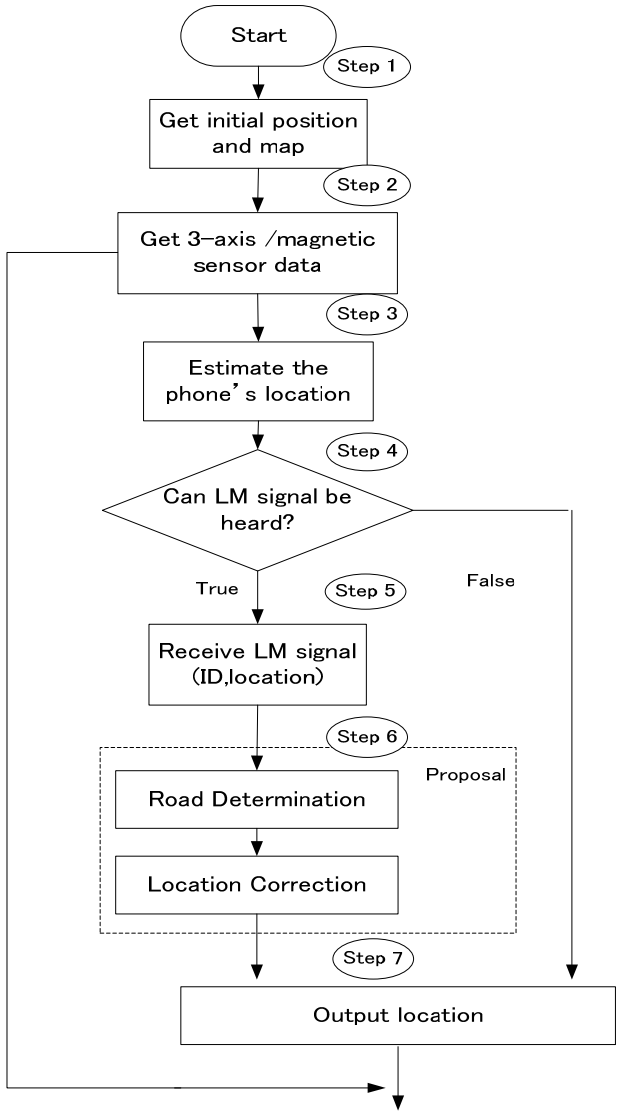


Figure 3. The flowchart of the proposed method

A. Overview of the Proposed Method

The following describes the detail steps of the overall method. The flowchart is illustrated in Figure 3.
[Step 1] Get initial location from the nearest LM.
[Step 2] Get 3-axis/magnetic/gyro sensor information from the mobile phone.
[Step 3] Estimate current mobile phone's location by using the method in a DR system [15].
[Step 4] Receive signals from LMs, and record LM's ID, and location.
[Step 5] Check if LM signals can be heard. If LM signal can be heard, go to step 5; if LM signals cannot be heard, go to step 6.
[Step 6] Stop location estimation by DR method. Estimate the mobile phone's location by using the following method, then go to step 7.
[Step 7] Location output.

 ISBN: 978-1-61208-183-0

B. Road Determination

As the LMs locate at the corners or the intersections of the roads, once the user receives TDOA signals, we need determine the direction of the user's road first. There are two cases for road determination:

-Case 1: (x_{est}, y_{est}) is on a single road
The signal road is determined as the current road.
e.g., (x_{est}, y_{est}) is on road 1-> road 1 is its current road.
-Case 2: (x_{est}, y_{est}) is on several roads (inside the transmission range of a LM)
e.g., (x_{est}, y_{est}) is on road 1,2,3
[step 1] Calculate the distance between the current location and the centroids of those roads.
e.g., (x_{est}, y_{est}) is on roads 1,2, and 3. Calculates the distance d_{11}, d_{21}, d_{31}, for (x_{est}, y_{est}) with (x_{11}, y_{11}), (x_{21}, y_{21}), and (x_{31}, y_{31}).
[step 2] Sort the calculated distance. The road with the shortest calculated distance is determined as current road.
e.g., $d_{11} < d_{21} < d_{31}$, then road 1 is current road.

C. Location Correction

Once the user receives TDOA signals from the reference LMs, the user should locate in the transmission range of the reference LMs. We have determined the user's road in the previous subsection. However, the user's location estimated by a DR system may be different from the transmission range of the heard LM, due to accumulated errors [17-18]. Therefore, we discuss the location correction based on the estimation differences between DR location and received LM ranges. Three cases are studied: within hearable TDOA ranges; within twice TDOA ranges; and out of twice TDOA ranges. Figure 4 shows the detail of the studied scenario.

-Case 1: $d_{est} < r_{LM}$
[Step 1] List the user's possible intersection(s) ($i=1,\ldots,N$) based on the past location (last three steps, t-3,t-2,t-1,t).
e.g.,
road 1 at time t-1, t-2, t-3 -> possible intersection is 1.
road 3 at time t-3, t-2, road 1 at time 1 -> possible intersection is 1,3.
[Step 2] Estimate the output location using the intersection's location obtained in step 1 and the estimated location (x_{est}, y_{est}).

$$x_{out} = \frac{1}{\mathrm{N}+1}\left(x_{est} + \sum_{\mathrm{i}=1}^{\mathrm{N}} x_{ik}\right) \quad (1)$$

$$y_{out} = \frac{1}{\mathrm{N}+1}\left(y_{est} + \sum_{\mathrm{i}=1}^{\mathrm{N}} y_{ik}\right) \quad (2)$$

[step 3] If (x_{out}, y_{out}) is on the map, then outputs it. If (x_{out}, y_{out}) is not on the map, then go to step 4.
[step 4] Calculate the distance between the location (x_{out}, y_{out}) and the centroids of those roads.
e.g., calculates the distance d_{11}, d_{21}, d_{31}, d_{41} between (x_{out}, y_{out}) and the (x_{11}, y_{11}), (x_{21}, y_{21}), (x_{31}, y_{31}) (x_{41}, y_{41}).

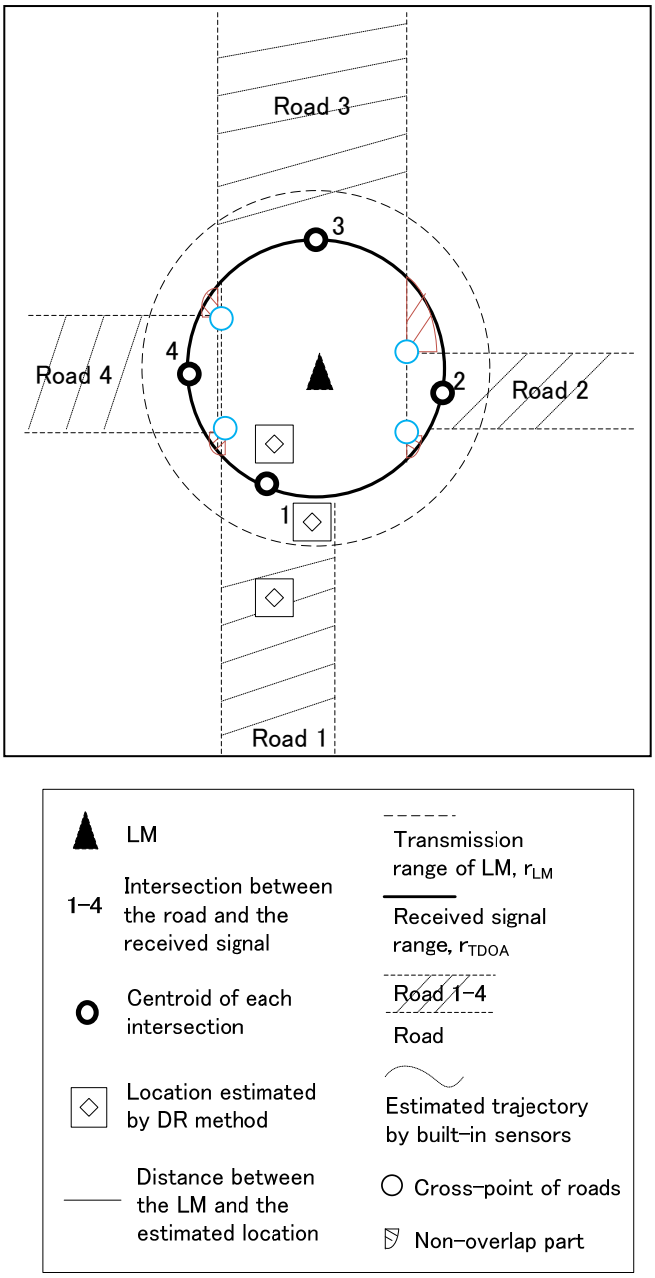


Figure 4. Study scenario

[step 5] Sort the calculated distances, and find the intersections with two shortest distances
e.g., the calculated result is as $d_{11} < d_{41} < d_{31} < d_{21}$, then intersections 1 and 4 are two intersections founded;
[step 6] Estimate and output the location by using the cross-point between the two intersections, $x_{out}=x_{cross,ij}$, $y_{out}=y_{cross,ij}$ e.g., $x_{out}=x_{cross,ij}$, $y_{out}=y_{cross,ij}$

-Case 2: $r_{LM} < d_{est} < 2\, r_{LM}$
[Step 1] Calculate the distance between the estimated location and the centroids of all the intersections, d_1,d_2,d_3,d_4;

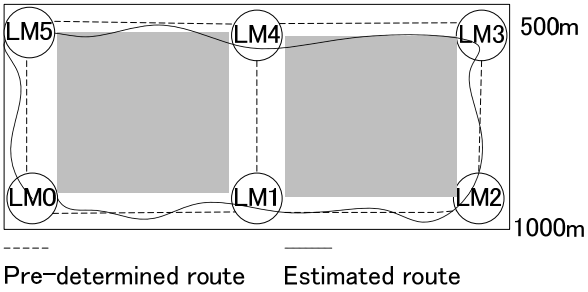


Figure 5. Simulation scenario

[Step 2] Sort the intersections in distance order, e.g., $d_1<d_4<d_2<d_3$, then 1<4<2<3;
[Step 3] Calculate the centroid on the arc between the shortest centroid and the second shortest centroid; and another centroid on the arc between the shortest centroid and the third shortest centroid,
e.g., (x_{141}, y_{141}) is the centroid on arc between (x_{11}, y_{11}) and (x_{41}, y_{41}). (x_{121}, y_{121}) is the centroid on arc between (x_{11}, y_{11}) and (x_{21}, y_{21}).
[Step 4] Calculate the centroid on the arc between the two centroids calculated from step 3; and output it, e.g., $(x_{out}\ y_{out}) =$ centroid on the arc between (x_{121}, y_{121}) and (x_{141}, y_{141}).

-Case 3: $d_{est}>2r_{LM}$
[Step 1] Determine the intersection based on the user's current road; e.g., user is in road 1, then the current intersection is 1;
[Step 2] Select a random location from the intersection determined in step 1, and output it, e.g., $(x_{random}, y_{random}) \in$ intersection(x_{mk}, y_{mk}), then $(x_{out}=x_{random}, y_{out}=y_{random})$.

V. SIMULATION

In this section, simulations are performed to evaluate the performance of the proposed method.

A. Simulation Setup

The simulation is evaluated in Matlab. We consider a simulation area, where six reference LMs are placed at the corners in a simulation area of 1000x500m, as shown in Figure 5. The coordinates of each LM are: (10, 10), (500, 10), (990, 10), (10, 490), (500, 490), (990, 490). The transmission distance of a LM is assumed to be 15m. Considering wireless radio wave propagation which may cause serious uncertainty while measuring signals from LMs, the noise is added to the simulated transmission distance with Gauss distribution$N(0, 1^2)$. In the simulations, we do not consider to adaptively change LM coverage.

We simulate the user's walking trajectory using a random walk model. The user's step length uses built-in sensor data from references [2, 3]. The user's walking speed is selected between 0.6m/s~1.4m/s. The user's walking distance is obtained at each second; and step size is sampled at each second. The walking distance is considered with $\pm 10\%$ error [15]. Walking direction detection is considered within $\pm 30\%$ error [15].

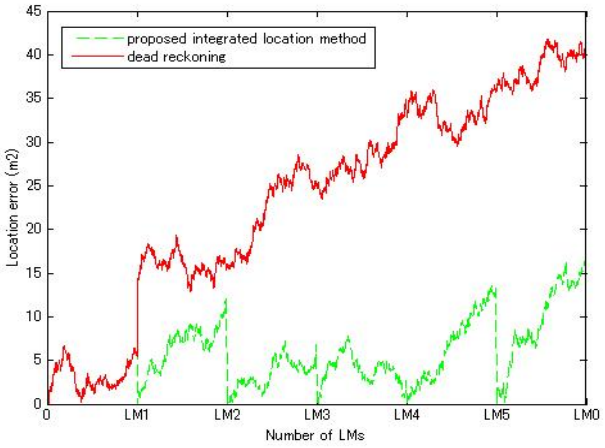


Figure 6(a). Location error, route 1

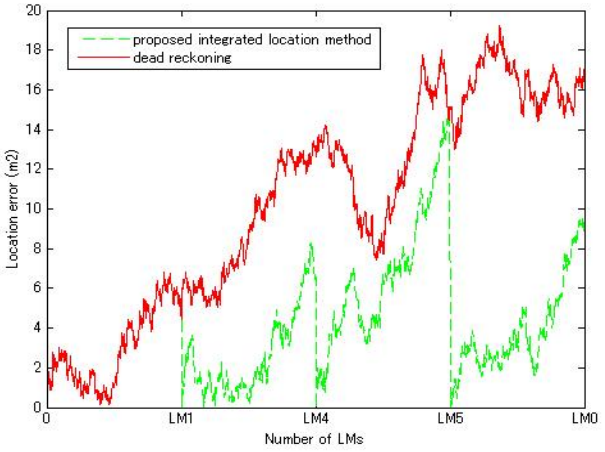


Figure 6(b). Location error, route 2

The goal of a wireless location system is to accurately locate a user. In the simulations, we evaluate the system performance using location error as the metric. Location error is defined as the difference between the user's pre-determined route and the estimated route. Two routes are studied in the simulation: route 1, LM0->LM1->LM2->LM3->LM4 ->LM5->LM0; and route 2, LM1->LM4->LM5->LM0. The location estimation method in the DR system [16] is used as a comparison method, represented as DR method. All the simulation results are the average of 10 runs.

B. Simulation Results

Figure 6 (a) shows the simulation results for location errors evaluated by two methods when the user is simulated to walk in route 1. From the figure, we see that DR method accumulates more location errors when the user walks in the direction of LM0->LM1->LM2->LM3->LM4 ->LM5->LM0. For instance, location error is about $15m^2$ at LM1, $23m^2$ at LM2, and about $40m^2$ when the user comes back to LM0. Obviously, this is due to accumulated errors with time being.

 ISBN: 978-1-61208-183-0

From the same figure, we see that using the proposed integrated location method, location error drops obviously, which means better location accuracy for the mobile phone user. For example, the user accumulates about 12 m^2 errors before coming to LM2, and drops about 6m^2 after receiving TDOA signals and re-estimating the location near LM2. Finally, the location error drops to 10 m^2 when the user goes back to LM0.

Figure 6 (a) shows the simulation results for location errors evaluated by two methods when the user is simulated to walk in route 2. Using DR method, location error is about 15 m^2 at LM4 and about 18 m^2 when the user goes back to LM0. Using the proposed integration method, location correction works when the user receives TDOA signal from each reference LM. The highest location error is about 14 m^2 that happens at LM5, and it quickly drops to 1~2 m^2. Finally location error is about 8~9 m^2 when the user comes back to LM0.

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a new method to integrate wireless TDOA LM signals (TDOA LMs locate at several fixed locations) as reference location information into dead reckoning system. We evaluated the location estimation error when different number of reference LMs is used to compensate the accumulated location errors. Performance of the proposed method is compared with the location method in a DR system. Simulation results demonstrate that average 9m^{2}~25m^2 location estimation error in a simulation area of 1000x500m can be obtained, verify the effectiveness of the proposed method. This level of location accuracy helps mothers locate their children more accurate and easier in a shopping mall. Map matching, implementation and evaluation of a practical system will be a part of our future work.

REFERENCES

[1] A.Smith, H.Balakrishnan, M. Goraczko, and N. Priyantha. “Tracking Moving Devices with the Cricket Location System”. *In Proceeding of Mobisys 2004.*(2004)

[2] M. Rudafshani, and S. Datta. “Localization in Wireless Sensor Networks”. *In Proceedings of IPSN 2007*, pp. 51-60.(2007)

[3] G.Zhang,S.Krishnan, F.Chin, and K.C.Chung. “UWB MultiCell Indoor Localization Experiment System with Adaptive TDOA Combination”. *In Proceeding of VTC 2008-fall*.(2008)

[4] Lui, K.W.K., So, H.C., and Ma, W.-K. “Maximum A Posteriori Approach to Time-of-Arrival-Based Localization in Non-Line-of-Sight Environment”. *IEEE Transactions on Vehicular Technology*,Vol. 59, Issue. 3, pp. 1517 - 1523, 2010 (2010).

[5] Koozyt, Inc.,: PlaceEngine, http://www.placeengine.com/, accessed May 17, (2010).

[6] M.J.Caruso, “Application of magnetoresistive sensors in navigation systems”. *Sensors and Actuators*, SAE SP-1220, pp. 15-21, 1997. (1997)

[7] B., N., Schmidt, G., “Inertial sensor technology trends”. *Sensors Journal, IEEE*, vol. 1, no. 4, pp. 332-339, (2001)

[8] G.Gartner, A.Frank, G.Retscher, “Pedestrian Navigation System in Mixed Indoor/Outdoor Environment”. *Journal of. CORP*. pp24-27, 2004 (2004)

[9] M.Kourogi and T.Kurata, “Personal positioning based on walking locomotion analysis with self-contained sensosrs and a wearable camera”. *In Proceeding of ISMAR 2003*, pp.103 (2003).

[10] R. Stirling, K.Fyfe, and G. Lachapelle. “Evaluation of a new method of heading estimation for pedestrian dead reckoning using shoe mounted sensors”. *Journal.Navigation. Royal Society of Navigation, vol.58, no.1*,, pp31-45 (2005).

[11] S. H. Shin, C.G. Park, J.W. Kim, and H.S. Hong and J.M. Lee, “Adaptive Step Length Estimation Algorithm Using Low-Cost MEMS Inertial Sensors”. *IEEE Sensors Applications System (SAS)*(2007).

[12] L.F, P.J. Antaklis, L.A. Montestruque, M.B. McMickell, M.Lemmon, Y.Sun, H.Fang, I.Koutroulis, M.Haenggi, M.Xie, and X.Xie. “Design of a wireless assisted pedestrian dead reckoning system”. *The NavMote experience. IEEE Trans. Instrum.* Meas.,vol. 54, no. 6, pp. 2342-2358, (2005).

[13] U. Steinhoff and B.Schiele, “Dead reckoning from the pocket- An experimental study”. *in Proceeding of PerCom 2010*, pp. 162-170, (2010).

[14] P.Pombinho, A.P. Afonso, and M.B.Carmo. “Indoor positioning using a mobile phone with an *integrated accelerometer and d*igitial compass”. *INForum*, (2010).

[15] D. K, S. Muramatsu, T. Iwamoto, and H. Yokoyama, “Design and Implementation of Pedestrian Dead Reckoning System on a Mobile *Phone”*. *IEICE Transactions on Information and Systems*, Vol. E94. No. 6, pp. 1137-1146, 2011.

[16] N. B. Priyantha, “The Cricket Indoor Location System PhD Thesis”. *Massachusetts Institute of Technology*, June 2005.

[17] Y. Wang, S.Goddard, and L.C.Perez “A Study on the Cricket Location-Support System Communication Protocols”. *in Proceeding of IEEE EIT 2007*, pp. 257-262, 2007.

[18] D. Etherington, “iCarte Turns the iPhone Into an RFID Reader”, *http://gigaom.com/apple/icarte-turns-the-iphone-into-an-rfid-reader/* , Nov, 2009

[19] X. C. Xu, N. S. V. Rao, and S. Sahni, “A Computational Geometry Method for Localization using Differences of Distances”. *ACM Trans. on Sensor networks*, Vol. 6, No. 2, pp. 10, Article. 10, 2010.

[20] S. Minamimoto, S. Fuji, H. Yamguchi, and T. Higashino, “Local Map Generation using Position and Communication History of Mobile node”. *in Proceeding of IEEE percom*, pp. 2-10, 2010.

[21] Beauregard, “Omnidirectional pedestrian navigation for first responders”. *in Proeedings of 4th Workshop on Positioning, Navigation and Communication*, pp. 33-36, 2007.

 ISBN: 978-1-61208-183-0

MANET with the Q-Routing Protocol

Ramzi A. Haraty and Badieh Traboulsi
Department of Computer Science and Mathematics
Lebanese American University
Beirut, Lebanon
Email: rharaty@lau.edu.lb, badieh.taboulsi@lau.edu.lb

***Abstract*--With ad hoc networks having much more advantages over other types of networks in a mobile world, this made it an attractive field for many protocols. In this paper, we propose an implementation of the Q-Routing protocol working over a mobile ad hoc network to enhance the performance of the packets sent and received. However, to implement a protocol in such an environment, many factors need to be taken into consideration, such as: exploration and learning over time to adapt to network changes. We used these factors in the protocol agents to update the routing tables found on each node accordingly. Our protocol has shown to perform well in MANETs as the load increased.**

Keywords-ad-hoc networks; exploration and learning; routing protocols.

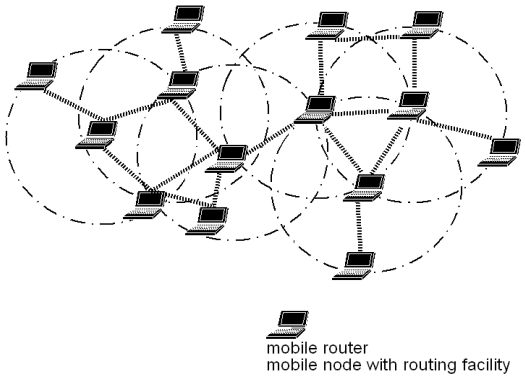
I. INTRODUCTION

As technology advances, people are gaining more interest towards mobile devices, which makes it an attractive source for many new enhancements. Mobile devices might range from mobile phones to portable design aids and laptops. The applications running on these devices most probably are not related to each other, and do not communicate. Nevertheless, there are scenarios where few devices move closer to each other forming a temporary network allowing transfer of different kinds of data and information. Such networks are known as *mobile ad hoc networks* (MANET) [1].

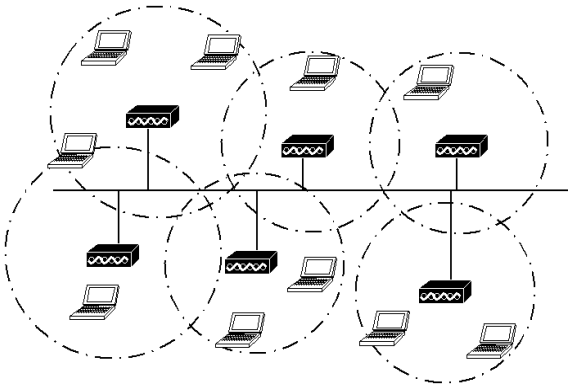
The main structure of the MANET, which distinguishes it from other networks, is that it can be formed without requiring any kind of infrastructure or administration. It contains mobile nodes that use an interface to communicate with each other wirelessly. These nodes can play the role of senders and receivers also known as *hosts*, or even *routers* that just forward the packets through. This gives a new ability to the network, nodes are no longer limited to the transmission range they got, and can connect to a mobile device, which is several hops away, and this is known as *multi-hop communication* [2].

Even though, as shown in Figure 1, ad hoc networks have different structure types, some networks share a hybrid of those two structures, such as the cellular network. Not to mention, the existence of a field of study to develop and simulate *wired-cum-wireless* protocols.

Ad hoc networks are not perfect, and they come with their own complications. Since the devices are mobile in nature, this means the network is ever changing. In other words, mobile nodes that were in range of each other, could leave the vicinity leading to disregard the direct link in between, and new nodes that were never near each other, come to be direct links. Direct link means it does not require multiple hops to reach the destination.



(a) Mobile ad hoc network



(b) Network with Infrastructure

Figure 1. Different network types of wireless structures.

In addition, different devices have different range capabilities, which mean a device might be able to reach another device, but this other device might not be able to reach the first one leading to an *asymmetric link*. A representation of an asymmetric link is shown in Figure 2. To solve this randomized nodes connectivity many different protocols were suggested [3] and will be subsequently discussed.

Q-Routing is the first routing algorithm to make use of reinforcement learning, called Q-learning [4][5]. Q-learning makes use of Q-values to perform updates and to estimate

 ISBN: 978-1-61208-183-0

how long it will take to send a packet to any particular destination through each one of the node's neighbors. In this paper, we propose an implementation of the Q-Routing protocol working over a mobile ad hoc network to enhance the performance of the packets sent and received.

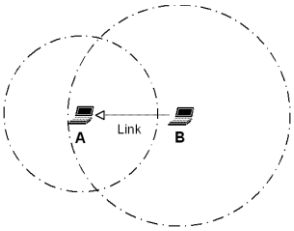


Figure 2. Asymmetric Link.

The rest of the paper is organized as follows: Section 2 discusses related work regarding mobile ad hoc networks and their associated routing protocols. Section 3 presents Q-Routing. Section 4 highlights the implementation. Section 5 presents the simulation results. Section 6 provides the conclusion.

II. RELATED WORK

Even though ad hoc networks do not require a specific structure, they still need a standardized way to base their communication at and make things work. For the nodes, in a mobile ad hoc network, to be able to decide on the route that the packets need to take to reach the specified destination, they need a convention, which is the *ad hoc routing protocol*. Moreover, at the start of the ad hoc network, the nodes are not aware of each other; thus, they need to discover each other by broadcasting to neighboring nodes their presence. Not only that, they also need to listen for other broadcasts in case a new node is added. The nodes might not only broadcast their own information, but are able to broadcast how to reach other nodes as well. Since there are many possibilities to design an ad hoc network, many types of related protocols specified for such network were discovered. Some of them are enhancements over others, and others are a combination of protocols.

A. Pro-active (Table-Driven) Routing

Just as its name suggests, this kind of routing maintains a table of the destinations and the paths to be taken, which is then broadcast to other nodes in the network. The time for convergence of the table varies between one and another depending on the algorithm used. Because of that time, it leads to a main disadvantage in case of a node dies or even a full restructuring of the network. The second disadvantage is that it needs loads of information for maintenance, which can scale up largely in big sized networks. However, while having a table of information, packets can be instantly sent according to the stored route reducing the latency for route discovery. The most common protocol used for this routing is *Highly Dynamic Destination-Sequenced Distance Vector Routing Protocol* (DSDV) [6].

B. Reactive (On-Demand) Routing

Unlike the pro-active routing, this type of routing only broadcasts for the destination when a send request is at hand. The broadcasted packets are basically Route Request packets, which are used to discover the path towards the destination. Since it does not store any information about the network, there will be a higher latency waiting for the path to be discovered. In case of a busy network, there might be too many broadcasts that can congest the network. On a highly mobile network, this routing works greatly, especially if nodes change frequently making any old route useless. The most known protocol used for this routing is *Ad hoc On Demand Vector* (AODV) [7].

C. Flow-Oriented Routing

Close to the reactive routing, this type of routing finds a route when demanded through a certain known flow. An example of this would be to consequently unicast whenever a new link is being advertised. With this comes its own consequences, which basically are taking too long while discovering totally new routes, and the probability of referencing to an existing traffic just to pay off for its lack of knowledge on the path. An example of this type of routing is the *Lightweight Mobile Routing protocol* (LMR) [8].

D. Hybrid (Both Pro-active and Reactive) Routing

This type of routing combines the best out of pro-active and reactive routing. It uses the pro-active routes stored when a routing is initialized and then uses the reactive broadcasting to deliver the packet to its destination. This gives the advantage of discovering better routes along the way. The disadvantage of this type is that its usefulness is related to how many active nodes are in the network. Also, it depends on how steep the traffic volume is. The *Temporally-Ordered Routing Algorithm* (TORA) [9] uses this type of routing.

E. Adaptive (Situation-aware) Routing

Just like the hybrid routing, this type uses both pro-active and reactive; however, the only difference is that it uses special metrics to decide which and when of the two models to be used. It shares the same disadvantages of the hybrid, and TORA is not just an example of hybrid routing, but also an adaptive routing.

F. Hierarchical Routing Protocols

This type of routing protocol designs hierarchical levels whereby pro-active or reactive routing is used depending on the level the node is in. It shares many of the Hybrid and adaptive properties. However, it differs in the decisions of which model to use depending on a specific attribute according to the level. This makes its advantages directly related to the depth of the levels drawn on the overall scheme. Another main disadvantage is that instead of depending on how steep the traffic volume, as is the case

 ISBN: 978-1-61208-183-0

with hybrid and adaptive routing protocols, it depends on the meshing parameters. One of the well-known protocols used is *Cluster Based Routing Protocol* (CBRP) [10].

III. Q-ROUTING

Q-Routing was proposed by Boyan and Littman [4][5]. They were able to come up with a new routing protocol that is based over reinforcement learning also known as adaptive routing over communicational networks. What most of other routing protocols focus on is finding the shortest route towards destination, but what many disregarded was taking into consideration the traffic load. Q-Routing was mainly developed as an enhancement to overcome the traffic load, which any network could fall into, through adapting to the current state the network is facing, and by that improving the performance. In Q-Routing, each node contains the algorithm for the decision making. They exchange information occasionally to update their stored data; one of which is the estimated time for delivery. Thus, the decision is made by taking the path that has the shortest delivery time instead of just the shortest path.

The Q-Routing policy works as follows: at first, every node stores an estimation of the time it will take the packet to reach all its neighbors. This is done by maintaining a routing table at every node making this protocol a pro-active one. The information that the table stores, is basically a combination of (y, d) where y is the neighbor and d is the destination. Their values imply the time taken for a packet to reach destination d passing by neighbor y. Assuming a node x wanted to send a packet to destination d. It will pass it to y since it has the lowest delivery time in its table, and then it will prompt y of the estimated time it takes the packet to reach d. After that, x updates the table information accordingly. Kardi Teknomo [11] wrote Q-learning pseudo code that can be used to enhance the Q-Routing algorithm when put in an ad hoc situation. If given a state diagram as an input with a defined goal, represented by a matrix R, the output is the minimal path from any state to its defined goal. This algorithm is mainly used by the agent that will learn through time, and by each state the agent either got a reward out of it, or none. This keeps going on until the agent reaches the given defined goal. The declared parameter γ has a range between 0 and 1. The closer it is to 1, the more the agent will delay the reward for a future better weight.

Since Q-Routing showed promising results with experimentations, other enhanced versions of Q-Routing where suggested such as the *Predictive Q-Routing* [12], *Dual Reinforcement Q-Routing* [13], and *Confidence-Based Q-Routing* [14].

IV. IMPLEMENTING A NEW MANET ROUTING PROTOCOL

There are numerous network simulators that can be used to help in simulating network scenarios and test performances - NS2/NS3, OPNET, NetSim, etc. In our work, we used Network Simulator 2, or NS2, for the features it gives, and since it is an open source which can be extended and modified. More than one version for the NS2 is out, but the one we used is ns-allinone-2.33 [15].

NS is an object-oriented, discrete network simulator, used primarily for research. It provides an important framework to simulate TCP, routing and multicasting protocols over wireless and wired networks [16].

Most of the studies made and developed around Q-Routing were involving wired networks; but, since the project is about Q-Routing over ad-hoc networks, things might need a few twists to make them work. Our implementation is based on work done by Francisco and Pedro [17].

As mentioned earlier, the main characteristic of mobile ad-hoc networks is that the nodes are in constant move, and so we should always adapt to the current new states of the network. To do so, we need a reinforcement learning procedure that displays the impact of a given action on the network. In our case, the action is the decision on which node the packet is to be sent next. The ideal situation is to have all of the packets sent from destination to source with a minimal cost. By cost we mean how much of the network resources were used.

Once a packet has been transmitted, it can either be delivered successfully or dropped. The packet is dropped when the *time to live* (TTL) has reached 0; it starts with a specific number and decreases at each hop from one node to the other. Once the packet reaches its destination, then it can be said that it was delivered successfully.

Q-Routing and reinforcement models require some value, say V, also known as a reward, which helps in the learning process. This value can be considered the sum of all the reinforcements starting at a specific state. Using the equations written by Bellman [18], the value of Q can be calculated based on the success or failure probability of transmitting a packet to the next hop. Therefore, every action done would affect the value of the V accordingly depending on how much the system did actually "learn" from the action done. So, if for example, the sent packet failed to be received by the target node, the value will decrease greatly; however, if it was received successfully, it will only drop a little. There are cases where the source becomes the destination at the same time. Doing so will not change the state of the value since nothing happened.

All the packets are built at the beginning state, and for each transition of the network state, estimation is stored at each node. Every node has two types of values, one is the optimal estimated V and the other is the V of its neighbors. This value will decrease as time passes from the time it started. The value of a node is not known for other nodes until a packet is sent. This is mainly to keep those nodes that do not interact with each other with a less value. Moreover, for each decision made by a routing table to transfer a packet through a specific node, disregarding others, will lead to a decreased value for those other nodes, which the packet did not pass through. The agents that are spread

 ISBN: 978-1-61208-183-0

through the network will act upon the information stored at the node they are currently residing. Once the agent arrives at the destined node, the value for that node will be broadcasted to the neighbors. Since it is always better to go backwards as well as forward with the updates, an update packet containing routing information will be sent from source to destination to keep both synchronized. With this bidirectional update, nodes that want to communicate with the source will have a better knowledge of it.

Exploring MANETs needs to have its own way, and the way we dealt with it is by using the Boltzmann distribution technique [19] to be able to identify as many of the actions as possible at first, and after that discovering whatever neighbors each of the nodes has to add them to the next-hops within the routing table. Since the network might scale up to become quite large, we do not want every agent to go exploring areas which it does not deem useful, and thus a greedy methodology could be used to identify these areas.

It is mainly impossible to set all of the media access control addresses as potential actions. Even if it was possible it would take a lot of unnecessary time. Thus, a new state needs to be added which explores the potential actions, and this is done by a broadcast. The action will be chosen using Boltzmann probabilistic technique. However, in case the agent does not have a specific action except to explore, then the exploration would be chosen as the action to be done, and it would be done by broadcasting to the neighboring nodes to discover them. Those nodes will most probably take it from there and continue to forward it to its destination. Since in mobile network exploration is the most important part of sending and receiving a packet, most of the network load should be dedicated for it. Therefore, there is no harm in using a greedy methodology to decide upon the next hops. The next hop is the node with a *V* that is greater than the one on the current node. Figure 3 demonstrates the idea.

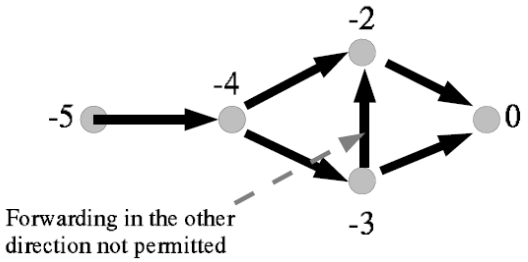


Figure 3. Greedy methodology.

The packets that are sent within the network go by the following format shown in Table I.

TABLE I. PACKET FORMAT

Field Name	Field Type	Field Contents
ORIGIN	IP Address	The original source of the packet
DESTINATION	IP Address	The final destination of the packet
SEQUENCENUMBER	Integer	Identifier generated from a counter at the source node
SOURCEVALUE	Floating-Point	$V_{src}(N)$
DESTINATIONVALUE	Floating-Point	$V_{dest}(N)$
HADERROR	Boolean	True if the packet's previous transmission failed
IPPACKET	Data Packet	IP Packet, or empty

The *routing agent table* has two variables, an id and a sequence number. The id is unique among all agents in the network. It is also able to produce distinctive indicator for the packets. The routing agent is connected to the routing entry and neighbor nodes in the sense that it keeps entries of all the interesting destinations, as well as, the neighboring nodes.

The *neighbor node table* has four variables, an id, a counter on how many sent attempts were made, how many were successful, and how many were received. Using this information, the estimation can be calculated.

The *routing entry table* has three variables, the id, the *V* for the destination, and the number of packets sent.

The *next hop route table* maintains two values whenever nodes advertise to each other a route. The values stored are the last known *V* and the time in addition to the table id.

The *forwarded packets table* stores the *V* of the node and the sequence number of the packets once it was forwarded. Since broadcasting packets to explore the network could lead to duplicates being generated, a unique sequence of numbers is maintained; therefore, if a node gets a packet twice, it will drop it.

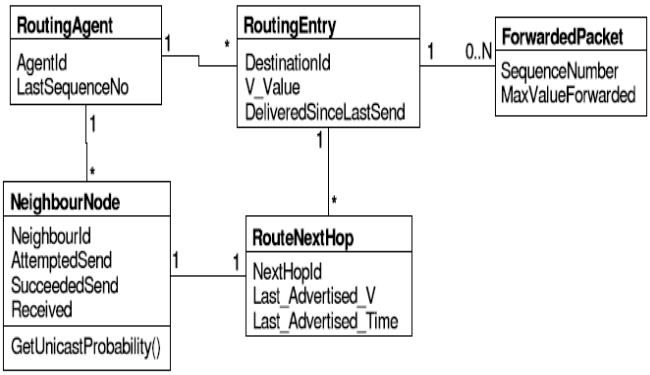


Figure 4. Table relations model.

The essential algorithm variables that we used are listed in Table II.

TABLE II. PROGRAM VARIABLES

Parameter Name	Units	Purpose
UNICASTSUCCESSREWARD	Reward	The Reinforcement received for successfully transmitting a packet. Fixed at -1.
UNICASTFAILUREREWARD	Reward	The Reinforcement received for a failed packet transmission. Fixed at -7.
EXPLORATIONUTILITY	Reward	The utility assigned to the exploration action
MINVALUE	Reward	The V-value of a broken route.
MINIMUMREWARD	Reward	Used to heuristically guide exploration
EVENTWINDOWSIZE	Time	The size of the window used to count events
EVENTWINDOWSAMPLES	Integer	Number of buckets to split event window into. Determines accuracy
PROBABILITYFROMRECEIVE	Float	Unicast Success Probability when have not attempted to transmit
RECEIVEWEIGHT	Float	How much Received Packets are weighted compared to Sent Packets
DECAYRATE	Float	How much the V-values grow every second that they are not advertised.
TEMPERATURE	Unitless	The temperature used in boltzmann action-selection
SEQUENCENUMBERMEMORY	Integer	Number of sequence numbers to record forwarded values for
MAXRECEIVESWITHOUTSEND	Integer	The number of packets that can be received on a flow without sending a response packet

V. SIMULATION RESULTS

We chose 50 nodes to be able to have enough fixed nodes while others are moving in different directions. Theoretically, the system should scale up as the number of nodes increases while still maintaining good results, but as the number of nodes rises up, the load on the network will increase in return; thus, affecting the total performance. For a detailed look at the simulation environment, users are referred to [20].

The trace file that was created out of the simulation was increasing greatly in terms of size due because the simulation included 50 nodes. The movement of the nodes was random. Since it was an ad hoc network, the 50 nodes were both clients and servers at the same time where they received, sent, and forwarded packets.

When the trace file was analyzed, we realized that there were a great percentage of packets received. The number of messages sent was relative to the time the simulation ran. Increasing the simulation time would increase the number of messages sent, but it should still maintain approximately 97% delivery rate, unless the network load increases gradually, then the delivery rate might start decreasing.

Messages Sent: 7385
Messages Received: 7233
Delivery Rate: 97.9417738659445

Next, the network load and the throughput of the Q-Routing protocol are calculated. The network load is basically the rate of the data sent by all the nodes. Whereas the throughput is the rate by which the nodes are receiving data. If we record the simulation to print out as it goes by, we end up with the graph, shown in Figure 5, for delivery ratio versus load.

Network Load = Packet size * Packets per Second * Number of Clients
Throughput = Network Load * Delivery Ratio

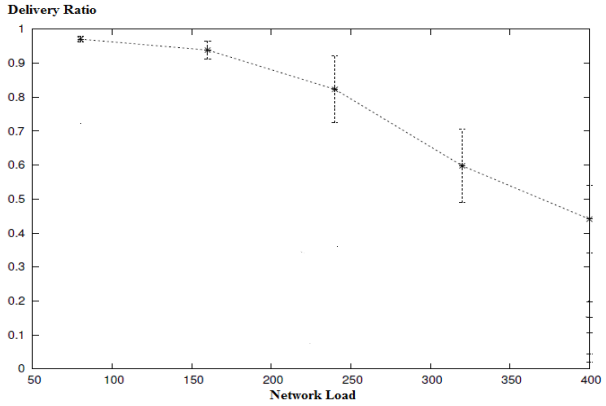


Figure 5. Delivery ratio versus network load.

As shown in Figure 5, while the network load increases, the delivery ratio decreases, and that is normal since more nodes will be congested leading to more packets being dropped. As for the throughput versus network load, we get the graph shown in Figure 6.

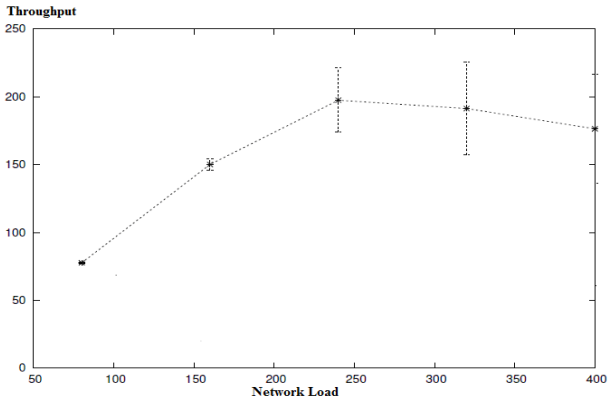


Figure 6. Throughput versus network load.

The throughput result shown in Figure 6 increases at the start, then slightly decreases as the network load starts to become large is due to the fact that at the beginning of the simulation, nodes start establishing connections with each other consecutively, and as the simulation time passes by, more nodes establish new connections while the previous ones still maintain their connection increasing the network load. After sometime, the network reaches a state where congestions occur at different nodes leading to a slight decrease in the overall throughput.

With the above results shown in Figures 5 and 6, we can say that the protocol was performing well not only when the

 ISBN: 978-1-61208-183-0

simulation started, but even later on when the load started increasing; thus, maintaining up to a 200 Kbps throughput. This is fairly good when compared to, for instance, what Jinyang Li *et al*. achieved in [21], where they stated that the maximum throughput achievable in an ad hoc network is 250 Kbps with packets having the size of 1500 bytes.

VI. CONCLUSION AND FUTURE WORK

When dealing with ad hoc networks, adding a little learning process to the way packets are sent and received will help out a lot, especially that the ad hoc networks are mobile making the exploration stage harder than usual. The mix between the Q-Routing and the wireless protocol led us to results which are promising. This protocol has proven to offer several advantages. For one, it is scalable as in no matter how many nodes we end up with, the number of agents will adjust accordingly. The second advantage is that there is not one main command, or central management system, where it leads the flows of the network; therefore, if any agent fails to do its work, it will not impact the overall reliability of the network. The third advantage is that agents are flexible in nature as in they can be modified according to any changes that the system may encounter.

As for future work, we plan to study other aspects of network performance such as latency and transmission rates. We also plan to implement the different types of Q-Routing and compare it with ours. In addition, the simulation we carried out was done only on 50 nodes, what we plan to do next is to analyze the protocol furthermore where more nodes are in place and more mobility around.

ACKNOWLEDGEMENT

This work was supported by the Lebanese American University.

REFERENCES

[1] S. Basagni, M. Conti, S. Goirdano, and I. Stojmenovic, Mobile Ad Hoc Networking, John Wiley and Sons, 2004, ISBN 0-471-37313-3.

[2] The MANET Web Page, Retrieved December 12, 2011, from http://www.ietf.org/html.charters/manet-charter.html.

[3] R. A. Haraty and W. Kdouh, "SDDSR: Sequence Driven Dynamic Source Routing for Ad Hoc Mobile Networks," Proc. of the World Automation Congress - International Symposium on Soft Computing for Industry. Budapest, Hungary, July 2006, pp. 1-8.

[4] J. A. Boyan and M. L. Littman, Packet Routing in Dynamically Changing Networks: A Reinforcement Learning Approach. Advances in Neural Information Processing Systems 6, San Francisco, CA, 1994. DOI: 10.1109/IPDPS.2005.323, pp. 671-678.

[5] J. A. Boyan and M. L. Littman, "A Distributed Reinforcement Learning Scheme for Network Routing," Proc. of the First International Workshop on Applications of Neural Networks to Telecommunications, 1993, pp. 45-51.

[6] C. E. Perkins and P. Bhagwat, Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers, *Comp. Commun. Rev.*, Oct. 1994, pp. 234–44.

[7] C. E. Perkins and E. M. Royer, "Ad-hoc On-Demand Distance Vector Routing," Proc. of the 2nd IEEE Workshop on Mobile Computing Systems and Applications, Feb. 1999, pp. 90–100.

[8] L. Ji and M. S. Corson, "A Lightweight Adaptive Multicast Algorithm," Proc. of GLOBECOM '98, Nov. 1998, pp. 1036–42.

[9] V. D. Park and M. S. Corson, "A Highly Adaptive Distributed Routing Algorithm for Mobile Wireless Networks," Proc. of the IEEE Conference on Computer Communications (INFOCOM '97), Apr. 1997, pp. 1405-1413.

[10] C. C. Chiang, "Routing in Clustered Multihop, Mobile Wireless Networks with Fading Channel," Proc. of the IEEE SICON '97, Apr. 1997, pp. 197–211.

[11] K. Teknomo, Q-Learning Algorithm, Retrieved December 12, 2011, from http://people.revoledu.com/kardi/tutorial/ReinforcementLearning/Q-Learning-Algorithm.htm.

[12] S.P.M. Choi, and D. Yeung, "Predictive Q-Routing: A Memory-based Reinforcement Learning Approach to Adaptive Traffic Control," Proc. of the Neural Information Processing Systems, 1995, pp. 945-951.

[13] S. Kumar and R. Miikkulainen, "Dual Reinforcement Q-Routing: An On-Line Adaptive Routing Algorithm," Proc. of the Artificial Neural Networks in Engineering Conference, 1997, vol. 7, pp. 231-238.

[14] S. Kumar and R. Miikkulainen, "Confidence Based Dual Reinforcement Q-Routing: An Adaptive Online Network Routing Algorithm," Proc. of 16th International Joint Conference on Artificial Intelligence, 1999, pp. 758-763.

[15] M. Greis, Tutorial for the Network Simulator "ns", Retrieved on December 12, 2011, from http://www.isi.edu/nsnam/ns/tutorial/index.html.

[16] The Network Simulator – ns-2, Retrieved on December 12, 2011, from http://www.isi.edu/nsnam/ns/.

[17] F. J. Ros and P. M. Ruiz, Implementing a New Manet Unicast Routing Protocol in NS2, Technical Report, University of Murcia, 2004.

[18] R. E. Bellman, Dynamic Programming. Princeton University Press, 1957, ISBN 0-486-42809-5.

[19] D. Lindley, Boltzmann's Atom: the Great Debate that Launched a Revolution in Physics, The Free Press, 2001, ISBN-10: 0684851865.

[20] B. Traboulsi, Manet with Q-Routing Protocol. Master's Thesis. Lebanese American University. 2011.

[21] J. Li, C. Blake, D. De Couto, H. Lee, and R. Morris, "Capacity of Ad Hoc Wireless Networks," Proc. of the 7th ACM International Conference on Mobile Computing and Networking, 2001, pp. 61-69.

 ISBN: 978-1-61208-183-0

Estimation of Collision Multiplicities in IEEE 802.11-based WLANs

Benoît Escrig
IRIT Laboratory
Université de Toulouse
Toulouse, France
E-mail: escrig@enseeiht.fr

***Abstract*—Estimating the collision multiplicity (CM), i.e. the number of users involved in a collision, is a key task in multi-packet reception (MPR) approaches and in collision resolution (CR) techniques. A new technique is proposed for IEEE 802.11 networks. The technique is based on recent advances in random matrix theory and rely on eigenvalue statistics. Provided that the eigenvalues of the covariance matrix of the observations are above a given threshold, signal eigenvalues can be separated from noise eigenvalues since their respective probability density functions are converging toward two different laws: a Gaussian law for the signal eigenvalues and a Tracy-Widom law for the noise eigenvalues. The proposed technique outperforms current estimation techniques in terms of underestimation rate. Moreover, this paper reveals that, contrary to what is generally assumed in current MPR techniques, a single observation of the colliding signals is far from being sufficient to perform a reliable CM estimation.**

***Index Terms*—multi-packet reception; collision multiplicity; model order selection; IEEE 802.11-based networks**

I. Introduction

The throughput of IEEE 802.11-based networks highly depends on the number of collisions. When the number of collisions increases, the throughput is degraded. Recent advances in multi-user detection (MUD) now allow the processing of collision signals, so data packets from collided user terminals can be successfully decoded at the access point even when a collision occurs. One [1], [2] or several [3], [4] transmissions from the colliding users are needed in order to achieve the decoding of the packets. The first step that is performed by these multi-packet reception (MPR) techniques consists of estimating the number of users involved in the collision: the collision multiplicity (CM). The estimation process implements a well-established Model Order Selection Technique (MOST) [5]. First, an eigenvalue decomposition is performed on the sample covariance matrix (SCM) of the observations ("snapshots"). Then, the well-known information criterion MDL (Minimum Description Length) is applied in order to perform the CM estimation. In the context of wireless local areas networks (WLANs), current CM estimation techniques (CMETs) are based on the following two assumptions: (i) the signal samples are uniformly distributed, i.e., signal samples are either PSK or QAM modulation symbols, and (ii) the number of snapshots is not much greater that the CM [1], [2], [6]. These two assumptions are rather questionable when applied to IEEE 802.11-based networks. First, the IEEE 802.11 standard relies on orthogonal frequency division multiplex (ODFM) transmissions so signal samples are not uniformly distributed but rather distributed according to a Gaussian law. Second, the assumption of a low number of observations (compared to the CM value) contradicts typical assumptions in MOSTs [7], [8].

In this paper, we propose a new CMET, denoted as TWIT (Tracy-Widom Inference Test). Simulation results show that the TWIT outperforms the classical MDL in the context of IEEE 802.11-based WLANs. Moreover, we show that the number of observations that are needed to perform the CM estimation is far much grater than the one that is used in CMETs.

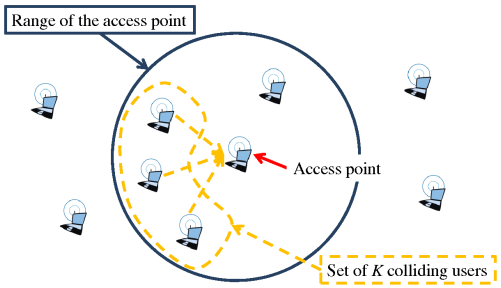


Fig. 1. Collision Scenario with $K = 3$ colliding users.

The rest of the paper is organized as follows. The system model is introduced in Section II and some results on eigenvalue statistics are stated in Section III. The CMETs are described in Section IV. Simulation results are presented in Section V and a conclusion is drawn in the last section.

II. System Model

In the proposed scenario, K users are simultaneously transmitting OFDM frames toward an AP. Each OFDM frame contains m symbols. The OFDM signal samples are Gaussian distributed and have a white power spectral density. For sake

 ISBN: 978-1-61208-183-0

of simplicity, the AP and the user terminals are assumed to be equipped with a single antenna. The AP can trigger retransmissions from the colliding users by transmitting a feedback frame. The signaling frame serves also as a synchronization flag for user transmissions. When T snapshots are available at the AP, the $T \times 1$ observation vector $\mathbf{y}_i$ is written as

$$\mathbf{y}_i = \mathbf{H}\mathbf{s}_i + \mathbf{w}_i, \qquad i = 0, 1, \ldots, m$$

where m denotes the number of samples per snapshot, $\mathbf{s}_i \sim \mathbb{C}\mathcal{N}_K(\mathbf{0}, \mathbf{R_s})$ are $K \times 1$ complex Gaussian vectors of OFDM samples with covariance matrix $\mathbf{R_s}$ and $\mathbf{w}_i \sim \mathbb{C}\mathcal{N}_T(\mathbf{0}, \mathbf{\Sigma})$ are $T \times 1$ complex Gaussian noise vectors with noise covariance matrix $\mathbf{\Sigma}$. In the white noise case, the covariance $\mathbf{\Sigma}$ is $\sigma^2\mathbf{I}_T$ where $\mathbf{I}_T$ is the $T \times T$ identity matrix. The channel matrix $\mathbf{H}$ is a $T \times K$ matrix with circularly symmetric Gaussian elements with power unity (Rayleigh fading). A block-fading wireless channel is considered here so the coefficients in $\mathbf{H}$ have constant values during an OFDM block of m samples, and then change randomly from one block to another. So the channel matrix $\mathbf{H}$ is considered as an unknown non-random matrix. When the noise covariance matrix $\mathbf{\Sigma}$ is known *a priori* and is nonsingular, the snapshots $\mathbf{y}_i$ can be "whitened" by the following transformation

$$\mathbf{y}_i^{\dagger} = \mathbf{\Sigma}^{-1/2}\mathbf{y}_i$$

where $\mathbf{\Sigma}^{-1/2}$ is the Hermitian nonnegative definite square root of $\mathbf{\Sigma}$. This transformation simply reduces to a normalization step in the case of a white Gaussian noise.

The signal and noise vectors being independent, the covariance matrix $\mathbf{R}$ of the snapshots $\mathbf{y}_i$ is given by

$$\mathbf{R} = \mathbf{H}\mathbf{R_s}\mathbf{H}^* + \mathbf{\Sigma}$$

with * denoting the complex conjugate. We assume that the matrix $\mathbf{H}$ is full rank and that the signal covariance matrix $\mathbf{R_s}$ is nonsingular. Hence the rank of $\mathbf{H}\mathbf{R_s}\mathbf{H}^*$ is $\min(K, T)$, i.e., $\mathbf{H}\mathbf{R_s}\mathbf{H}^*$ has exactly T non-zero eigenvalues when $T \leq K$ and K non-zero eigenvalues when $T > K$. When the whitening transformation is applied, the covariance matrix $\mathbf{R}^{\dagger}$ of the whitened snapshots is defined as

$$\mathbf{R}^{\dagger} = \mathbf{\Sigma}^{-1/2}\mathbf{R}\mathbf{\Sigma}^{1/2} = \mathbf{\Sigma}^{-1/2}\mathbf{H}\mathbf{R_s}\mathbf{H}^*\mathbf{\Sigma}^{1/2} + \mathbf{I}_T$$

Let $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_T$ denote the population eigenvalues of $\mathbf{R}^{\dagger}$. We have that

$$\lambda_i > 1 \quad \text{for } 1 \leq i \leq \min(K, T) \tag{1}$$

$$\lambda_i = 1 \quad \text{for } \min(K, T) < i \leq T \tag{2}$$

When $\mathbf{R}$ and $\mathbf{\Sigma}$ are known, and when the rank of $\mathbf{\Sigma}^{-1/2}\mathbf{H}\mathbf{R_s}\mathbf{H}^*$ is K, the CM estimation can be easily performed from the multiplicity of the λ_i equalling one. When $\mathbf{R}$ and $\mathbf{\Sigma}$ are unknown and have to be estimated, another approach must be used. We defined the sample covariance matrix (SCM) of the snapshots $\mathbf{y}_i$, denoted $\widehat{\mathbf{R}}$, by

$$\widehat{\mathbf{R}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{y}_i\mathbf{y}_i^*$$

and the SCM $\widehat{\mathbf{\Sigma}}$ of the noise by

$$\widehat{\mathbf{\Sigma}} = \frac{1}{N}\sum_{j=1}^{N}\mathbf{w}_j\mathbf{w}_j^*$$

where the $\mathbf{w}_j$, $1 \leq j \leq N$ are independent noise-only samples. We assume that the noise variance σ^2 can be estimated by different other means at the AP when σ^2 is the only parameter that is needed. Empty time slots can provide the N samples that are needed to compute the SCM $\widehat{\mathbf{\Sigma}}$. When the $\mathbf{y}_i$ are constituted of simultaneous transmissions from K users, we aim at estimating K based on the eigenvalues of $\widehat{\mathbf{R}}^{\dagger}$

$$\widehat{\mathbf{R}}^{\dagger} = \widehat{\mathbf{\Sigma}}^{-1/2}\widehat{\mathbf{R}}\widehat{\mathbf{\Sigma}}^{1/2}$$

Let $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq \hat{\lambda}_T$ denote the sample eigenvalues of $\widehat{\mathbf{R}}^{\dagger}$.

III. Some Results on Eigenvalue Statistics

Eigenvalue-based MOSTs rely on hypothesis tests on eigenvalues or on the computation of information criteria. In this paper, we concentrate on the population eigenvalues of $\widehat{\mathbf{R}}^{\dagger}$. According to (1) and (2), signal eigenvalues ($\hat{\lambda}_i > 1$) and noise eigenvalues ($\hat{\lambda}_i = 1$) could be separated by counting the number of eigenvalues strictly above one. Recent advances in RMT have shown that it exists a threshold below, which it is not possible to separate signal eigenvalues from noise eigenvalues. More precisely, this threshold being strictly higher than one, each signal eigenvalue between the threshold and one will be considered as being a noise eigenvalue, thus misleading typical eigenvalue-based MOSTs. Moreover, increasing the number of observations will exacerbate the phenomenon since the threshold increases with the number of observations. In this section, we present several properties on the eigenvalues and characterize the threshold issue.

Property 3.1: Let $\mathbf{X}$ be a $T \times N$ matrix constituted of N samples. The samples are drawn from a T-dimensional complex Gaussian law $\mathbb{C}\mathcal{N}_T(\mathbf{0}, \mathbf{\Sigma})$. Let $\mathbf{A} = \mathbf{X}\mathbf{X}^H$. Let $\mathbb{C}W_T(N, \mathbf{\Sigma})$ denote the T-variate complex Wishart law with N degrees of freedom. From [9], we have that

$$\mathbf{A} \sim \mathbb{C}W_T(N, \mathbf{\Sigma})$$

Property 3.2: Let $\mathbf{A} \sim \mathbb{C}W_T(N, \mathbf{\Sigma})$ be independent of $\mathbf{B} \sim \mathbb{C}W_T(m, \mathbf{\Sigma})$ where $m \geq T$. In the case of a double Wishart setting, we search for the eigenvalues θ that satisfy

$$\mathbf{A}v = \theta(\mathbf{A} + \mathbf{B})v \tag{3}$$

where v denotes the eigenvector corresponding to the eigenvalue θ. Let $\lambda_1^{[(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}]}$ be the largest eigenvalue satisfying (3). When $m, N \to \infty$ as $T \to \infty$ with $m > T$, we have that

$$\mathrm{P}[\frac{W(\lambda_1^{[(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}]}) - \mu(T, m, N)}{\sigma(T, m, N)} \leq x] \to TW_{\mathbb{C}}(x)$$

where $TW_{\mathbb{C}}(x)$ is the Tracy-Widom distribution function for complex data and $W(\theta)$ denotes the logit transformation of θ,

 ISBN: 978-1-61208-183-0

i.e., $W(\theta) = \log[\theta/(1-\theta)]$ and

$$\begin{aligned}
\beta &= \min(m,T) \\
\gamma &= N-T \\
\delta &= |m-T| \\
\mu(T,m,N) &= (\frac{u_\beta}{\tau_\beta}+\frac{u_{\beta-1}}{\tau_{\beta-1}})(\frac{1}{\tau_\beta}+\frac{1}{\tau_{\beta-1}})^{-1} \\
\sigma(T,m,N) &= 2(\frac{1}{\tau_\beta}+\frac{1}{\tau_{\beta-1}})^{-1} \\
\sin^2(\gamma_\beta/2) &= (\beta+1/2) \\
&\times (2\beta+\gamma+\delta+1)^{-1} \\
\sin^2(\phi_\beta/2) &= (\beta+\delta+1/2) \\
&\times (2\beta+\gamma+\delta+1)^{-1} \\
\tau_\beta^3 &= 16(2\beta+\gamma+\delta+1)^{-2} \\
&\times \sin^{-2}(\phi_\beta+\gamma_\beta) \\
&\times \sin^{-1}(\phi_\beta)\sin^{-1}(\gamma_\beta) \\
u_\beta &= 2\log[\tan(\frac{\phi_\beta+\gamma_\beta}{2})]
\end{aligned}$$

Property 3.2 has been proved for $\mathbf{\Sigma} = \mathbf{I}_T$ but the property can be applied to any $\mathbf{\Sigma}$ since the covariance matrix has no effect on the distribution of the eigenvalues [10].

Rewriting (3) for $\theta = \lambda_1^{[(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}]}$, we have that

$$\mathbf{A}^{-1}\mathbf{B}v = \frac{\lambda_1^{[(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}]}}{1-\lambda_1^{[(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}]}}v$$

Property 3.3: Let $\mathbf{A} \sim \mathbb{C}W_T(N, \mathbf{R_X})$ be independent of $\mathbf{B} \sim \mathbb{C}W_T(m, \mathbf{R_Y})$ where $m > T$. The largest eigenvalue of $\mathbf{A}^{-1}\mathbf{B}$, denoted $\lambda_1^{(\mathbf{A}^{-1}\mathbf{B})}$, satisfies

$$\mathrm{P}\{\frac{\log[\lambda^{(\mathbf{A}^{-1}\mathbf{B})}] - \mu(T,m,N)}{\sigma(T,m,N)} \leq x\} \to TW_{\mathbb{C}}(x)$$

when $m, N \to \infty$ as $T \to \infty$.

A. Signal-free Case

In the signal-free case, no user is transmitting, so $K = 0$. As a consequence, $\mathbf{R} = \mathbf{\Sigma}$, $\mathbf{R}^\dagger = \mathbf{I}_T$, and all the population eigenvalues λ_i are equal to 1. Moreover, when the number of observations T is fixed and when $m, N \to \infty$, the sample eigenvalues $\hat{\lambda}_i$ are symmetrically centered around the population eigenvalues λ_i for $i = 1, \ldots, T$. In the $T, m \to \infty$ asymptotic regime, the spreading of the sample eigenvalues can be characterized by the empirical distribution function (edf) [9], [11]–[13].

Property 3.4: In the signal-free case, when all the population eigenvalues λ_i are equal to 1, when $T, m \to \infty$ such that $T/m \to c \in (0, \infty)$, the limiting edf of the sample eigenvalues $\hat{\lambda}_i$ is given by

$$1/T\#\{\hat{\lambda}_i : \hat{\lambda}_i \leq x\} \to H(x)$$

where

$$\begin{aligned}
dH(x) &= \max(0, (1-\frac{1}{c}))\delta(x) \\
&+ \frac{1}{2\pi xc}\sqrt{(b-x)(x-a)}\mathbf{1}_{a,b}(x)dx
\end{aligned}$$

with $a = (1-\sqrt{c})^2$, $b = (1+\sqrt{c})^2$, and $\mathbf{1}_{a,b}(x) = 1$ when $a \leq x \leq b$.

The probability density function (pdf) $dH(x)$ is the Marčenko-Pastur density. From this property, a first characterization of the $\hat{\lambda}_1$ distribution can be inferred [9], [14].

Property 3.5: In the signal-free case, the whitened snapshots $\mathbf{y}_i^\dagger$ are $\mathcal{N}_T(\mathbf{0}, \mathbf{I}_T)$ and the largest eigenvalue $\hat{\lambda}_1$ of the SCM $\widehat{\mathbf{R}}^\dagger$ is Tracy-Widom distributed. When $T, m \to \infty$ such that $T/m \to c \in (0, \infty)$,

$$\mathrm{P}[\frac{m\hat{\lambda}_1 - \mu_{T,m}}{\sigma_{T,m}} \leq x] \to TW_{\mathbb{C}}(x)$$

where

$$\begin{aligned}
\mu_{T,m} &= (\sqrt{T}+\sqrt{m})^2 \\
\sigma_{T,m} &= (\sqrt{T}+\sqrt{m})(\frac{1}{\sqrt{T}}+\frac{1}{\sqrt{m}})^{1/3}
\end{aligned}$$

When the snapshots $\mathbf{y}_i^\dagger$ are $\mathcal{N}_T(\mathbf{0}, \sigma^2\mathbf{I}_T)$, the convergence limit of $m\hat{\lambda}_1$ is $\sigma^2(\sqrt{T}+\sqrt{m})^2$. This corresponds to the non-normalized case. Note that the convergence rate to the $TW_{\mathbb{C}}(x)$ distribution function is $\mathcal{O}(T^{-1/3})$. When the parameters m and T are not so large, which is practically the case when we want to reduce the number of observations, the convergence rate to the Tracy-Widom distribution is rather $\mathcal{O}(T^{-2/3})$ provided that the mean and standard deviation have been modified appropriately [9], [10].

From Property 3.2, we have that $N\widehat{\mathbf{\Sigma}} \sim \mathbb{C}W_T(N, \mathbf{\Sigma})$ and $m\widehat{\mathbf{R}} \sim \mathbb{C}W_T(m, \mathbf{R})$. We search for the largest eigenvalue $\hat{\lambda}_1$ that satisfies

$$\widehat{\mathbf{R}}v = \hat{\lambda}_1\widehat{\mathbf{\Sigma}}v$$

or, equivalently

$$(m\widehat{\mathbf{R}})v = \frac{m}{N}\hat{\lambda}_1(N\widehat{\mathbf{\Sigma}})v$$

So, using Property 3.3, we have that

$$\mathrm{P}\{\frac{\log(\frac{m}{N}\hat{\lambda}_1) - \mu(T,m,N)}{\sigma(T,m,N)} \leq x\} \to TW_{\mathbb{C}}(x)$$

B. Signal Bearing Case

When there are K signals and when $T \to \infty$, the limiting edf of $\widehat{\mathbf{R}}^\dagger$ still converges to a Marčenko-Pastur distribution. Moreover, the i^{th} largest eigenvalue $\hat{\lambda}_i$ converges to a limit different from that in the signal-free case if and only if the signal eigenvalue is above a certain threshold [12].

Property 3.6: In the signal bearing case, when $T \to \infty$,

$$\hat{\lambda}_i = \begin{cases} \lambda_i(1+\frac{c}{\lambda_i-1}) & \text{if } \lambda_i > (1+\sqrt{c}) \\ (1+\sqrt{c})^2 & \text{if } \lambda_i \leq (1+\sqrt{c}) \end{cases}$$

When $K \ll T$, the signal eigenvalues strictly below the threshold $(1+\sqrt{c})$ exhibit, on rescaling, fluctuations described by the Tracy-Widom distributions, i.e., the noise eigenvalues are closely approximated by the distributions obtained for the signal-free case ($K = 0$). For signal eigenvalues above the

threshold [11], the fluctuations about the asymptotic limit are Gaussian distributed

$$\mathrm{P}[\frac{\hat{\lambda}_i - \mu_i}{\sigma_i} \leq x] \rightarrow G(x)$$

where

$$\begin{aligned} \mu_i &= \lambda_i(1 + \frac{c}{\lambda_i - 1}) \\ \sigma_i &= \frac{\lambda_i}{\sqrt{m}}\sqrt{1 - \frac{c}{(\lambda_i - 1)^2}} \end{aligned}$$

and

$$G(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp{(-\frac{y^2}{2})} dy$$

In the $T, m \rightarrow \infty$ asymptotic regime with $T/m \rightarrow c \in (0, \infty)$, if signal eigenvalues are below the threshold, then reliable sample-eigenvalue-based detection is not possible. Note that adding more observations does not solve the problem since the threshold is increasing with T. Inversely, when signal eigenvalues are above the threshold, then reliable detection is possible. Note that these properties hold for large T and relatively large m. For lower T and m, the eigenvalues are more fluctuating, so the CM estimation is less reliable.

IV. Collision Multiplicity Estimation Techniques

We assume that the AP has collected T snapshots. First, the SCM of the whitened snapshots is computed. Then, the eigenvalue decomposition of the SCM is performed and the eigenvalues are sorted. The eigenvalues are processed with two CMETs: the first one is based on eigenvalue statistics and the second one relies on information criteria.

A. CMET based on eigenvalue statistics

The proposed algorithm relies on the eigenvalue statistics that have been stated in the previous section. Basically the estimate $\hat{K}_{\text{TWIT}}$ is initialized to zero and incremented by one for each iteration as long as the eigenvalue $\hat{\lambda}_{\hat{K}_{\text{TWIT}}+1}$ is not considered as being an eigenvalue of the noise subspace, i.e., as being Tracy-Widom distributed. The TWIT algorithm is detailed in Algorithm 1. The mean $\mu(x, y, z)$ and the standard deviation $\sigma(x, y, z)$ in the algorithm are defined in Property 3.2. Similar expressions can be found for the case of real-valued data. The threshold τ_α is defined as $TW_{\mathbb{C}}^{-1}(1 - \alpha)$ where α is some significance level. Note that this criterion has been originally designed for arbitrary (or colored) noise [13]. That is the reason why the algorithm uses the eigenvalues of $\widehat{\mathbf{R}}^{\dagger}$.

Algorithm 1 TWIT algorithm

```
Compute R̂†
Compute and sort the eigenvalues λ̂_i, i = 1, ..., T of R̂†
K̂_TWIT ← 0 and Test ← False
while Test = False and K̂_TWIT < T do
  μ ← μ(T − K̂_TWIT, m, N)
  σ ← σ(T − K̂_TWIT, m, N)
  Test ← {σ^{-1}[log(mλ̂_{K̂_TWIT+1}/N) − μ] < τ_α}
  if Test = False then
    K̂_TWIT ← K̂_TWIT + 1
  else
    break
  end if
end while
```

B. CMET based on information criteria

Information criteria, such as the MDL or the Akaïke's information criterion (AIC), have been originally designed in order to avoid subjective threshold settings in MOSTs [7]. The MDL has been widely used over the past two decades and is still used in current CMETs. We shall not refer to the AIC hereafter since the criterion has been proven to be inconsistent in the $m \rightarrow \infty$ sense [7]. The MDL criterion is defined as

$$\hat{K}_{\text{MDL}} = \underset{k=1,\ldots,T}{\operatorname{argmin}} \{\text{MDL}(k)\}$$

where

$$\text{MDL}(k) = -m(T-k)\log[\frac{g(k)}{a(k)}] + \frac{1}{2}k(2T-k)\log(m)$$

where

$$\begin{aligned} g(k) &= \prod_{i=k+1}^{T} \hat{\lambda}_i^{\frac{1}{T-k}} \\ a(k) &= \frac{1}{T-k}\sum_{i=k+1}^{T} \hat{\lambda}_i \end{aligned}$$

where the $\hat{\lambda}_i$ denote the eigenvalues of $\widehat{\mathbf{R}}^{\dagger}$ with $1 \leq i \leq T$. This estimator is consistent in the $m \rightarrow \infty$ sense. One of the reason why the MDL criterion has been widely used over the past two decades comes from its robustness to model mismatch, in particular when the underlying assumptions of snapshots and noise Gaussianity can be relaxed [15]–[17].

V. Simulation Results

CMETs are compared in the context of Rayleigh fading channels. User stations are transmitting OFDM signals that are built according to the IEEE 802.11 standard [18]. The signals are composed of 1024 sub-carriers ($N_{\text{sub}} = 1024$) and use BPSK modulation, the guard interval is 1/4 of the total symbol period (GI $= 1/4$). There are 48 OFDM symbols per OFDM block ($N_{\text{OFDM}} = 48$), so the total number of samples per snapshot is $m = 61440$. For sake of simplicity, we have chosen the same number for N, i.e., $N = 61440$. The performance of CMETs have been evaluated over 10000 Monte Carlo trials.

Figures 2 and 3 show the simulations results for two CMETs: the MDL criterion and the TWIT. The results have been obtained for a fixed number of signals $K = 4$, a variable number of snapshots $4 \leq T \leq 32$, and two typical values of the signal to noise ratio SNR: a low value (10 dB) and a high value (30 dB). The significance level α for the TWIT is set to 0.01 [13]. The first figure shows the estimated values of K. The second figure shows the underestimation rate, i.e.,

 ISBN: 978-1-61208-183-0

$P[\hat{K} < K]$. The TWIT outperforms the MDL criterion since it provides similar results with less observations, and so for any value of SNR. Another important result is that T should be significantly larger than K in order to achieve relevant performance levels, i.e. underestimation rate much lower than 10 %.

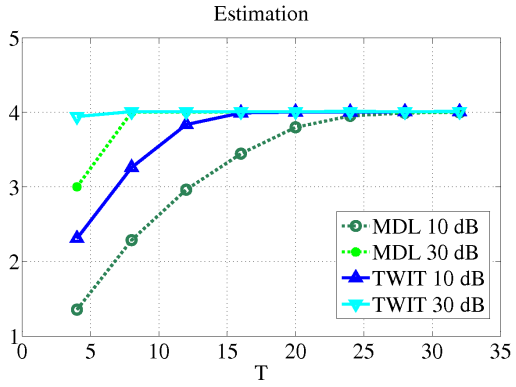


Fig. 2. Estimates of $K = 4$ for a variable number of snapshots T, $4 \leq T \leq 32$.

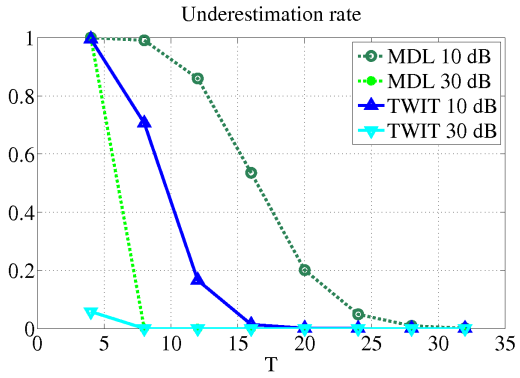


Fig. 3. Underestimation rate for $K = 4$ and a variable number of snapshots T, $4 \leq T \leq 32$.

Figures 4 to 6 show the pdfs of the eigenvalues $\hat{\lambda}_K$ and $\hat{\lambda}_{K+1}$ for $K = 4$. These statistics have been obtained with 10000 draws. The pdf of $\hat{\lambda}_K$ represents the pdf of the lowest "signal" eigenvalue whereas the pdf of $\hat{\lambda}_{K+1}$ represents the pdf of the largest "noise" eigenvalue.

Figures 4 and 5 show the eigenvalue statistics for $T = 10$ and two values of the signal to noise ratio, 10 and 20 dB. On these two figures, the lowest signal eigenvalue is always higher than the detectability threshold so all the signal eigenvalues are detectable. However, the pdfs of the signal and the noise eigenvalues are overlapping for $SNR = 10$ dB. That explains the degradation on the underestimation rate on Fig. 3.

Figures 4 and 6 show the eigenvalue statistics for $SNR = 10$ dB and two values for T: 10 and 30. The spreading of the pdfs depends on T. The higher is T, the shaper are the density curves. Here again, the performance of the CMET is improving when the number of observations is increasing.

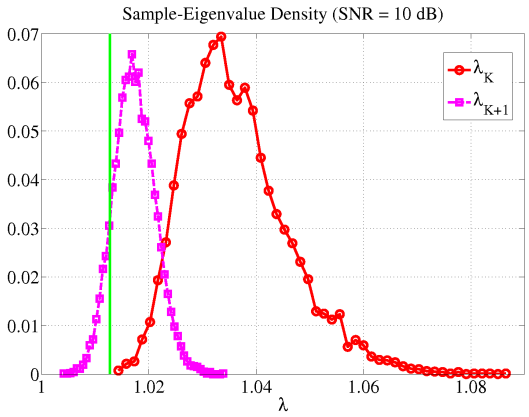


Fig. 4. Estimated probability density functions of the eigenvalues $\hat{\lambda}_K$ and $\hat{\lambda}_{K+1}$ for $K = 4$, $T = 10$ and $SNR = 10$ dB, given that $\hat{\lambda}_K \sim \widehat{G}(\lambda)$ and $\hat{\lambda}_{K+1} \sim \widehat{TW}(\lambda)$. The vertical line depicts the position of the detectability threshold $1 + \sqrt{c}$.

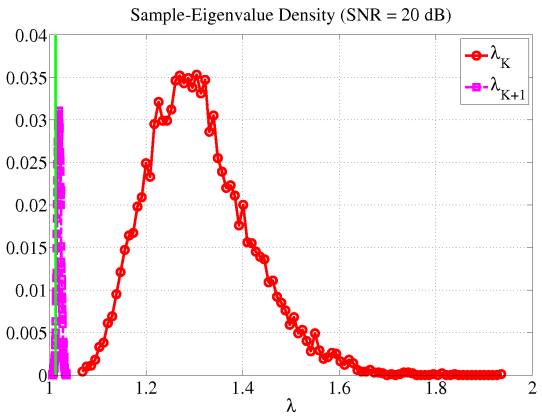


Fig. 5. Estimated probability density functions of the eigenvalues $\hat{\lambda}_K$ and $\hat{\lambda}_{K+1}$ for $K = 4$, $T = 10$ and $SNR = 20$ dB, given that $\hat{\lambda}_K \sim \widehat{G}(\lambda)$ and $\hat{\lambda}_{K+1} \sim \widehat{TW}(\lambda)$. The vertical line depicts the position of the detectability threshold $1 + \sqrt{c}$.

A. Discussion

The simulation results suggest that the CMETs perform well when the number of snapshots is much larger than the number of signals. However, in many current CMETs, the number of snapshots is set to a value close to K. In [3], [4], T is set to a value not greater than $K + 1$ or $K + 2$. More surprisingly, in [1], [2], a single transmission of the colliding users is required to proceed to the user separation, using the blind separation technique in [6]. Note that, in [6], the number of sources (users) is assumed to be known or to have been estimated using information criteria such as the MDL criterion. The simulation results presented in this paper are in stark contrast with the settings that are used in these papers. Note also that typical MOSTs rely on similar settings, i.e., $T \gg K$ (see [12] and the reference therein).

 ISBN: 978-1-61208-183-0

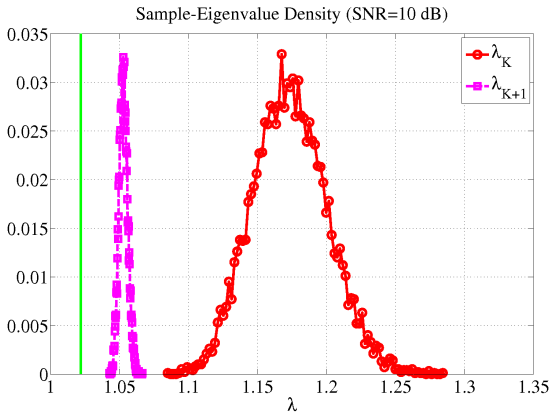


Fig. 6. Estimated probability density functions of the eigenvalues $\hat{\lambda}_K$ and $\hat{\lambda}_{K+1}$ for $K = 4$, $T = 30$ and $SNR = 10$ dB, given that $\hat{\lambda}_K \sim \widehat{G}(\lambda)$ and $\hat{\lambda}_{K+1} \sim \widehat{TW}(\lambda)$. The vertical line depicts the position of the detectability threshold $1 + \sqrt{c}$.

VI. Conclusion

In this paper, a new CMET, denoted TWIT, has been proposed. The method is based on eigenvalue statistics. Eigenvalues are tested in descending order, from the largest to the lowest. The first eigenvalue $\hat{\lambda}_q$ that is considered as being Tracy-Widom distributed allows the CM estimation by $\hat{K} = q - 1$. This CMET has been shown to outperform the typical MDL criterion. Moreover, simulation results have shown that a large number of snapshots T is needed in order to allow a good estimation of K in terms of underestimation rates. Furthermore, the number of snapshots must be significantly higher than the number of colliding users K ($T \gg K$). These settings are similar to the settings that are used in MOSTs for signal array processing.

The impact of these results is twofold. First, some CR techniques such as the network-assisted diversity multiple access (NDMA) [3], [4] cannot be implemented in IEEE 802.11 networks notably because these CR techniques are based on the assumption that T can be made as small as $K + 1$ or $K + 2$. Second, some MPR protocols for IEEE 802.11 networks that use the blind user separation in [6] appear to be rather questionable since they assume that the CM estimation can rely on a single observation of collided request-to-send (RTS) frames. Even if the AP is equipped with four antennas ($T = 4$), our simulation results have shown that the receiver at the AP needs many more snapshots in order to provide a good estimation of K. This paper has pointed out a strong constraint in the design of MPR techniques. It revealed that a single observation of the colliding signals is far from providing enough information to estimated the number of colliding nodes.

Further investigations are now needed in order to fully characterized the performance of the proposed CMETs in typical operating conditions. The obtained results will allow the implementation of these CMETs in current or future standards.

References

[1] P. X. Zheng, Y. J. Zhang, and S. C. Liew, "Multipacket Reception in Wireless Local Area Networks," in *Proc. IEEE International Conference on Communications (ICC)*, 2006.

[2] W. L. Huang, K. B. Letaief, and Y. J. Zhang, "Cross-Layer Multi-Packet Reception Based Medium Access Control and Resource Allocation for Space-Time Coded MIMO/OFDM," *IEEE Transactions on Wireless Communications*, vol. 7, no. 9, pp. 3372–3384, 2008.

[3] R. Zhang, N. D. Sidiropoulos, and M. K. Tsatsanis, "Collision Resolution in Packet Radio Networks Using Rotational Invariance Techniques," *IEEE Transactions on Communications*, vol. 50, no. 1, pp. 146–155, 2002.

[4] B. Özgül and H. Deliç, "Wireless Access with Blind Collision-Multiplicity Detection and Retransmission Diversity for Quasi-Static Channels," *IEEE Transactions on Communications*, vol. 54, no. 5, pp. 858–867, 2006.

[5] P. Stoica and Y. Selèn, "Model-Order Selection : A review of information criterion rules," *IEEE Signal Processing Magazine*, vol. 21, no. 4, pp. 36–47, 2004.

[6] S. Talwar, M. Viberg, and A. Paulraj, "Blind Separation of Synchronous Co-Channel Digital Signals Using an Antenna Array. Part I. Algorithms," *IEEE Transactions on Signal Processing*, vol. 44, no. 5, pp. 1184–1197, 1995.

[7] M. Wax and T. Kailath, "Detection of Signals by Information Theoretic Criteria," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 387–392, 1985.

[8] C. Xu and S. Kay, "Source Enumeration via the EEF Criterion," *IEEE Signal Processing Letters*, vol. 15, pp. 569–572, 2008.

[9] I. M. Johnstone, "High Dimensional Statistical Inference and Random Matrices," in *Proceeding of the International Congress of Mathematicians*, 2006.

[10] ——, "Multivariate analysis and Jacobi ensembles: Largest eigenvalue, TracyWidom limits and rates of convergence," vol. 36, no. 6, pp. 2638–2716, 2008.

[11] J. Baik, G. B. Arous, and S. Péché, "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices," *The Annals of Probability*, vol. 33, no. 5, pp. 1643–1697, 2005.

[12] R. R. Nadakuditi and A. Edelman, "Sample Eigenvalue Based Detection of High-Dimensional Signals in White Noise Using Relatively Few Samples," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 2625–2638, 2008.

[13] R. R. Nadakuditi and J. W. Silverstein, "Fundamental Limit of Sample Generalized Eigenvalue Based Detection of Signals in Noise Using Relatively Few Signal-Bearing and Noise-Only Samples," *IEEE Transactions on Signal Processing*, vol. 4, no. 3, pp. 468–480, 2010.

[14] P. O. Perry and P. J. Wolfe, "Minimax Rank Estimation for Subspace Tracking," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 3, pp. 504–513, 2010.

[15] G. Xu, R. H. Roy, and T. Kailath, "Detection of number of sources via exploitation of centro-symmetry property," *IEEE Transactions on Signal Processing*, vol. 42, no. 1, pp. 102–112, january 1994.

[16] P. Stoica and M. Cedervall, "Detection tests for array procesing in unknown correlated noise fields," *IEEE Transactions on Signal Processing*, vol. 45, no. 9, september 1997.

[17] E. Fishler and H. V. Poor, "Estimation of the number of sources in unbalanced arrays via information theoretic criteria," *IEEE Transactions on Signal Processing*, vol. 53, no. 9, pp. 3543–3553, september 2005.

[18] "IEEE 802.11n standard, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: Amendment 5: Enhancements for Higher Throughput," IEEE Computer Society, Tech. Rep., 2009.

 ISBN: 978-1-61208-183-0

CTC Turbo Decoding Architecture for LTE Systems Implemented on FPGA

Cristian Anghel, Valentin Stanciu, Cristian Stanciu, and Constantin Paleologu
Telecommunications Department
University Politehnica of Bucharest
Romania
canghel@comm.pub.ro, svl117@yahoo.com, cristian@comm.pub.ro, pale@comm.pub.ro

***Abstract*— This paper describes a turbo decoder for Long Term Evolution (LTE) standard, release 8, using a Max Log MAP algorithm. The Forward Error Correction (FEC) block dimensions, as indicated in the standard, are inside a range of 40 to 6144 bits. The coding rate is 1/3, the puncturing block not being taken into discussion here. The number of turbo iterations is variable, but in this study it was usually set to 3. The turbo decoder is implemented on a Xilinx Virtex-5 XC5VFX70T Field Programmable Gate Array (FPGA).**

Keywords- turbo codes; Max Log MAP decoder; FPGA implementation; LTE standard.

I. INTRODUCTION

The discussions around the channel coding theory were intense in the last decades, but even more interest around this topic was added once the turbo codes were found by Berrou, Glavieux, and Thitimajshima [1][2][3].

At the beginning of their life, after proving the obtained decoding performances, the turbo codes were introduced in different standards as recommendations, while convolutional codes were still mandatory. The reason behind this decision was especially the high complexity of turbo decoder implementation. But the turbo codes became more attractive once the supports for digital processing, like Digital Signal Processor (DSP) or Field Programmable Gate Array (FPGA), were extended more and more in terms of processing capacity. Today the chips include dedicated hardware accelerators for different types of turbo decoders, but this approach makes them standard dependent.

The Third-Generation Partnership Project (3GPP) [4] is an organization, which adopted early these advanced coding techniques. Turbo codes were standardized from the first version of Universal Mobile Telecommunications System (UMTS) technology, in 1999. The next UMTS releases (after High Speed Packet Access was introduced) added support for new and interesting features, while turbo coding remained still unchanged. Some modifications were introduced by the Long Term Evolution (LTE) standard [5][6], not significant as volume, but important as concept. While keeping exactly the same coding structure as in UMTS, 3GPP proposed for LTE a new interleaver scheme.

Valenti and Sun presented in [7] a UMTS dedicated turbo decoding scheme. Due to the new LTE interleaver, the decoding performances are improved compared with the ones corresponding to UMTS standard. Moreover, the new LTE interleaver provides support for the parallelization of the decoding process inside the algorithm, taking advantage on the main principle introduced by turbo decoding, i.e., the usage of extrinsic values from one turbo iteration to another.

This paper presents an efficient solution for the hardware implementation of a Convolutional Turbo Code (CTC) LTE decoder. The optimization indicators refer to the used logic area and to the obtained decoding speed. Also the level of performances degradation introduced by the finite precision representation is taken into account when selecting the final implementation solution.

The paper is organized as follows. Section II describes the LTE coding scheme with the new introduced interleaver. Section III presents the decoding algorithm. In Section IV, the implementation solutions and the proposed decoding scheme are discussed. Section V presents area and speed results obtained when targeting a XC5VFX70T [8] chip on Xilinx ML507 [9] board; it also provides simulation curves comparing the results obtained when varying the most important decoding parameters. Section VI presents the final conclusions and the future perspective of this study.

II. LTE CODING SCHEME

The coding scheme presented in 3GPP LTE specification is a classic turbo coding scheme, including two constituent encoders and one interleaver module. It is described in Fig. 1. One can observe at the input of the LTE turbo encoder the data block C_k. The K bits corresponding to this block are sent as systematic bits at the output in the steam X_k. In the same time, the data block is processed by the first constituent encoder resulting parity bits Z_k, while the interleaved data block C'_k is processed by the second constituent encoder resulting parity bits Z'_k. Combining the systematic bits and the two streams of parity bits, the following sequence is obtained at the output of the encoder: $X_1, Z_1, Z'_1, X_2, Z_2, Z'_2, \ldots, X_k, Z_k, Z'_k$.

At the end of the coding process, in order to drive back the constituent encoders to the initial state, the switches from Fig. 1 are moved from position A to B. Since the final states of the two constituent encoders are different, depending on the input data block, this switching procedure will generate tail bits for each encoder. These tail bits have to be transmitted together with the systematic and parity bits resulting the following final sequence: $X_{k+1}, Z_{k+1}, X_{k+2}, Z_{k+2}, X_{k+3}, Z_{k+3}, X'_{k+1}, Z'_{k+1}, X'_{k+2}, Z'_{k+2}, X'_{k+3}, Z'_{k+3}$.

 ISBN: 978-1-61208-183-0

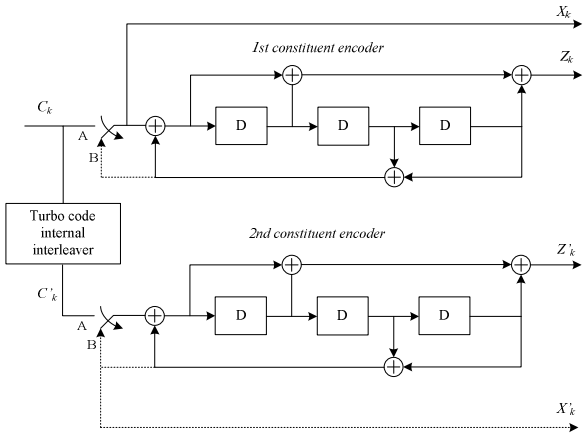


Figure 1. LTE CTC encoder.

As mentioned before, the novelty introduced by the LTE standard in terms of turbo coding is the interleaver module. The output bits are reorganized using

$$C_i^{'} = C_{\pi(i)}, \ i = 1, 2, ..., K, \tag{1}$$

where the interliving function π applied over the output index i is defined as

$$\pi(i) = (f_1 \cdot i + f_2 \cdot i^2) \bmod K. \tag{2}$$

The length K of the input data block and the parameters f_1 and f_2 are provided in Table 5.1.3-3 in [5].

III. DECODING ALGORITHM

The LTE turbo decoding scheme is depicted in Fig. 2. The two Recursive Systematic Convolutional (RSC) decoders are using in theory the Maximum A Posteriori (MAP) algorithm. This classic algorithm provides the best decoding performances, but it suffers from very high implementation complexity and it can lead to large dynamic range for its variables. For these reasons the MAP algorithm is used as a reference for targeted decoding performances, while for real implementation new sub-optimal algorithms have been studied: Logarithmic MAP (Log MAP) [10], Maximum Log MAP (Max Log MAP), Constant Log MAP (Const Log MAP) [11], and Linear Log MAP (Lin Log MAP) [12].

For the proposed decoding scheme, the Max Log MAP algorithm is selected. This algorithm reduces the implementation complexity and controls the dynamic range problem with the cost of acceptable performances degradation, compared to classic MAP algorithm. The Max Log MAP algorithm keeps from Jacobi logarithm only the first term, i.e.,

$$\begin{aligned} &\max^*(x, y) = \ln(e^x + e^y) = \\ &\max(x, y) + \ln(1 + e^{-|y-x|}) \approx \max(x, y). \end{aligned} \tag{3}$$

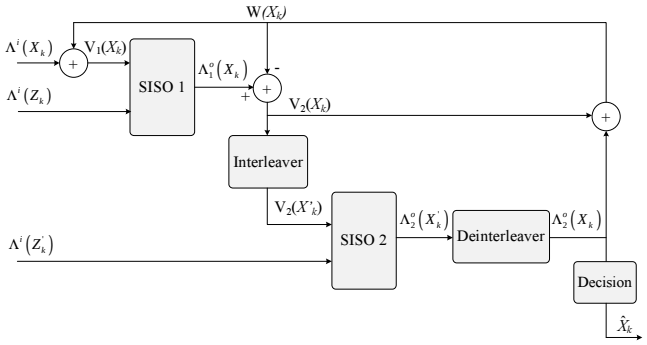


Figure 2. LTE turbo decoder.

The LTE turbo decoder trellis diagram contains 8 states. Each diagram state permits 2 inputs and 2 outputs. The branch metric between the states S_i and S_j is

$$\gamma_{ij} = V(X_k) X(i, j) + \Lambda^i(Z_k) Z(i, j), \tag{4}$$

where $X(i,j)$ represents the data bit and $Z(i,j)$ is the parity bit, both associated to one branch. Also $\Lambda^i(Z_k)$ is the Log Likelihood Ratio (LLR) for the input parity bit. When Soft Input Soft Output (SISO) 1 decoder is taken into discussion this input LLR is $\Lambda^i(Z_k)$, while for SISO 2 it becomes $\Lambda^i(Z_k^{'})$; $V(X_k)=V_1(X_k)$ represents the sum between $\Lambda^i(X_k)$ and $W(X_k)$ for SISO 1 and $V(X_k)=V_2(X'_k)$ represents the interleaved version of the difference between $\Lambda_1^o(X_k)$ and $W(X_k)$ for SISO 2. In Fig. 2, $W(X_k)$ is the *extrinsic information* and $\Lambda_1^o(X_k)$ and $\Lambda_2^o(X_k^{'})$ are the output LLRs generated by the two SISOs.

In the LTE turbo encoder case, there are 4 possible values for the branch metrics between 2 states in the trellis:

$$\begin{aligned} &\gamma_0 = 0 \\ &\gamma_1 = V(X_k) \\ &\gamma_2 = \Lambda^i(Z_k) \\ &\gamma_2 = V(X_k) + \Lambda^i(Z_k). \end{aligned} \tag{5}$$

The decoding process is based on going forward and backward through the trellis.

A. Backward recursion

The trellis is covered backward and the computed metrics are stored in a normalized form at each node of the trellis. These stored values are used for the LLR computation at the trellis forward recursion. The backward metric for the S_i state at the k^{th} stage is $\beta_k(S_i)$, where $2 \le k \le K+3$ and $0 \le i \le 7$. The backward recursion is initialized with $\beta_{K+3}(S_0) = 0$ and $\beta_{K+3}(S_i) = 0, \forall i > 0$.

Starting from the stage k=K+2 and continuing through the trellis until stage k=2, the computed backward metrics are

$$\hat{\beta}_k(S_i) = \max\left\{(\beta_{k+1}(S_{j1}) + \gamma_{ij1}), (\beta_{k+1}(S_{j2}) + \gamma_{ij2})\right\}, \quad (6)$$

where $\hat{\beta}_k(S_i)$ represents the un-normalized metric and S_{j1} and S_{j2} are the two states from stage k+1 connected to the state S_i from stage k. After the computation of $\hat{\beta}_k(S_0)$ value, the rest of the backward metrics are normalized as

$$\beta_k(S_i) = \hat{\beta}_k(S_i) - \hat{\beta}_k(S_0) \quad (7)$$

and then stored in the dedicated memory.

B. Forward recursion

During the forward recursion, the trellis is covered in the normal direction, this process being similar with the one specific for Viterbi algorithm. Now only the forward metrics from the last stage (k-1) have to be stored, in order to allow the computation of the current stage (k) metrics. The forward metric for the state S_i at the stage k is $\alpha_k(S_i)$ with $0 \le k \le K-1$ and $0 \le i \le 7$. The forward recursion is initialized with $\alpha_0(S_0) = 0$ and $\alpha_0(S_i) = 0, \forall i > 0$. Starting from the stage k=1 and continuing through the trellis until the last stage k=K, the un-normalized forward metrics are given by

$$\hat{\alpha}_k(S_j) = \max\left\{(\alpha_{k-1}(S_{i1}) + \gamma_{i1j}), (\alpha_{k-1}(S_{i2}) + \gamma_{i2j})\right\}, \quad (8)$$

where S_{i1} and S_{i2} are the two states from stage k-1 connected to the state S_j from stage k. After the computation of $\hat{\alpha}_k(S_0)$ value, the rest of the forward metrics are normalized as

$$\alpha_k(S_i) = \hat{\alpha}_k(S_i) - \hat{\alpha}_k(S_0). \quad (9)$$

Because the forward metrics α are computed for the stage k, the decoding algorithm can obtain in the same time a LLR estimated for the data bits X_k. This LLR is found the first time by considering that the likelihood of the connection between the state S_i at k-1 stage and the state S_j at k stage is

$$\lambda_k(i,j) = \alpha_{k-1}(S_i) + \gamma_{ij} + \beta_k(S_j). \quad (10)$$

The likelihood of having a bit equal to 1 (or 0) is when the Jacobi logarithm of all the branch likelihoods corresponds to 1 (or 0) and thus:

$$\Lambda^o(X_k) = \max_{(S_i \to S_j): X_i = 1} \{\lambda_k(i,j)\} - \max_{(S_i \to S_j): X_i = 0} \{\lambda_k(i,j)\}, \quad (11)$$

where "max" operator is recursively computed over the branches, which have at the input a bit of 1 $\{(S_i \to S_j): X_i = 1\}$ or a bit of 0 $\{(S_i \to S_j): X_i = 0\}$.

IV. Proposed Decoding Scheme

A. Block Scheme

Since one constituent decoder extrinsic outputs are inputs for the other, and because the interleaving or deinterleaving procedure is applied over data blocks, the operating periods for the two constituent decoders are not overlapped. Thus, the decoding scheme can use a single constituent decoder, which operates time-multiplexed. The proposed scheme is depicted in Fig. 3 and it is based on the previous work presented in [13] for a WiMAX CTC decoder. The memory blocks are used for storing data from one semi-iteration to another and from one iteration to another. SISO 1 reads the memory locations corresponding to $V_1(X_k)$ and $\Lambda^i(Z_k)$ vectors. The reading process is performed forward and backward and it serves the first semi-iteration. At the end of this process, SISO 2 reads forward and backward from the memory blocks corresponding to $V_2(X'_k)$ and $\Lambda^i(Z'_k)$ vectors in order to perform the second semi-iteration.

Vector $V_1(X_k)$ is obtained by adding the input vector $\Lambda^i(X_k)$ with the extrinsic information vector $W(X_k)$. After having the input data ready, SISO 1 starts the decoding process. At the output, the LLRs are available sequentially, at 8 clock periods distance. Performing the subtraction between these LLRs and the extrinsic values $W(X_k)$, the vector $V_2(X_k)$ is computed and then stored into its corresponding memory. The interleaving process is started and the re-ordered LLRs $V_2(X'_k)$ are stored in their memory, where the corresponding values for the 3 tail bits X'_{k+1}, X'_{k+2}, X'_{k+3} are also added on the last memory locations. The second semi-iteration can start at this point. The same SISO unit is used, but reading this time data inputs from the other memory blocks. As one can see from Fig. 3, two switching mechanisms are included in the scheme. When in position 1, the memory blocks for $V_1(X_k)$ and $\Lambda^i(Z_k)$ are used, while in position 2 the memory blocks for $V_2(X'_k)$ and $\Lambda^i(Z'_k)$ become active.

At the output of the SISO unit, after each semi-iteration, K LLRs are obtained. The ones corresponding to the second semi-iteration are stored in the $\Lambda_2^o(X'_k)$ memory, then they are deinterleaved and finally they are stored in the $\Lambda_2^o(X_k)$ memory. Subtracting from these deinterleaved LLRs the values of $V_2(X_k)$ vector, the extrinsic information $W(X_k)$ is obtained. Also, if the decoder performs the last

 ISBN: 978-1-61208-183-0

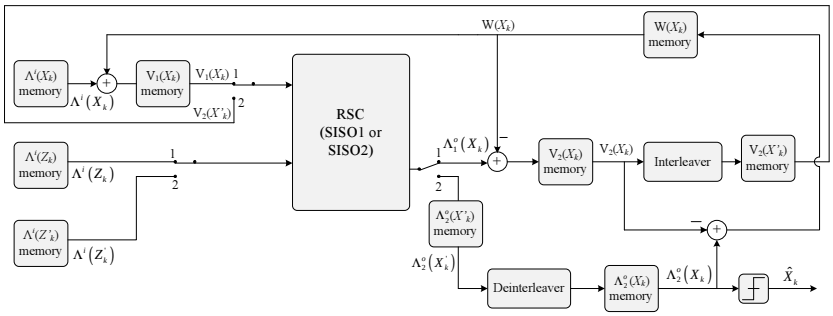


Figure 3. Proposed turbo decoder block scheme.

second semi-iteration, the hard decision is made over these deinterleaved LLRs, resulting this way the decoded bits.

In order to be able to handle all the data block dimensions, the used memory blocks have 6144 locations (this is the maximum data block length), except the ones storing the input data for RSCs, which have 6144 + 3 locations, including here also the tail bits. Each memory locations is 10 bits wide, the first bit being used for the sign, the next 6 bits representing the integer part and the last 3 bits indicating the fractional part. This format was decided studying the dynamic range of the variables (for the integer part) and the variations of the decoding performances (for the fractional part).

B. The Interleaver

The interleaver module is used both for interleaving and deinterleaving. The interleaved index is obtained based on a modified form of (2), i.e.,

$$\pi(i) = \{[(f_1 + f_2 \cdot i) \bmod K] \cdot i\} \bmod K. \qquad (12)$$

In order to obtain both functions, either the input data is stored in the memory in natural order and then it is read in interleaved order, either the input data is stored in the interleaved order and then it is read in natural order. Fig. 4 depicts the implementation solution for this module.

As one can observe from Fig. 4, the interleaved index computation is performed in three steps. First the value for $(f_1 + f_2 \cdot i) \bmod K$ is computed. This partial result is multiplied by natural order index *i* and then a new modulo *K* function is applied. In the first stage of this process, the remark that the formula is increased with f_2 for consecutive values of index *i* is used. This way, a register value is increased with f_2 at each new index *i*. If the resulted value is bigger than *K*, the value of *K* is subtracted from the register value. This processing is one clock period long, this being the reason why data is generated in a continuous manner.

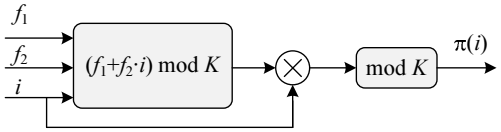


Figure 4. Proposed interleaver logic scheme.

In the second stage, a pipe-line multiplier is used for obtaining the result of the multiplication between index *i* and the first stage resulted value. The product result is obtained after 13 clock periods and it is 26 bits wide. In the third stage this result is compared with values 2^nK, with *n* between 13 and 0. Less subtraction for computing modulo *K* function are performed this way, the total number of clock periods being reduced from 6124 to 13. At the end of this third stage the interleaved indexes are obtained.

C. The SISO module

The internal SISO scheme is presented in Fig. 5. One can notice both the un-normalized metric computing blocks ALPHA (forward) and BETA (backward), and the transition metric computing block GAMMA, which in addition includes the normalization function (subtract the metrics for the first state from all the other metrics). The L block computes the output LLRs, which are normalized by the NORM block. The MUX-MAX block selects inputs corresponding to the forward or backward recursion and computes the maximum function. The MEM BETA block stores the backward metrics, which are computed before forward metrics. The metric normalization is required to preserve the dynamic range. Without normalization, the forward and backward metric width should be wider in order to avoid saturation, which means more memory blocks, more complex arithmetic (i.e., more used resources), and lower frequency (as an overall consequence). Hence, reducing the logic levels by eliminating the normalizing procedure does not increase the system performances.

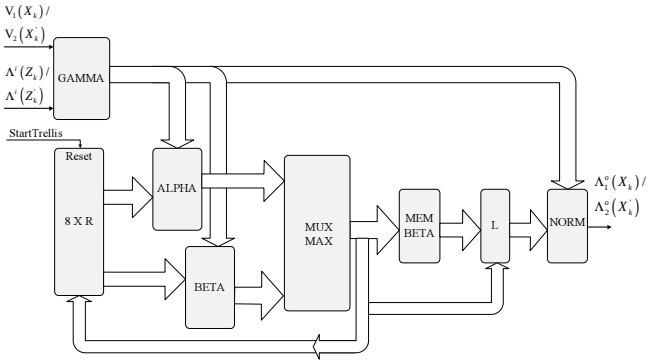


Figure 5. Proposed SISO block scheme.

The ALPHA, BETA, and GAMMA blocks are implemented in a dedicated way. Each metric corresponding to each state is computed separately, not using the same function with different input parameters.

Consequently, 16 equations should be used for transition metric computation (2 possible transitions for each of the 8 states from a stage). In fact, only 4 equations are needed [as indicated in (5)]; moreover, from these 4 equations one of them leads to zero value, so that the computational effort is minimized for this implementation solution.

V. IMPLEMENTATION RESULTS

A. Performances

The used hardware programming language is Very High Speed Hardware Description Language (VHDL). For the generation of RAM/ ROM memory blocks Xilinx Core Generator 11.1 was used. The simulations were performed with ModelSIM 6.5. The synthesis process was done using Xilinx XST from Xilinx ISE 11.1. Using these tools, the obtained system frequency when implementing the decoding structure on a Xilinx XC5VFX70T-FFG1136 chip is around 210 MHz. The occupied area is around 1000 (8.92%) slices from a total of 11200, while the used 18Kb memory blocks number is 32 from a total of 296.

B. Simulations

The following performance curves were obtained using a finite precision Matlab simulator. This approach was selected because the Matlab simulator produces exactly the same outputs as the ModelSIM simulator, while the simulation time is smaller.

All the simulation results are using the Max Log MAP algorithm, and the results are presented for different types of decoding parameters variations. All pictures describe the Bit Error Rate (BER) versus Signal-to-Noise Ratio (SNR) expressed as the ratio between the energy per bit and the noise power spectral density.

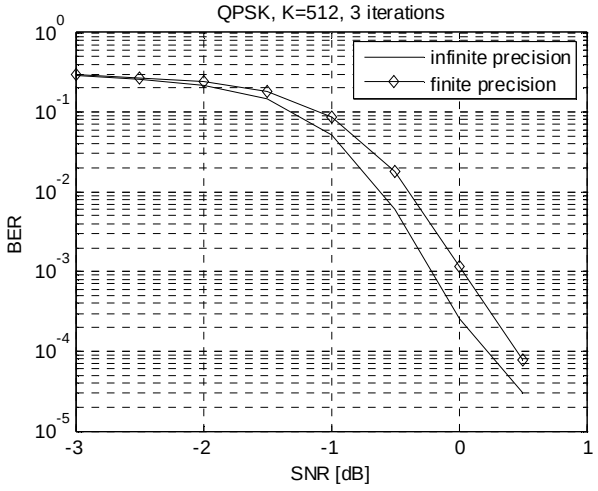


Figure 6. Finite precision vs. infinite precision.

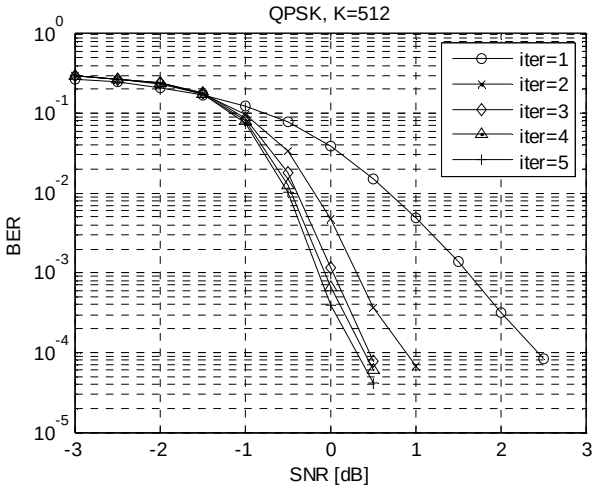


Figure 7. Decoding performances vs. number of iterations.

Fig. 6 depicts the obtained performances when executing the decoding process of the same input data, in infinite precision and in finite precision. For finite precision, as mentioned before, a 10 bit format was used, one bit for the sign, 6 bits for the integer part and 3 bits for the fractional part. In these simulations, K=512 bits, the used modulation is QPSK, and the number of turbo iterations is set to 3.

Fig. 7 depicts the performances improvement when the number of turbo iterations is increased. One can observe that after a certain number of turbo iterations the decoding improvement is not significant anymore and thus the added decoded latency is not justified. In these simulations, K=512 bits, the used modulation is QPSK, and the number of turbo iterations is increased from 1 to 5.

Finally, Fig. 8 describes the decoding performances improvement when the data block size increases. For these simulations the used modulation is QPSK, the number of turbo iterations is 3, and the data block lengths are K=40,

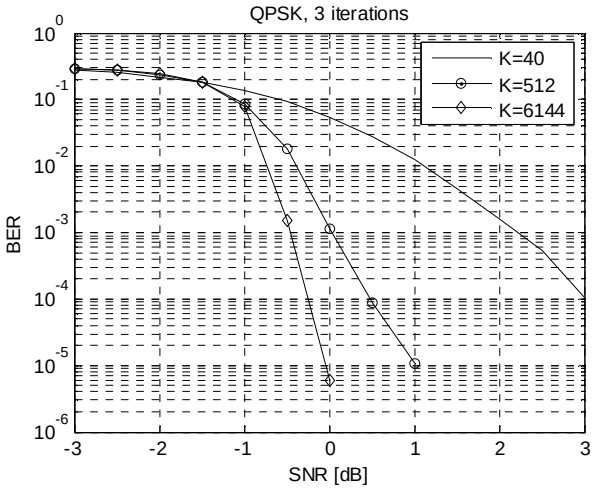


Figure 8. Decoding performances vs. block dimension.

 ISBN: 978-1-61208-183-0

K=512, and *K*=6144. One can observe an improvement of about 1.8 dB at BER = 10^{-2} between the smallest and the biggest block size defined by standard (*K*=40 and *K*=6144).

VI. CONCLUSIONS AND FUTURE WORKS

The most important aspects regarding the FPGA implementation of a CTC decoder for LTE systems were presented in this paper. Area and speed optimization solutions have been proposed based on the specific decoding scheme. A very efficient method of increasing the clock frequency was proposed, i.e., the normalization operation from the ALPHA/BETA updating loop was removed from that loop and distributed into the GAMMA block and also into the LLR computing block. Simulation and implementation results were given for different data block sizes and for different number of turbo iterations.

The perspective for a future work is to implement a stop criterion in order to reduce the decoding latency. A possible solution is the stop the decoding iterations when some indicators are not changing from one iteration to another.

ACKNOWLEDGMENTS

This work was supported under the Grant UEFISCDI PN-II-RU-TE no. 7/5.08.2010.

REFERENCES

[1] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon Limit Error-Correcting Coding and Decoding: Turbo Codes," *IEEE Proceedings of the Int. Conf. on Communications*, Geneva, Switzerland, pp. 1064-1070, May 1993.

[2] C. Berrou and A. Glavieux, "Near Optimum Error Correcting Coding and Decoding: Turbo-Codes," *IEEE Trans. Communications*, vol. 44, no. 10, pp. 1261-1271, Oct. 1996.

[3] C. Berrou and M. Jézéquel, "Non binary convolutional codes for turbo coding," *Electronics Letters*, vol. 35, no. 1, pp. 9-40, Jan. 1999.

[4] Third Generation Partnership Project. 3GPP home page. www.3gpp.org, last accessed on November 2011.

[5] 3GPP TS 36.212 V8.7.0 (2009-05) Technical Specification, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding (Release 8)."

[6] F. Khan, *LTE for 4G Mobile Broadband*, Cambridge University Press, New York, 2009.

[7] M. C. Valenti and J. Sun, "The UMTS Turbo Code and an Efficient Decoder Implementation Suitable for Software-Defined Radios," *International Journal of Wireless Information Networks*, Vol. 8, No. 4, pp. 203-216, October 2001.

[8] "Xilinx Virtex 5 family user guide," retrieved from www.xilinx.com on January 2011.

[9] "Xilinx ML507 evaluation platform user guide," retrieved from www.xilinx.com on January 2011.

[10] P. Robertson, E. Villebrun, and P. Hoeher, "A Comparison of Optimal and Sub-Optimal MAP Decoding Algorithms Operating in the Log Domain," *Proc. IEEE International Conference on Communications* (ICC'95), Seattle, pp. 1009-1013, June 1995.

[11] S. Papaharalabos, P. Sweeney, and B. G. Evans, "Constant log-MAP decoding algorithm for duo-binary turbo codes," *Electronics Letters Volume 42*, Issue 12, pp. 709 – 710, June 2006.

[12] J. F. Cheng and T. Ottosson, "Linearly approximated log-MAP algorithms for turbo decoding," *Vehicular Technology Conference Proceedings*, 2000. VTC 2000-Spring Tokyo. 2000 IEEE 51st Volume 3, pp. 2252 – 2256, May 2000.

[13] C. Anghel, A. A. Enescu, C. Paleologu, and S. Ciochina, "CTC Turbo Decoding Architecture for H-ARQ Capable WiMAX Systems Implemented on FPGA," "*Ninth International Conference on Networks*" ICN 2010, Menuires, France, April 2010.

 ISBN: 978-1-61208-183-0

Multi-Relay Cooperative NB-LDPC Coding with Non-Binary Repetition Codes

David Declercq
ETIS ENSEA/UCP/CNRS UMR 8051
95000 Cergy-Pontoise, France
email: declercq@ensea.fr

Valentin Savin
CEA-LETI, MINATEC
38054 Grenoble, France
email: valentin.savin@cea.fr

Stephan Pfletschinger
Centre Tecnolgic de Telecom. de Catalunya
7, Av. Carl Friedrich Gauss, 08860 Castelldefels, Spain
email: stephan.pfletschinger@cttc.es

***Abstract*—In this paper, we propose a system based on non-binary Low-Density Parity-Check (LDPC) codes to communicate efficiently over the multiple-relay fading channels, with a simple joint decoding strategy at the receiver end. The particularity of our approach is to rely on non-binary LDPC codes at the source, coupled with multiplicative non-binary local codes at the relays, such that the joint decoding complexity is not increased compared to a system without relays, while preserving the coding gain brought by the re-encoding of the sequence at the relays. We show by simulations on simple configurations that this cooperative scheme is superior to other techniques proposed in the literature, and close to the Gaussian relay channel capacity, even at moderate codeword lengths.**

***Index Terms*—Non-binary LDPC; Coopertive coding; Relay networks.**

I. Introduction

In wireless communication systems, the spatial diversity brought by the existence of relays, which can broadcast modified re-encoded versions of the source streams, helps to improve greatly the global information throughput and its error rates. Those improvements are impacted by the use of *cooperative diversity* [1], [2], which has been proposed in the literature for wireless relay channels and their multi-terminal extensions. A relay channel is a multi-terminal network consisting of a source, a destination, and a collection of relays which could be of different nature, as depicted on Figure 1 for the case of two relays. The communication system acts as follows: the source broadcasts a message to both relays and destination, while the relays forward the message or modified versions of it to the destination. Subsequently, different authors have proposed cooperation protocols for the relay channel, which can be classified into three major categories, namely the amplify-and-forward (AF) relays, the compress-and-forward (CF) relays and finally the decode-and-forward (DF) relays [3]. In AF protocols, the relays simply amplify the received signals and forward them to the destination, while in CF protocols the received noisy signals at the relays are quantized and forwarded. The DF protocol allows each relay to decode the received signal, re-encode it, and forward it to the destination. The forwarded message can either be identical to, or part of the initial transmission (repetition coding), or it can be obtained by using a dedicated coding scheme at the relays (cooperative coding). In the repetition coding case the destination combines received signals from both source and relays, which results in an improved signal-to-noise ratio (SNR) on the received transmission. In the cooperative coding case, the receiver at the destination uses the global knowledge of the cooperative coding (namely all code structures corresponding to the source and the relays), to jointly decode the received signals from both source and relays.

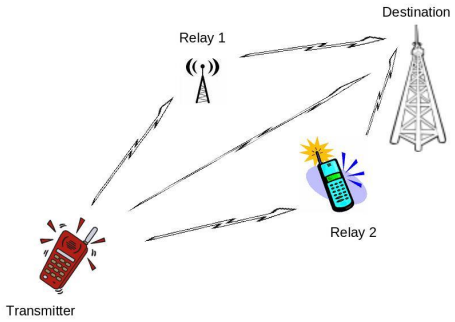


Fig. 1. Example of wireless relay channel with two relays.

In the simple case of repetition coding, there is no extra coding gain brought by the relay transmissions since the protocol does not change the Forward Error Correcting (FEC) code, and impacts only on the SNR improvement. On the other hand the receiver at the destination has the same low complexity as if no relays where used. In the case of distribution coding however, a proper design of the system aims at maximizing the coding gain brought by the relays to get closer to the relay channel capacity. This comes however at the cost of an extra decoding complexity at the receiver end, while joint decoding of the source and the relays are necessary to take advantage of the cooperative coding. In this paper, we propose a scheme which aims at having both advantages, namely an extra coding gain at no extra decoding cost. Distributed coding using parallel turbo-codes [4] or binary LDPC codes [5], [6], [7], [8], [9], has already been proposed in the literature. The existing approaches are either based on serial or parallel code concatenation, such that the graph of the LDPC code broadcasted from the source is a only subgraph of the destination decoding graph, or based on punctured rate-compatible LDPC codes. In a recent publication a cooperative LDPC code design which uses a turbo-like decoder at the receiver to jointly

 ISBN: 978-1-61208-183-0

decode the different sub-graphs of the source and the relays is proposed [10]. All these methods suffer from the large increase of decoding complexity at the receiver and the fact that they are not robust when the number of relays is larger than one, *i.e.* the coding gain using several relays is less and less important when the number of relays increases.

In this paper, we propose a new approach to the problem of distributed coding for cooperation in the case of multiple relays. The approach is based on non-binary LDPC (NB-LDPC) codes and the recently introduced technique of multiplicative non-binary coding [11], [12], which will be referred to as non-binary repetition coding. In our setting, the source transmits a codeword issued from a NB-LDPC code to the destination and the relays. When the relays successfully decode the received codeword, extra parity symbols are computed at the relays through optimized non-binary repetition codes, and are broadcasted to the destination. The receiver then collects the original received codeword from the source and the non-binary extra symbols from the relays and combines them before the iterative decoding. The iterative decoding complexity is the same in the presence or the absence of relays, while the combining of the codeword and the additional non-binary repetition symbols from the relays brings an effective coding gain. The paper is organized as follows. In Section II, we recall the basics about NB-LDPC codes and decoders and we present the concept of non-binary repetition coding coupled with NB-LDPC codes. In Section III, we describe our proposed cooperative system and discuss its advantages. We also propose in Section IV an optimization for the design of NB-LDPC codes and non-binary repetition codes for maximizing the coding gain, and finally, we present some simulation results on simple relay channels in Section V.

II. Non-Binary LDPC Codes and Decoding

A. Non-Binary LDPC codes

A Low Density Parity Check (LDPC) code is defined by a very sparse random parity check matrix H, which consists of $N-K$ rows and N columns, where K is the information block length and N is the codeword length; the code rate is defined by $R \leq \frac{K}{N}$. LDPC codes are nowadays used and proposed for a large number of communication and storage applications and standards, as their performance under low complexity iterative decoding approach the capacity for a large variety of channels. Binary LDPC codes can be generalized to non binary LDPC codes (NB-LDPC). The parity-check equations are written using symbols in a Galois field of order q, denoted GF(q), where $q = 2$ is the particular binary case. Throughout the paper, the Galois field elements will be denoted $\left\{0, \alpha^0, \alpha^1, \ldots, \alpha^{(q-2)}\right\}$, where α is a primitive element of the Galois field. The parity check matrix defining a NB-LDPC code has only a few nonzero coordinates h_{ij} which belong to GF(q), and a single parity equation involving d_c codeword symbols follows:

$$\sum_{j=1}^{d_c} h_{i,j}.c_j = 0 \qquad (1)$$

where $\{h_{i,j}\}$ are the nonzero values of the $i-th$ row of H, and $\mathbf{c} = \{c_1, \ldots, c_N\}$ is the notation used for the NB-LDPC codeword.

NB-LDPC codes are usually preferred to their binary counterparts when the blocklength is small to moderate [13], [14], or when the order of the symbols sent through channel are not binary [15], which is the case for high-order modulation (M-QAM) or for Multiple-antennas channels [16]. As a matter of fact, when the LDPC code is build in a field with order q equal or higher than the modulation order M, the non-binary LDPC decoder is initialized with uncorrelated vector messages, which helps the decoder to be closer to Maximum Likelihood Decoding than in the binary case. Recently, another advantage of NB-LDPC codes has been identified [17], [12]. The authors have shown in these papers that one can design flexible coding transmission in a very simple, though efficient way. The proposed approach is to concatenate non-binary multiplicative codes to a mother NB-LDPC, which leads to extra redundancy built from non-binary repetition symbols. When the repetition coding is properly designed, it results that the coding gain is greatly increased compared to binary repetition coding, expecially when the field order is sufficiently large GF(q), with $q \geq 64$. In this paper, we make use of the concatenation of NB-LDPC codes and non-binary repetition codes to design our distributed coding scheme, as presented in Section III.

B. Brief presentation of NB-LDPC decoders

The performance improvement of NB-LDPC codes is achieved at the expense of increased decoding complexity. As in all practical coding schemes, an important feature is the complexity/performance tradeoff, it is very important to try to reduce the decoding complexity of NB-LDPC codes, especially for high order fields GF(q) with $q \geq 64$. The base decoder of NB-LDPC codes is the Belief-Propagation (BP) decoder over the Tanner graph representation of the code [18]. The Tanner graph of an NB-LDPC code is drawn on Figure 2. The nonzero values of the parity-check matrix are put as *labels* for the edges connected to the non-binary parity check nodes. In this figure, we have represented all four parity-check nodes with the same labels $\{h_1, h_2, h_3, h_4, h_5\}$, and the information symbols are represented in red (left side of the codeword) while the redundancy symbols are drawn in blue (right side of the codeword). The number of edges connected to the nodes is constant throughout the Tanner graph, and furthermore the number of edges for the symbol nodes is minimum, equal to $d_v = 2$. This Tanner graph corresponds to a *regular and ultra-sparse* NB-LDPC code, with code rate $R = 1 - \frac{d_v}{d_c} = \frac{3}{5}$.

The main difference with the binary BP decoder is that for GF(q) LDPC codes, the messages from variable nodes to check nodes and from check nodes to variable nodes are defined by q probability weights, or $q-1$ log-density-ratios. As a result, the complexity of NB-LDPC decoders scales as $\mathcal{O}(q^2)$ per check node [19], which prohibits the use of codes build in high order fields.

 ISBN: 978-1-61208-183-0

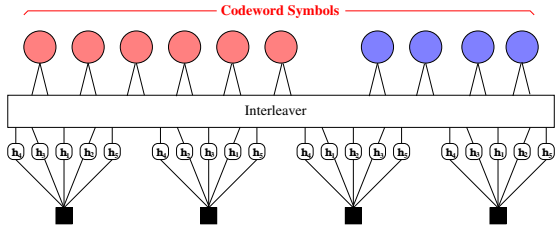


Fig. 2. Tanner graph of an ultra-sparse NB-LDPC code, with $(d_v, d_c) = (2, 5)$ and $R = 3/5$.

Sub-optimum decoders based on generalization of the min-sum decoder have been developed [19], [20], with the goal of reducing the decoding complexity at the check-node side. In particular, the EMS algorithm presented in [21] proposes the best complexity/performance tradeoff found in the literature, as the complexity scales as $\mathcal{O}(n_m.q)$ with $n_m << q$. We do not present with more details the EMS non-binary decoder in this paper and refer to the cited article for a complete description and analysis. Only the computation of the Log-Likelihood ratios (LLR) used for the initialization of the decoder are presented in the next section.

C. Computation of the LLR vectors

For transmission over a general wireless channel, the code symbols defined in GF(q) have to be mapped to $M-$QAM symbols, where M is a power of two, including $M \in \{2, 4\}$ for BPSK and QPSK. The information message $\mathbf{u} \in \text{GF}(q)^K$ is encoded into a codeword $\mathbf{c} \in \text{GF}(q)^N$, which is passed to the modulator and then transmitted over a continuous-valued fading channel. At the receiver, the soft demapper computes log-likelihood values (LLR-values), which constitute a sufficient statistic of the received signal $\mathbf{y}$ and form the initialization of the channel decoder.

To obtain a bijective mapping, we have to map m_1 code symbols to m_2 QAM symbols such that $q^{m_1} = M^{m_2}$. We denote the QAM alphabet by χ_M, and the mapping function by $\boldsymbol{\mu}$, i.e.

$$\boldsymbol{\mu} : \text{GF}(q)^{m_1} \rightarrow \chi_M^{m_2} \tag{2}$$

The code symbols $\mathbf{b} = (b_1, b_2, \ldots, b_{m_1})$ belong to the same codeword $\mathbf{c} = (c_1, c_2, \ldots, c_N)$ and are mapped to a vector of QAM symbols,

$$\mathbf{x} = [x_1, \ldots, x_{m_2}] = \boldsymbol{\mu}(\mathbf{b}) = [\mu_1(\mathbf{b}), \ldots, \mu_{m_2}(\mathbf{b})] \tag{3}$$

For binary codes (i.e. for q = 2), we always have $m_2 = 1$, while for codes in higher order Galois fields, for many modulations $m_1 = 1$, which is quite beneficial for the demapping, as we will see below. The soft demapper computes the LLR-vector $\mathbf{L}_i = [L_{i,0}, L_{i,1}, \ldots, L_{i,q-1}]^{\text{T}}$, which corresponds to the code symbol b_i , and whose components are given by, for $i = 1, \ldots, m_1$ and $g \in \text{GF}(q)$

$$L_{i,g} \triangleq \ln \frac{P[b_i = g|\mathbf{y}]}{P[b_i = 0|\mathbf{y}]}$$

where we identify, with a slight abuse of notation, the elements of GF(q) by their indices $g = 0, 1, \ldots, q-1$.

For a memoryless channel and assuming that all code symbols are equiprobable, we obtain

$$L_{i,g} = \ln \frac{\sum_{\mathbf{b} \in \mathcal{B}_i^g} \prod_{j=1}^{m_2} p\left(y_j|\mathbf{b}\right)}{\sum_{\mathbf{b} \in \mathcal{B}_i^0} \prod_{j=1}^{m_2} p\left(y_j|\mathbf{b}\right)}, \quad \begin{matrix} i = 1, \ldots, m_1 \\ g = 0, \ldots, q-1 \end{matrix} \tag{4}$$

where $\mathcal{B}_i^g \triangleq \{\mathbf{b} \in \text{GF}(q)^{m_1} : b_i = g\}$ is the set of all code symbol vectors whose i-th component is fixed to g.

The mapping and in particular the demapping simplifies significantly for $m_1 = 1$, which means that exactly *one* code symbol $b \in \text{GF}(q)$ is mapped to a vector of QAM symbols. In this case, we can drop the index i in (4), and since the sets $\mathcal{B}_i^g$ reduce to one element, we can write

$$L_g = \ln \frac{\prod_{j=1}^{m_2} p\left(y_j|b = g\right)}{\prod_{j=1}^{m_2} p\left(y_j|b = 0\right)} \tag{5}$$

For a flat fading channel given by $y_j = a_j \cdot x_j + w_j$ with $w_j \sim \mathcal{CN}(0, N_0)$, the conditional pdf is given by $p(y_j|b = g) = \frac{1}{\pi N_0} \exp\left(-\frac{|y_j - a_j \mu_j(g)|^2}{N_0}\right)$. With this, and noting that a BP decoder is typically insensitive to additive constants of the LLR vectors, we can further simplify (5) to

$$L_g = -\frac{1}{N_0} \sum_{j=1}^{m_2} |y_j - a_j \mu_j(g)|^2 + \ell_0 \tag{6}$$

where ℓ_0 is an arbitrary additive constant which does not depend on g.

As we can see from the LLR computations, using NB-LDPC codes with $m_1 = 1$ results in a significant complexity reduction of the demodulator (without any approximation) with respect to binary demappers, since no marginalization is required. This is the case if the number of bits per code symbol is a multiple of the number of bits per QAM symbol, i.e. $\log_2(q) = m_2 \cdot \log_2(M)$. For instance, codes in GF(64) allow a simple LLR computation for $\log_2(M) \in \{1, 2, 3, 6\}$, which corresponds to BPSK, QPSK, 8-QAM and 64-QAM. Note that more options are possible by mapping code symbols separately to the I or Q component, *i.e.* by considering real-valued PAM constellations. This does not incur any performance penalty, but allows e.g. to easily combine 16-QAM with a GF(64) code by mapping the 6 bits of one code symbol to three 4-PAM symbols.

III. Proposed Cooperative Transmission Scheme

A. Channel Model and System Description

Throughout the paper, we will assume that the source broadcasts a NB-LDPC codeword to the destination and a given number N_r of relays. All wireless channels in the system are either Rayleigh fading channels, or memoryless additive white Gaussian noise (AWGN) channels, depending on the type of relay (fixed or mobile). For sake of simplicity in the presentation, we restrict the model description to AWGN channels, but without loss of generality since the LLR computation presented in the preceding section does not change for AWGN or Rayleigh fading. The different channels will make use of the following notations:

 ISBN: 978-1-61208-183-0

- The link between source and destination uses $M_{SD}-QAM$ signalling with signal-to-noise ratio equals to γ_{SD},
- The link between source and the i-th relay uses $M_{SR_i}-QAM$ constellations with signal-to-noise ratio equals to γ_{SR_i},
- Finally, the link between the i-th relay and the destination uses $M_{R_iD}-QAM$ constellations with signal-to-noise ratio γ_{R_iD}.

Note that since relays and destination receive the same modulated signals, we have by construction $M_{SD}-QAM = M_{SR_i}-QAM$, although the SNRs could be different.

Now, let us discuss the channel and transmission protocol assumptions that we use in our work. First we assume that the direct link between the source and the destination is weak and that the relays links are stronger, both from source to relay and from relay to destination, which is a usual assumption in relay channels. It follows that $\gamma_{SR_i} \geq \gamma_{SD}$ and $\gamma_{R_iD} \geq \gamma_{SD}$, $\forall i$. The improved SNRs of the relay channels implies that either higher-order modulations would be used for the relay-to-destination channel, or higher rate cooperative coding. The optimization of the modulation order or of the cooperative coding rates requires that channel fading estimation and a link adaptation strategy is performed on the relay channel. We leave this issue for future research, and in this paper, we will assume fixed values for the modulation orders and coding rates, and measure the performance by the gap of error rates to the capacity of the relay channel.

Now, we must make an assumption regarding the non-propagation of errors through the relays. Since the relays are decode-and-forward nodes in the network, we can reasonably assume that the relay can detect if it decodes the received codeword from the source or not. We will then assume that when the relay fails to decode to a valid codeword, it does not transmit any information to the destination, which prevents unavoidable decoding failures at the destination.

Given this model and assumption, our proposed cooperative coding scheme can be described as follows:

- The source encodes the packet of information bits, generating a NB-LDPC codeword $\mathbf{c}$ of the parity check matrix H, with rate R. The source modulates $\mathbf{c}$ with the $M_{SD}-QAM$ constellation and broadcasts the modulated symbols $\mathbf{x}$ to both relays and destination.
- Each relay i decodes the received signal, correcting the transmission errors on $\mathbf{c}$. The relays then generate a new sequence $\mathbf{c}^{(i)}$ of non-binary symbols using the repetition coding, as depicted on Figure 3 in the case of 2 relays. Note that the size of the vectors $\mathbf{c}^{(i)}$ are not necessarily the same as the original codeword $\mathbf{c}$, since the coding rates for the links relay-destination are typically higher. The encoding procedure and the optimization of the repetition codes are presented in the next section. The vectors $\mathbf{c}^{(i)}$ of non-binary repetition symbols are then transmitted from the relays to the destination using $M_{R_iD}-QAM$ constellations.
- Thus, the destination receives noisy versions of $\mathbf{x}$ and $\mathbf{x}^{(i)}$ (from both the source and the relay), which can be jointly decoded using the only the matrix H, and the LLR computation presented in Section II-C.

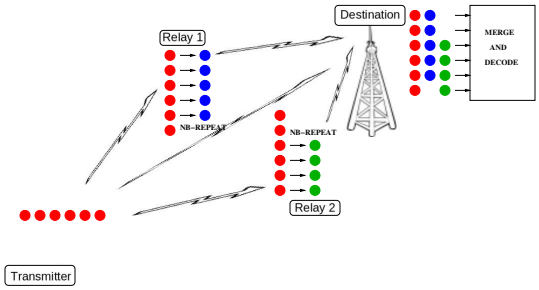


Fig. 3. Cooperative coding scheme using non-binary repetition coding.

The proposed distributed scheme relies mainly on the way the repetition symbols are generated and taken into account in the joint decoding at the destination. We explain in details in the next section why the non-binary repetition symbols bring a significant coding gain, at no extra encoding or decoding cost.

B. Non-Binary Repetition coding and Joint Decoding

As mentioned in the previous section, we assume in this paper that the parameters of the transmission system for each link are fixed, namely the constellation orders $\{M_{SD}, M_{R_iD}\}$ and the coding rates $\{R_{SD}, R_{R_iD}\}$ are fixed *a priori*. Now let us present how the non-binary repetition symbols are generated.

The $i-th$ relay is supposed to receive and decode correctly the broadcasted codeword $\mathbf{c}$ (otherwise, relay i does not transmit anything). From the N symbols in $\mathbf{c}$, the relay needs to build $N_i = N.\frac{R_{SD}}{R_{R_iD}}$ repetition symbols, which represents the coded block that relay i has to send to the destination. For example, if $R_{SD} = 1/2$ and $R_{R_iD} = 3/4$, $N_i = N.\frac{2}{3}$ repetition symbols have to be encoded at relay i. The repetition encoding is performed as follows for relay i:

- Select N_i non-binary symbols $\left\{c_{k_l^{(i)}}\right\}_{l=1...N_i}$ inside the codeword $\mathbf{c}$, the N_i symbols could be chosen arbitrarily, so either a random selection or a selection based on the knowledge of the transmitted NB-LDPC code are possible,
- For each selected symbol $c_{k_l^{(i)}} \in \mathrm{GF}(q)$, generate a new symbol $c_l^{(i)} = h_l^{(i)}.c_{k_l^{(i)}}$, with $h_l^{(i)} \in \mathrm{GF}(q)$ being the non-zero field value corresponding to the local repetition code. The vector $\mathbf{c}^{(i)}$ of size N_i is then sent from relay i to the destination.

We can easily see that this repetition encoding procedure is extremely simple, as it requires only N_i Galois field operations after a successful decoding at the relay. Note also that a sub-case of the proposed scheme corresponds to the particular choice of $h_l^{(i)} = 1$, $\forall l$, and which reduces to the usual decode-and-forward strategy, where the same codeword is sent both

 ISBN: 978-1-61208-183-0

from the source and the relays. In our case, with a very limited extra complexity, we allow the use of non-binary repetitions with $h_l^{(i)} \neq 1$, which provides a non-negligible coding gain, as explained in Section IV.

Now let us discuss how the collection of received symbols are jointly treated at the destination. For some particular code symbol $c \in \mathrm{GF}(q)$, we denote by $\mathbf{x}$ the QAM symbols build from c transmitted by the source and by $\mathbf{y}^{(0)}$ the corresponding received value at the destination. We also denote by $\mathbf{x}^{(i)}$ the QAM symbols transmitted by the relays corresponding to the same code symbol c, and $\mathbf{y}^{(i)}$ the corresponding channel outputs. Note that here we dropped the index of the symbol in the codewords to simplify the notations, and we just assume that the received values correspond indeed to the same symbol c. So, the symbol c receives the channel values $\left\{\mathbf{y}^{(0)}, \mathbf{y}_l^{(1)}, \ldots, \mathbf{y}_l^{(I)}\right\}$, from the source and I active relays according to one row in Figure 3 at the destination. The destination needs to compute the joint-LLR vector $\mathbf{L} = \{L_g\}_{g=1\ldots q}$, which merges the sufficient statistics of all active links. Like in Section II-C, we define the LLR vector up to an additive constant as

$$L_g \triangleq \ln P[c = g|\mathbf{y}^{(0)} \ldots \mathbf{y}^{(I)}] + \ell_1 \qquad \forall g \qquad (7)$$

Using Bayes' theorem and the fact that the source-destination and relay-destination channels are conditionally independent, we obtain

$$P[c = g|\mathbf{y}^{(0)} \ldots \mathbf{y}^{(I)}] \propto p\left(\mathbf{y}^{(0)}|c = g\right) \prod_{i=1}^{I} p\left(\mathbf{y}^{(i)}|c^{(i)} = h^{(i)}.g\right) \qquad (8)$$

where $h^{(i)}$ is the non-zero value used for the non-binary repetition encoding of symbol c at relay i.

We define then the LLR vectors corresponding to each separated channel, for the source ($i = 0$) and for the relay transmissions ($i > 0$) as

$$\lambda_g^{(i)} \triangleq \ln p\left(\mathbf{y}^{(i)}|c^{(i)} = g\right) + \ell_2 = -\frac{1}{N_0} \sum_{j=1}^{m_2} \left| y_j^{(i)} - a_j^{(i)} \mu_j(g) \right|^2 \qquad (9)$$

With $h^{(0)} = 1$ and (8), we finally obtain the joint-LLR vector components as the sum of the partial L-values:

$$L_g = \sum_{i=0}^{I} \lambda_{h^{(i)}.g}^{(i)} \qquad \forall g \qquad (10)$$

We thus combine the L-values of the main transmission and the relay transmissions into one joint-LLR vector per code symbol and feed the joint-LLR vectors to the decoder. In other words, the repetition scheme is transparent to the decoder and therefore does not affect the decoding complexity.

The process is depicted on Figure 4, which shows the factor graph used for the joint decoding of the NB-LDPC code from the source and the repetition codes from the relays. We considered on this figure the case of 3 relays, each of them sending $N_i = N.\frac{2}{3}$ extra repetition symbols. The repetition symbols are equally distributed among the codeword, *i.e.*, we have selected the repetition locations $\{c_{k_l}\}_{l=1\ldots N_i}$ at each relay, such that the destination receives 3 LLR measures for each coded symbol: one from the source, and two from the relays. This is a completely arbitrary choice and shows only the case of 3 relays with coding rates $R_{R_i D} = 3/4$. Our cooperative scheme is more general than the example of Figure 4 as the decoder can be initialized with joint-LLR vectors build from a different number of channel measurements for each coded symbol. We will discuss this issue in more details in the optimization Section IV.

Since the non-binary repetition codes impact only the the joint-LLR computation, it follows that the decoder complexity is the same, for any number of relays, which is a great feature of our cooperative coding scheme. Indeed, most of the cooperative coding schemes proposed in the literature require a joint decoding of the source and relays codes in an iterative turbo-decoding fashion [4], [10]. So the existing approaches firstly increase the receiver decoding complexity, but also prevents the use of multiple relays since the turbo-decoders with more than 2 component codes are very difficult to design so that they approach the capacity of the channel [22].

In our scheme, the decoding complexity does not depend on the number of relays — only the computation of joint-LLR vectors depends linearly on the number of relays — and more importantly, if the non-zero values for the repetition codes are well designed at the different relays, each relay brings an extra coding gain.

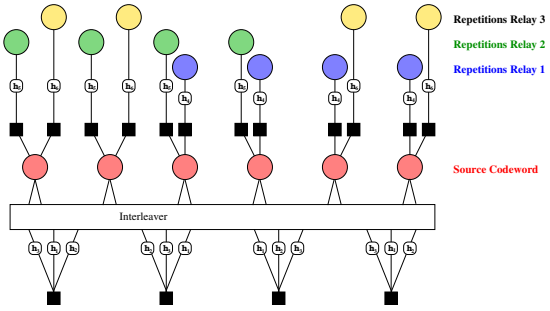


Fig. 4. Tanner graph of the joint-receiver at the destination. Case of 3 relays.

IV. Optimization of the Cooperative NB-LDPC Codes

In this section, we propose a fine optimization of the NB-LDPC cooperative coding scheme, which aims at having the best possible performance for the practical case of short to moderate codeword lengths. Both the NB-LDPC code used at the transmitter and the non-binary repetition codes at the relays have to be properly optimized. The source NB-LDPC code will be chosen so as to have the best performance in the waterfall region, to ensure that the successful decoding rates at the relays are large enough (remember that the relays are transmitting to the destination only if they successfully decode the word received from the source). Also, we propose a specific quasi-cyclic protograph construction of the Tanner graph of the NB-LDPC such that the relays can chose efficiently the locations at

 ISBN: 978-1-61208-183-0

which they should build the repetition symbols. As for the NB-repetition codes at the relays, we will propose the optimization of the non-zero field values such that the coding gain at the receiver is maximized.

A. NB-LDPC code Optimization at the Transmitter: Component Codes

For codes defined over GF(q), when addressing finite length design, it has been shown in [14] that selecting carefully the non binary entries of the parity-check matrix can improve the overall performance of the code when compared to randomly chosen coefficients. The selection of the non zero values can impact both on the waterfall and the on error floor. The observed performance gains are dependent of both the field order and the code rate.

In the waterfall region, selecting the edges label row-wise is critical. It is shown in [14] that *best* rows are selected according to their equivalent binary minimum distance and multiplicity of the minimum distance. The Binary Component Code of a non-binary parity check is build from the transpose of the companion matrices H_{ij} of the non-zero values h_{ij} composing the parity check. Using binary matrix images for the non-zero values of the check and binary vector images $\underline{c}_j$ for the codeword symbols, one get the following parity-check equation in a vector form, corresponding to the non-binary parity-check equation (1):

$$\sum_{j=1}^{d_c} H_{ij}.\underline{c}_j = \underline{0}_p \qquad \text{in GF}(2)^p$$

where $p = \log_2(q)$ is the number of bits per symbol of the Galois field.

The binary image of a non-binary parity-check in GF(q) for $q = 64$ is explained in Figure 5, and it can be easily seen that it acts as a binary component code of size $(N - K, N) = (p, p.d_c)$. The better is the component code in terms of minimum distance, the better will be the error decoding performance in the waterfall region.

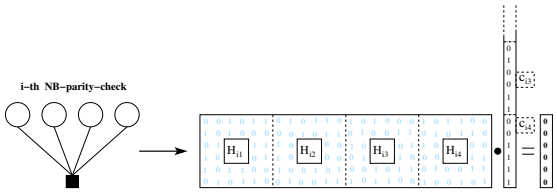


Fig. 5. Binary image of a non-binary parity-check equation in GF(64)

In this paper, we generalize the approach proposed in [14], and propose a new criterion selection for the non-zero values which compose a parity-check. In existing approaches, it is advised to maximize the strength of the component code, and then choose the non-zero field values such that the binary image has the maximum *minimum Hamming distance* (D_{min}), together with the minimum multiplicity of codewords with Hamming weight D_{min}. Although locally optimal, this strategy is not optimal when used in a message passing iterative decoder, where extrinsic vector messages are propagated along edges, *i.e.* between component codes. A better strategy, which is especially efficient when the code is a strictly regular ultra-sparse code with $d_v = 2$, is to optimize the balance between sub-codes of the component code.

Let us describe here this new idea. Since the message-passing decoder will propagate d_c extrinsic messages computed from the incoming message at each iteration, it is better to build extrinsic messages which statistically behave equally. In other words, the extrinsic messages should have their quantity of mutual information as close as possible to their average. Indeed, increasing the mutual information of one particular extrinsic output will result in decreasing the mutual information for other extrinsic messages, therefore propagating worse messages to the rest of the Tanner graph. Note that this property of *equal balance* between the extrinsic messages is verified on average if the non-zero values are taken uniformly at random. However, when one wants to optimize only a limited number of non-zero values to improve the performance, then the equal-balance property is lost, and we propose here a design technique to compensate for it.

The new optimization criterion for component code selection is described in the following algorithm.

Algorithm 1 Component Code Optimization

1) Let us a non-binary parity check of degree d_c with non-zero values $\{h_1 \dots h_{d_c}\}$
2) Consider the d_c binary subcodes $\mathcal{S}_{cc}(k)$ formed from the combination of the $d_c - 1$ values in $\{h_1 \dots h_{d_c}\}$ except h_k.
3) We choose for $\{h_1 \dots h_{d_c}\}$ the field values in $GF(q)$ such that:

$$\{h_1 \dots h_{d_c}\} = \max_{\{h_1 \dots h_{d_c}\}} \left(\sum_{k=1}^{d_c} \mathrm{D}_{min}(\mathcal{S}_{cc}(k)) \right)$$

constrained to $\left| \mathrm{D}_{min}(\mathcal{S}_{cc}(k) - \mathrm{D}_{min}(\mathcal{S}_{cc}(k') \right| \leq 1$

This criterion ensures that both the component code and all the sub-codes of the components codes have good and equally distributed error correction capability. This new optimization criterion is indeed interesting since we saw slight improvement in the waterfall region compared to codes that use existing sets of non-zero values. We give below the best sets of field coefficients for GF(64) and GF(256) that have been optimized with the new criterion, and that we can use for the source NB-LDPC code design. Recall that the notations used for the field elements are $\left\{0, \alpha^0, \alpha^1, \dots, \alpha^{(q-2)}\right\}$. For $d_c = 4$ and $d_c = 6$, four sets of values were found to have the exact same performance with respect to the criterion of the optimization algorithm.

- best rows for GF(64) and $d_c = 4$

$$(\alpha^0, \alpha^9, \alpha^{26}, \alpha^{46}) \quad (\alpha^0, \alpha^{17}, \alpha^{26}, \alpha^{43})$$
$$(\alpha^0, \alpha^{17}, \alpha^{37}, \alpha^{54}) \quad (\alpha^0, \alpha^{20}, \alpha^{37}, \alpha^{46})$$

- best rows for GF(256) and $d_c = 4$

$$(\alpha^0, \alpha^8, \alpha^{173}, \alpha^{183}) \quad (\alpha^0, \alpha^{10}, \alpha^{82}, \alpha^{90})$$
$$(\alpha^0, \alpha^{72}, \alpha^{80}, \alpha^{245}) \quad (\alpha^0, \alpha^{165}, \alpha^{175}, \alpha^{247})$$

 ISBN: 978-1-61208-183-0

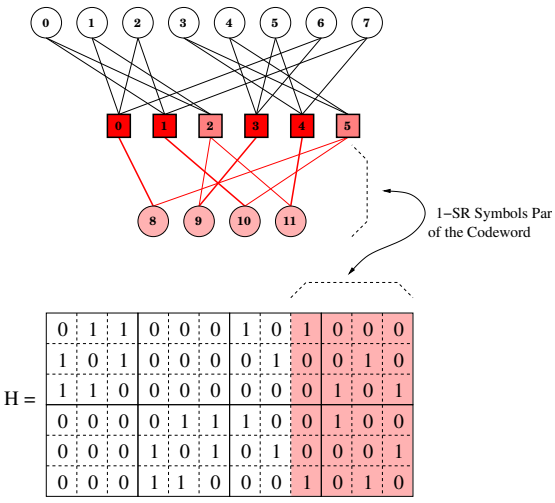
B. NB-LDPC code Optimization at the Transmitter: Global Tanner Graph

In this section, we describe our NB-LDPC code design, based on protographs. First introduced by [23], a binary protograph is defined as a small bipartite graph from which a larger graph is obtained, by the so-called *lifting* technique. The protograph itself is generally described using its *adjacency matrix* H_B also called base matrix [24], where the coefficients $H_B(i,j)$ represent the number of edges between the $i-th$ check node C_i of the protograph and the $j-th$ variable node V_j. The base matrix H_B is then a small matrix containing small integer values. The lifting process is then to expand the base matrix by replacing each non-zero entry $H_B(i,j) > 0$ by the same number of non-overlapping circulant matrices. Circulant matrices are usually preferred for practical purposes since it reduces the descriptional complexity (ie. storage) of the parity check matrix in the hardware realizations of the LDPC encoder and decoder. If L is the size of the circulant matrices, we obtain — after lifting — a Tanner graph with L times more nodes and edges than the protograph. The last step for non-binary LDPC codes is then to assign non-zero values to the edges of the lifted Tanner graph. The nonzero values are randomly assigned from the optimized subsets presented in the previous section. Note that an additional optimization step could be performed with the objective of improving the performance in the error floor, as described in [14]. We have performed this optimization technique in our code design, but we do not present it in this paper, and refer to [14] for a complete description.

On Figure 6, we show the protograph which has been chosen for the coding rate $R = 1/2$. Similar protographs have been build for higher rates, but we limit the discussion to the rate $R = 1/2$ in this paper. The structure of the protograph has been chosen so as to maximize the number of *1-SR survivors* [25]. In [25], the authors show that under iterative decoding, all the codeword symbols do not have an equal protection, in the situation that some symbols in their direct neighborhood in the computational tree are erased or very noisy. They introduce the concept of $k-SR$ survivor symbols (k-steps recovery), which is defined as a symbol which can be recovered from the other symbols after k iterations of the message passing decoder — assuming that the other symbols are either correctly decoded, or have a large likelihood. The authors use this property to design puncturing patterns: $1-SR$ survivors can be preferably punctured, since they are less sensitive than other symbols and can be recovered easily under iterative decoding. Note that the $k-SR$ survivor property does not change if the code is a binary LDPC code or a NB-LDPC code.

Here we will use this concept to indicate to the relays where it is preferable to add non-binary repetition symbols in the codeword. The reasoning is the dual of the puncturing problem. If we know where the 1-SR symbols are in the codeword, then the relay will preferably build repetition symbols in the part of the codeword which does not contain the 1-SR symbols. This way, after merging the LLR vectors into a joint-LLR vector, the symbols which have the 1-SR property will receive less information — on average — than the other symbols. This is not a problem since the rest of the symbols, with better joint-LLR values, will be able to retrieve the 1-SR symbols. In the extreme case where the main link is so noisy that the received LLR corresponds almost to a complete erasure of the codeword, if we assume that the relays have transmitted only the symbols without the 1-SR property, then the receiver which uses the joint-LLR will successfully recover the entire codeword.

It is obvious, from this discussion, that the NB-LDPC code with the maximum of 1-SR symbols would be the best choice. This way, it helps the relays to concentrate only on the remaining part of the codeword to build the repetition symbols. This is the approach that we used in this paper for the design of the protograph. Figure 6 presents the obtained protograph for the case of a coding rate $R = 1/2$, which is the protograph with the maximum number of 1-SR symbols. In this structure, 4 checks are connected to the bottom 1-SR symbol (connected with a bold/strong link) while 2 checks are connected to 2 of the bottom symbols (indicated with a weak link). However, each one of the bottom symbol is connected exactly to one strong link and one weak link. The 1-SR condition is ensured when each and every symbol is connected to at least one strong link, i.e. at least one check node from which this symbol can be recovered in 1 iterative step. With this protograph we get 4 symbols with the 1-SR property, which is the maximum number for coding rate $R = 1/2$, and we would get $4\,L$ symbols with the 1-SR property after the lifting step. Note also that this protograph has girth 6 (size of the minimum cycle), which is also a good feature in order to get good Tanner graphs after the lifting step [26].



0	1	1	0	0	0	1	0	1	0	0	0
1	0	1	0	0	0	0	1	0	0	1	0
1	1	0	0	0	0	0	0	0	1	0	1
0	0	0	0	1	1	1	0	0	1	0	0
0	0	0	1	0	1	0	1	0	0	0	1
0	0	0	1	1	0	0	0	1	0	1	0

Fig. 6. Detailed protograph for the source NB-LDPC code design. This protograph has the property to maximize the number of symbols with the 1-SR property.

Let us have a look at the computational tree seen from one of the 1-SR symbol node, which is drawn on Figure 7. Symbol #8 is 1-SR from check node #0, and 2-SR from check node #5 since the symbol #10 will be recovered after the first decoding

 ISBN: 978-1-61208-183-0

iteration. Note that all the symbols in the 1-SR region have this property of being 1-SR from one of their edge, and 2-SR from the other edge, which indicates that the number of 1-SR symbols in maximum and there is no protograph with these dimensions having more 1-SR symbols.

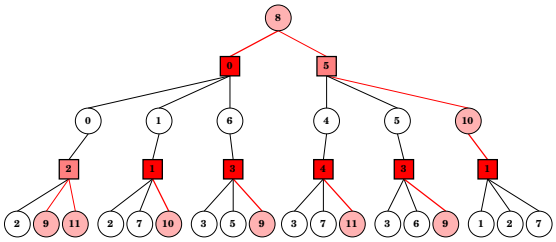


Fig. 7. Computational tree of the proposed protograph expanded from symbol node #8.

To conclude this section, we have designed source NB-LDPC codes based on a protograph approach, with connexion degrees $(d_v = 2, d_c)$. The considered protographs have both a good girth $g = 6$ and the property of localized (*a priori* known locations) maximum number of 1-SR symbols. The good girth of the protograph ensures that very large girths can be obtained for the lifted-graph, for example we obtained a girth of $g = 16$ for a NB-LDPC Tanner graph of length $N_s = 480$ coded symbols. The localized 1-SR symbols, known at the relays, are used to select the preferred locations of the non-binary repetition symbols.

C. Repetition code Optimization at the Relays

We now discuss the impact of the non-binary repetition symbols build by the relays and used in the joint-LLR computation at the destination. Usually, repetition coding is thought as having no real coding gain, but is employed to reduce the amount of noise in the received noisy symbols. Indeed, repetition coding is used in many transmission schemes, such as H-ARQ transmissions with *Chase combining* or in cooperative coding with DF or CF strategies. In the case of non-binary repetition codes however, it can be shown that the simple repetition of a symbol, weighted by a non-zero value $h^{(i)} \in \mathrm{GF}(q)$ with $h^{(i)} \neq 1$, results in a non-negligible coding gain [11].

Let us first concentrate on the case of a single repetition. We can simply explain the coding gain the following way: let c be the symbol to be repeated and $h^{(i)}.c$ being the repeated Galois field value. The receiver receives both noisy values on c and $h^{(i)}.c$, corresponding to the same codeword symbol. It follows that the demodulation actually acts as a maximum-a-posteriori decoder of the repetition code, which is build from the concatenation of the two Galois field values $[1, h^{(i)}]$. Now the coding gain is increasing with the minimum distance of the binary image of $[1, h^{(i)}]$. In the case of simple a copy — regular repetition with $h^{(i)} = 1$ — the binary minimum distance is $\mathrm{D}_{min} = 2$ and no coding gain can be achieved, while for non-binary repetitions, this minimum distance is typically larger $\mathrm{D}_{min} \geq 3$ when the field size q is sufficiently large. Additionally, the non-zero repetition values $h_l^{(i)}$ need to be optimized with the knowledge of the non-zero values which have been used in the source NB-LDPC code. Indeed, during the iterative decoding algorithm, the extrinsic vector messages will be computed using the joint-LLRs, that is, with the modified parity-check nodes, including the repetition nodes as well, as depicted on Figure 8. The modified parity-check nodes act then as the new *component codes* of the joint coding scheme. Following the discussion of Section IV-A, it is the minimum distance of this modified parity-check nodes that need to be optimized in order to have the best performance in the waterfall region.

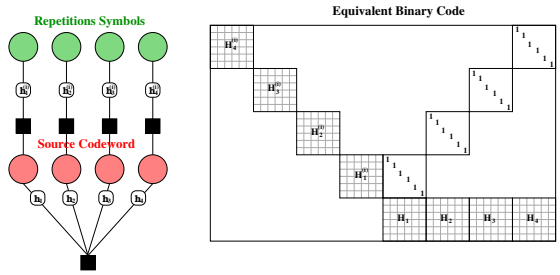


Fig. 8. NB-LDPC parity-check node with concatenated repetition codes.

We now present the optimization of non-binary repetition codes, with the objective of using only a small number of non-binary field values. We advice in particular to use the same non-zero value $h^{(i)}$ for all the repeated symbols at the relay i. By proper optimization, the coding gain is not reduced compared to a relay which would use different non-zero values $\{h_l^{(i)}\}_l$, and a single value per relay reduces greatly the complexity of re-encoding at the relay. We then use the following optimization procedure to optimize the values $h^{(i)}$, $\forall i = 1 \ldots I$.

We proceed as follows. As described in the Section IV-A, each and every check node of degree d_c will be labeled with the same set of non-zero values $\{h_1, h_2, \ldots, h_{d_c}\}$. So each location chosen by the relay in order to build a repetition symbol will *see* two of the non-zero values in this set (since $d_v = 2$). As a consequence, the non-zero values $h^{(i)}$ needs to be optimized jointly with all the values in $\{h_1, h_2, \ldots, h_{d_c}\}$. We have chosen to fix the set corresponding to the check-node values, and optimize the repetition values, conditionally to this set. The optimization is described by the following algorithm:

The optimization algorithm is stopped when the maximum number of potential relays I has been reached. We give as an example the optimized repetition values for the case of a $d_c = 4$ NB-LDPC code in $\mathrm{GF}(64)$ and $\mathrm{GF}(256)$ using the non-zero values sets presented in Section IV-A. The values that we obtained with our algorithm are indicated in table I. We have also indicated the minimum distance corresponding to the equivalent binary code (parity-check plus repetition codes). It can be seen that the minimum distance of equivalent codes grows linearly with the number of relays, which shows that the

 ISBN: 978-1-61208-183-0

Algorithm 2 Non-Binary Repetition Code Optimization

1) Let a parity-check equation have fixed non-zero values corresponding to the set $\{h_1, h_2, \ldots, h_{d_c}\}$. Let H_0 be the binary image of the equivalent code. Let $i = 1$.
2) Consider the modified binary code H_i, build from H_{i-1} and the repetition codes with the same $h^{(i)}$ on all the d_c symbols,
3) Choose $h^{(i)} \in$ GF(q) such that the the minimum distance of H_i is maximum. If several values $h^{(i)}$ have the same minimum distance, choose the one with minimum multiplicity,
4) $i = i + 1$, goto step 2).

Relay #	1	2	3	4	5	6	7	8
GF(64)	α^{26}	α^{41}	α^{52}	α^{6}	α^{56}	α^{17}	α^{50}	α^{11}
D_{min}	8	14	20	25	31	37	43	49
GF(256)	α^{15}	α^{165}	α^{71}	α^{150}	α^{128}	α^{122}	α^{113}	α^{104}
D_{min}	10	17	24	32	39	46	54	62

TABLE I
OPTIMUM NONZERO VALUES USED AT THE RELAYS FOR REPETITION CODING.

coding gain at the receiver increases as well with the number of relays.

In order to measure the performance gain in terms of frame error rate, brought by our optimized repetition scheme, we have performed Monte-Carlo simulations over a QPSK-AWGN channel for 2 different schemes using GF(256) NB-LDPC codes. The results are plotted on Figure 9. The direct link is indicated in black, and uses a rate $R = 1/2$ source NB-LDPC code. Then, we assume that the receiver receives gradually other channel values from the relays, with in this situation, the case of 4 relays. Each additional channel measurements lowers the overall coding rate, and in this figure, we assumed that 25% of the codeword length have been sent each time by the 4 relays. When the receiver has received the information from the 4 relays, the overall coding rate is indicated as $R = 1/4$. In our experiment, we have compared the simple repetition scheme with our optimized repetition scheme presented in this paper. One can see that the coding gain, when using the optimized repetition codes is non-negligible, between $0.3dB$ to $0.8dB$, with no extra decoding complexity at the receiver. It can also be seen on these curves, that the coding gain increases with the number of relays.

V. SIMULATION RESULTS IN A SIMPLE COOPERATIVE SITUATION

In order to evaluate the performance of our cooperative scheme and compare it to other works proposed in the literature, we have chosen to take the example of the simplest case of a single relay channel. The performance will be measured by the distance to the *capacity function*, inferred from the channel capacity. Capacities of various relaying strategies in case of a single relay have been computed in [27], [28], [29], and depend on the capacities of the three links. Since we assumed that source-to-relay transmission is error free, as the relay only propagates signals in case of successful decoding, we only consider the capacities of the two other links.

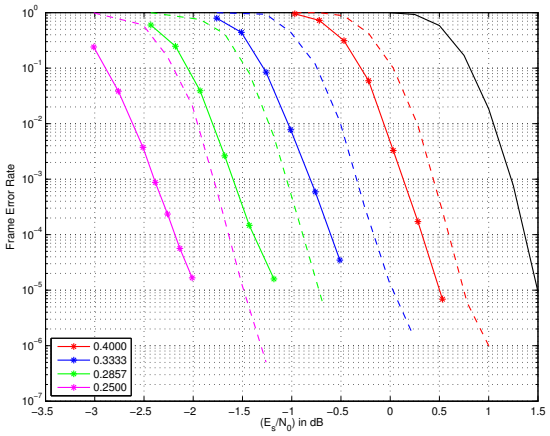


Fig. 9. Performance comparison of simple repetition scheme at the relays, and the optimized repetition scheme.

The meaning of the capacity function is the following. Assume that we want to transmit information with distributed rate (R_{SD}, R_{RD}) over a single relay channel. The rate R_{SD} is chosen according to the quality of the channel between source and relay, such that to ensure error free transmission between them. The rate R_{RD} is generally chosen according to the delay constraints of the cooperation system. Recall that γ_{SD} and γ_{RD} represent the signal-to-noise ratios from source-to-destination and respectively relay-to-destination. We define the relay *channel discrepancy* by the following quantity $\delta = \dfrac{\gamma_{\text{RD}}}{\gamma_{\text{SD}}} \geq 1$, which represents the relative quality of the relay link compared to the main link.

We have plotted on Figure 10 the SNR $(\gamma_{\text{RD}})_{\text{dB}}$ of the RD link with respect to the discrepancy $\Delta = 10\log_{10}(\delta)$ in the case of two scenarii, corresponding to $(R_{SD}, R_{RD}) = (0.5, 0.5)$ and $(R_{SD}, R_{RD}) = (0.5, 1.0)$. The second scenario is especially difficult since we assume that the relay transmit only K symbols to the destination. We have compared our cooperative scheme with the binary *split-and-extend* LDPC codes, which have been optimized for infinite length using density evolution techniques [10]. For the binary LDPC curves, we have plotted the minimum SNR $(\gamma_{\text{RD}})_{\text{dB}}$ for which the binary split-and-extend LDPC families converge to a zero error probability, when the codeword size grows to $+\infty$. As for our NB-LDPC cooperative coding scheme, we have indicated with symbols (circles and triangles) the SNR $(\gamma_{\text{RD}})_{\text{dB}}$ at which a Frame Error Rate of 10^{-5} has been reached with Monte-Carlo simulations. For our NB-LDPC coding scheme, GF(256) codes and repetition codes have been used, with a codeword length of $N = 720$ coded symbols. For the modulations, QPSK have been used for all links.

As we can see on these curves, the NB-LDPC cooperative scheme is close to the capacity curves in all cases, and shows especially a better robustness than the binary LDPC codes for the $(R_{SD}, R_{RD}) = (0.5, 1.0)$ scenario, when the discrepancy

 ISBN: 978-1-61208-183-0

becomes large. Additionnally, the capacity curves and the binary LDPC curve correspond to asymptotic performance, while our simulations are performed at relatively small lengths, corresponding to $N = 720$ coded symbols in GF(256). We then expect an axtra performance gain of our scheme by considering either longer codeword lengths or irregular mother NB-LDPC codes.

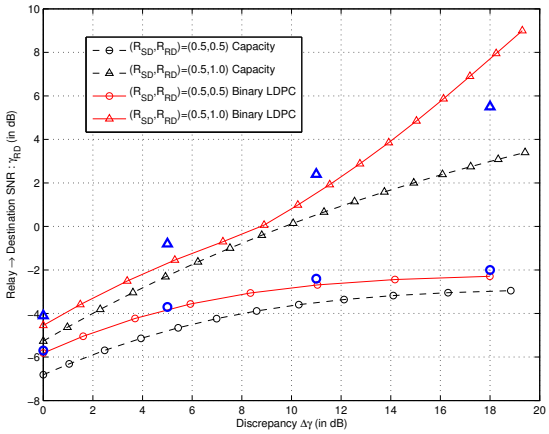


Fig. 10. Comparison of our cooperative scheme with existing binary LDPC cooperative scheme and the relay channel capacity. Our NB-LDPC scheme is indicated with the symbols (circles and triangles) in blue.

VI. Conclusions

In this paper, we have introduced and optimized a new cooperative coding scheme based on non-binary LDPC codes and the concept of non-binary repetition coding at the relays. We have shown that our scheme can reconciliate the problems usually encountered in decode-and-forward strategies, by ensuring a non-negligible coding gain at the receiver, while the joint-decoding complexity stays constant with the number of relays in the system. Additionnally, the cooperative coding scheme is independant on the channel model or on the order of the modulation used for each link in the network, which allows to keep all advantages shown in this paper with advanced link-adaptation and channel estimation techniques. This will be the purpose of a future work.

References

[1] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity. Part I. System description," *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1927–1938, 2003.

[2] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity. Part II. Implementation aspects and performance analysis," *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1939–1948, 2003.

[3] J.N. Laneman, D.N. Tse, and G.W. Wornell, "Cooperative diversity in wireless networks: efficient protocols and outage behaviour," *IEEE Trans. on Information Theory*, vol. 50, no. 12, pp. 3062–3080, 2004.

[4] M. C. Valenti and B. Zhao, "Distributed turbo codes: towards the capacity of the relay channel," in *IEEE Vehicular Technology Conference (VTC)*, 2003, pp. 322–326.

[5] M. A. Khojastepour, N. Ahmed, and B. Aazhang, "Code design for the relay channel and factor graph decoding," in *Asilomar Conf. on Signals, Systems and Computers*, 2004, pp. 2000–2004.

[6] P. Razaghi and W. Yu, "Bilayer LDPC codes for the relay channel," in *IEEE Inter. Conf. on Communications (ICC)*, 2006, pp. 1574–1579.

[7] P. Razaghi and W. Yu, "Bilayer low-density parity-check codes for decode-and-forward in relay channels," *IEEE Trans. on Information Theory*, vol. 53, no. 10, pp. 3723–3739, 2007.

[8] A. Chakrabarti, A. De Baynast, A. Sabharwal, and B. Aazhang, "Low-density parity-check codes for the relay channels," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 2, pp. 280–291, 2007.

[9] J. Hu and T. M. Duman, "Low density parity check codes over wireless relay channels," *IEEE Trans. on Wireless Communications*, vol. 6, no. 9, pp. 3384–3394, 2007.

[10] V. Savin, "Split-extended LDPC codes for coded cooperation," in *Information Theory and its Applications (ISITA), 2010 International Symposium*, October 2010.

[11] C. Poulliat K. Kasai, D. Declercq and K. Sakaniwa, "Multiplicatively repeated non-binary LDPC codes," *IEEE Trans. Infomation Theory*, vol. 57, no. 10, pp. 6788–6795, October 2011.

[12] D. Declercq K. Kasai and K. Sakaniwa, "Fountain coding via multiplicatively repeated non-binary LDPC codes," *to appear in IEEE Trans. Communications*, 2011.

[13] X. Y. Hu, E. Eleftheriou, and D. M. Arnold, "Regular and irregular progressive edge-growth tanner graphs," *IEEE Trans. Inform. Theory*, vol. 51, no. 1, pp. 386–398, 2005.

[14] M. Fossorier C. Poulliat and D. Declercq, "Design of regular (2,dc)-LDPC codes over GF(q) using their binary images," *IEEE Trans. Communications*, vol. 56, no. 10, pp. 1626–1635, 2008.

[15] A. Bennatan and D. Burshtein, "Design and analysis of nonbinary ldpc codes for arbitrary discrete-memoryless channels," *IEEE Trans. Information Theory*, vol. 52, no. 2, pp. 549–583, 2006.

[16] S. Pfletschinger and D. Declercq, "Non-binary coding for vector channels," in *SPAWC'11*, San Fransisco, CA, USA, June 2011.

[17] S. Pfletschinger and M. Navarro, "Link adaptation with retransmissions for non-binary ldpc codes," in *Future Network and Mobile Summit*, Florence, Italy, June 2010.

[18] R. M. Tanner, "A recursive approach to low complexity codes," *IEEE Transactions on Information Theory*, vol. 27, no. 5, pp. 533–547, 1981.

[19] H. Steendam H. Wymeersch and M. Moeneclaey, "Log-domain decoding of ldpc codes over GF(q)," in *ICC'04*, Paris, France, June 2004.

[20] D. Declercq and M. Fossorier, "Decoding algorithms for non-binary LDPC codes over GF(q)," *IEEE Trans. Commun.*, vol. 55, no. 4, pp. 633–643, 2007.

[21] F. Verdier M. Fossorier A. Voicila, D. Declercq and P. Urard, "Low-complexity decoding for non-binary LDPC codes in high order fields," *IEEE Trans. Communications*, vol. 58, no. 5, pp. 1365–1375, May 2010.

[22] D. Divsalar and F. Pollara, "On the design of turbo codes," *The JPL TDA Progress Report*, pp. 42–123, November 1995.

[23] J. Thorpe, "Low-density parity-check (ldpc) codes constructed from protographs," *JPL INP, Tech. Rep.*, August 2003.

[24] Lan Lan Yifei Zhang Shu Lin William Ryan Gianluigi Liva, Shumei Song, "Design of ldpc codes: A survey and new results," *Journal of Communications Software and Systems*, 2006.

[25] D. Klinc J. Ha, J. Kim and S. W. McLaughlin, "Rate-compatible punctured low-density parity-check codes with short block lengths," *IEEE, Trans. Inform. Theory*, vol. 52, pp. 728–738, 2006.

[26] D. Declercq A. Venkiah and C. Poulliat, "Design of cages with a randomized progressive edge growth algorithm," *IEEE Commun. Lett.*, vol. 12, no. 4, pp. 301–303, April 2008.

[27] T. Cover and A. E. Gamal, "Capacity theorems for the relay channel," *IEEE Trans. on Information Theory*, vol. 25, no. 5, pp. 572–584, 1979.

[28] M. A. Khojastepour, A. Sabharwal, and B. Aazhang, "On capacity of Gaussian 'cheap' relay channel," in *IEEE Global Telecom. Conference (GLOBCOM)*, 2003, vol. 3, pp. 1776–1780.

[29] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Trans. on Information Theory*, vol. 51, no. 9, pp. 3037–3063, 2005.

 ISBN: 978-1-61208-183-0

A Redundancy Information Protocol for P2P Networks in Ubiquitous Computing Environments: Design and Implementation

Rafael Dias Araújo, Hiran Nonato Macedo Ferreira, Pedro Frosi Rosa, Renan Gonçalves Cattelan
Universidade Federal de Uberlândia
Uberlândia, MG, Brazil
{*rdaraujox,hirannonato*}*@gmail.com,* {*frosi,renan*}*@facom.ufu.br*

***Abstract*—The ubiquitous computing vision brings high computational and communication demands. In this paper, we propose a high availability protocol for information replication in ubiquitous computing environments. Running at the application layer, on top of a P2P overlay network based on JXTA, our protocol allows the transfer of multimedia contents automatically generated by multimedia capture systems. We formalize the characteristics of the protocol and present its design rules and procedures.**

Keywords-Redundancy protocol; P2P networks; ubiquitous computing; multimedia content sharing.

I. INTRODUCTION

The concept of ubiquitous computing [15] has been widely used nowadays. In particular, automated systems that allow the capture of live experiences are a recurring research theme [1]. Classrooms instrumented with electronic whiteboards, microphones and video cameras produce multimedia artifacts that reconstruct the captured experience for future use and review. In order to reach true ubiquity, the produced contents should be available to users in a transparent way, i.e., independent of their physical location.

In instrumented settings for multimedia capture, we face the problem of high computational and communication demands [8]. Traditional implementations reported in the literature have scalability problems caused by centralized entities that become the system bottlenecks [3]. Satyanarayanan [10] notes that issues like remote access, high availability, power management, mobile information access have increased and, in parallel, ubiquitous and pervasive computing take advantage of distributed and mobile computing. Thereby, systems must provide high availability [13] to ensure the transparency requirements. The main aspect of high availability is the redundancy of information that must be transmitted to different points of the network by reliable communication channels.

To solve these issues, we developed a peer-to-peer (P2P) architecture for the storage and distribution of captured multimedia content. P2P networks have the potential to make the process of sharing information much easier. Studies show P2P are responsible for more than 50% of the overall Internet traffic in some regions [12]. Another key advantage of P2P networks is the direct availability of resources to the network participants, without the need for any central coordination by servers or stable hosts [11]. Furthermore, its robustness is increased because it removes the single failure point commonly observed in a client-server based solution [7].

In order to contribute to the proper capture and storage of multimedia content, it becomes indispensable the creation of a protocol that ensures communication and interoperability features for heterogeneous devices. Thus, in this paper, we propose a P2P based protocol that aims at minimizing the aforementioned problems of large data transfers and ensures the availability of the information captured from the environment. We based our approach in the general idea of cooperation among user devices joining a P2P network.

The remaining of the paper is structured as follows: in Section II, we present a real multimedia capture scenario which inspired the design of our protocol; in Section III, we detail the abstraction layer defined for the protocol; in Section IV, we describe our protocol specification and design, i.e., its environment, encoding, vocabulary, services and procedure rules; in Section V, we present the implementation details behind our approach; in Section VI, we present related works; and finally, in Section VII, we make our final remarks.

II. CLASSROOM INFORMATION CAPTURE

Consider classrooms equipped with electronic devices (e.g., mobile phones, notebooks, tablets, electronic whiteboards, video cameras, etc.) and responsible for capturing multimedia raw data from the environment. The resulting captured data, in the form of multimedia artifacts, may be useful both for instructors that need to reuse them later and for students to review what was presented.

Take, for example, iClass [9], an open-source capture platform for ubiquitous learning environments. iClass comprises a federation of capture clients and an access servers. Capture clients are software components for a particular capture device and generates a corresponding multimedia artifact (e.g., an audio capture component monitors a microphone and generates audio streams). Access servers are daemon applications which collects multimedia artifacts sent by capture clients and merges then into a single, synchronized

 ISBN: 978-1-61208-183-0

document. Users have access to those documents by using an integrated Webserver. iClass infrastructure is thus predominant based on a client-server approach: clients produce content which is sent to a server for user access.

iClass presents characteristics that provide some interesting insights for our research context:

- The captured data have widely varying formats such as video, audio, image and text.
- Data reaches large volumes over time, thus requiring scalable and high availability software and hardware infrastructures.
- Moreover, it is important that storage do not be centralized and be made in a reliable way to protect users' personal annotations.
- Finally, it is also worth to observe that, once created and stored, such data artifacts usually do not change. This characteristic allows a simpler data replication policy and reduces consistency problems.

With these characteristics in mind, we extended iClass original architecture by adapting it to the P2P paradigm. We conceived the concept of capture agents, software components with well defined interfaces that capture information and distribute it not through a single, a priori known server, but through an access service run by cooperative peers on a P2P overlay.

III. Content Abstraction Layer

In order to abstract the content transferring services, we created a layer using the widely available JXTA open-source P2P protocol [14]. The main goals of JXTA are: operating system independence, language independence and providing services and infrastructure for P2P applications.

We use JXTA services to make content transferring transparent to the application. Such approach creates an abstraction capable of aggregating and providing many services for content storage and synchronization. The proposed layer was named CAL (Content Abstraction Layer).

Thus, capture peers implement the JXTA protocol to communicate among themselves and intermediate peers may be used to route messages to external peers without direct connection. For this, there is a concept that should be explored: peer grouping. In JXTA, a peer group is defined as a collection of peers that have agreed upon a common set of services. Each peer group is identified by a unique peer group ID and each one can determine its own membership policy from open to highly secure and protected (credentials are required to join). Peers may belong to more than one peer group simultaneously.

When devices finish a capture session, they can search for an available storage service running on some peer group by using the *discovery* service. Devices need to join the network only when they wish to transfer content, i.e., they can work offline while capturing content.

IV. Protocol Specification and Design

The five essential protocol elements (environment, vocabulary, encoding, services and procedure rules) [4] are described in this session. The environment takes into account the physical and logical characteristics where the protocol is used, as architecture, computer's organization, etc. The vocabulary is the set of events (name of messages) used to specify the state transitions of protocol and the encoding the format of each message. The service is an abstract element that defines a feature and its behavior is defined by procedure rules (automaton).

CAL was designed to be used over P2P networks and each peer can perform both roles of producer or consumer. The producer is the one that captures the real world data through user devices. The consumer is the one that stores the data received from the producer and also shares them with other consumers. Thus, peers can share their contents to provide redundancy and high availability on the network.

Another aspect of this layer that should be considered is the connection-oriented networking, i.e., the upper layer must first establish a communication session with the other node and, after that, it becomes capable to deliver data in the same order it was sent [4].

After these considerations about the environment, we formalize the protocol vocabulary, which defines the semantic of messages used in communication [4]. Our protocol's set of messages consists of:

- LIST_STATUS_REQUEST: message sent to a peer to request a list of its contents. This message can be sent only by consumer peers. This is the first step for two peers to synchronize content between themselves;
- SEND_SEG_REQUEST: message used to send a content segment. When the content to be sent is to large, then they have to be broken into smaller pieces before being sent;
- SEND_SEG_RESPONSE-: this message represents a SEND_SEG_REQUEST unacknowledgement. It is sent when the last SEND_SEG_REQUEST was not recognized. Note that the positive SEND_SEG_RESPONSE is not adopted because this message is used by a service with negative acknowledgment [4];
- FT_GET_REQUEST: message used to request a specific content using its identifier obtained by calling the *list_status* service. This message can be sent only by the consumer;
- FT_PUT_REQUEST: message used to send a content that was just created. This message can be sent only by the producer;
- SEND_MSG_RESPONSE+: this message represents an acknowledgment. It is used to indicate that the last SEND_MSG_RESPONSE primitive was properly recognized. This message represents the positive response for the JXTA SEND_MSG service;

 ISBN: 978-1-61208-183-0

- SEND_MSG_RESPONSE-: this message represents an unacknowledgement. It is used to indicate that the last SEND_MSG_RESPONSE primitive was not properly recognized.

These messages are encoded by a character-oriented method and have the same format and comprising header, body and trailer.

In the message header, there is a flag named *MORE_BIT* that is used to indicate whether there are more segments to be transferred. This situation can occur when a consumer wants to synchronize to another and the second has so many files that it exceeds the maximum primitive's length, or when a producer wants to send a content so large that it need to be broken into smaller pieces. The *SERVICE* field represents the called service's name that is encoded in a binary number. As CAL has seven available services, then three bits are enough for this field in order to represent all services (2^3 = 8). Finally, there is a field to store the length of the current content and, hence, the data field can be variable. The trailer stores the data hash code for further error checking.

Now, we can define the services provided by CAL and present its behavior. As previously mentioned, the layer contains eight services: *publish*, *start_session*, *end_session*, *list_status*, *ft_get*, *ft_put*, *send_seg* and *reject*.

The first one, *publish*, is used to make a new consumer peer available on the network. It means the new peer wishes to provide its disk space to store content for other peers (acting as a server). The peer uses this service to become available for connection to other peers. This is not a confirmed service, i.e., the peer sends this primitive and do not expect a confirmation.

The *start_session* service represents the connection establishment phase from a peer to another. Regardless of the role played by the peer in the network, this phase is mandatory. Fig. 1 shows the automaton for this service. The application that uses CAL requires a connection by calling the service mentioned above. This service starts the connection establishment process. The first action is to find some available peer in the network. For this, CAL uses the *discovery* service from JXTA platform and waits the *discovery_confirmation* with the needed peer's information to allow the communication. If timeout is reached, it retries *n* times, where *n* is a previously defined constant. When the service receives a positive *discovery_confirmation*, it calls the *connect* service from JXTA and waits a confirmation, which can be positive or negative. In the first case, i.e., if it receives a positive *connect_confirmation*, it goes to the *CONNECTED* state. Otherwise, or if timeout, it tries to discover and to connect to other peer during the pre-established number of attempts.

As the previous one, the *end_session* is also mandatory and represents the disconnection phase. When the application wants to close the connection with the other peer, it calls this service and, then, CAL instantiates the *disconnect*

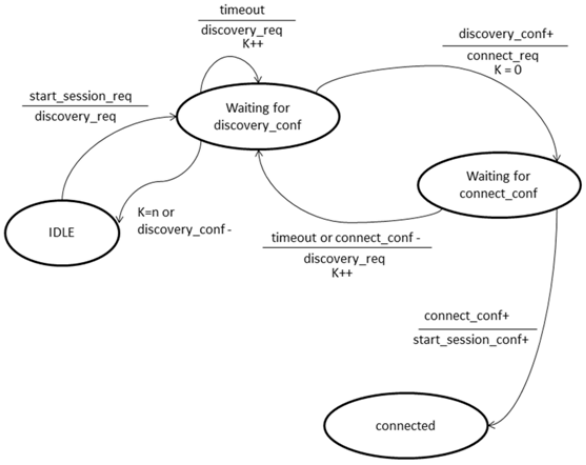


Figure 1. *start_session* service behavior

service from JTXA, returns to IDLE state and become ready to establish a new connection to another peer. This behavior is shown in Fig. 2.

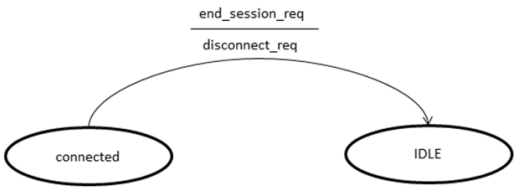


Figure 2. *end_session* service behavior

The *list_status* service is available only to the consumer peer. This service is responsible for asking the other connected peer what contents it has stored. As a result, the application obtains a list of contents identifiers. Once connected, the application sends a *ls_request* (*ls* is an abbreviation for *list_status*) and CAL sends a *send_msg_request* with the *ls_request* inside its data field. CAL waits for the confirmation that can be positive or negative. If the incoming message is a negative confirmation, it returns to *CONNECTED* state. If the incoming message is a positive confirmation, CAL sends a *send_msg_response+* to confirm the message. If the received message has any error, the *send_msg_response-* is sent to refuse the message. This confirmation message is sent inside the *send_msg_request* data field. After this message, CAL becomes ready to receive the list. There are cases that the peer has so many files that the list must be partitioned in segments to be sent as small pieces, as shown in Fig. 3. In this case, each segment is sent with the *MORE_BIT* flag equals to 1 until there is no more segments to be sent. This automaton also considers the timeout while waiting for the *ls_confirmation* message, while waiting for a segment or while receiving a segment.

For clarity, Fig. 4 shows the temporal order diagram for

 ISBN: 978-1-61208-183-0

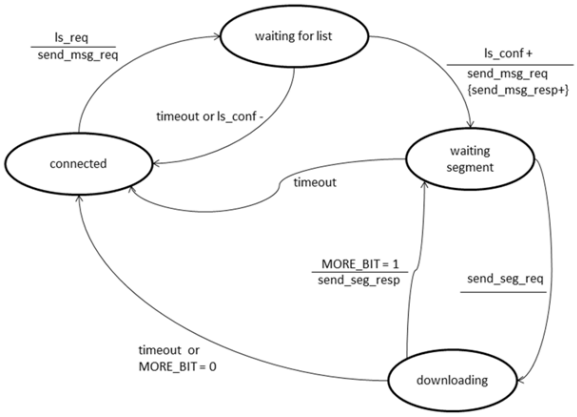


Figure 3. *list_status* service behavior

the *list_status* service. One can observe that the incoming messages at CAL are transmitted inside the data field of the *send_msg* primitive of JXTA. This behavior can also be observed in other services, once they were designed to have grater abstraction from those provided by JXTA.

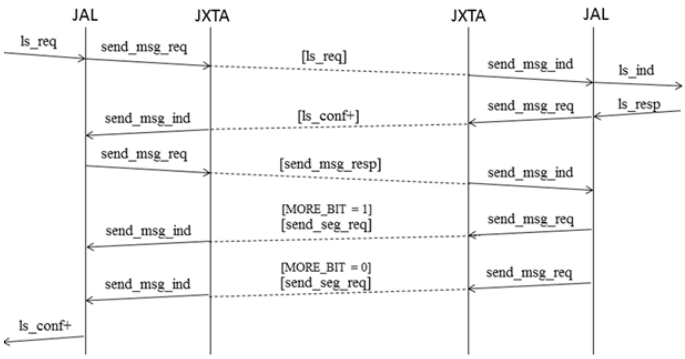


Figure 4. Temporal order diagram for *list_status*

The *ft_get* service is also only available to the consumer peer. It is used to request a specific content transfer, using the identifiers that were obtained by calling the *list_status* service. Fig. 5 shows the *ft_get* service behavior. When the application is already connected, it can request the transfer of some content by sending the *ft_get_request* primitive and, then, CAL sends a *ft_get_req* within the *send_msg_request* primitive data field. If it receives a negative *ft_get_confirmation*, it returns to the *CONNECTED* state, otherwise it sends a *send_msg_response+* message to confirm the receipt of the primitive and indicates that it is ready to receive the segments of the required content. Thus, the segments are sent until the whole content is transferred. Note the flag *MORE_BIT* to indicate that there are more segments to be sent. When the transfer is finished, CAL sends a confirmation message to the application indicating the content transferring is complete.

The *ft_put* service is available only to the producer peer.

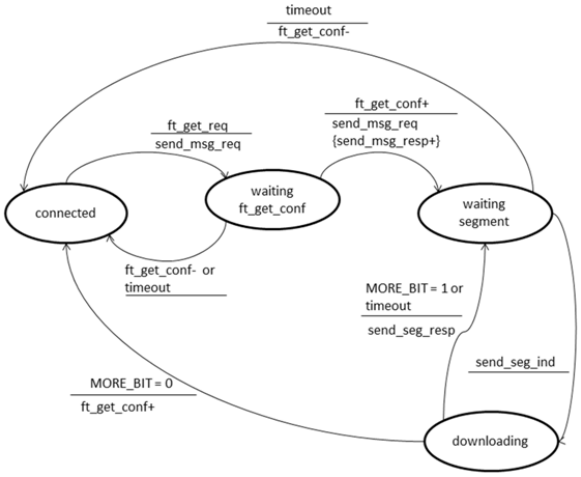


Figure 5. *ft_get* service behavior

This service is used to send contents to be stored in other peers (those who are consumers). When the application is already connected, it sends a *ft_put_request* message to CAL and waits confirmation from the consumer. If CAL receives a negative confirmation, it returns to the *CONNECTED* state and indicates to the application that the transfer has failed. This happens to let the decision of trying to resend the content to the same peer or connect to another peer as a responsibility of the upper layer. If the received message is a positive confirmation, it becomes ready to receive the content segments. Once finished the segments transferring (i.e. *MORE_BIT* = 0), CAL goes to the *CONNECTED* state and sends a positive *ft_put_confirmation* message to the application, indicating that the content upload was done. This service automaton is shown in Fig. 6.

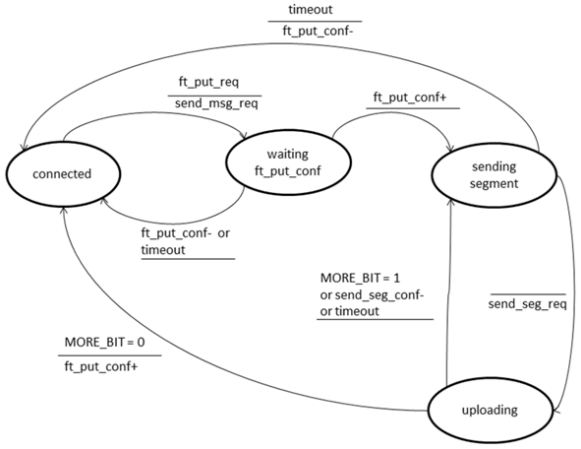


Figure 6. *ft_put* service behavior

The *send_seg* is available to both producer and consumer peers. It is responsible for sending segments of contents. In this scenario of multimedia content sharing, the data are

usually too large to be sent in a single transfer. So, data must be broken and sent in small segments, as indicated by the *MORE_BIT* flag. Its behavior is shown inside the automata of the previously presented *list_status*, *ft_get* and *ft_put* services.

An interesting behavior observed in *list_status*, *ft_get* and *ft_put* services refers to the message confirmation technique used while sending the content segments. We opted for a negative acknowledgment approach, which means that CAL do not wait for a positive confirmation for each *send_seg_request*. Instead, CAL receives a *send_seg_response*- when an error is found in the transmitted message.

Finally, the last service named *reject* is also available to both producer and consumer peers. CAL can receive faulty primitives or worse, it may receive some primitive whose service should not be recognized by the peer in question, i.e., those messages that are only recognized by peers who play a particular role. In these cases, this service must be used to refuse the mistaken incoming primitives. Eventually, these errors may cause side effects in CAL because there could be burst errors that the link layer is not able to handle in consequence of false positives. The service behavior is very simple. In every state of all services mentioned before, the layer may receive mistaken primitives from the application. Then, if it happens, CAL sends a message with the *reject* service and goes back to the same state and waits another message. Therefore, it ensures the consistency of communication between the application and CAL.

V. Implementation Details

Our initial implementation was written in Java. One of the main reasons for this choice is due to its portability and wide use by the community. In this section, we describe the prototype implementation of our P2P protocol and discuss several implementation decisions that were taken during its and design and development.

As previously mentioned, CAL sits between the JXTA platform and the P2P application. Our implementation was proposed this way in order to give full autonomy for the P2P application to create its own content transfer policies. Thus, the transferring information rules are not defined by CAL but by the P2P application.

Each service in CAL (*publish*, *start_session*, *end_session*, *list_status*, *ft_get*, *ft_put*, *send_seg* and *reject*) was implemented through a well-defined interface, allowing applications to use its functionality with the least possible effort. The implementation details of each service are shown below.

publish: This interface is invoked to provide a storage service. It uses native JXTA primitives to publish an advertisement. Advertisements are XML documents that describe network resources [14]. Our publish interface consists of only one method named `publish()` that receives the provided service's name.

start_session: This interface is used to establish a connection to a storage service available on the network. Among the various available JXTA services, this interface uses `discovery` and `connect`. One of the greatest potential abstractions of our protocol is achieved through this service, because the network search happens completely transparently to the application. It is optional to the peer querying the network to inform or not the storage service it wants to connect to. If the connection is successfully established, a JxtaBiDiPipe is created between the two peers to perform the various connections about to come. The JxtaBiDiPipe uses the core JXTA uni-directional pipes (InputPipe and OutputPipe) to simulate bi-directional pipes in the J2SE binding [14].

end_session: This interface is responsible for finalizing a connection with a storage service. The `end_session()` method is responsible for disconnecting the pipe created in `start_session()` and releasing the peer to create new connections or simply cease to exist.

list_status: The consumer peer uses this interface to know the contents of a peer storage service. The communication between the peers happens through the messaging service provided by JXTA, which is implemented by the `StringMessageElement()` method. The main methods used for this service are: `getMessageElement()`, `addMessageElement()` and `sendMessage()`. In this interface the data are trafficked through content segments. A *MORE_BIT* flag is used to indicate if there are more threads for each request. After all segments are received, the information is collated and reported to the application as previously requested.

ft_get: This interface is called by a consumer peer that wishes some content that is unavailable in its repository. Content is sent through the network in segments, thus employing the same methods used by the `list_status` service. Among the various methods used by this interface, we can mention the `check_free_space()`, which is used to check whether there is enough free disk space to receive such file.

ft_put: This interface is used by producers to send content to a storage service. As in *ls_status* and *ft_get*, contents are sent into segments in the network, using always the *MORE_BIT* flag to identify the last segment.

send_seg: This interface is responsible for segmenting the content and sends it over the network. As previously mentioned, it is used by various services of this protocol, as: `list_status`, `ft_get` and `ft_put`.

reject: This interface is called whenever a new packet is received. it is used to inform the sender that the received package has some error and then it was rejected. This interface, as well as some mentioned above, also uses the messaging services provided by JXTA to send it inside its data field.

 ISBN: 978-1-61208-183-0

VI. Related Work

The use of P2P platforms in ubiquitous computing systems is not a completely new idea. For example, eComP [6] focuses on designing P2P networks for everyday objects. eComP is a decentralized XML-based messaging system that abstracts the underlying network and communication protocol and provides services through well-defined interfaces. Similar to our JXTA-based approach, the underlying network infrastructure requires no fixed infrastructure or the support of any other entity except computing peers. However, a user-defined ID (like a familiar, personal or rational textual name) is deemed necessary to identify resources.

Hong et al. [5] developed an effective scheme to manage multimedia sharing based on specially designed profiles and a virtual community. Their multimedia sharing layer is responsible for sharing multimedia contents, and is constructed as a specially designed scheme based on locality. They focuses on an effective community construction scheme performed by community construction layer.

Barolli and Xhafa present JXTA-Overlay [2], a JXTA-based P2P middleware for distributed and collaborative systems. JXTA-Overlay allows the integration of end devices, such as sensors and personal/mobile computers, providing transparency and security for sharing, contributing and controlling available resources. JXTA-Overlay comprise a set of primitive operations: peer discovery, resource allocation, file/data sharing, discovery and transmission, instant communication, among other services. Such primitive operations can be exchanged between connected peers and support different types of applications related to collaborative activities. Besides presenting a similar layered structure, differently from JXTA-Overlay, where the application must know the identity of peers to which it desires to connect, our protocol does not require any *a priori* knowledgement of peer identities. In our proposal, CAL provides such network abstraction and still ensures the reliability and security properties offered by JXTA during content transfers.

VII. Conclusion and Future Work

We presented CAL, a P2P-based protocol designed to transfer multimedia information captured by different types of devices installed in instrumented environments. As a result, devices that produce multimedia contents do not need to have any previous knowledgement of the network to be able to transfer contents to other devices. CAL creates abstractions for discovering peers offering storage capabilities with safety and reliability properties.

Complementing our approach and as future work, we are currently developing an effective access mechanism for retrieval of the multimedia information stored in our platform. Its query approach is based on contextual preferences that uses information available about users, devices and the environment in order to automatically recommend and personalize returned content.

Acknowledgments

The authors would like to thank the Brazilian Research Agencies CNPq and FAPEMIG for supporting this work.

References

[1] G. D. Abowd and E. D. Mynatt. Charting past, present, and future research in ubiquitous computing. *ACM Trans. Comput.-Hum. Interact.*, 7:29–58, 2000.

[2] L. Barolli and F. Xhafa. Jxta-overlay: A p2p platform for distributed, collaborative, and ubiquitous computing. *IEEE Trans. on Industrial Electronics*, 58(6):2163–2172, 2011.

[3] R. Hasan, Z. Anwar, W. Yurcik, L. Brumbaugh, and R. Campbell. A survey of peer-to-peer storage techniques for distributed file systems. In *Proc. of the Intl. Conf. on Information Technology: Coding and Computing*, pages 205–213, 2005.

[4] G. J. Holzmann. *Design and validation of computer protocols*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1991.

[5] C.-P. Hong, E.-H. Lee, and S.-D. Kim. An efficient scheme to construct virtual community for multimedia content sharing based on profile in a ubiquitous computing environment. In *Proc. of the Intl. Joint Conf. on INC, IMS and IDC*, pages 1271–1276, 2009.

[6] A. Kameas, I. Mavrommati, D. Ringas, and P. Wason. ecomp: An architecture that supports p2p networking among ubiquitous computing devices. In *Proc. of the Intl. Conf. on Peer-to-Peer Computing*, pages 57–, 2002.

[7] E. K. Lua, J. Crowcroft, M. Pias, R. Sharma, and S. Lim. A survey and comparison of peer-to-peer overlay network schemes. *Commun. Surveys Tuts.*, 7(2):72–93, 2005.

[8] F. Perich, A. Joshi, T. Finin, and Y. Yesha. On data management in pervasive computing environments. *IEEE Trans. on Knowl. and Data Eng.*, 16:621–634, 2004.

[9] M. Pimentel, L. A. Baldochi Jr., and R. G. Cattelan. Prototyping applications to document human experiences. *IEEE Pervasive Computing*, 6:93–100, 2007.

[10] M. Satyanarayanan. Pervasive computing: vision and challenges. *IEEE Pers. Commun.*, 8(4):10 –17, 2001.

[11] R. Schollmeier. A definition of peer-to-peer networking for the classification of peer-to-peer architectures and applications. In *Proc. of the Intl. Conf. on Peer-to-Peer Computing*, pages 101–102, 2001.

[12] H. Schulze and K. Mochalski. Internet study 2008/2009, 2009. http://www.ipoque.com/en/resources/internet-studies. Retrieved: Dec. 2011.

[13] Sun BluePrints Online. High availability fundamentals, 2000. http://www.sun.com.br/blueprints. Retrieved: Dec. 2011.

[14] Sun Microsystems, Inc. Project JXTA v2.5: Java programmer's guide, 2007.

[15] M. Weiser. The computer for the 21st century. *Scientific American*, 265(3):94–104, 1991.

 ISBN: 978-1-61208-183-0

Kernel Module Implementation of IPv4/IPv6 Translation Mechanisms for IPv4-oriented applications

Katsuhiro Naito, Kazuo Mori, and Hideo Kobayashi
Department of Electrical and Electronic Engineering, Mie University,
1577 Kurimamachiya, Tsu, 514-8507, Japan
Email: {naito, kmori, koba}@elec.mie-u.ac.jp

***Abstract*—Only IPv6 addresses are currently being assigned to hosts because IPv4 addresses will be exhausted in the near future. However, almost all network applications still lack support for IPv6 communication. Therefore, users will suffer from the unavailability of IPv6 oriented applications. Bump-In-the-Stack (BIS) mechanisms can allow hosts to communicate with other hosts through IPv6 networks using existing IPv4-oriented applications. These mechanisms will be required to achieve a smooth transition from IPv4 to IPv6 networks in the near future. However, detailed implementation schemes are dependent upon the operating system. Additionally, since conventional network address translation mechanisms usually perform in a user space, throughput performance degrades as a result of the memory copy between kernel space and user space. Recently, Session Initiation Protocol (SIP) has been used to achieve multimedia communication. However, BIS does not support address translation mechanisms for embedded IP addresses in packet payload, such as in SIP messages. This paper presents a specially developed Linux kernel module for IPv4/IPv6 address translation supporting SIP messages. The kernel module can hook all packets in a Linux network socket using Linux netfilter mechanisms. The advantages are high throughput, as the memory copy is limited to a socket buffer in a Linux network stack, and flexible installation to an original Linux kernel. Thus, the kernel module allows users to achieve IPv4/IPv6 address translation by installing it in a generic Linux kernel, without modifying the kernel source.**

Keywords-Bump-In-the-Stack; Session Initiation Protocol; Kernel module; Address translation; Linux.

I. INTRODUCTION

The Internet will soon exhaust another IPv4 address range. Recently, the Asia-Pacific Network Information Center (APNIC) announced that the APNIC pool had reached its final /8 IPv4 address block [1]. Hence, only the IPv6 address range will be assigned to new networks in the near future.

IPv6 is the network layer protocol for the next generation Internet and offers a much larger address space. Since networking equipment vendors have developed IPv6 implementation, almost all networking equipment for enterprise networks already supports IPv6 communication. However, IPv6 is a different protocol from IPv4. Furthermore, there is still a great deal of IPv4 content on the Internet. While IPv4 devices and services continue to be widespread, it is difficult to replace IPv4 with IPv6. Therefore, we are entering a transition period during which network address translation (NAT) mechanisms will be required to communicate between IPv4 and IPv6 networks [2], [3].

Various translation mechanisms for IPv4/IPv6 have been proposed to facilitate the interoperation and coexistence of both protocols [4]. Dual-stack lite requires an IPv6 access network and tunnels between a host and a Network Address Port Translation (NAPT) 44 device, which is operated by service providers [5]. The dual-stack host, which has both an IPv4 and an IPv6 address, sends its IPv4 traffic through a NAPT44 device even though the service provider's access network is IPv6. Additionally, the host can send its IPv6 traffic routed normally. In dual-stack lite, hosts require an IPv4 address and an IPv6 address. Therefore, it will be difficult to apply in the near future because new IPv4 addresses will have been exhausted.

The other candidate mechanisms are NAT64 and NAT-PT [6], [7], [8], which are translation mechanisms where the host runs only IPv6. They are called large-scale NATs (LSNs) or carrier grade NATs. Recent NAT64 devices can serve a translation function to 10,000 subscribers, but their scalability will be limited due to the expansion of network traffic. In addition, the translations have several technical issues [9].

These mechanisms can translate between IPv4 and IPv6 packets. However, applications also need to support IPv6 addresses in order to use IPv6 networks. Moreover, since modification of the source code is required before IPv4-oriented applications can support IPv6 addresses, almost all these applications still cannot do so.

Bump-In-the-Stack (BIS) allows hosts to communicate with other IPv6 hosts using existing IPv4 applications [10]. However, the BIS implementation is unable to provide high throughput and flexible installation, while application protocols that embed IP addresses in the packet payload are not supported. Since Session Initiation Protocol (SIP) messages include host's IP addresses, translators need to modify the IP address part in such messages if they are to support SIP [11], [12]. Session Traversal Utilities for NAT (STUN) [13] and Universal Plug and Play (UPnP) [14] are well known tools in the context of a NAT traversal solution. However, they are difficult to apply for translation between IPv4 and IPv6 addresses because we have to assume that SIP client applications do not support IPv6 addresses. Therefore, the SIP Application Level Gateway (ALG) is a better solution for translation between IPv4 and IPv6 addresses.

In this paper, we develop a kernel module for Linux

 ISBN: 978-1-61208-183-0

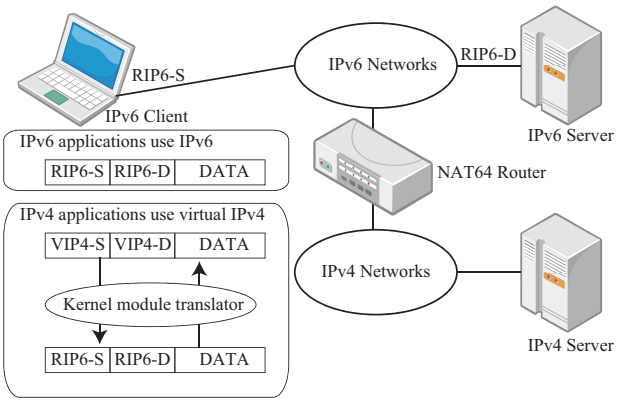


Figure 1. Overview network.

netfilter [15]. The developed kernel module can translate between IPv4 and IPv6 addresses in a header and modify addresses in SIP messages. Since the developed kernel module performs packet manipulation in a kernel space, we can achieve high throughput performance, even with IPv4/IPv6 address translation, by reducing the memory copy of packet data. Furthermore, the developed kernel module can be implemented in Linux OS without kernel modification.

II. IPv6/IPv4 Translation Mechanisms

Figure 1 shows an overview of the network presented in this paper. Here, we focus on hosts with IPv6 addresses, which represent the reality in the near future, as discussed above. As it is difficult for users to modify the source code of applications to support IPv6 networks, these hosts also have IPv4 oriented applications. However, IPv4 oriented applications cannot establish connections because the hosts do not have IPv4 addresses.

The proposed implementation provides two virtual IPv4 addresses: a source IPv4 address for a virtual network interface and a destination IPv4 address corresponding to a destination host for IPv4 applications. Thus, IPv4 applications can establish connections using virtual IPv4 addresses. During real communication, these virtual IPv4 addresses are translated to corresponding real IPv6 addresses. As a result, IPv6 applications can communicate using real IPv6 addresses, while IPv4-oriented application can communicate using virtual IPv4 addresses.

The fundamentals of IPv6/IPv4 translation mechanisms are discussed in BIS. But the implementation method is not described because it is specific to the operating system. Additionally, SIP is usually used for multimedia communications, such as voice or video conference applications. However, as these SIP applications depend on service providers, it is difficult for the user to select optimum SIP applications that will support IPv6 communication. In this paper, we extend the BIS mechanisms to support SIP applications, clarify the design for implementation, and develop a special kernel module for Linux OS.

Figure 2 shows the system model for packet manipulation in the developed kernel module. The functions of this module are classified into address translation function, payload modification function, and DNS message handling function. The kernel module uses the Linux netfilter function to handle a socket buffer for each packet. Therefore, modification of the original Linux kernel is not required in order to use the developed kernel module.

A. Virtual Interface

In the developed kernel module, instead of a real IPv6 address, IPv4-oriented applications use a virtual IPv4 address that is allocated in the network interface. Therefore, some network interface for the virtual IPv4 address is required to transmit packets with the virtual IPv4 address as a source address.

In the proposed implementation, we create a virtual network interface for the virtual IPv4 address using tun/tap interfaces. Tun is software emulation of ethernet devices and tap is software emulation of a network layer. Usually, tap is used for creating a bridge interface and tun is used for creating tunnels. However, since in our proposed implementation the virtual interface is used to assign the virtual IPv4 address, both mechanisms are available. Additionally, the developed kernel module can hook all packets from IPv4 oriented applications. Therefore, the virtual interface does not receive any packets from IPv4 oriented applications.

B. Packet hook in Linux netfilter

Netfilter provides a packet manipulation framework inside the Linux 2.4.x and 2.6.x kernel series, and it is also a set of hooks inside the Linux kernel. Therefore, kernel modules can register their callback function with the Linux network stack and the function is called when packets traverse the respective hook points. As netfilter also allows kernel modules to send the hooked packets back to the network stack, these modules can modify packet information without modification of the original Linux kernel.

In the developed kernel module, outbound packets from both IPv4 and IPv6 applications are hooked at the point NF_INET_LOCAL_OUT. In the Linux network stack, IPv4 and IPv6 are processed separately. Therefore, the developed kernel module receives both IPv4 and IPv6 packets separately from the point NF_INET_LOCAL_OUT. Whereas IPv6 packets from IPv6 applications are sent back to the Linux network stack immediately, at the point NF_INET_POST_ROUTING, IPv4 packets from IPv4-oriented applications undergo some manipulation, in respect of address translation and payload modification, before being sent back to the latter point. A similar differentiation is made for inbound packets, where the respective stack points are NF_INET_POST_ROUTING and NF_INET_LOCAL_IN. Thus, IPv6 applications engage in real IPv6 communication in the normal way, while IPv4 oriented applications perform virtual IPv4 communication through IPv6 networks.

 ISBN: 978-1-61208-183-0

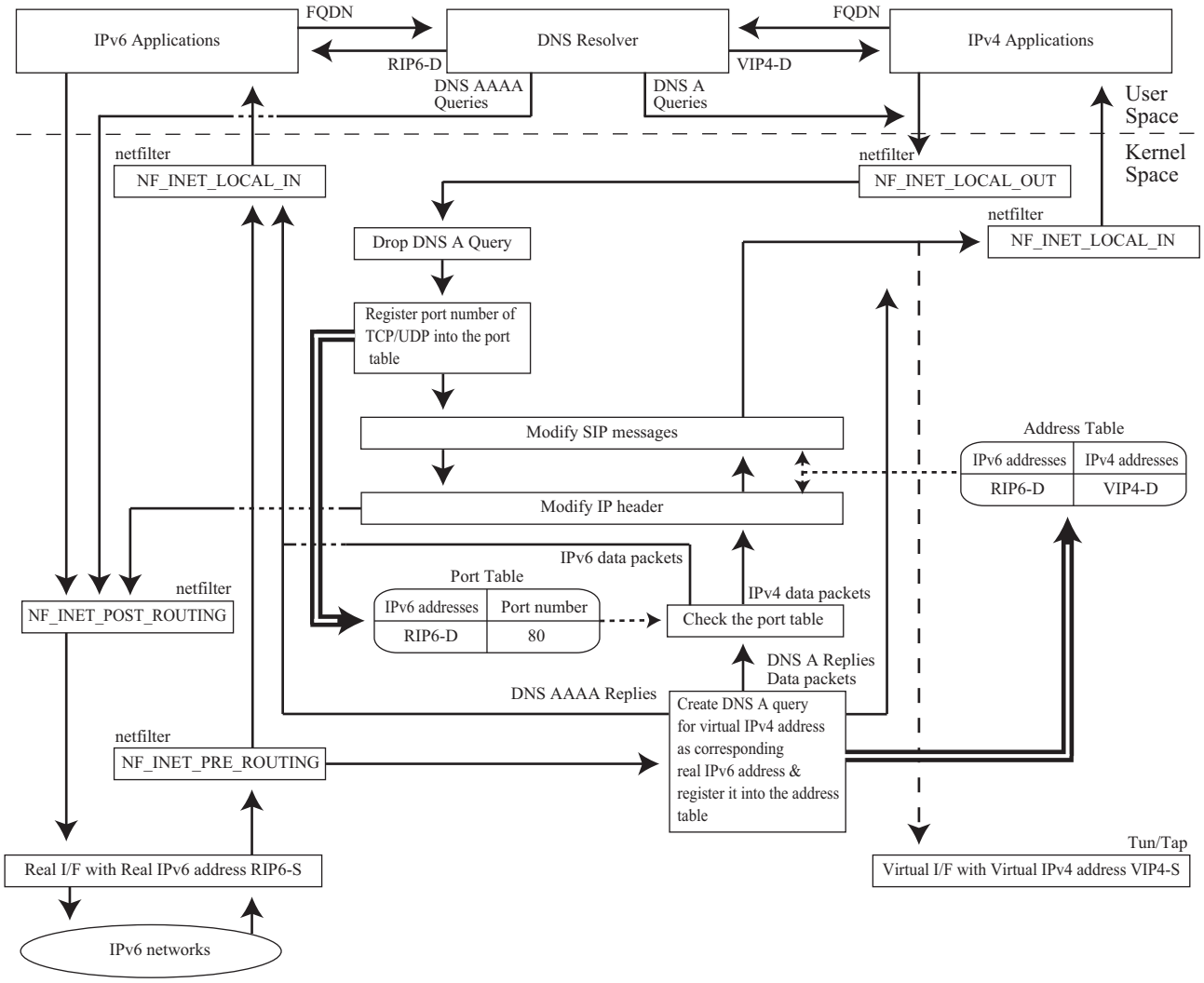


Figure 2. Design of packet manipulation in kernel module.

III. Assignment of virtual IPv4 address

In the proposed implementation, a source IPv4 address for IPv4-oriented applications is allocated as a predefined virtual IPv4 address, $VIP4-S$, and a source IPv6 address for physical network interface is allocated as a predefined real IPv6 address, $RIP6-S$. Additionally, a destination IPv4 address corresponding to a real IPv6 address is assigned dynamically when a DNS reply message is received. These IPv4 addresses consist of private addresses and are used internally in the operating system; thus, the address assignments do not negatively affect other hosts. This subsection describes the procedure for virtual IPv4 assignment.

A. Translation of DNS messages

- Resolution of Fully Qualified Domain Name (FQDN)
  When IPv4 oriented applications try to communicate with a server host, the DNS resolver transmits a DNS AAAA query and a DNS A query to find an IP address corresponding to the FQDN.
- Discard of DNS A query
  The DNS A query is meaningless, because the host does not have a real IPv4 address. Therefore, the transmitted DNS A query is dropped in the kernel module.
- Creation of a virtual IPv4 address
  A new virtual IPv4 address corresponding to the real IPv6 address is required as a destination IP address for the IPv4-oriented application. Therefore, the DNS AAAA reply corresponding to the transmitted DNS AAAA query is hooked by the kernel module when it is received from the physical interface. The kernel module creates a new virtual IPv4 address, $VIP4-D$, corresponding to the real IPv6 address, $RIP6-D$, in the DNS
- Registration of the IPv4/v6 address pair
  Since information about the pair of virtual IPv4 address

 ISBN: 978-1-61208-183-0

and the real IPv6 address is required in order to modify the IP header, the kernel module registers the pair of them in the address table.

- Response of virtual IPv4 address
  The DNS resolver returns the virtual IPv4 address corresponding to the FQDN by creating a DNS A reply. As a result, the IPv4-oriented application can communicate with the IPv6 host using the source virtual IPv4 address, $VIP4-S$, paired with the destination virtual IPv4 address, $VIP4-D$, while the host can communicate with the IPv6 server host using the source real IPv6 address, $VIP6-S$, paired with the destination real IPv6 address, $VIP6-D$.

B. Translation of IPv4/IPv6 addresses

This paper assumes that both IPv4-oriented applications and IPv6 applications communicate via IPv6 networks. This subsection describes the process for translation of IPv4/IPv6 addresses in the developed kernel module.

- Registration of IPv4 applications
  The developed kernel module receives IPv6 packets for both IPv4-oriented applications and IPv6 applications. However, the IPv6 packets received do not have information corresponding to the IP version of the destination application. Accordingly, the kernel module employs a port table, where a destination IPv6 address and a port number are registered when the kernel module receives IPv4 packets from IPv4-oriented applications.
- Modification of transmitted packets
  The kernel module handles transmitted IPv4 packets as a socket buffer in the Linux network stack. Since the header size of IPv4 is different from that of IPv6, the kernel module extends the header space in the socket buffer and modifies the header information of IPv4 to conform to that of IPv6. As in the case of BIS mechanisms, the developed kernel module cannot be used with IPv4 applications that use any IPv4-specific option.
- Selection of received packets
  The kernel module checks the port table to determine whether the received packets are destined for IPv6 applications or IPv4 applications. Packets for IPv6 applications are sent directly back to the point NF_INET_LOCAL_IN. For IPv4 applications, however, before being sent back to that point, the packets undergo a process that reduces their header space in the socket buffer and modifies the IPv6 header information to conform with that of IPv4, according to the address table.

IV. Translation mechanisms for SIP

In usual network address translation, only IP addresses included in header information are modified. However, complete IP conversion also requires the translation of IP addresses embedded in application layer protocols, such as those found in File Transfer Protocol (FTP) and SIP. Since some implementations of FTP already support IPv6, FTP applications will be available in IPv6 networks. However, almost all SIP applications still do not support IPv6. Additionally, since the profiles of SIP applications depend on those of the service provider, it is difficult to modify SIP applications to suit a user's preferences.

The developed kernel module also supports translation mechanisms for SIP applications. The general application level 2 gateway for SIP applications only converts IP addresses within private networks, whereas the mechanisms proposed here need to convert IP addresses from virtual IPv4 addresses to real IPv6 addresses. Therefore, the conversion point in the packet payload is different from the usual cases.

Since messages from networks include real IPv6 addresses in the packet payload, the kernel module needs to convert real IPv6 addresses to virtual IPv4 addresses. In addition, messages from applications include virtual IPv4 addresses in the packet payload. Therefore, the kernel module also needs to convert virtual IPv4 addresses to real IPv6 addresses.

The following fields are converted in the kernel module.

- Via header
  A via header includes a client's host name or an IP address, and a port number at which it wishes to receive responses.
- Record-Route header
  A Record-Route header field includes a host name or an IP address of a proxy. It is usually used to log SIP traffic so as to charge a usage fee.
- Contact header
  A Contact header field value may contain a display name, a URI with URI parameters, and header parameters. It indicates a response host.
- Body
  A Session Description Protocol (SDP) field includes a client's host name or an IP address. It provides information about session identification and the types of data communication used in the session.

Table I
PERFORMANCE EVALUATION PARAMETERS.

OS	Linux
Distribution	Ubuntu 10.04
Kernel version	linux-2.6.32-24-generic
CPU	Intel Pentium 4 2.40GHz
Memory	512 MBytes
Application	iperf, nuttcp
Size of transferred data	200 MBytes
Transport protocol	TCP
Number of measurement	10

V. Numerical results

In order to evaluate the design of the developed kernel module, we measured throughput, standard deviation of throughput, and round trip time by changing the size of the maximum segment size (MSS) of the Transmission

 ISBN: 978-1-61208-183-0

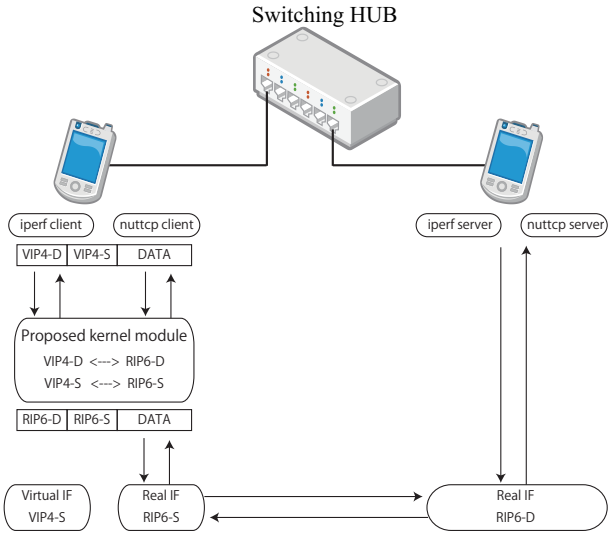


Figure 3. Evaluation model.

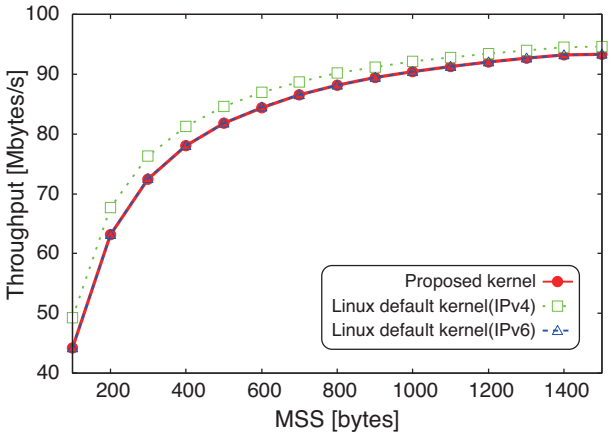


Figure 4. Throughput performance(iperf).

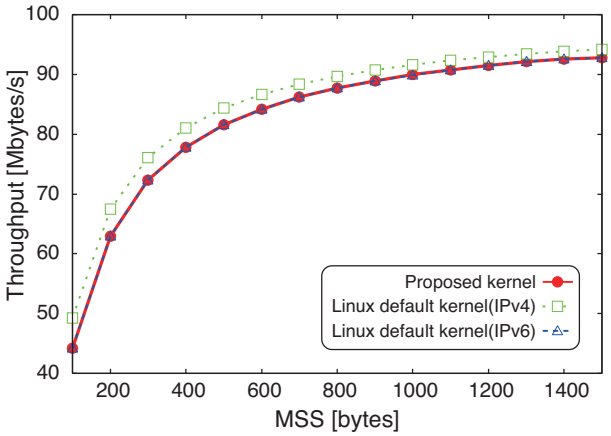


Figure 5. Throughput performance(nuttcp).

Control Protocol (TCP). The measurements were made using iperf [16] and nuttcp [17], which are well-known network benchmark tools. The purpose of this evaluation was to confirm the packet manipulation overhead, because extension or shortening of a header may result in a big overhead. Figure 3 shows the evaluation model, in which two hosts communicated 50 with each other through the developed kernel module. The virtual interface is constructed by tun during the measurements. In addition, throughput overhead may depend on MSS size, because the ratio of header size to total packet size will be larger when the MSS size decreases. Hence, we evaluated the throughput of certain sizes of MSS. From this evaluation, we were able to determine the packet manipulation overhead in the proposed implementation. Details of the evaluation parameters are given in Table I.

Figures 4 and 5 show the throughput performance as the size of the MSS changes. The results are an average of ten measurements and show that the throughput of the developed kernel module has almost the same level of performance as the general Linux kernel. The reason that the throughput of the IPv4 general Linux kernel is slightly better than that of the IPv6 general Linux kernel is the difference in total packet length due to the header sizes of IPv4 and IPv6. In addition, because the ratio of header size to total packet size increases as MSS size decreases, the deficit also increases for smaller MSS.

Figures 6 and 7 show the standard deviation of throughput performance as the size of the MSS changes. The results show that the performance of the developed kernel module is similar to that of the general Linux kernel. This means that the load of the developed kernel module does not affect the network performance.

Figure 8 shows the Round Trip Time (RTT) as the size of the MSS changes. The results show that the developed kernel module has almost the same round trip time as the general Linux kernel. This means that the packet manipulation mechanism in the developed kernel module does not take much time and does not affect the transmission delay for communication.

These results indicate that using the Linux kernel module for netfilter can achieve high throughput and short processing delay.

VI. Conclusion

This paper presents a newly developed kernel module that performs IPv4/IPv6 address translation for IPv4-oriented applications. This kernel module provides virtual IPv4 addresses to IPv4-oriented applications, enabling them to communicate with IPv6 hosts through IPv6 networks. Since the kernel module can be implemented without modification of the general Linux kernel, it can easily be used to support IPv4 applications in IPv6 networks. The developed kernel module also supports an application level gateway for SIP messages. The packet manipulation of the proposed implementation takes place in Linux kernel space, thus achieving

 ISBN: 978-1-61208-183-0

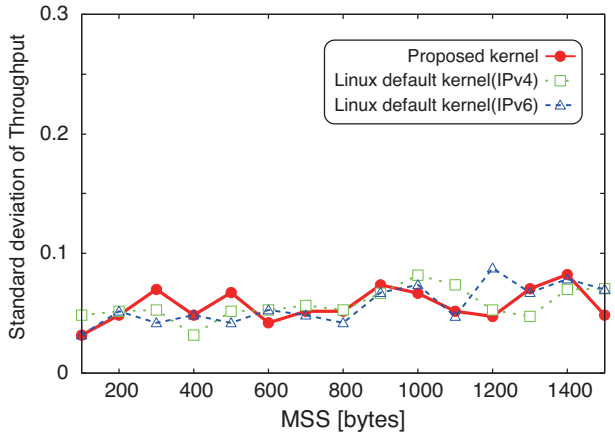


Figure 6. Standard deviation of throughput(iperf).

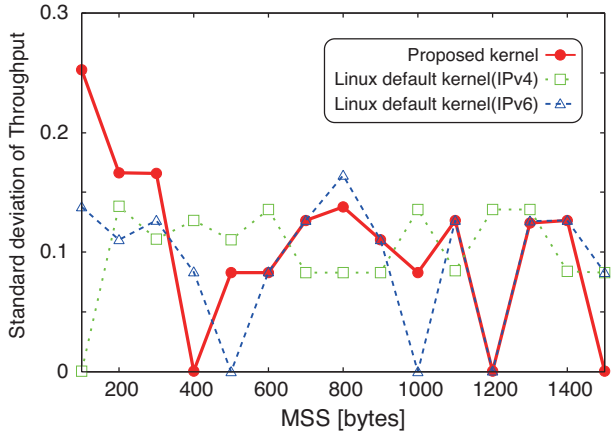


Figure 7. Standard deviation of throughput(nuttcp).

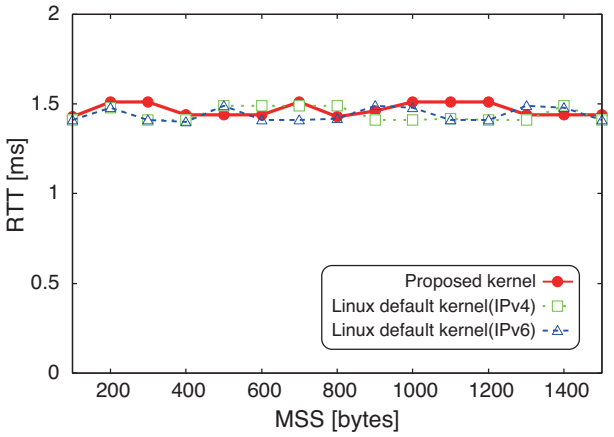


Figure 8. Round Trip Time(nuttcp).

a high throughput performance and short processing delay by reducing memory copy in operating systems.

ACKNOWLEDGMENT

The authors thank Tokyo System House Co., LTD. for the valuable comments. A part of this research was supported by Japan Science and Technology Agency.

REFERENCES

[1] http://www.apnic.net/publications/news/2011/final-8, retrieved: Dec., 2011.

[2] D. Wing, "Network Address Translation: Extending the Internet Address Space," IEEE Internet Computing, Vol. 14, No. 4, pp. 66 – 70, July-Aug. 2010.

[3] E. Nordmark and R. Gilligan, "Basic Transition Mechanisms for IPv6 Hosts and Routers," IETF RFC 4213, Oct. 2005.

[4] Y. Xia, B. S. Lee, C. K. Yeo, and V. L. S. Seng, "An IPv6 Translation Scheme for Small and Medium Scale Deployment," 2010 Second International Conference on Advances in Future Internet (AFIN), pp. 108 – 112, Jul. 2010.

[5] A. Durand, R. Droms, J. Woodyatt, and Y. Lee, "Dual-Stack Lite Broadband Deployments Following IPv4 Exhaustion," IETF draft, May 2011.

[6] M. Bagnulo, P. Matthews, and I. van Beijnum, "NAT64: Network Address and Protocol Translation from IPv6 Clients to IPv4 Servers," IETF draft, Mar. 2009.

[7] G. Tsirtsis and P. Srisuresh, "Network Address Translation - Protocol Translation (NAT-PT)," IETF RFC 2766 Feb. 2000.

[8] F. Baker, X. Li, C. Bao, and K. Yin, "Framework for IPv4/IPv6 Translation," IETF draft Aug. 2010.

[9] C. Aoun and E. Davies, "Reasons to Move the Network Address Translator - Protocol Translator (NAT-PT) to Historic Status," IETF RFC 4966, Jul. 2007.

[10] K. Tsuchiya, H. Higuchi, and Y. Atarashi, "Dual Stack Hosts using the "Bump-In-the-Stack" Technique (BIS)," IETF RFC 2767, Feb. 2000.

[11] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler, "SIP: Session Initiation Protocol," IETF RFC 3261, Jun. 2002.

[12] W. Chen, C. Su, and J. Weng, "Development of IPv6-IPv4 translation mechanisms for SIP-based VoIP applications," 19th International Conference on Advanced Information Networking and Applications, AINA 2005. vol. 2 pp. 819 – 823, Mar. 2005.

[13] J. Rosenberg, R. Mahy, P. Matthews, and D. Wing, "Session Traversal Utilities for NAT (STUN)," IETF RFC 5389, Oct. 2008.

[14] ISO/IEC 29341-1:2008, UPnP Device Architecture – Part 1: UPnP Device Architecture.

[15] http://www.netfilter.org, retrieved: Dec., 2011.

[16] http://iperf.sourceforge.net, retrieved: Dec., 2011.

[17] http://www.lcp.nrl.navy.mil/nuttcp, retrieved: Dec., 2011.

 ISBN: 978-1-61208-183-0

A Complementary Approach for Transparent NAT Connectivity

Lucas Clemente Vella, Lásaro Camargos, and Pedro Frosi Rosa
Computing Faculty
Federal University of Uberlândia
Minas Gerais, Brazil
Email: lvella@comp.ufu.br, {lasaro, frosi}@facom.ufu.br

***Abstract*—NAT has been responsible for the survival of IPv4 and in essence should not be left out in IPv6. NAT are virtually transparent to client-server applications that generally do not require special configuration to work properly. However, P2P applications are responsible for generating about half of Internet traffic and require special settings on home routers to support outside connections. This paper presents a complement to the UPnP IGD protocol, including changes in the core of the Linux operating system, to make NAT traversal transparent to home and small office users in the use of P2P applications or in providing services to the outside world. Our approach overcomes some of the major limitations of NAT solutions, by extending existing standard behaviours. In the proposed solution, current applications need no changes once the transparency is provided through the improvement made in the network related system calls. Tests using a reference implementation and network applications supports the feasibility of the approach.**

Keywords-Network Address Translation; NAT traversal; UPnP; home networks.

I. INTRODUCTION

Network Address Translation (NAT) is a commonly used tool to bridge Local Area Networks (LAN) to the Internet, effectively allowing multiple clients to share one valid Internet link. In the usual setup, there is a single public IP address, which can be reached from outside the LAN, and multiple private IP addresses used by the local clients to communicate among themselves. When a local client tries to reach an Internet host, the device in the role of Internet gateway (usually, a small router) performs the necessary address translations, so the message is transparently forwarded via the single public address.

For the network usage pattern of conventional client applications, where the client always initiates the communication with a server, NAT is transparent. Once the first contact is made from inside the LAN, the Internet gateway is able to automatically handle the responses from the server, turning the translation step invisible to most ordinary TCP/IP clients.

While restricted to servers and datacenters in the dawn of the Internet, programs who wait for incoming connections are increasingly more frequent to the users of NAT, namely the home and office Internet users, specially with the great popularity of Peer-to-Peer (P2P) software. To these programs, NAT is not totally transparent and requires explicit network configuration in the gateway to be able to receive external requests.

In order to let P2P and server applications receive connection from the Internet in the presence of NAT, a technique known as NAT traversal must be implemented. The burden of implementing NAT traversal, however, is placed either on the user, that must have the expertise to configure his equipment, or in the software developer, who must support the protocol to configure the router, increasing the development cost.

We argue that NAT transparency should be taken one step further, and the role of traversing NAT should be pushed inside the network stack, being performed by the operating system. In this paper we present a proof of concept extension to the Linux operating system to provide transparent NAT traversal through the standard network API.

In Section II some related work is reviewed. In Section III the proposed integration of NAT traversal with the network stack is explained. In Section IV the reference implementation is detailed along with its protocols. Section V enumerates the test cases for the implementation and its results. Finally, Section VI draws some conclusions and proposes directions on which this work might be improved.

II. BACKGROUND

With the usage of NAT in home and office gateways, services provided internally are not readily accessible from the outside. This problem has spawn a number of solutions, that targets various levels of abstraction in the network stack.

A. Other Home Network Solutions

The HomeDNS [1] approach works on the level of resource names, with emphasis on HTTP services, which are common for multimedia streaming applications. It provides a dynamic Domain Name Service (DNS) solution that is able to reference, from the external network, the multiple services available inside the home network, which in turn are used to build the URLs for the HTTP requests. This work concerns itself with augmenting to the Internet the reach of HTTP services targeted at the LAN. Differently, ours addresses connectivity issues of any TCP or UDP applications already targeted to the Internet.

Next to HomeDNS, the solution presented in [2] provides means to expose local services of home networks in remote

 ISBN: 978-1-61208-183-0

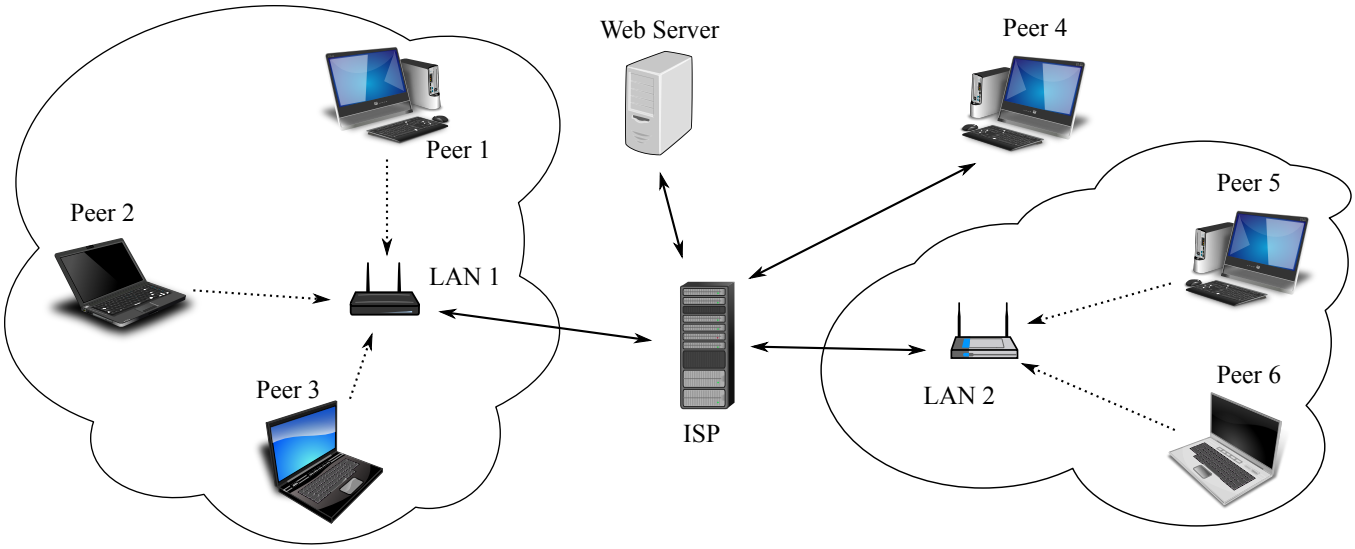


Figure 1. Reachability among peers behind NAT

guest networks. The work is specially focused on UPnP media servers and services targeted to the inside network, which access is intentionally restricted by the local network and requires tight control on its exposure. The difference of our solution it that ours is targeted to services that *should* be externally exposed, but are not due to the network topology.

In [3], some requirements of home gateways are identified. The work tries to address the issues found at home networks by designing a new home gateway, that is, a replacement for the gateways that we currently find in the market. It would provide the means to access internal services from the Internet, but would not dive into details on how it should be done with existing applications behind NAT, except that saying that UPnP IGD protocol could be, in the future, used to perform NAT traversal. Our approach is more pragmatic in that we solve only one problem, namely, NAT traversal, and do so without pushing for new equipment.

B. Internet Protocol v.6 (IPv6)

It has become a common practice among Internet Service Providers (ISP) to assign a single IP addresses to small customers, such as households and small offices. While the practice is somewhat justified by the IPv4 address exhaustion problem, it is possible that even with the eventual widespread adoption of IPv6 [4] (which eliminates the exhaustion problem) the ISP may still assign single addresses to their Small Office/Home Office (SOHO) customers. The Brazilian Internet Steering Committee (CGI.br) states, in a website dedicated to IPv6 adoption, that this is an acceptable practice [5]. In such a scenario, if multiple clients are to connect to the Internet, then there must be a way to share one valid Internet link with these clients; the Network Address Translation has become the *de facto* standard for doing so.

If NAT is still to be useful in an IPv6 world, then NAT drawbacks will persist, and the solution presented in this work will still be relevant, even without the address exhaustion problem.

C. NAT Transparency

NAT is an abstraction. One network element, the Internet gateway, is aware of the abstraction and hides the address translation complexity behind the standard interface of UDP/TCP and IP. Network applications unaware of NAT are functional as long as they do not need to expose any service to the external network. Upon this need, the effect of being behind NAT is felt.

Consider the Fig. 1, depicting a directed graph whose reachability from one node to another means the ability of this node to initiate a communication with the other via the Internet. While Peer 1 needs are nothing but to access the Web Server, the NAT taking place on LAN 1 router will not be felt. There is no problem, either, if Peer 2 wants to download a file via some P2P application from Peer 4, because, since it is a initiative from Peer 2, it will be able to open the needed TCP connection to Peer 4. The problem occurs on the opposite. By its own initiative, Peer 4 will not be able to establish the TCP connection of the P2P application to Peer 2, because the farthest Peer 4 can address is the LAN 1 router. Worse, Peers 1, 2 and 3 are invisible to Peer 5 and 6 (and *vice-versa*) on a P2P environment. Same problem if Peer 3 tries to join a game session hosted by Peer 6, because it has nowhere to send the first UDP packet that would let it into the game, once Peer 6 IP address is masqueraded by LAN 2 router.

To counter these problems, the router must be configured to follow-up TCP connection attempts or unknown UDP packets to specific hosts and ports inside the network, and

 ISBN: 978-1-61208-183-0

this host is the one running the application responsible to deal with the request. One should note that this is not simply IP routing, since the destination address of the IP datagram is the one of the gateway itself (which is public), not the one of the application's host (which is private).

Listening applications are those that either expect for the incoming of TCP connections or the first contact from the remote UDP peers. The need of some listening applications to be externally reachable started to weaken the NAT abstraction. The widespread use of NAT made it a concern to users and developers of those applications. It is now common to find applications implementing protocols to configure automatically gateway equipments on regarding NAT. It is also common to find advanced P2P users aware of the issue and experienced in manual NAT traversing setup.

The current scenario is that there are NAT-aware users running NAT-aware applications on top of *unaware* operating systems communicating through an Internet gateway implementing the NAT technique; whose design goal is to be invisible. As stated by NAT's RFC:

> "Basic Network Address Translation or Basic NAT is a method by which IP addresses are mapped from one group to another, transparent to end users." [6]

D. UPnP and NAT Traversing

Listening applications' developers found in the protocol commonly known as *UPnP* the means to hide the complexity of NAT traversing from their users. UPnP, which stands for Universal Plug and Play, is a set of protocols for discovery and automatic configuration of home networks. Initially developed by Microsoft [7], it is now maintained by the industry consortium named UPnP Forum [8].

UPnP protocols are built on top of HTTP and its UDP version, HTTPU, where its messages are XML based. As such, UPnP protocols are application level protocols, with relatively high overhead and complexity. This design choice renders unpractical the implementation of the protocols in low-level software, like operating system kernels, because the software stack providing those base technologies are often unavailable at this level.

Among the provided protocols, there is the UPnP Internet Gateway Device Protocol (UPnP IGD), whose goal is to control and configure small network gateways. Despite numerous security flaws in many and popular implementations [7], [9], [10], it became the most well supported NAT traversing mechanism by applications and routers. The competing NAT-PMP [11] protocol, despite being much simpler, is young and still does not have the availability of UPnP IGD among off-the-shelf devices.

UPnP IGD Protocol plays an important role in SOHO local networks, because it is the protocol being simply referred as UPnP by the listening applications implementing it, rendering it one of the most common NAT traversal techniques available.

The term UPnP is also commonly used to refer to a functionality present in networked multimedia applications. This usage of the term is a shorthand for UPnP Audio/Video, which is another set of protocols developed by the same UPnP initiative targeted to multimedia streaming, but is otherwise unrelated to the UPnP IGD, which is the one relevant to this work.

When using UPnP, the listening applications inside the network shares the same TCP or UDP address space, allocated in the gateway. Thus, if one application exposes one TCP port to the Internet, another LAN application willing to listen on the Internet must choose another port. This usually does not results in a port scarcity issue, since UPnP is meant to be used in small home and office networks. The 16 bit address of a port is often enough to serve all the Internet applications of these small networks.

The same port can not be shared between applications, as it is done with port multiplexing by some NAT devices. Port multiplexing uses extra previously known information, such as source port and address, to demultiplex an incoming message. When listening to the new connections, there is no such previous information available when an unknown packet arrives, thus the only mean to identify the intended receiver inside the network is the port.

Not only UPnP IGD, the *de facto* standard technique employed as NAT traversal, but also NAT-PMP and other techniques have the drawback of needing support in a per-application basis, a cost paid by the developer. Not all applications have the UPnP feature, especially the legacy ones. Developers may lack the resources to implement the feature in a project, but even if they do, it is an extra functional requirement to be taken into account.

III. The Proposed Approach

In order to make NAT transparent to listening applications, they must be able to use the bare TCP/IP interface of the operating system to wait for contact from the outside network, in the same way client applications may just connect. There should be no extra cost in the development of listening application related specifically to NAT traversal. Users and developers of client applications need not to worry if the host is behind a NAT, neither should listening applications' developers and users.

To achieve this goal, operating systems must be aware of the NAT issue. Since they are already responsible for interfacing with the *sockets* API, the one used by the applications to reach the TCP/IP functionality, it has all the means to automatically manage the gateway in place of the application or, in worse cases, the user.

A. Connection and Disconnection

Upon a `bind()` system call made on a TCP or UDP socket, the operating system may use the same protocols that applications explicitly use to forward the ports on the

 ISBN: 978-1-61208-183-0

gateway to themselves. Since UPnP or other NAT traversal technique is to be implemented by the operating system, this burden is then taken from the application. In the same way the operating system abstracts away the complexities of TCP/IP, it shall also take care of any NAT traversal employed.

When the execution flow is returned to the application by `bind()`, the operating system shall have already attempted to forward the requested port on the router. The return value is dependant on the outcome of this attempt. In case the port is being used by another host, port forwarding operation fails and the operating system must also fail the system call. *This way* the application can perform its default behavior in case of port already in use.

Upon the closing of the socket, either explicit or by process exit, the operating system must automatically remove the port association in the gateway. Unlike the creation of a port forwarding, the removal operation shall be performed asynchronously. Since port forwarding removal can not affect the outcome of the `close()` or `exit()` operation, there is no need to synchronize them.

Applications that automatically forward ports may fail to cleanup their associations on the gateway when no longer needed. It may be so either in case of application crash or because of bad implementation of the port forwarding protocol. A beneficial side effect of our approach is that, since the port forwarding is automatically managed by the network infrastructure, the needed cleanup is performed as long as the system is running, even if the application crashes.

B. Security

The proposed automatic management of NAT traversal targets end-user applications. Due to security concerns, it is important that system daemons and servers which require fine administrative control, such as FTP, HTTP, Telnet and SSH are *not* automatically exposed to the external network. For this reason, associations made by processes on UDP or TCP ports bellow 1024 should be filtered and not configured on the router. The services previously mentioned are by default bound to these ports, and privileges given by the system administrator are needed to use them. Since no common user's application shall use the privileged ports, the activity on them is out of our scope.

Applications may also choose what IP address available in the system to use when binding a socket. As a placeholder meaning *any address available*, an application may use the fake address 0.0.0.0 (aliased as `INADDR_ANY` in POSIX systems). Applications that choose to bind to specific interfaces usually know their intended peers and hold fine control of the network topology, often being manually configured on what IP interfaces to use. Upon this case, we consider that association not generic enough to be automatically forwarded by the gateway, even if the specified address is the route to the gateway. Internet applications do not try to restrict their reachability. If they are to be seen in the Internet, their logical choice of IP interface is *any* (or 0.0.0.0). Should a program or user try to control the connectivity by choosing what interface to use, then we shall not take this control by automatically exposing it on the external network.

One may question that, being the task of opening ports on the router automatic, the system would be more vulnerable to viruses and malicious software. With our approach implemented, a virus would be able to expose the system to the external network, when otherwise the system would only be exposed to the internal network. This is not indeed the case, and a virus might well find its route through a NAT in the same way a legitimate application could do, simply by implementing the same protocol that we use, with no different clearance.

IV. The Implementation

Our reference implementation consists of an extension to the Linux kernel, together with an ancillary user space daemon to handle the gateway configuring protocol [12], resembling a microkernel architecture where system functions are performed by isolated special processes. It only affects applications using sockets API to access to either UDP or TCP on top of IPv4.

This is fundamentally different from other UPnP IGD solutions for GNU/Linux, such as LinuxIGD [13], PseudoICSD [14] and MiniUPnPd [15], in a way that these packages are just plain userspace applications that implements the server part of the protocol, *i.e.* they are used to turn a GNU/Linux NAT router into a UPnP enabled gateway. The focus here is to implement the client side operated by the kernel, and UPnP server implementation is out of scope.

We choose Linux because of its popularity and its source code availability that allows us to do the kind of low-level modification needed. The protocol we use to traverse NAT is the UPnP IGD Protocol, because it is the most well supported by small routers.

A. Changes to Kernel

Inside the kernel, every call to the POSIX system call `bind()` performed on a TCP/UDP IPv4 socket is intercepted. The calls made to privileged ports or to specific IP interfaces are ignored by the automatic port forwarding mechanism. Otherwise, packets to the given port arriving at the Internet gateway must be forwarded to our host. The kernel delivers the bind request to the helper user space daemon responsible for setting up the forwarding. The calling process is put into a sleep state while awaiting the answer from the user space daemon. When the answer arrives, the process is awaken and deals with it. The `bind()` system call may then resume, failing or succeeding in according to the answer received.

To avoid a race condition, the port is preallocated internally before the control is given to the daemon. Otherwise, at least one scenario could lead to an inconsistent state. Consider it: process A tries to bind to address 0.0.0.0 on TCP port

 ISBN: 978-1-61208-183-0

6881, and is put to sleep while awaits the answer from the daemon, then process B tries to bind to address 127.0.0.1 on TCP port 6881. If the TCP port 6881 was not preallocated to process A, the bind on process B succeeds before process A receives its answer from userspace. When A receives the answer, the `bind()` call will no longer be able to succeed as the port is already in use, but it would have already been successfully configured on the gateway for a process that can no longer answer on it.

Linux provides a number of different ways of communicating between kernel space and user space. In order to pass the bind request on to the daemon, we choose to use the Netlink protocol. Netlink is a Linux specific protocol on top of sockets API and network stack, meaning the applications can use it through the usual socket related system calls. It is not a true network protocol, processes may use it only to communicate with the kernel or other processes in the same host.

Netlink was chosen for the sake of simplicity. It is very easy to add a custom protocol on top of Netlink, providing a well definite interface for user level programs, as well as for kernel code. For our purpose, we defined a new protocol called `NETLINK_NAT_PASS`.

Other kernel $\Leftrightarrow$ userspace communication methods are not as fit for our purpose as it is Netlink. For instance: we can not use system calls because they are unidirectional, and unlike Netlink, requests can not be sent from kernel to userspace as required by our architecture. Also, unlike *procfs* [16], *sysfs* [17] and other similar file based interfaces, Netlink require no changes on the filesystem, since it have its own namespace.

The processes awaiting for the daemon are placed in a linked list, that is traversed when an answer is issued by the daemon. There is no explicit guarantee that the first made request will be the first answered by the daemon, but that is the likely scenario, with no indications on how it could be otherwise. Since process are queued in the list in the same order they are sent to the daemon, when an answer arrives it will probably be referent to the first process in the list, making the search practically constant.

B. User Space Daemon

Because of the complexity of UPnP IGD Protocol, being a high-level application protocol on top of web services and HTTP, we choose to use it from user space instead of directly inside the kernel. The daemon we called *Natbinder* is responsible for controling the gateway via UPnP IGD. To build this daemon we used the UPnP IGD Protocol implementation from the *MiniUPnPc* routines library [15].

Upon startup, the daemon searches the network for some UPnP enabled Internet Gateway Devices and gets its local host private IP address. This address is used to construct the UPnP requests.

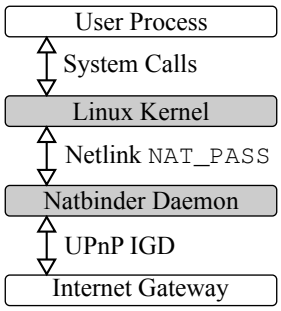


Figure 2. Logical communication stack

After verifying that it is able to reach the gateway, the daemon registers itself on the broadcast channel 0 of the `NETLINK_NAT_PASS` protocol with a Netlink socket. In this channel it will listen for kernel messages regarding IPv4 binding activities of the processes.

On a `bind()` attempt by a process, the action *AddPortMapping* is issued to the gateway. Among the responses specified by the UPnP standard [18] we may receive, two are of particular interest: code 0, meaning success and code 718 (*ConflictInMappingEntry*), meaning that the port requested by the process was already in use by another host. In those cases, the answers given to the kernel are, respectively, to proceed successfully or to fail the bind.

In case we receive a different answer, it is treated as an exceptional condition, which the system is unprepared to handle and unable to further help the binding process or the user. In this case, the message sent back to kernel is that the daemon ignored the bind request. The practical effect is the same as processed successfully, since forbidding the process to use the port will do no good in this case. The exceptional condition is logged by the daemon for manual investigation of the system administrator.

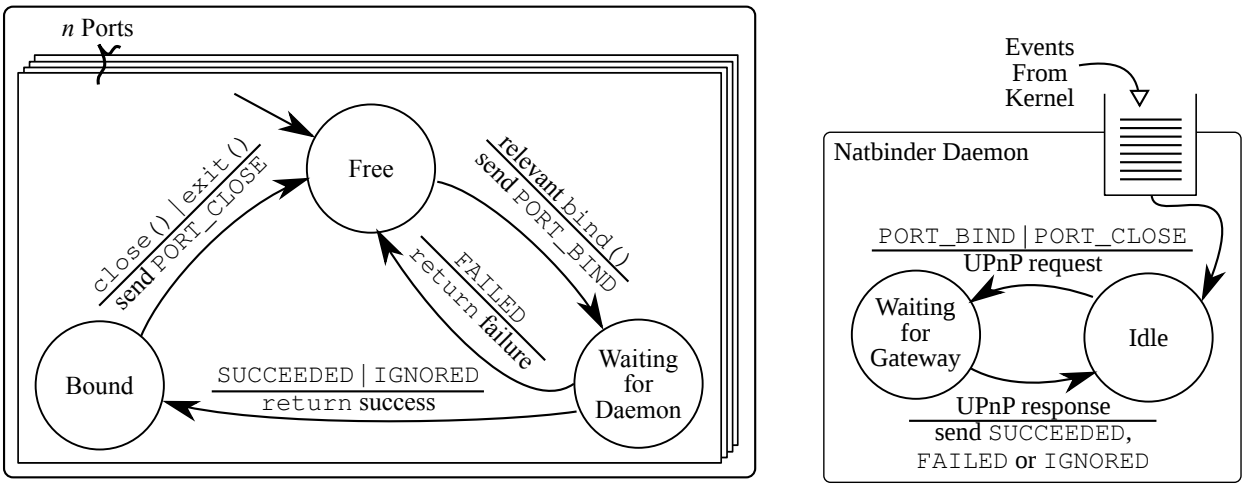
C. The Protocol

The definition of a new protocol on top of Netlink was fairly simple, being a matter of picking a free protocol number in the `netlink.h` header file and aliasing the name `NETLINK_NAT_PASS` to this number. Thus, most of the work on creating the protocol lies in defining its vocabulary and the behavioural interaction between kernel and daemon.

Messages of `NETLINK_NAT_PASS` can be split into two categories: one can be either *request* or *response*. The two *request* type messages, which are always sent by the kernel, are `PORT_BIND` and `PORT_CLOSE`. Each message takes four parameters: a sequence number, that will identify the request within the kernel; the IP address of the request (must be 0.0.0.0 to be relevant); the requested port number (greater than 1023 to be relevant) and the transport protocol used (either TCP or UDP).

The *response* type messages are always issued by the daemon to the kernel in response to a *request* message. They

 ISBN: 978-1-61208-183-0



(a) Kernel ports' state machine

(b) Natbinder daemon's state machine

Figure 3. Protocols' behavioural *automata*

can be either SUCCEEDED, FAILED or IGNORED. Each *response* takes as parameter the sequence number of the *request* that originated it, so the kernel can match each received *response* with a pending *request*.

The whole communication dynamics can be seen as a layered architecture, as shown in Fig. 2. In this perspective, the top layer are the user processes, which are served by the kernel. The kernel provides services to the user processes via system calls, and in turn is served by the Natbinder daemon, layered beneath it. In this layer our newly defined protocol NETLINK_NAT_PASS is used as the interface the daemon uses to provide the services to the kernel. In the lowest layer, the Internet gateway serves the daemon via the UPnP IGD protocol.

Inside the kernel, there is a three-state *automaton* for every port (see Fig. 3a). This *automaton* reacts from system calls events relevant to port forwarding (as discussed in Subsection III-B), and requests services from the daemon. The bind() event triggers a PORT_BIND request for the daemon to perform the port forwarding, and the *automaton* is hold on state "Waiting for Daemon" until receiveis a *response* message (or upon timeout). When terminating a port usage with PORT_CLOSE, the *automaton* waits for no response, resuming immediately.

On the daemon, the requests from the kernel are serialized by the Netlink protocol, being handled sequentially. As show in Fig. 3b, for every *request* message arriving, there is one request made via UPnP to the gateway. Upon each response received, a corresponding *response* type message is issued back to the kernel, using the same sequence number given in the *request*.

D. Error Handling

It is very important that an application do not hang too long while waiting for the bind() to conclude. Since the kernel have no control on the status of processes listening on the broadcast channel of NETLINK_NAT_PASS, two precautions are taken inside the kernel. First, before issuing the PORT_BIND request, kernel checks if there are any listeners on the broadcast channel. If there are none, the ordinary procedure of port binding is resumed.

As a system daemon, Natbinder is expected to run on system startup, but if for some reason it is not running, maybe because of a bug or because it was explicitly shutdown by the user, that check on kernel side will ensure bind() calls will be served normally (*i.e.* without the automatic NAT traversal feature).

There may also be the case the PORT_BIND request was already issued by the kernel, but the daemon stopped answering due to a bug or network error. To avoid letting the user process blocked indefinitely, inside the kernel there is a timeout of half second on waiting for the response. After that, bind() will resume as if it had received the IGNORE message as response.

Having the daemon up and running, we cannot fully trust on the gateway reliability. Configurations made by Natbinder on the gateway are volatile, so if the user simply manually restart a modem working as the Internet gateway, all NAT configurations previously performed are lost. Also, UPnP IGD implementation on devices might have its own issues. For instance, the gateway used during the development of this work had a maximum limit of 32 ports simultaneously forwarded via UPnP IDG. Port mapping requests after this limit would fail with an unknown error code, what will generate an IGNORED response to the kernel.

To counter this kind of problems, concerning the gateway reliability, the daemon holds the ideal state of all port mappings managed by itself. Periodically, it checks the gateway state and compares it with its own ideal state. If

 ISBN: 978-1-61208-183-0

they are divergent, the daemon tries to make the necessary changes to match the gateway state with the ideal state. This way, if a port mapping was not possible because of an unknown error, ideally it should have been mapped, so the ideal state will hold this port mapping. Upon the periodical check on the gateway, this pending port mapping will be tried again. Considering the previously mentioned case where the gateway was restarted, the daemon will find no port mapping entries on the gateway, so it will try to register all of them.

V. Use Cases

To test the approach some existing applications that would benefit from it were chosen to build test case scenarios. The first test was performed with the P2P application rTorrent [19], a BitTorrent client working over TCP. The second test was performed with the game OpenArena [20], whose netwoking multiplayer is done via UDP. Both of the previusly mentioned application are not prepared to handle NAT automatically. The third test was performed with Transmission [21], another BitTorrent client which is able to perform automatic port forwarding.

In the first test, with rTorrent, the program was configured to use ports in the range from 10000 to 10009, and five instances of it were executed. The odd port numbers of this range were already taken in the gateway by another host in the network, so only the 5 remaining even ports were available to be used externally. All five instances were able to execute and properly bind to each one of the remaining ports. All five were able to receive incoming connections from the Internet. As expected, if a sixth instance is executed, it is unable to find a suitable port and exits with the error message: *"Could not open/bind port for listening: Address already in use."*

There is a mechanism inside rTorrent, similarly to other applications, including OpenArena, that was designed to find an available port on what to operate. The approach implemented in this work was conceived so not to disrupt such behavior. In this case, some ports where already taken in the gateway, but since they were presented to the application as ports already used by the TCP stack, its own port finding mechanism was able to devise an usable port.

The second test was performed by running the standalone server of OpenArena, which by default wait for players on UDP port 27960, and then opening the client and creating another multiplayer game rood, which will use the next available UDP port: 27961. Both ports were correctly and automatically exposed to the Internet, and both removed when the application closed. External clients were able to connect.

The Transmission test was performed by executing it with our automatic port forwarding mechanism disabled, then the program was killed, simulating a crash. Since it implemented the UPnP protocol, it was able to automatically forward its port on the gateway, but when it was killed, it was not given time to cleanup, so its entry persisted on the gateway. With the automatic port forwarding, the port was forwarded all the same, despite the redundant work performed both by the application and our daemon, but when it was killed, the daemon still cleaned it up on the gateway.

VI. Conclusion and Future Works

In this work, we presented a new approach to NAT traversal, including its architecture and reference implementation. The architectural choices are closely related to the technologies used. Were we using a less common but simpler protocol like NAT-PMP, our architecture would also be simpler.

It is hard to measure the real benefit of our approach, being it subjective when concerning user experience and useful to software developers mostly after its widespread adoption. In any case, as a future work we expect to survey the benefits of its usage with a group of volunteering users.

To reach a public of users and probable volunteers, our implementation will be integrated with Ubuntu, a popular GNU/Linux distribution. At first via a third party package maintained in Ubuntu's Personal Package Archives (PPA). Eventually, depending on community acceptance, into mainline Ubuntu.

As a Linux kernel modification, the implementation shall be submitted for inclusion into the official kernel distribution. Among other factors, the inclusion will be subject on the community acceptance of the concept idea, and the code stability asserted by the early testers.

Using the reference implementation, any OS based on Linux could easily support the approach, even Android, which is a mobile OS but still subject to the common networks behind NAT.

To make our implementation more user friendly, we plan to include a Graphical User Interface (GUI) for the daemon integrated with the system shell. Being a separated application, each system could have its own GUI that provides the better integration with its environment. In Ubuntu, such GUI would be and extension of the NetworkManager application, providing seamless desktop integration. Through this GUI, the user would be able to monitor the status of the automatically forwarded ports on the router.

The approach is not limited to our architecture or specific implementation, and could be done by vendors or third party software providers of other platforms and operating systems, including Windows, MacOS, Playstation, *&c*. The approach could also be implemented over other protocols, like IPv6, should NAT prove to still be useful with it.

In matters of network transparency in the operating systems, network interfaces are virtual enough abstraction. On Linux, besides the real Ethernet devices with associated IP addresses, there are virtual interfaces for loopback, VPN's, PPP, tunnels, *&c*. In a future work, the approach proposed here could be generalised as another virtual network interface,

 ISBN: 978-1-61208-183-0

managed by a NAT client network driver, which would have as address the shared public IP. A bind to it would be automatically forwarded. An interface like this seems more naturally fit.

ACKNOWLEDGMENT

This research was sponsored in part by a grant from CAPES. L. C. V. would like to thank Leonardo de Sá Alt and Rodrigo Queiroz Saramago for their invaluable help in understanding the inner workings of the Linux kernel.

REFERENCES

[1] P. Belimpasakis, A. Saaranen, and R. Walsh, "Home dns: Experiences with seamless remote access to home services," in *World of Wireless, Mobile and Multimedia Networks, 2007. WoWMoM 2007. IEEE International Symposium on a*, June 2007, pp. 1 –8.

[2] A. Haber, J. De Mier, and F. Reichert, "Virtualization of remote devices and services in residential networks," in *Next Generation Mobile Applications, Services and Technologies, 2009. NGMAST '09. Third International Conference on*, Sept. 2009, pp. 182 –186.

[3] V. Pankakoski, "Experimental design for a next generation residential gateway," Master's thesis, Aalto University, School of Science and Technology, 2010, retrieved: Dec., 2011. [Online]. Available: http://lib.tkk.fi/Dipl/2010/urn100389.pdf

[4] S. Deering and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification," RFC 2460 (Draft Standard), Internet Engineering Task Force, Dec. 1998, updated by RFCs 5095, 5722, 5871.

[5] Ipv6.br faq. Núcleo de Informação e Coordenação do Ponto BR. Retrieved: Dec., 2011. [Online]. Available: http://www.ipv6.br/IPV6/MenuIPV6FAQ#Que_tamanho_de_bloco_IPv6_devo_f

[6] P. Srisuresh and K. Egevang, "Traditional IP Network Address Translator (Traditional NAT)," RFC 3022 (Informational), Internet Engineering Task Force, Jan. 2001.

[7] A. Hemel, "Universal plug and play: Dead simple or simply deadly?" in *System Administration Network Engineering, 2006*, May 2006.

[8] Upnp forum. Retrieved: Dec., 2011. [Online]. Available: http://upnp.org/

[9] J. Blokhuis, "Universal plug and play vulnerabilities in eventing," University of Amsterdam, Tech. Rep., 2009.

[10] Upnp hacks: Hacking universal plug and play. Retrieved: Dec., 2011. [Online]. Available: http://www.upnp-hacks.org/

[11] S. Cheshire, M. Krochmal, and K. Sekar, "NAT Port Mapping Protocol (NAT-PMP)," Internet-Draft, Internet Engineering Task Force, 2008, retrieved: Dec., 2011. [Online]. Available: http://tools.ietf.org/id/draft-cheshire-nat-pmp-03.txt

[12] Natbinder. Retrieved: Dec., 2011. [Online]. Available: http://www.gitorious.org/natbinder

[13] Linux upnp internet gateway device. Retrieved: Dec., 2011. [Online]. Available: http://linux-igd.sourceforge.net/

[14] Pseudo ics daemon. Retrieved: Dec., 2011. [Online]. Available: http://pseudoicsd.sourceforge.net/

[15] Miniupnp project homepage. Retrieved: Dec., 2011. [Online]. Available: http://miniupnp.free.fr/

[16] *The `/proc` filesystem*, Documentation/filesystems/proc.txt, contained in Linux source code distribution.

[17] *sysfs – The filesystem for exporting kernel objects*, Documentation/filesystems/sysfs.txt, contained in Linux source code distribution.

[18] *UPnP IGD WANIPConnection*, UPnP Forum Std., Rev. 2.0, Sept. 2010, retrieved: Dec., 2011. [Online]. Available: http://upnp.org/specs/gw/igd2/

[19] The libtorrent and rtorrent project. Retrieved: Dec., 2011. [Online]. Available: http://libtorrent.rakshasa.no/

[20] Openarena. Retrieved: Dec., 2011. [Online]. Available: http://www.openarena.ws/

[21] Transmission. Retrieved: Dec., 2011. [Online]. Available: http://www.transmissionbt.com/

 ISBN: 978-1-61208-183-0

Minimization of Branching in the Optical Trees with Constraints on the Degree of Nodes

Massinissa Merabet
LIRMM
University Montpellier 2
Montpellier, France
massinissa.merabet@lirmm.fr

Sylvain Durand
LIRMM
University Montpellier 3
Montpellier, France
sylvain.durand@lirmm.fr

Miklos Molnar
LIRMM
University Montpellier 2
Montpellier, France
miklos.molnar@lirmm.fr

***Abstract*—Multicast routing applied to optical networks provide several research problems on spanning tree. In optical networks, the ability of dividing the light signal is limited. Two recently problems try to take into account this constraint: looking for spanning trees with minimum number of branching vertices (vertices of degree strictly greater than 2) and looking for spanning trees such that the sum of branch vertices degrees is minimal. There are two kinds of optical nodes: nodes equipped with splitters, able to divide the input light signal, and nodes without splitters, unable to split the signal. The two problems mentioned above do not distinguish between the type of nodes. In this study, we discuss the relationship between the two problems, we thus prove that the two previous problems are not necessarily linked. We also propose two variants of them, taking into account this additional constraint in the construction of the spanning tree, and we find an experimental upper bound on the number of nodes to equip with splitters in an optical network.**

***Keywords*—optical network; multicast routing; spars splitters; degree bounded spanning tree.**

I. INTRODUCTION

Wavelength-Division Multiplexing (WDM) is an effective technique to exploit the large bandwidth of optical fiber to meet the explosive growth of bandwidth demand in the Internet [1].

Multicast consists in simultaneously transmit information from one source to multiple destinations [1] in a bandwidth efficient way (it duplicates the information only when necessary). From the computational point of view, multicast routing protocols are mainly based on spanning tree structure. When the cost of communications has to be minimized, finding such a tree is NP-complete [2] and is known as the Steiner problem. However, the classical Steiner problem does not take into account the physical constraints of the network needed to perform successfully the multicast routing. Indeed, in order to divide the light signal, some nodes must be equipped with optical splitters. In the optical networks, a node which has the ability to replicating any input signal on any wavelength to any subset of output fibers is referred to as a Multicast-Capable (MC) node [3]. On the other hand, a node which has the ability to tap into the signal and forward it to only one output is called a Multicast-Incapable (MI) node [3]. Optical networks will have a limited number of MC nodes, and these nodes should be positioned such as the multicast routing is feasible. In addition to that, the light power in optical networks should be controlled because of the power loss. Indeed if a light signal is splitted into m copies, the signal power of one copy will be reduced with a factor of $1/m$ of the original signal power [4]. For this reason, it is useful to find a spanning tree such that the number of branching nodes (nodes of degree strictly greater than 2) is limited [5]. To better take into account this constraints it is necessary to find a spanning tree such that the sum of the degrees of nodes dividing the light signal is limited.

Although the two previous problems aim at satisfying reel constraints, they do not take into consideration the ability of an optical node to divide the light signal. They consider that all nodes can be branching nodes in the spanning tree. furthermore, in the examples given in the literature, often the same optimal spanning tree is used for both the first and the second problem. In this study, we introduce two variants of the previous problems that take into consideration the type of an optical node (MC or MI), so that all nodes connecting the spanning tree are effectively able to divide the light signal, and we prove that the two previous problems are not necessarily linked.

The rest of the paper is organized as follows. Section II contains basic definitions and formal statements of the problems considered in this paper. Section III proves that the problems MBV and MDC are not necessary linked. Section IV provides the ILP formulations of MBV-DC and MDS-DC. In Section IV, we analyse the experimental results about MBV-DC and MDS-DC on a set of scenarios. In that Section, we also found an experimental upper bound on the number of nodes to equip with splitters in an optical network. Conclusions are object of Section VI.

II. DEFINITIONS AND FORMULATIONS

Let the topology of an optical network be modelized by a connected graph $G = (V, E)$, where V is set of the vertices (corresponding to optical nodes) and E the set of edges (corresponding to optical links). For each vertex $v \in V$ we denote by $d_G(v)$ the degree of v in G. We denote by $CC(G)$ the number of connected components of the graph G. We denote by $MC(G)$ the set of multicast-capable vertices in the

 ISBN: 978-1-61208-183-0

graph G and $MI(G)$ the set of multicast-incapable vertices. Let $T = (V_T, E_T)$ be a spanning tree of G. A vertex $v \in V_T$ is a branch vertex in T iff $d_T(v) > 2$. Let $NB(T)$ be the set of branching vertices of the tree T. We denote by $s(T)$ the size of $NB(T)$ and by $q(T)$ sum of the of branching nodes degrees of the tree T ($q(T) = \sum_{v \in NB(T)} d_T(v)$).

We denote by $s^*(G)$ the smallest number of branching nodes of all spanning trees of G and $q^*(G)$ the smallest sum of branching nodes degrees of all spanning trees of G.

The two problems initially proposed in [5] have been defined as follows:

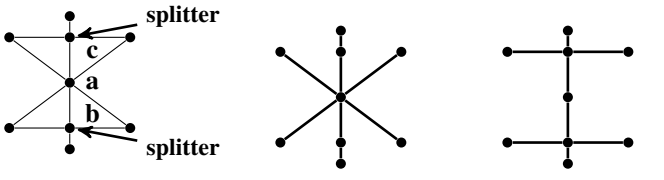
Problem II.1. *The problem MBV (Minimum Branch Vertices spanning tree) consists in finding a spanning tree of G which has the minimum number of branch vertices.*

Problem II.2. *The problem MDS (Minimum Degrees Sum branch vertices spanning tree) consists in finding a spanning tree of G which has the minimum sum of branching nodes degrees.*

We propose the modification of MDS and MBV, such that they take into account the additional constraint of ability of an optical node to divide the light signal. These two new problems allow a network node to be a branch node in the corresponding spanning tree if and only if this node is multicast-capable.

Problem II.3. *The problem MBV-DC (minimum branch vertices spanning tree with degree contraints) consists in finding a spanning tree T of G which has the minimum number of branch vertices such that $NB(T) \subseteq MC(G)$.*

Problem II.4. *The problem MDS-DC (minimum degrees sum of branch vertices spanning tree with degree contraints) consists in finding a spanning tree T of G which has the minimum sum of branch vertices degrees, such that $NB(T) \subseteq MC(G)$.*



(a) **Starting network** (b) **Optimal spanning tree for the MBV** (c) **Optimal spanning tree for the MBV-DC**

Figure 1. Example of the difference between the MBV and MBV-DC

Suppose that the network shown in Figure $1(a)$ contains two multicast-capable nodes : $MC(G) = \{b, c\}$. The optimal solution (Figure $1(b)$) for the problem MBV does not take into account this constraint and selects the node a as a branching node. This tree is not feasible in the optical network. On contrary the optimal solution (Figure $1(c)$) for the problem MBV-DC is greater (two branching nodes) but feasible.

III. Relation between MBV and MDS

In all examples shown in the literature, there is an optimal spanning tree for both the MBV and the MDS. However, the MBV and the MDS are two different problems. In this section, we present an example where the set of optimal spanning trees for MBV and the set of optimal spanning trees for MDS are disjoint.

Remember that $s(T)$ denotes the number of branching vertices of the tree T and $q(T)$ the sum of branching nodes degrees of T. We denote by $s^*(G)$ the smallest number of branching nodes of all spanning trees of G and $q^*(G)$ the smallest sum of the degrees of branching nodes of all spanning trees of G.

Proposition III.1. *The MDS problem and MBV are not linked:*
There exists a graph G such that:
For all spanning tree T of G:

1) *If T is optimal for the MBV problem, it is not optimal for the MDS.*
   That is: if $s(T) = s^*(G)$ then $q(T) \neq q^*(G)$,
2) *If T is optimal for the MDS problem, it is not optimal for the MBV.*
   That is: if $q(T) = q^*(G)$ then $s(T) \neq s^*(G)$.

Proof: Figure 2 presents a graph $G = (V, E)$ which respects conditions of Proposition III.1:

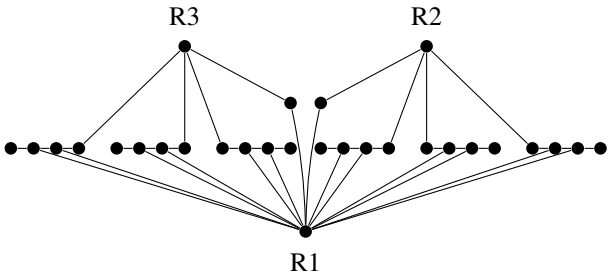


Figure 2. Instance proving the proposition III.1

If G in Figure 2 is Hamiltonian, then any optimal solution for one problem will also be an optimal one for the other one since $s^*(G) = q^*(G) = 0$. Thus, we must first prove that G does not contain Hamiltonian path. We use the following result of [6]:

Proposition III.2. *[6] Let $G(V, E)$ be a graph, if G has a Hamiltonian path, then for all $S \subseteq V$, the graph $(G - S)$ has at most $\mid S + 1 \mid$ connected components.*

Using the contrapositive of proposition III.2 on G with $S = \{R1, R2, R3\}$, we conclude that G is not Hamiltonian.

Thus we have $s^*(G) \geq 1$. The tree $T1$ in Figure 3 is a spanning tree of the graph G and $s(T1) = 1$. Therefore $s^*(G) = 1$.

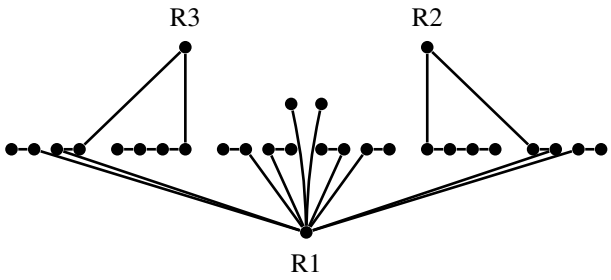


Figure 3. Spanning tree $T1$ of G

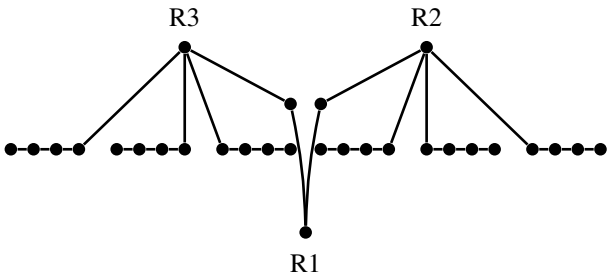


Figure 4. Spanning tree $T2$ of G

Let G' be the sub graph of G induced by $V-\{R1, R2, R3\}$. G' is composed of 8 connected components. We conclude that for all spanning tree T of G, T contains at least $R1$ or $R2$ or $R3$ as a branching node. Note that for all spanning tree T of G such that $s(T) = s^*(G) = 1$, $R1$ is the only possible branching node.

We now prove that $q^*(G) = 8$: G is not Hamiltonian so $q^*(G) \geq 3$. In the spanning tree $T2$ of Figure 4 we have $q(T2) = 8$, therefore $8 \geq q^*(G) \geq 3$.

Suppose that $q^*(G) < 6$. Let T be a spanning tree such that $q(T) < 6$. T has a single branching node. But, if $d_T(R1) < 6$ then $CC(T) \geq 2$. So T has at least two branching nodes, which is in contradiction with the hypothesis $q^*(G) < 6$. Therefore, $8 \geq q^*(G) \geq 6$.

Let T be a spanning tree of G such that $8 \geq q(T) \geq 6$, $q(T) = q^*(G)$, T contains 2 branching nodes, and at least $R1$ or $R2$ or $R3$ are branching nodes in T.

If $R1$ is a branching node in T, then there must be at least two other branching nodes so that T is connected. So $R1$ is not a branching node in T (otherwise $q(T) \geq 9$).

Since $d_G(R2) = 4$, $R2$ is a branching node in T, then only $R3$ has a large enough degree in G so that T is connected ($R1$ is already eliminated). Thus $R2$ and $R3$ are the only branching nodes in T.
Symmetrically if $R3$ is a branching node, $R2$ must be the only other one.

For all spanning tree T of G with $R3$ and $R2$ as the only branching nodes, we must have $d_T(R2) = d_G(R2) = 4$ and $d_T(R3) = d_G(R3) = 4$, which implies that $q^*(G) = 8$.

Conclusion: For all spanning tree T of G such that $q(T) = q^*(G) = 8$, $s(T) > 1$, so $s(T) \neq s^*(G)$. For all spanning tree T of G such that $s(T) = s^*(G) = 1$, $q(R1) > 8$, so $q(T) \neq q^*(G)$.

■

IV. ILP FORMULATION

In this section, we resume from [5] the ILP formulations of MBV and MDS problems, and we modify them in order to take into account the capacity of an optical node to divide the input light signal.

In order to define a spanning tree T of G, we can send from a source vertex $S \in V$ one flow unit to every other vertices $v \in V \setminus \{S\}$. Although edges of G are undirected, we define two variables for each edge $e = \{u, v\} \in E$: f_{uv} and f_{vu} define respectively the flow going from u to v and the flow going from v to u along $\{u, v\}$. For each edge $e = \{u, v\} \in E$, we consider a binary decision variable x_e such that $x_e = 1$ when e belongs to T and $x_e = 0$ otherwise. Finally, for each $v \in V$, we have a decision variable y_v that is equal to 1 if v is a branching node, and 0 otherwise.

Let us denote by $\omega(v) = \{w \in V \mid \{v, w\} \in E\}$ the set of neighbours of v. The mathematical formulation of MBV given in [5] is the following:

$$
\begin{cases}
min\ s^* = \sum_{v \in V} y_v & \text{(1a)} \\
\sum_{e \in E} x_e = n - 1 & \text{(1b)} \\
\sum_{v \in \omega(S)} f_{Sv} - \sum_{v \in \omega(S)} f_{vS} = n - 1 & \text{(1c)} \\
\sum_{u \in \omega(v)} f_{vu} - \sum_{u \in \omega(v)} f_{uv} = -1, \quad \forall v \in V \setminus \{S\} & \text{(1d)} \\
f_{uv} \leq (n-1)x_e, \quad \forall e = \{u, v\} \in E & \text{(1e)} \\
f_{vu} \leq (n-1)x_e, \quad \forall e = \{u, v\} \in E & \text{(1f)} \\
\sum_{e=(u,v) \mid u \in \omega(v)} x_e - 2 \leq (n-1)y_v, \ \forall v \in V & \text{(1g)} \\
x_e \in \{0, 1\}, \quad \forall e \in E & \text{(1h)} \\
y_v \in \{0, 1\}, \quad \forall v \in V & \text{(1i)} \\
f_{uv} \geq 0, \quad \forall e = \{u, v\} \in E & \text{(1j)} \\
f_{vu} \geq 0, \quad \forall e = \{u, v\} \in E & \text{(1k)}
\end{cases}
$$

The mathematical model for MDS [5] requires additional integer decision variables counting the degree of branch vertices in the solution:

$$z_v = \begin{cases} d_T(v), & \text{if } v \text{ is a branching node,} \\ 0, & \text{otherwise.} \end{cases}$$

The objective function is then:

$$min\ q^* = \sum_{v \in V} z_v$$

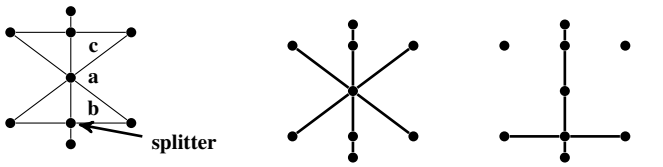
There is an additional constraint:

$$\sum_{e=(u,v)|u \in \omega(v)} x_e - 2 + y_v \leq z_v \qquad \forall v \in V$$

In our problems, we want to satisfy optical constraints imposed by the presence / absence of splitters in nodes. The mathematical formulation of the MBV-DC, resp. MDS-DC, is the same as the MBV, respectively MDS, but we must add the following constraint:

$$y_v = 0 \text{ if } v \notin MC(G)$$

An important difference between the two problems has to be analyzed. For all undirected connected graph input of MBV and MDS, it is guaranteed to have a feasible solution (every connected graph admits a spanning tree). On the contrary, the existence of a feasible solution for the MBV-DC and DC-MDS depends strongly on the positioning of splitters in the network. In Figure 5, only vertex b has a splitter ($MC(G) = \{b\}$), this instance does not have a feasible solution.



(a) **Starting network** (b) **Optimal spanning tree of MBV** (c) **No feasible solution for the MBV-DC**

Figure 5. Example of instance for which there are no feasible solutions of MBV-DC

V. EXPERIMENTAL RESULTS

We measured solutions of MBV-DC (and MDS-DC) according to the proportion nbv of nodes (of degree strictly greater than 2) equipped with splitters in the network. When the proportion is 100% the solution is the same as for MBV (and MDS).

Instances of MBV-DC and MDS-DC are undirected and connected graphs. To produce such graphs, the NetGen random graph generator was used. NetGen is a powerful tool dedicated specifically to the generation of random transport networks [7]. NetGen is used in most experiments on the MBV and MDS already done (especially in [5]). If parameters dedicated to capacities of arcs are set to zero, the generator will produce non-valued connected random graphs. The input files used by NetGen to generate instances follow the format given in Table 1. According to the table, the only parameters that can vary are the seed for the random number generator and the number of vertices and edges of the output graph.

TABLE I. NETGEN PARAMETERS FOR INPUT FILES

parameters	Input	Parameter description
SEED	variable	Random numbers seed
NODES	variable	Number of nodes
SOURCES	1	Number of sources (including transshipment)
SINKS	1	Number of sinks (including transshipment)
DENSITY	variable	Number of (requested) edges
MINCOST	0	Minimum cost of edges
MAXCOST	0	Maximum cost of edges
SUPPLY	1	Total supply
TSOURCES	0	Transshipment sources
TSINKS	0	Transshipment sinks
HICOST	0	Percent of skeleton edges given maximum cost
CAPACITED	0	Percent of edges to be capacitated
MINCAP	0	Minimum capacity for capacitated edges
MAXCAP	0	Maximum capacity for capacitated edges

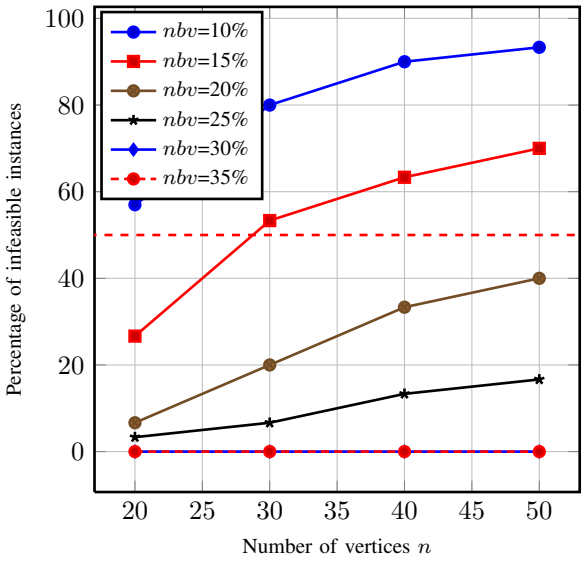
In order to solve the problems MBV-DC and MDS-DC, we used the linear program solver GLPK [8]. We consider five different values for the number of vertices of random graph: $n \in \{20,\ 30,\ 40,\ 50\}$. For each value of n, we consider a single density value (ratio between the number of edges and the number of vertices) d = 1.5. We have chosen this density because it allows to have a significant number of branching nodes in the solutions. This makes the comparison between the MBV (resp. MDS), and MBV-DC (resp. MDS-DC) be more relevant. We consider seven values for the percentage of nodes equipped with splitters among the nodes of degree strictly greater than 2: $nbv \in \{10\%,\ 15\%,\ 20\%,\ 25\%,\ 30\%,\ 35\%,\ 100\%\}$. If a node has degree smaller or equal to 2, it can not be a branching node whatever the constraints.

A random graph associated with a fixed number of vertices, and a fixed percentage of nodes equipped with splitters is called scenario. In order to have a set of significant test, thirty instances are generated for each scenario.

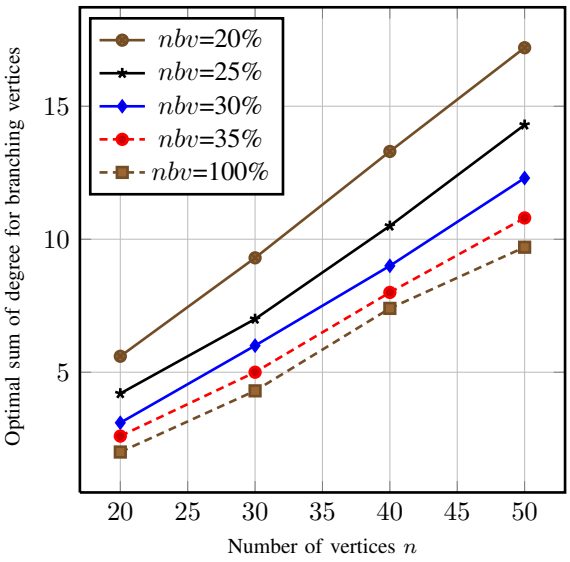
To analyse results in a meaningful way, it is imperative to consider the percentage of infeasible instances for a given scenario. Note that, if an instance is infeasible for MBV-DC then it is infeasible for MDS-DC, and conversely. Therefore, the proportion of infeasible instances is the same for both problems. We consider that, if this proportion is strictly greater than 50% then the value of MBV-DC and MDS-DC on this scenario is not significant. The Figure 6(a) shows the proportion of infeasible instances for MBV-DC. The curves representing $nbv = 10\%$ et $nbv = 15\%$ are above the threshold of 50%. We therefore consider that the comparison between MBV (resp. MDS) and MBV-DC (resp. MDS-DC) is not significant for $nbv < 20\%$.

Figures 6(b) and 6(c) represent average values of solutions of k feasible instances generated for each scenario, such that k less or equal to 30. Note that if nbv is high, then it approaches the solution of problems without constraints (MBV or MDS), which is represented by $nbv = 100\%$.

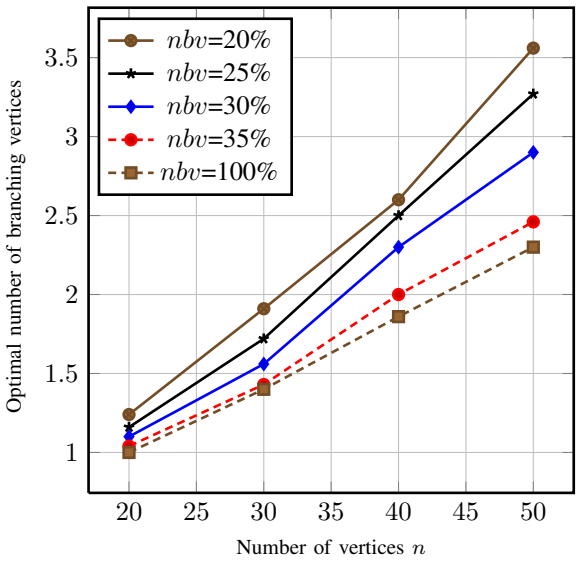
 ISBN: 978-1-61208-183-0



(a) Proportion of infeasible instances

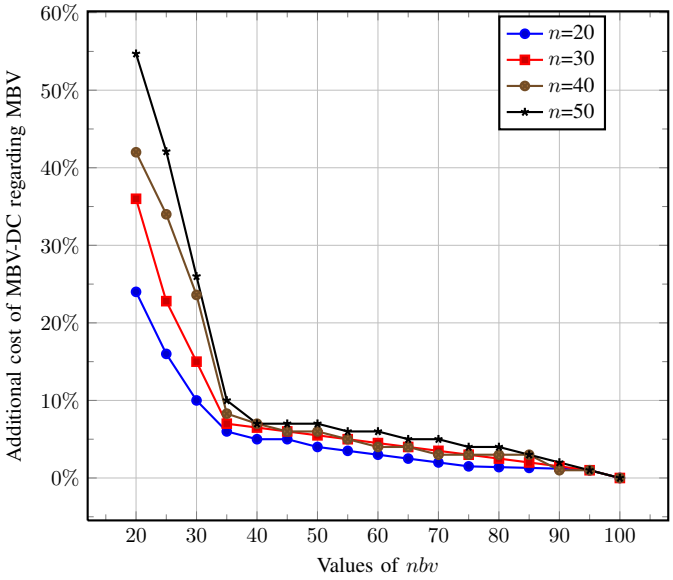


(b) Results for MDS-DC

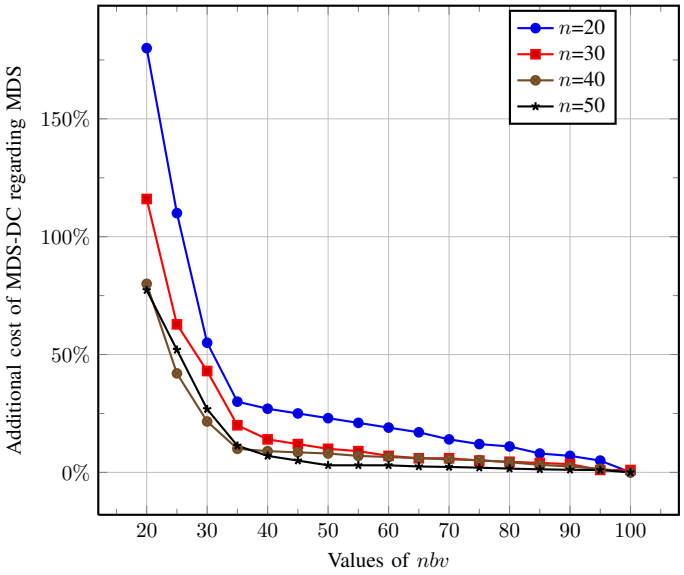


(c) Results for MBV-DC

Figure 6. Results of MBV-DC and MDS-DC



(a) Variation of ratio between MBV-DC and MBV



(b) Variation of ratio between MDS-DC and MDS

Figure 7. Comparison between solutions with or without degree constraints

Moreover, if nbv is high then the percentage of infeasible instances is low. We also observe that, from $nbv = 35\%$, the solution of MBV-DC significantly approach the solution of MBV. For $nbv \geqslant 30\%$, we see that the percentage of infeasible instances is equal to zero.

In Figure 7, we show the influence of the degree constraint for the two studied problems, the percentage of additional cost due to the degree constraint regarding the value of nbv is given for different sizes of networks.

The threshold $nbv = 35\%$ can be considered as an experimental bound about constraints on degrees of nodes problems MBV-DC and MDS-DC. Beyond $nbv = 35\%$,

 ISBN: 978-1-61208-183-0

this constraint has little impact on the optimal solution of MBV-DC and MDS-DC: when more than 35% of the nodes of degree higher than 2 are randomly designed as MC nodes, the cost of MBV-DC solution is only 10% larger than the cost of the MBV solution. This result is also true for MDS providing that the size of the network is large enough (more than 40 nodes).

The interest of this result in practice is important: From 35% of nodes equipped with splitters, the constraint on the number of nodes equipped with splitters has little effect on the value of the optimal solution. Specially, through this bound we can say that in an optical network, we can position the splitters randomly on 35% of nodes of degree strictly greater than 2, and have a high probability of ensuring that the cost of multicast connection will be weakly influenced.

Note that for MBV-DC (and MDS-DC), the feasibility of an instance can not however be guaranteed only by the proportion of MC nodes. There are infeasible instances such that only one single node is not equipped with splitter (see Figure 8).

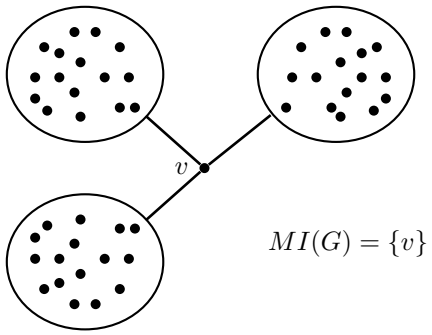


Figure 8. Graph G containing only one single node multicast-incapable, but no feasible solution.

VI. CONCLUSION AND FUTURE WORKS

Two problems have been the source of our study. MBV, which practical interest is to minimize the number of splitters in an optical network, but does not limit the degree of optical nodes, and MDS which practical interest is to minimize the sum of degrees of these splitters nodes in the solution. Both problems consider that all nodes of a network are equipped with optical splitters, and can therefore make divisions of light, which is not true in practice. Indeed, in an optical network, only a subset of the nodes is equipped with splitters (MC nodes). Therefore, only the MC nodes are able to duplicate the light, and to be branching nodes in the spanning tree corresponding to the network, while the other nodes (MI nodes) may only crossed or reached.

The respect of this requirement, it is essential that these theoretical issues best reflect the reality of optical networks. This is why we have introduced two variants of the two problems (problems MBV-DC and MDS-DC) which take into account this constraint in the construction of the spanning tree. Following the resolution of these problems by integer linear programming, and tests on random graphs, we found an experimental upper bound on the number of nodes to equip with splitters in an optical network. Over 35% of nodes equipped with splitters, this constraint has little effect on the corresponding optimal spanning tree. Indeed, beyond this threshold the additional cost due to the degree constraint is less than 10% for the problem MBV-DC. This assumption is also true for MDS-DC provided that the number of nodes is greater than 40.

In problems treated, there is no real limit on the degree of branching nodes because we consider that their degree can be as large as needed in the optimal tree, thus the degree constraint on the nodes in a spanning tree is either 2 (MI nodes) or its degree in the original graph (MC nodes). Knowing that splitters has limited capacity to divide the light signal, consideration may be given to improve the modelling of our problems by setting an upper bound on the degree of the nodes of the spanning tree varying between 1 and the degree of the node in the original graph.

REFERENCES

[1] J. He, S.-H. G. Chan, and D. H. K. Tsang, "Multicasting in WDM Networks," *IEEE Communications Surveys and Tutorials*, pp. 2–20, 2002.

[2] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York, NY, USA: W. H. Freeman & Co, 1979.

[3] R. Malli, X. Zhang, and C. Qiao, "Benefits of multicasting in all-optical networks," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, ser. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, J. M. Senior & C. Qiao, Ed., vol. 3531, Oct. 1998, pp. 209–220.

[4] M. Ali and J. S. Deogun, "Power-Efficient Design of Multicast Wavelength-Routed Networks," in *IEEE Journal of Selected areas in communication*, 2000, pp. 1852–1862.

[5] R. Cerulli, M. Gentili, and A. Iossa, "Bounded-degree spanning tree problems: models and new algorithms," *Comput. Optim. Appl*, vol. 42, pp. 353–370, April 2009.

[6] D. B. West, *Introduction to Graph Theory*. University of Illinois-Urbana: Prentice-Hall, 1996.

[7] D. Klingman, A. Napier, and J. Stutz, "A program for generating large scale capacitated assignment, transportation, and minimum cost flow network problems," *Management Science*, vol. 20, no. 5, pp. 814–821, 1974.

[8] A. O. Makhorin, *GNU Linear Programming Kit (GLPK) v 4.38*, gnu project ed., May 2009.

[9] L. Gargano, P. Hell, L. Stacho, and U. Vaccaro, "Spanning Trees with Bounded Number of Branch Vertices," in *Proceedings of the 29th International Colloquium on Automata, Languages and Programming*, ser. ICALP '02. London, UK, UK: Springer-Verlag, 2002, pp. 355–365.

 ISBN: 978-1-61208-183-0

Mitigating Spoofing Attacks in MPLS-VPNs using Label-hopping

Shankar Raman*, Gaurav Raina†
India-UK Advanced Technology Centre of Excellence in Next Generation Networks
**Department of Computer Science and Engineering, †Department of Electrical Engineering*
Indian Institute of Technology - Madras
Chennai - 600 036, India
Email: mjsraman@cse.iitm.ac.in, gaurav@ee.iitm.ac.in

***Abstract*—In certain models of inter-provider Multi-Protocol Label Switching (MPLS) based Virtual Private Networks (VPNs) spoofing attack against VPN sites is a key concern. For example, MPLS-based VPN inter-provider model "C" is not favoured, owing to security concerns in the data-plane, even though it can scale with respect to maintenance of routing state. Since the inner labels associated with VPN sites are not encrypted during transmission, a man-in-the-middle attacker can spoof packets to a specific VPN site. In this paper, we propose a label-hopping technique which uses a set of randomized labels and a method for hopping amongst these labels using the payload of the packet. To prevent the attacker from identifying the labels in polynomial time, we also use an additional label. The proposed technique can be applied to other variants of inter-provider MPLS based VPNs where Multi-Protocol exterior-BGP (MP-eBGP) multi-hop is used. As we address a key security concern, we can make a case for the deployment of MPLS based VPN inter-provider model "C".**

***Keywords*-MPLS; VPN; Model C; Spoofing attacks; Label-hopping;**

I. INTRODUCTION

Multi-Protocol Label Switching (MPLS) [6] technology uses fixed size labels to forward data packets between routers. By stacking labels, specific customer services such as Layer 3 Virtual Private Networks (L3-VPNs) based on Border Gateway Protocol (BGP) extensions are widely deployed in the Internet. BGP-based MPLS L3-VPN services are provided either on a single Internet Service Provider (ISP) core or across multiple ISP cores. The latter cases are known as inter-provider MPLS VPNs which are broadly categorized and referred to as models: "A", "B" and "C" [10].

Model "A" uses back-to-back VPN Routing and Forwarding (VRF) connections between Autonomous System Border Routers (ASBRs). Model "B" uses eBGP redistribution of labelled VPN-IPv4 routes from Autonomous Systems (AS) to neighbouring AS. Model "C" uses multi-hop MP-eBGP redistribution of labelled VPN-IPv4 routes and eBGP redistribution of IPv4 routes from an AS to a neighbouring AS. Model "C" is more scalable for maintaining routing states and hence preferred for deployment in the Internet; refer to [2] for more details. Security issues in MPLS, especially MPLS-based VPNs has attracted attention [1].

The security of model "A" matches the single-AS standard proposed in [9]. Model "B" can be secured well on the control-plane, but on the data-plane the validity of the outer-most label (Label Distribution or Resource Reservation Protocol label) is not checked. This weakness could be exploited to inject crafted packets from inside an MPLS network core. A solution for this problem is proposed in [2]. Model "C" can be secured on the control-plane but has a security weakness on the data-plane. The Autonomous System Border Routers (ASBRs) do not have any VPN information and hence the inner-most label cannot be validated. In this case, the solution used for Model "B" cannot be applied. An attacker can exploit this weakness to send unidirectional packets into the VPN sites connected to the other AS. Therefore, ISPs using model "C" must either trust each other or not deploy it [4].

Control plane security issue in model "C" can be resolved by using IPSec. If IPSec is used in the data-plane then configuring and maintaining key associations could be extremely cumbersome. Even though model "C" is highly scalable for carrying VPN Routing and Forwarding (VRF) routes, the vulnerability of the data-plane renders it unusable. The current recommendation is that model "C" must not be used. A simple solution to this problem is to filter all IP traffic with the exception of the required eBGP peering between the ASBRs, thereby preventing a large number of potential IP traffic-related attacks. However, controlling labelled packets is difficult. In model "C", there are at least two labels for each packet: the Provider Edge (PE) label, which defines the Label Switched Path (LSP) to the egress PE, and the VPN label, which defines the VPN associated with the packet on the PE.

In [5], the authors propose encryption techniques, such as IPSec, for securing the provider edge (PE) of the network. The authors also highlight that the processing capacity could be over-burdened. Further, if an attacker is located at the core of the network, or in the network between the providers that constitute an inter-provider MPLS VPN, then spoofing attacks are possible. The vulnerability of MPLS against spoofing attacks and performance impact of IPSec has been discussed in [3]. If the inner labels that identify packets going towards a L3 VPN site are spoofed, then

 ISBN: 978-1-61208-183-0

sensitive information related to services available within the organizational servers can be compromised. As far as we know, there is no scheme available for installing an anti-spoofing mechanism for these VPN service labels.

This paper outlines a label-hopping technique that helps to alleviate the data-plane security problem in model "C". We propose a scheme that changes the inner VPN labels dynamically based on the payload. By using a mix of algorithms and randomized labels, we can guard against spoofing and related attacks. The advantage of our scheme is that it can be used wherever Multiprotocol-external BGP (MP-eBGP) multi-hop scenarios arise.

The rest of the paper is organized as follows. In Section II, we discuss the pre-requisites of our proposed scheme. In Section III, we discuss the label-hopping technique and some implementation issues. In Section IV, we discuss the preliminary simulation and implementation issues. We present our conclusions and provide avenues for future work in Section V.

II. Pre-requisites for the label-hopping scheme

In this section, we briefly review the network topology for model "C", the PE configuration and the control-plane exchanges needed for our proposed scheme.

A. MPLS VPN model "C"

The reference MPLS-eBGP based VPN network for model "C" as described in [11] is shown in Figure 1, which also shows the control plane exchanges. The near-end PE (PE_{ne}) and far-end PE (PE_{fa}) are connected through the inter-provider MPLS core. The VPN connectivity is established through a set of routers from different Autonomous Systems (AS) and their ASBRs. In the VPN, MP-eBGP updates are exchanged for a set of Forward Equivalence Classes (FECs). These FECs, which have to be protected, originate from the prefixes behind PE_{ne} in a VPN site or a set of VPN sites.

B. PE configuration

Various configurations are needed on the PEs to implement the label hopping scheme. A set of "m" algorithms that generate collision-free labels (universal hashing algorithms) are initially implemented in the PEs. Each algorithm is mapped to an index $A = (a_1, a_2, \cdots a_m), m \geq 1$. The bit-selection pattern used by the PEs for generating the additional label is also configured. PE_{ne} must be configured for a FEC or a set of FECs represented by an aggregate label (per VRF label) which will use the label-hopping scheme. For each FEC or a set of FECs, a set of valid labels used for hopping, $K = (k_1, k_2, k_3, \cdots k_n), n > 1$ and, $k_i \neq k_j$ if $i \neq j$, is configured in PE_{ne}. In the case of bi-directional security, the roles of the PEs can be reversed.

C. Control and data-plane flow

Initially, set K and the bit-selection pattern used by the PEs are exchanged securely over the control-plane. Optionally an index from A, representing a hash-algorithm, could also be exchanged. We propose that only the index is exchanged between the PEs, as it enhances the security, for two reasons. First, the algorithm itself is masked from the attacker. Second, the algorithm can be changed frequently, and it would be difficult for the attacker to identify the final mapping that generates the label to be used for a packet. Figure 1 depicts this unidirectional exchange from PE_{ne} to PE_{fa}.

Once the secure control-plane exchanges are completed, we apply the label-hopping technique, and PE_{fa} forwards the labelled traffic towards PE_{ne} through the intermediate routers using the label-stacking technique (Figure 2). The stacked labels along with the payload are transferred between the AS and ASBRs before they reach PE_{ne}. Using the label-hopping algorithm PE_{ne} verifies the integrity of labels. Upon validation, PE_{ne} uses the label information to forward the packets to the appropriate VPN service instance or site. This data-plane exchange from PE_{fa} and PE_{ne} is depicted in Figure 3. We now present the label-hopping scheme.

III. Label-hopping technique

In this section, we describe the label-hopping technique and discuss some implementation aspects.

Once a data packet destined to the PE_{ne} arrives at the PE_{fa} a selected number of bytes from the payload is chosen as input to the hashing algorithm. The hash-digest obtained as a result is used to obtain the first label for the packet. The agreed bit-selection pattern is then applied on the hash-digest to obtain an additional label, which is then concatenated with the first label. Once PE_{ne} receives these packets it verifies both the labels.

The implementation steps for the control-plane at the PE_{ne} and PE_{fa} are given by Algorithms 1 and 2. The implementation steps for the data-plane at the PE_{fa} and PE_{ne} are given by Algorithms 3 and 4.

Note: The values in K need not be contiguous and can be randomly chosen from a pool of labels to remove coherence

Algorithm 1 Control-plane PE_{ne} algorithm

Require: FEC[] Forward Equivalence Classes, K[] valid labels, A[i] hash algorithm instance, I[] the bit-selection pattern chosen for the inner label.

```
Begin
packet = makepacket(FEC,K, A[i], I);
CP-SendPacket(PE_fa, MP-eBGP, packet);
End
```

 ISBN: 978-1-61208-183-0

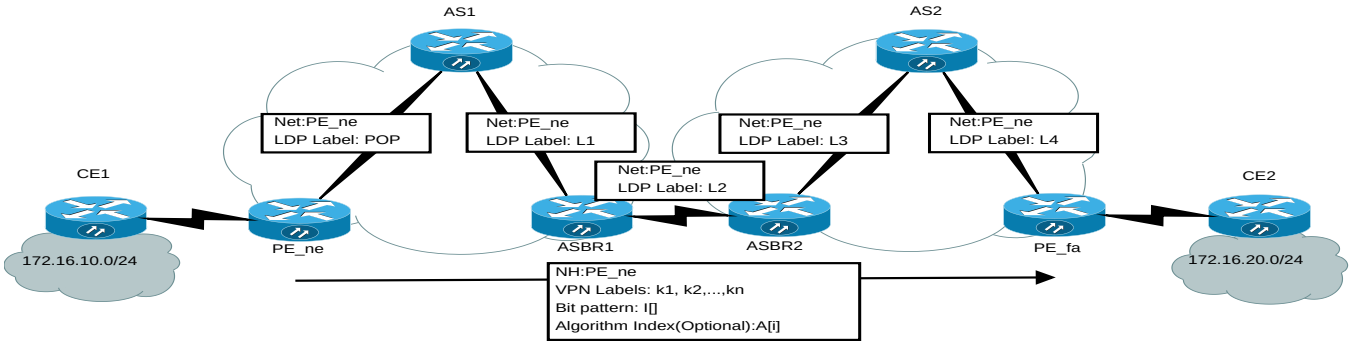


Figure 1: Control-plane exchanges for model C [11]

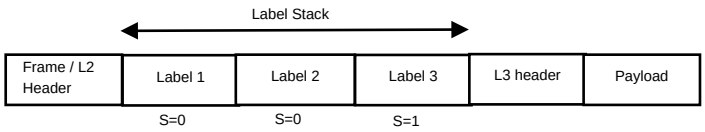


Figure 2: Label stack using scheme outlined for Model "C"

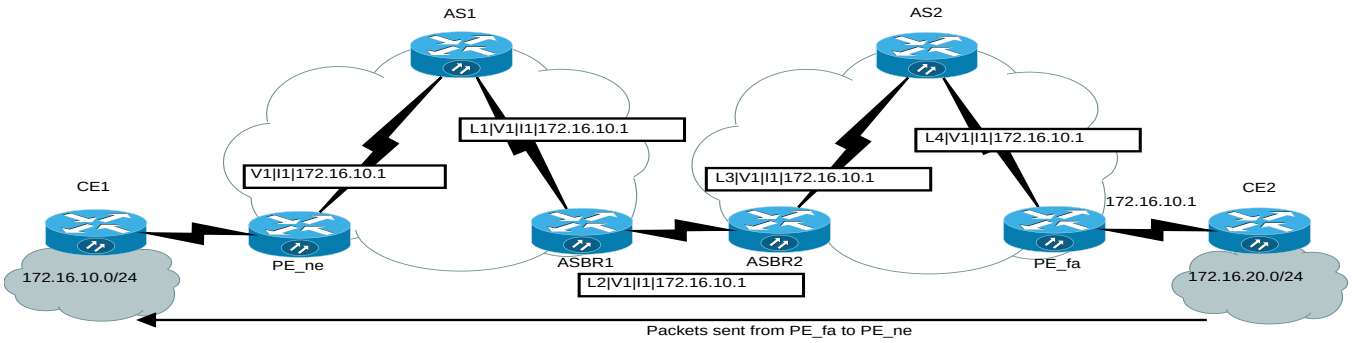


Figure 3: Data-plane flow for model C [11]

Algorithm 2 Control-plane PE_{fa} algorithm

Require: None

Begin
packet = CP-ReceivePacket(PE_{ne}); // from PE_{ne}
FEC[] = ExtractFEC(packet); // extract FECs
K[] = ExtractLabels(packet); // extract the labels
selectHashAlgorithm(A[i]); // hash algorithm to use
RecordValues(FEC); // information for PE_{fa}
RecordValues(K);
RecordValues(I); // bit-selection pattern to be used
End

Algorithm 3 Data-plane PE_{fa} algorithm

Require: None

Begin
packet = DP-ReceivePacket(Interface);
match = CheckFEC(packet); // Is the algorithm enabled?
if $match == 0$ **then**
  return; // no match
end if
hash-digest = calculateHash(A[i],packet);
first-label = hash-digest % $|K|$;
additional-label = process(hash-digest,I)
DP-SendPacket(PE_{ne}, first-label, additional-label, packet);
End

in the label space. Also the algorithms used could be either vendor dependent or a set of standard algorithms mapped the same way by the PE_{ne} and PE_{fa}. If the two PEs involved are from different vendors we assume that a set of standard algorithms are used. In order to avoid too many processing cycles in the line cards of PE_{ne} and PE_{fa}, the hash-digest is calculated over a predefined size of the payload. An additional inner label is further added to enhance protection against spoofing attacks. With an increased label size, an

Algorithm 4 Data-plane PE_{ne} algorithm

Require: None

```
Begin
packet = DP-ReceivePacket(Interface);
match = CheckFEC(packet);
if match == 0 then
   return; //no match
end if
label-in-packet=extractPacket(packet, LABEL);
inner-label=extractPacket(packet, INNER-LABEL);
hash-digest=calculateHash(A[i],packet);
first-label=hash-digest % |K|;
additional-label = process(hash-digest,I)
if label-in-packet ≠ first-label then
   error(); return;
end if
if inner-label ≠ additional-label then
   error(); return;
end if
DP-SendPacket(CE1, NULL, NULL, packet);
End
```

attacker spends more than polynomial time to guess the VPN instance label for the site behind PE_{ne}. There could be two hash-digests that generate the same label. In this case, the two hash-digests is differentiated using the additional label. Collisions can be avoided by re-hashing or any other suitable techniques that are proposed in the literature [8]. If collisions exceed a certain number, then Algorithms 1 and 2 can be executed with a set of new labels.

Illustration: We now briefly illustrate the label-hopping scheme. In Figure 1, using Algorithms 1 and 2, a set of labels are forwarded from PE_{ne} to PE_{fa}. The roles of PE_{ne} and PE_{fa} are interchanged for reverse traffic. Figure 2 shows a packet from the data-plane for model "C", with the proposed scheme. In the figure, "Label 1" refers to the outermost label, while "Label 2" refers to the label generated from the hash-digest and "Label 3" refers to an additional label generated as in Algorithm 3. This additional label has bottom of stack bit (denoted by S in Figure 2) set. These labels are stacked immediately onto the packet and the path labels for routing the packets to appropriate intermediary PEs are added. Figure 3 also shows these path labels used by the data packet to reach PE_{ne}. When the packet passes through the core of an intermediary AS involved in model "C", or through the network connecting the intermediary AS, the intruder or the attacker has the capability to inspect the labels and the payload. However, the proposed scheme prevents the attacker from guessing the right combination of the labels. We can increase the size of the additional inner-labels thereby reducing threats from polynomial time attacks.

IV. Simulation and Implementation

In this section, we present the preliminary simulation results on performance, comparing the label-hopping technique with deep packet inspection where we encrypt and decrypt the complete packet. We also briefly highlight some implementation issues.

A. Simulation

Implementing the label-hopping scheme for all set of FECs belonging to any or all VPN service instances may cause throughput degradation. This is because the hash-digest computation and derivation of the inner-label / additional inner label calculation can be computation intensive. We therefore compared our technique by choosing a part of the payload as input to our hashing algorithm.

We simulated our algorithm on a 2.5 GHz processor Intel dual processor quad core machine. We compared the performance of the label-hopping technique with a deep packet inspection technique where the complete packet was encrypted before transmission and decrypted on reception. The performance figures are shown in Figure 4. These simulation figures indicate that we were able to process 10 million packets per second when we used 64-byte for hashing on a payload of size 1024 bytes. For a hash using 128-byte, we were able to process about 6.3 million packets per second. However with a deep packet inspection where we encrypted and decrypted the complete packet, we were able to process only about 1 million packets per second.

In cases where performance becomes a bottleneck, this label-hopping scheme can be applied to specific traffic which are mission-critical, sensitive and most likely need to be protected as they travel from the PE_{fa} to the PE_{ne}. Selective application of this service which could be offered as a premium for a selected set of FECs is a suitable option, there by protecting the traffic of organizations that are paranoid about the integrity of the switched traffic into their VPN sites.

B. Implementation

We are modifying the open source Quagga router software on Linux to implement our scheme. One of the concerns in the scheme is the use of payload for generating the random source. If the payload does not vary between two packets then the control-plane exchanges have to be renegotiated with a different set of labels for the second packet. The other concern in the scheme is to tackle the problem of fragmentation that can occur along the path from PE_{fa} to PE_{ne}. We can fragment the packet at PE_{fa} and ensure that the size of the packet is fixed before transmission. We could also employ the Path Maximum Transfer Unit (Path-MTU) discovery process so that packets do not get

 ISBN: 978-1-61208-183-0

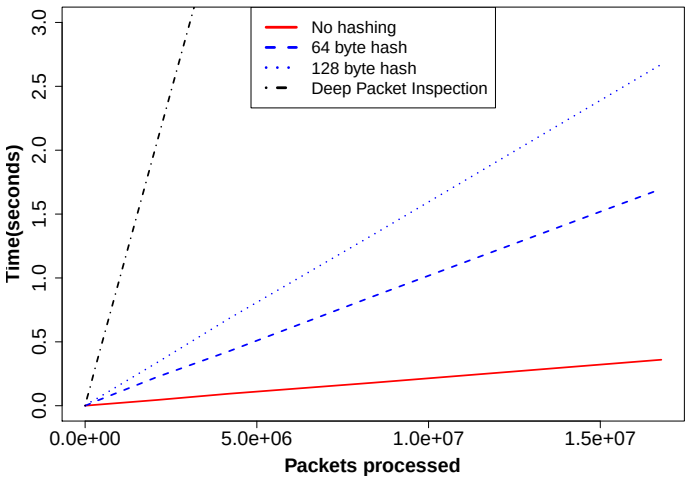


Figure 4: Performance comparison of complete packet encryption and decryption with a 64, 128 byte hash on a payload of size 1024 bytes.

split into multiple fragments. If packets are fragmented this scheme fails. However, networks usually employ the Path-MTU discovery process to prevent fragmentation and hence this problem may not occur.

V. Conclusion and Future Work

In this paper, we proposed a label-hopping scheme for inter-provider BGP-based MPLS VPNs that employ MP e-BGP multi-hop control-plane exchanges. In such an environment, without label-hopping, the data-plane is subject to spoofing attacks.

The technique proposed uses a payload-based label-hopping scheme to prevent attackers from easily deciphering labels and their respective VPNs. The scheme is less computationally intensive than encryption-based methods. It prevents the spoofed packets from getting into a VPN site even if the attacker is in the core or at an intervening link between ISPs. In our scheme, we chose the payload of the packet as the variable component since the use of encryption or IPSec to secure the inner labels are time intensive strategies. Instead of using the payload as a random source, other options like time-of-the-day could be used. This requires the use of time synchronization mechanism. Such mechanisms like "Timing over IP Connection and Transfer of Clock (TicToc)" are receiving much attention from the IETF. This will be the subject of our future study.

Acknowledgements

The authors would like to acknowledge the UK EP-SRC Digital Economy Programme and the Government of India Department of Science and Technology (DST) for funding given to the IU-ATC.

References

[1] S. Alouneh, A. En-Nouaary and A. Agarwal, "MPLS security: an approach for unicast and multicast environments", Annals of Telecommunications, Springer, vol. 64, no. 5, June 2009, pp. 391–400, doi:10.1007/s12243-009-0089-y.

[2] M. H. Behringer and M. J. Morrow, "MPLS VPN security", Cisco Press, June 2005, ISBN-10: 1587051834.

[3] B. Daugherty and C. Metz, "Multiprotocol Label Switching and IP, Part 1, MPLS VPNS over IP Tunnels", IEEE Internet Computing, May–June 2005, pp. 68–72, doi: 10.1109/MIC.2005.61.

[4] L. Fang, N. Bita, J. L. Le Roux and J. Miles, "Interprovider IP-MPLS services: requirements, implementations, and challenges", IEEE Communications Magazine, vol. 43, no. 6, June 2005, pp. 119–128, doi: 10.1109/MCOM.2005.1452840.

[5] C. Lin and W. Guowei, "Security research of VPN technology based on MPLS", Proceedings of the Third International Symposium on Computer Science and Computational Technology (ISCSCT 10), August 2010, pp. 168–170, ISBN-13:9789525726107.

[6] Y. Rekhter, B. Davie, E. Rosen, G. Swallow, D. Farinacci and D. Katz, "Tag switching architecture overview", Proceedings of the IEEE, vol. 85, no. 12, December 1997, pp. 1973–1983, doi:10.1109/5.650179.

[7] E. Rosen and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, Standard Track, February, 2006.

[8] T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein, "Introduction to algorithms", 3rd edition, MIT Press, September 2009, ISBN-10:0262033844.

[9] C. Semeria, "RFC 2547bis: BGP/MPLS VPN fundamentals", Juniper Networks white paper, March 2001.

[10] Advance MPLS VPN Security Tutorials [Online], Available: "http://etutorials.org/Networking/MPLS+VPN+security/ Part+II+Advanced+MPLS+VPN+Security+Issues/", [Accessed: 10th December 2011]

[11] Inter-provider MPLS VPN models [Online], Available: "http://mpls-configuration-on-cisco-ios-software.org.ua/1587051990/ ch07lev1sec4.html", [Accessed 10th December 2011]

 ISBN: 978-1-61208-183-0

A Proposal of Dynamic RWA Using Ant Colony in Optical Burst Switched Networks

Erick A. Donato, Joaquim Celestino Júnior, Antônio S. S. Vieira
Computer Networks & Security Laboratory (LARCES)
State University of Ceará (UECE)
Fortaleza, Ceará , Brazil
{aguiardonato, celestino, sergiosvieira}@larces.uece.br

Ahmed Patel
Software Technology & Management Research Center
Faculty of Information Science & Technology,
University Kebangsaan Malaysia
UKM Bangi, Sengalor, Malaysia
whinchat2010@gmail.com

***Abstract* - One of the main problems in optical networks concerns routing and wavelength assignment for establishing optical circuits. Burst switching is a viable alternative to overcome the problem of idle circuit waste, but it also prevents data to pass to the electrical domain where conversion is necessary, thus making this switching purely optical. A solution to the problem of dynamic RWA in Optical Burst Switching (OBS) networks, called AntOBS, is inspired by ant colony behavior. It is used to decrease the blocking probability during the routes establishment phase. The dynamicity and self-organization are the main features of AntOBS, which through various experiments shows a reduction of the blocking probability of requests from the cross connect electro-optical network.**

Keywords-Optical Networks; Ant Colony; Burst Switching; RWA.

I. INTRODUCTION

Many technologies and techniques have emerged to enhance the use of optical fibers, aiming to enjoy its unique qualities such as very low transmission error, high transmission speed and capacity, overall reliability, and much longer distance ranges before the use of repeaters are required.

Some multimedia applications require certain minimum requirements for good functioning and near real-time source to destination delivery. For example, these applications require that some network resources like buffer space, high payload capacity delivery, throughput and minimum transmission delays are guaranteed as part of a Service Level Agreement (SLA) with stated Quality of Service (QoS) parameters and their values at an offered price. There are some features like very low propagation distortion and transmission error rates in optical networks that can meet the needs of these applications better than non-optical networks.

Optical networks provide high transmission rates with very low interference between virtual channels and also immunity from electromagnetic interference [1]. Only where the merging of an optical network cross connects with non-optical networks, interference and higher error rates and lower capacities are still a problem [1, 2]. Other than this, the optimization of Routing and Wavelength Assignment (RWA) can be made as close as possible to their maximum limits. Any failures of these functions can result in wastage and effectively reduce reaching the upper threshold value limits. Given this fact, it is asserted that optical network may provide the ability to handle failures through smart routing algorithms which take into account the current network state to find the best path between the source and destination nodes considering various bottlenecks and restrictions. In doing so, such algorithms must assign the wavelength so that they maximize the use of network resources by minimizing the blocking probability.

The current problems of RWA algorithms in optical networking have the goal of choosing the path between two nodes in the network and set the wavelength to be used in communication. The performance of these algorithms directly compromises the performance of these networks.

In solving the above mentioned RWA problems, it is further asserted that applying the technique of Ant Colony Optimization (ACO) [3] in proposing a new dynamic RWA algorithm for optical burst switching against a defined set of QoS and autonomic self-organization learned knowledge would result in a more satisfactorily solution compared to the performance of the current RWA algorithm.

ACO is a bio-inspired system that is based on ants' social organization and behavior patterns [3]. Ant societies have division of labour, communication between individuals, and an ability to solve complex problems [4]. Ants communicate with each other using *pheromones* [4]. These chemical signals are more developed in ants than in other similar groups. Like other insects, ants perceive smells with their long, thin and mobile antenna. The paired antennae provide information about the direction and intensity of scents. Since most ants live on the ground, they use the soil surface to leave pheromone trails that can be followed by other ants. In species that forage in groups, a forager that finds food marks a trail on the way back to the colony; this trail is followed by other ants, these ants then reinforce the trail when they head back with food to the colony. When the food source is exhausted, no new trails

 ISBN: 978-1-61208-183-0

are marked by returning ants and the scent slowly dissipates. This behavior helps ants deal with changes in their environment. For instance, when an established path to a food source is blocked by an obstacle, the foragers leave the path to explore new routes. If an ant is successful, it leaves a new trail marking the shortest route on its return. Successful trails are followed by more ants, reinforcing better routes and gradually finding the best path [5]. The colony organization and how the internal ants' communication structure through pheromones is made makes this meta-heuristics functional and appropriate to find the best solution for the problem [6]. The ants come down paths of the "nest" until the "food source" and place the pheromone there. This enables other ants to be induced to paths in which there is the largest amount of pheromone. This induction is probabilistic, what makes a possible variety in the path choice by the ant. This scenario allows the meta-heuristics to converge for the best solutions within a set of viable alternatives.

Ant colony optimization algorithm (ACO) is a probabilistic technique for solving computational problems which can be reduced to finding good paths through graphs. This algorithm is a member of ant colony algorithms family, in swarm intelligence methods, and it constitutes some meta-heuristic optimizations. Initially proposed by Marco Dorigo in 1992 in his PhD thesis [2], the first algorithm was aiming to search for an optimal path in a graph, based on the behavior of ants seeking a path between their colony and a source of food. The original idea has since diversified to solve a wider class of numerical problems, and as a result, several problems have emerged, drawing on various aspects of the behavior of ants as *distributed optimization* [7].

This article is organized as follows: the next section outlines relevant related work; Section 3 describes the main AntOBS features and QoS; Section 4 explains the setup of the simulation and gives the evaluation of the results of from it by comparing and discussing the different approaches of the RWA problem. Section 5 concludes this research study and presents suggestions for future works.

II. ANTOBS

In Optical Burst Switching (OBS) networks, data is stored in the network edge node, waiting for the burst to be mounted. With the ready burst, a wavelength (λ) and the path to be covered are assigned to the burst. This problem is called Routing and Wavelength Assignment (RWA) [8, 9, 10]. Then, a burst control packet (BCP) is sent to the target node. BCP uses a channel independent of the data channel (out-of-band), which goes through the path off the plan in a specific wavelength, using a signaling protocol. The main advantage in switching is the data and control plain separation, which enables a good network management and avoids the resources waste because it does not establish a connection. This fact reduces latency and improves network efficiency, i.e., increases the use of network resources, besides using a viable technology.

On the routing problem, routes are calculated according to some heuristics applied to data network. The goal is to find routes that can satisfy the requests that arrive on the network. In WDM networks, each link has multiple wavelengths, and each one of them can carry different data. Therefore, besides setting the route, the wavelength to be used must be set. On the assignment of wavelength problem, one of the available wavelengths must be chosen for a given route.

The behavior of the ACO and the problems of RWA in OBS networks motivated this study to propose the AntOBS. This is an algorithm based on ACO that treats the routing problem and wavelength assignment in OBS networks dynamically. It is also important to mention that in this work, wavelength converters and buffers are not used.

The signaling protocols are responsible for determining how network resources will be allocated and deallocated. In this work the Just-Enough-Time (JET) signaling protocol [8] was chosen due its main feature: delayed reservation, one-way reservation and implicit release. In JET, after the BCP is sent, the burst is sent without confirmation that it has accomplished its task, i.e., reserve network resources along the route of the burst. The BCP contains the information of the burst size and setup time. This enables the reservation is made only during the time of the burst [8].

In the approach proposed in this paper, the ants are characterized by packets that feed the routing tables storing routes pheromone levels that represent the burst success probability in a route, i.e., the higher the level of pheromone, the lower the burst blocking probability. Ants are generated in the nodes and are sent to targets randomly. As the ants follow the path to a particular target, they update the pheromone level in each node.

Two types of ants are proposed: IAnt and MAnt. The first is responsible for creating the routing tables and the second for maintaining up-to-date pheromone levels. The details of each type of behavior are explained in the following sections. The pathway of ants is covered out-of-band. The ants sending frequency is a system parameter.

The wavelength assignment problem is treated by first-fit algorithm and the nodes OBS are not equipped with wavelength conversion capabilities. The choice of wavelength is independent of the choice of the route. In this algorithm, the wavelengths are put in a list of fixed order, determined in advance. In the search for an available wavelength, the first on the list is chosen. If this is already allocated to another request, the second is tested and it continues until you find one available. Global information is not required. The choice of this algorithm was made by the simplicity of it.

A. Main Features

The model used can be defined as a graph G = (V, E), in which V represents the set of network nodes and E

 ISBN: 978-1-61208-183-0

represents the set of edges or links in the network. Each link e_{ij} represents a connection between node i and node j.

The IAnt ant is released during the optical networking startup to communicate to other network nodes the node source release. Initially, the node knows only its immediate neighbors, so it is necessary to meet all other nodes in the network so that the protocol can work correctly. This ant works based on the behavior of a broadcast packet. The IAnt has the following structure: identifier code, the source node and the number of hops. This phase is called Initialization Phase.

The next phase is the Maintenance Phase. In this, the MAnt ants keep are released by the network nodes. MAnt must maintain and find out new routes to the source node, being launched with a destination chosen randomly. The launching frequency of is a system parameter. The MAnt carries data about the path and the target with the goal of selecting edges with pheromone along its way. When it reaches to the node, this node interprets and renders the MAnt source address as the destination address, i.e., the node will update the routing table in the opposite ant path. Thus, the routing table of the current node is updated to the MAnt source node.

The data conveyed by an Ant MAnt is composed of the following fields: identifier code, source node, target node, stack, Time-To-Live (TTL) and bitmask. The TTL is used to prevent the ant to be on the network indefinitely. The stack is used to store the path covered by the ant. At each hop, the node processes data from the ant and checks if there is a cycle in the path.

The bitmask, in its turn, has the task of bringing the available wavelengths for that route. In this route, each bit represents wavelength availability.

The probability of the ant MAnt, being at node i, choose the node j as next hop path is given by Equation 1.

$$p_{ij}^{k} = \begin{cases} r \cdot \dfrac{[\tau_{ij}]^{\alpha}}{[d_{ij}]^{\beta}}, & \text{se } j \in N_i^k \ ; \\ 0, & \text{se } j \ni N_i^k \ . \end{cases} \qquad (1)$$

In this equation, r is a constant chosen randomly that aims to give diversity to the solution and k is a number of neighbors of node i. N_i^k represents the set of neighbors of node i, and τ_{ij} and d_{ij} represent the pheromone levels and the cost associated to the link e_{ij}. In this study, the cost associated to each link represents the number of collisions that occurred when the node j was chosen. This metric is important because it takes into account the history of collisions in that link. The values of α and β are constant responsible for deciding the importance of the equation terms.

To avoid stagnation in suboptimal solutions, we have proposed a second transition rule for Ant MAnt. So, the Ant MAnt will have two options for transitional rules, in which the Equation 1 is the first one. The second option is a randomly choice, in which the Ant decides the next hop without taking into account the link pheromone level, i.e., all possibilities have the same chance of being chosen. Therefore, before applying the transitional rule, the Ant may decide which rule to use. This first choice is also done statistically, where each rule has the same chance of being chosen.

Equation 2 below illustrates the reinforcement of pheromone [l], i.e., illustrates how the level of pheromone is updated. When an ant reaches the node, it updates the pheromone level. Assume that $\Delta\tau^k$ is a constant and represents the pheromone update. This is one of the system parameters.

$$\tau_{ij} \leftarrow \tau_{ij} + \Delta\tau^{k} \qquad (2)$$

To update the table, the current node assumes the Ant MAnt source node as target and changes the registry corresponding to it. If the Ant arrives to the current node it implies that there is a route from the current node to the source node of the Ant Mant.

Based on the natural behavior of ants, periodically, the pheromone levels change because the evaporation. This allows the choices to be more diversified and that new solutions are found as shown in Equation 3 [l]. The pheromone level is reduced by ρ percent.

$$\tau_{ij} \leftarrow (1-\rho) \cdot \tau_{ij} \ , \ \forall (i,j) \in A \qquad (3)$$

where ρ is a constant responsible for the pheromone evaporation level in the links. Along with α and β, 1 equation coefficients, ρ is also a parameter of the AntOBS algorithm presented below.

B. Initialization Phase

Table I below illustrates the structure of a routing table protocol. The nodes have their probabilities of being selected, represented by the column pheromone and, according to the target column, the next hop is chosen. Initially, only the neighbors are known. Ants add and update new rows in the table, thus new routes become available which will be used by control packets and by bursts of data.

TABLE I. PHEROMONE ROUTING TABLE PROTOCOL

Target	Next Hop	Hops Number	Pheromone
N_1	N_1	0	P^i_1
N_2	N_2	0	P^i_2
N_3	N_1	1	P^i_3
...	...	...	...
N_n	N_2	2	P^i_n

 ISBN: 978-1-61208-183-0

After the simulation start, each node i send an ant IAnt through broadcast to its neighbors. This operation creates new registries and values of the nodes routing tables. A node j, intermediate between the source and target Ant nodes, creates a new registry in the routing table. The node j interprets the IAnt source address as the target address and the previous node's address as the next hop, and then initializes the value of the pheromone in the link e_{ji}. Finally, the node updates the value of the field number of hops and the Ant IAnt is forwarded to its neighbors.

Fig. 1 below illustrates the IAnts behavior. In Fig. 1 (a), the node 0 launches two ants IAnt, one for each neighbor. Then they create a new entry in the routing table with the node 0 as a target and also save the link of the Ant arrival as outbound link to node 0. When nodes 3 and 4 receive Ant respectively from node 2 and 1, Fig. 1 (b), they will create a new entry in the routing table with the node 0 as a target with the respective node as an output.

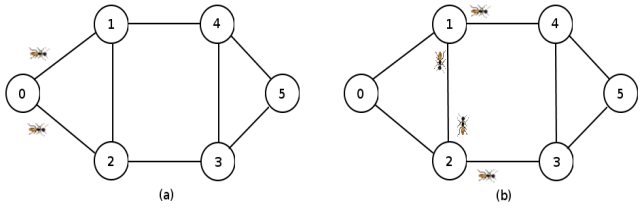


Figure 1. IAnt search and routing behavior

After the end of this phase, all routes have the same probability; therefore, they have the same level of pheromone. Thus, it is necessary that the routes are optimized. This is accomplished through the ants MAnt behavior.

C. Maintenance phase

MAnt ants are responsible for maintaining the routes updated. They circulated through the network updating pheromone levels of links. Periodically, the nodes send the MAnt ants with random targets. The frequency of release of MAnt is a parameter and will be shown later.

When a given node i receives an ant MAnt it can be either the ant target or an intermediate path node to the target. The steps in each of the cases are described in the algorithm below.

Algorithm1 ReceiveMAnt (pkt)
1: **if** *myaddress != pkt.dst address* **then**
2: **if** *pkt.T T L == 0* **then**
3: *exclude (pkt)*
4: **end** if
5: *add entry route table (pkt)*
6: *update pheromones ()*
7: *next hop ← transition rule ()*
8: *new pkt ← pkt*
9: *new pkt.hops++*
10: *new pkt.T T L −−*
11: *new pkt.dst address ← next hop*
12: *send (new pkt)*
13: **else**
14: *update pheromones ()*
15: *exclude (pkt)*
16: **end if**

The MAnt behavior differs from the real ant behavior, because the real ant traverses the path nest-food and goes back using the same path. In the case of ant MAnt, this only makes a one-trip path. This decreases the overhead caused by the use of ant on a network.

In AntOBS, the BCP is only a user of the data generated by the ants, it means, ants create and update the routing tables and the BCP uses the data in these tables.

The process of routes maintenance by the ants happens along the normal use of the network. The main benefit of this use is to ensure that the routes are always in accordance with the current state of the network.

D. Resources saving and Burst sent

The ants start and update the routing tables of the network. However, the BCP uses these routes designed to reserve themselves for the burst sent. The behavior of the BCP is shown below.

On the edge router, the BCP is created. According to the level of pheromone, the next hop is chosen. In this case, the choice is probabilistic, which gives a possibility of diversification in the solution. Soon after, an available wavelength is chosen through First-Fit algorithm [1].

When the BCP leaves the edge node, it acts according to the steps below:

1. Choose the next hop in the routing table taking into account the level of pheromone. This choice is not probabilistic, is based on the highest level of pheromone. Due to the calculation of tuning time, the BCP should not change the number of hops in the path. This could cause a burst blocking;
2. According to the next hop, the possibility of saving the wavelength set in the edge node:
   - If possible:

Set the saving;
Send the BCP;

   - If not possible:
     - Return to step 1 to select a second option to the next hop;

E. Example of burst blocking

Based on Figure 1 and Table 1 structure, suppose there is a burst to be sent from node N_0 to node N_4. There are two possible ways. The first possibility is N_0-N_1-N_4 and the second path is N_0-N_2-N_3-N_4.

As described in the previous section, assume that the BCP choose N_1 as the next hop. After that, choose the wavelength λ_1. Then the BCP is sent from the N_0 to N_1.

 ISBN: 978-1-61208-183-0

If N_1 is checked that no wavelength is available for N_4, then the reservation of wavelength λ_1 is not performed and therefore burst blocking will occur.

III. Related Work

Many research proposals apply computational intelligence techniques to solve the routing problem in networks. Among these techniques there are: genetic algorithms (GA) [11], ant colonies optimization (ACO) [3, 4, 12], particle swarms optimization (PSO) [13].

Techniques using ACO to solve problems in optical networks, including OBS networks, have been proposed in the literature, but there are some deficiencies which in some cases may compromise the network operation. Some studies are discussed below.

In [12], ACO was used to solve the problems of RWA and recovery dynamically. In this work, the source node sends the BCP to the target with the goal of reserving resources for the next coming burst. Initially, as the levels of pheromone (routing tables) has not been started, each hop path is chosen randomly. When the target node receives the BCP, it means, when the reserves for the burst are effective, the target node responds to the source one with a message that travels the opposite way, updating the pheromone levels of intermediate. Thus, all other control packets take into account the level of pheromone for choosing a path to the recipient.

To achieve this stage, two types of control packets are used: BCP-REQ and BCP-ACK. The first packet type BCP-REQ reserves resources along the path from source to destination. The second packet type BCP-ACK returns from target destination to the source updating pheromone levels of intermediate switches. BCP-ACK travels the same path as BCP-REQ, changing only its way.

This solution results in unnecessary overheads and, according to the simulations made in the study, in some cases the problem is not solved satisfactorily. For example, when the network goes down or it does not send bursts successfully, even for a moment, this kind of problem leads to a situation in which the levels of pheromone does not represent the current network state. In this solution, the routing tables are well set only if there is traffic in the network, since the control packet plays the role of setting. It is necessary, therefore to have, the existence of bursts in order to have a smooth optical network operation and configuration. If the routing tables are not close to the ideal setup, many bursts may be lost.

This problem does not happen in the purpose of this study. There is no connection with the ants, the burst and the control packet, that is, the routing tables do not depend on the existence of bursts to be setup. The ants are responsible for configuring routes and are independent of the data plan and control package.

In [14], a solution to the problems of routing and wavelength assignment in WDM optical dynamic networks using ACO is presented. In this work, the ants act separately from the control packet and feed two types of routing tables: one table stores the complete routes in the border and another table stores the routing data in core nodes that have the pheromone levels.

The approach proposed in this article differs from the previous work in some aspects. The first one is that the AntOBS uses a single routing table. Another difference is that the AntOBS stores and considers the number of collisions on the node to update of pheromone, to ensure routing table is updated according to real data from the network operation.

IV. Simulations and Results

This section presents the results obtained in simulations to test the performance of routing AntOBS algorithm, comparing to the algorithm called Pure OBS, a routing algorithm in OBS networks that is not based on Ant Colony and uses the shortest path strategy. This algorithm is based on Dijkstra's algorithm [15]. The simulations were made using the Ubuntu operating system and the Network Simulator 2 (NS-2). The NS-2 is a discrete event based simulator, widely used in research on computer networks. The process of generating traffic is stochastic and follows the Pareto distribution. The algorithm parameters were set according to Table II below and were chosen taking into account the simulations made during the work.

TABLE II. ALGORITHM PARAMETERS USED TO TEST ANTOBS' ROUTING PERFORMANCE

Parameter	Value
α (Importance of pheromone level)	0.5
β (Importance of number of collisions)	0.5
ρ (Evaporation level)	0.4
Δ (Increase pheromone update)	0.6
Mant Ant Creation frequency	0.6 s

During the simulations two topologies were used. Fig. 2 shows the topology of a small test network with six nodes, called topology 1. Fig. 3 shows a topology similar to the network NSFNET with fourteen nodes.

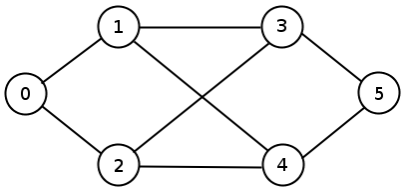


Figure 2. Simulation network topology of test 1

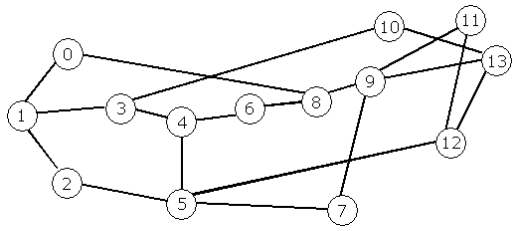


Figure 3. Simulation network topology of test 2

A. Topology of test 1

In the first tests, the network represented by Fig. 2 is used, where two scenarios were used. The first scenario has 6 available wavelengths on each link and the second has 10 available wavelengths on each link. The amounts of wavelength in each scenario were determined taking consideration the simulations done.

Fig. 4 shows the graph of the probability of blocking performance behavior of new AntOBS versus Pure OBS algorithms as a function of load on the network using six wavelengths, with. Not surprisingly, with low loads, the blocking probability rate is substantially lower than in high loads. Due to the reduced number of wavelengths, with high loads the network enables a high burst probability of blocking.

It is also possible to realize in the graph that the probability of blocking of the algorithm AntOBS is smaller than to the Pure OBS algorithm, a routing algorithm in OBS networks that uses the shortest path strategy. The probability of blocking is shorter, mainly from 18 Erlangs on. The congestion generated by the load increasing on the network is the main cause of this difference. Although with this increased workload, the dynamicity of the AntOBS ensures that the best routes will be chosen taking into account the current network status, preventing the block bursts.

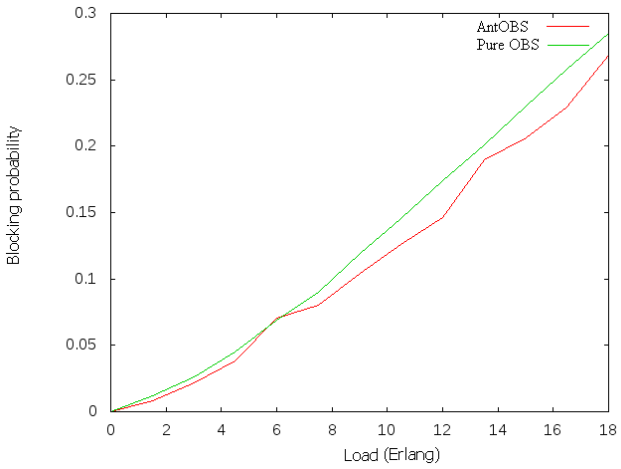


Figure 4. Blocking burst probability performance behavior of new AntOBS versus pure OBS algorithms in topology 1 test with 6 wavelengths.

Fig. 5 shows the graph of the blocking probability as a function of load of the optical Network using topology 1 (see Fig. 2), in which 10 wavelengths are available. The increase in the number of wavelengths obviously decreases the blocking probability in both cases, but the AntOBS continues with the fall of it in both scenarios.

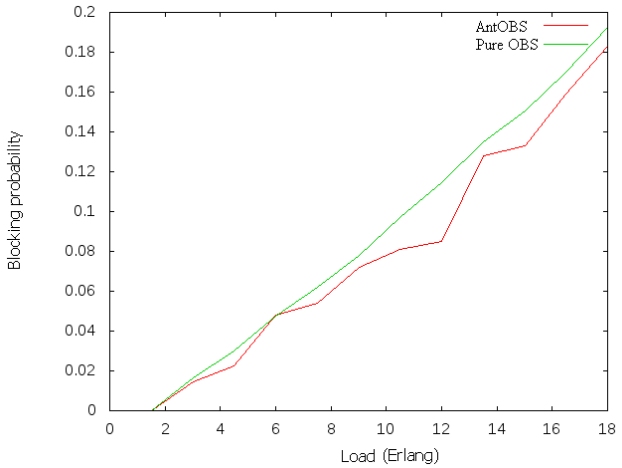


Figure 5. Blocking burst probability performance behavior of new AntOBS versus pure OBS algorithms in topology 1 test with 10 wavelengths.

B. Topology of test 2

Two scenarios have been defined for the second topology again: with 8 and 12 wavelengths, respectively.

Fig. 6 shows the comparison between the AntOBS and Pure OBS algorithms, illustrating the blocking graph as a function of the network load with 8 wavelengths available.

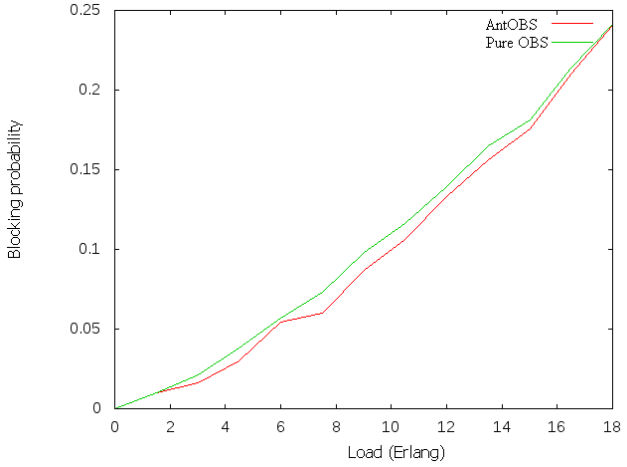


Figure 6. Blocking burst probability performance behavior of new AntOBS versus pure OBS algorithms in topology 2 test with 8 wavelengths.

 ISBN: 978-1-61208-183-0

The graph behavior shows again a performance improvement of AntOBS over Pure OBS, allowing a decrease in the probability of blocking, in the same scenario of 18 Erlangs load.

Fig. 7 compares the algorithms AntOBS and Pure OBS, using a topology similar to the network NSFNET with 12 wavelengths. In this scenario, once again the AntOBS had the blocking probability less than the Pure OBS.

It is possible to notice that in simulations in this topology, when in high loads, the behavior of algorithms is very close. However, despite having similar performances it is possible to see that in almost all scenarios the AntOBS had better performance, and the efficiency of AntOBS in relation to the Pure OBS could be proved, since any blocking on optical networks can lead to loss of high amounts of data, mainly in OBS networks which work in bursts.

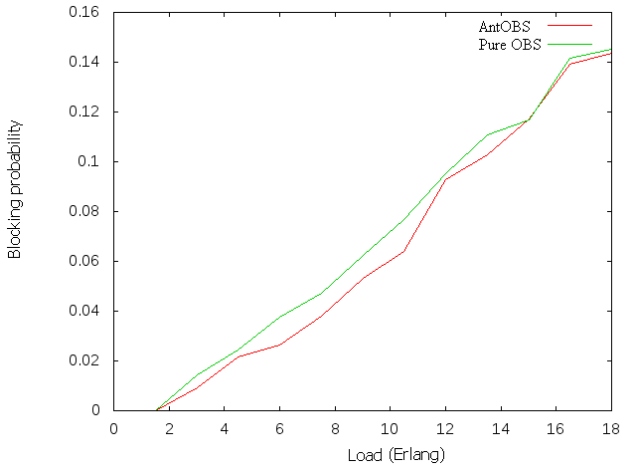


Figure 7. Blocking burst probability performance behavior of new AntOBS versus pure OBS algorithms in topology 2 with 12 wavelengths.

Comparing the results of the topology simulations, topology 2 (see Fig. 3) results showed a minor difference between the performances of two algorithms. This is caused by the greater number of nodes of topology 2, which requires a greater number of updates of routing tables, i.e., it decreases the AntOBS performance.

V. Conclusion and Future Works

This work proposed an algorithm, called AntOBS, aimed in decreasing the blocking probability in OBS networks, taking into account that traffic in this type of network is done in bursts and that blocking can lead to loss of an excessive amount of data. This solution is based on adaptive behavior of an Ant Colony.

One important solution presented concerns the possibility of AntOBS to treat the problem of Adaptive RWA, through changes in the routing tables provided by ants, a route can be changed in case of a link failure.

In terms of signaling, the cost of AntObs Algorithm was small. During the simulations no ant or BCP were lost.

For a future work, another type of Ant can be inserted into the network to do new tasks in the optical network, for example, to check the link status periodically. Other signaling protocols can be tested such as JIT, which explores the immediate saving; in order to compare to the results of this work which used the JET.

References

[1] R. Ramaswami and K. N. Sivarajan, Optical Networks: A Practical Perspective. 2 ed., San Francisco, CA: Morgan Kaufmann Publishers 2002, ISBN 1558606556.

[2] M. Dorigo, Optimization, Learning and Natural Algorithms, PhD thesis, Politecnico di Milano, Italy, 1992.

[3] M. Dorigo and G. D. Caro, “AntNet: distributed stigmergetic control for communications networks,” Journal of Artificial Intelligence Research, 1998, pp. 317-365.

[4] D. E. Jackson and F. L. Ratnieks, "Communication in ants," Current Biology, August 2006, vol. 16 (15), pp. R570–R574, doi:10.1016/j.cub.2006.07.015. PMID 16890508.

[5] S. Goss, S. Aron, J. L. Deneubourg, and J. M. Pasteels, "Self-organized shortcuts in the Argentine ant," Naturwissenschaften, vol. 76, pp. 579–581, 1989, doi:10.1007/BF00462870.

[6] E. Dicke, A. Byde, D. Cliff, and P. Layzell, "An ant-inspired technique for storage area network design," in A. J. Ispeert, M. Murata & N. Wakamiya. Proceedings of Biologically Inspired Approaches to Advanced Information Technology: First International Workshop, BioADIT 2004 LNCS 3141, pp. 364–379.

[7] A. Colorni, M. Dorigo, and V. Maniezzo, “Distributed Optimization by Ant Colonies,” Actes de la première conférence européenne sur la vie artificielle, Paris, France, Elsevier Publishing, 1991, pp. 134-142.

[8] M. Yoo and C. Qiao, “Just-enough-time (JET): A high speed protocol for bursty traffic in optical networks,” in Proceeding of IEEE/LEOS Conf. on Technologies For a Global Information Infrastructure, 1997, pp. 26-27.

[9] J. Y. Wei and R. I. McFarland, “Just-In-Time signaling for WDM optical burst switching networks,” Journal of Lightwave Technology, 2000, pp. 2019-2037.

[10] B. Mukherjee, Optical WDM Networks. Davis, CA: Springer, 2006, ISBN 0387290559.

[11] D. Bisbal, “Dynamic Routing and Wavelength Assignment in Optical Networks by Means of Genetic Algorithms,” in Photonic Network Communications, vol. 7, nº 1, pp. 43–58, 2004.

[12] Z. Shi, Y. TinJin, and Z. Bing, “Ant algorithm in OBS RWA,” Proc. of SPIE, Optical Transmission, Switching and Subsystems II, vol. 5625, pp. 705-713, Feb. 2005.

[13] A. W. Mohemmed, N. C. Sahoo, and K. G. Tan, “Solving Shortest Path Problem Using Particle Swarm Optimization,” in Applied Soft Computing, Jan. 2008, pp. 1643-1653.

[14] S. H. Ngo, X. Jiang, and S. Horiguchi, “An ant-based approach for dynamic RWA in optical WDM networks,” Photonic Network Communications, vol. 11, no. 1, pp. 39-48, Jan. 2006.

[15] E. W. Dijkstra, "A note on two problems in connexion with graphs," Numerische Mathematik 1, 1959, pp. 269–271, doi:10.1007/BF01386390.

 ISBN: 978-1-61208-183-0