JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

1/2016

(*Contents Continued on Back Cover*)

Editorial Board

Editorial Advisory Board

ISSN 1509-4553 on-line: ISSN 1899-8852

Circulation: 300 copies

Sowa – Druk na życzenie, www.sowadruk.pl, tel. 22 431-81-40

JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

Preface

In this issue of the *Journal of Telecommunications and Information Technology* we have collected a set of fourteen papers documenting a broad range of topics related to modern wire and wireless telecommunications networks. Rapid advances in wire and wireless technologies have significantly accelerated the introduction of new technologies and services, defining and building the rules to ensure the transfer of huge volumes of data.

Quality of Service is becoming one of the most important criteria for the selection of old and new services and technologies. Two papers in this issue are related to this topic. In the first one, *QoS Requirements as Factor of Trust to 5G Network*, the authors V. Tikhvinskiy, G. Bochechka, and A. Gryazev discuss the role of QoS in the formation of trust in both consumers and regulators in the case of the 5G network. The second article, entitled *Digital Audio Broadcasting or Webcasting: A Network Quality Perspective*, describes the advantages and disadvantages of digital audio broadcasting and webcasting transmission techniques from a network quality perspective. The authors P. Gilski and J. Stefański present a case study of user expectations with respect to the perceived quality of real digital broadcasted and webcasted radio stations.

The next three articles concern the topic of network monitoring. The paper *Monitoring of a Cloud-Based Environment for Resilient Telecommunication Services* by G. Wilczewski presents a tool designed for Data Center resources monitoring. The results presented in the paper allow a high-level resiliency analysis of telecommunication services. The problem of ensuring the quality of service of video is discussed in the paper entitled *Study of No-Reference Video Quality Metrics for HEVC Compression*. The authors K. Rouis, M. Leszczuk, L. Janowski, Z. Papir, and J. B. H. Tahar propose the application of a No-Reference (NR) quality assessment measurement for High Efficiency Video Coding (HEVC). In the last paper concerning this topic, entitled *Quantifying the Suitability of Reference Signals for the Video Streaming Analysis for IPTV*, the authors C. Hoppe, R. Manzke, M. Rompf, and T. Uhl, present the assessment of the quality of video streaming in IPTV based on PEVQ and VQuad-HD algorithms. The conducted measurements provide information that may be valuable for determining the QoS of IPTV services in practice.

The need to introduce new services directly affects the development of ICT operator networks. However, network expansion entails substantial investments. Therefore, the optimization process is very important for telecommunication networks. This problem is addressed

in the next six papers. The first three present analytical methods that may be used for modeling and dimensioning of elements of modern multi-service ICT networks. The next two discuss algorithms enhancing the efficiency of wireless networks. The last paper introduces a new algorithm for the optimization of the expected costs of project implementation. The first paper, *Properties of the Multiservice Erlang's Ideal Gradings* by S. Hanczewski and D. Kmiecik, discusses the conditions for the application of the Erlang's Ideal Grading (EIG) for modeling of the multiservice systems. In the second paper, entitled *Call Blocking Probabilities of Multirate Elastic and Adaptive Traffic under the Threshold and Bandwidth Reservation Policies*, authors I. D. Moscholios, M. D. Logothetis, A. C. Boucouvalas, and Vassilios G. Vassilakis propose an analytical model of a multi-service system in which threshold and reservation traffic management mechanisms have been applied. The next paper, *Estimation of Network Disordering Effects by In-depth Analysis of the Resequencing Buffer Contents in Steady-state* by A. Pechinkin and R. Razumchik, presents a calculation method that allows the analysis of the resequencing problem in the buffers of packet networks. In the fourth article, entitled *Multicast Connections in Wireless Sensor Networks with Topology Control*, the authors M. Piechowiak, K. Stachowiak, and T. Bartczak discuss the performance analysis of multicast trees constructed by heuristic routing algorithms in relation to protocols of topology control for wireless sensor networks. The fifth paper in this group, *LDAOR – Location and Direction Aware Opportunistic Routing in Vehicular Ad hoc Networks* by M. Barootkar, A. Ghaffarpour Rahbar, and M. Sabaei, proposes an opportunistic routing mechanism called Location and Direction Aware Opportunistic Routing (LDAOR) for Vehicular Ad hoc Networks. The algorithm finds the best neighbor node based on, i.e., vehicle positions and directions, and prioritization of messages from buffers. The investigations conducted by the authors show that LDAOR not only increases the delivery rate, but also reduces network overhead, traffic loss, and number of aborted messages. The last article in this group concerning optimization of telecommunication networks is entitled *A Novel Technique of Optimization for the COCOMO II Model Parameters using Teaching-learning based Optimization Algorithm*. The authors, T. T. Khuat and M. Hanh Le, propose a novel technique to optimize the estimation of project cost. In the paper, the teaching-learning-based optimization (TLBO) algorithm for the COCOMO II model is presented. The results indicate that the proposed TLBO algorithm allows for obtaining better estimation capabilities compared to the original COCOMO II model.

One of the best-developed research areas of wire and wireless networks are broadband wireless networks. This is the subject of the next paper, *100 Gb/s Data Link Layer – from a Simulation to FPGA Implementation* by Ł. Łopaciński, M. Brzozowski, R. Kraemer, S. Buechner, and J. Nolte. The paper presents a simulation and hardware implementation of a data link layer for 100 Gb/s terahertz wireless communications. The investigations show that uncoded transmissions are most influenced by the change of the segment size and that the FPGA memory footprint can be reduced when the hybrid automatic repeat request type II is replaced by type I with link adaptation.

The introduction of new services and technologies leads to increased energy consumption. Therefore, it is important to develop technologies and algorithms enabling the reduction of this consumption. In the paper *DS-UWB and TH-UWB Energy Consumption Comparison*, the authors A. Elabboubi, F. Elbahhar, M. Heddebaut, and Y. Elhillali study the energy efficiency of multi-user techniques for UWB systems. The elaborated energetic model can be used as a green communication tool in order to determine the best multiuser techniques. Threats to modern ICT systems also arise from giving access to data connected with users' customs and activities. One of the easiest methods of obtaining information about users' customs is monitoring the consumption of electricity. It is made possible by automatic systems monitoring electricity consumption. Ensuring security in such systems is the topic of the next paper, *Monetary Fair Battery-based Load Hiding Scheme for Multiple Households in Automatic Meter Reading System* by R. Negishi, S. Haruta, C. Inamura, K. Toyoda, and I. Sasase. In the paper, the authors proposed Battery-based Load Hiding (BLH) algorithms to obfuscate the actual user's energy consumption profile by charging and discharging. The proposed BLH algorithms are discussed in the case of multiple households where one battery is shared among them due to its high cost.

Sławomir Kula, PhD, Warsaw University of Technology
Piotr Zwierzykowski, PhD, Poznan University of Technology
Guest Editors

Paper

QoS Requirements as Factor of Trust to 5G Network

Valery Tikhvinskiy[1], Grigory Bochechka[1], and Andrey Gryazev[2]

[1] *LLC Icominvest, Moscow, Russia*

[2] *Federal State Unitary Enterprise Central Science Research Telecommunication Institute, Moscow, Russia*

Abstract—Trust to modern telecommunications networks plays an important role as a driver of technological and market success of any technology or telecommunication services. Most of the technological approaches to this problem are focused only on network security and do not include such a factor as the quality of service (QoS), which also plays an important role in the formation of trust both from the consumers and the regulator. The future 5G mobile technology will be the engine of development of telecommunications until 2020 and the formation of trust to the 5G networks is one of the main tasks for developers. The authors present the view on the trust to 5G networks in the plane of QoS requirements formation and QoS management. QoS requirements to 5G networks were determined on the basis of three main business models of services: xMBB, M-MTC and U-MTC and the need to ensure user trust to networks. Infrastructure requirements for QoS control and spectrum management network entities which are based on Network Function Virtualization (NFV) principles have been formed.

Keywords—network performance, network security, QoE, trusted network.

1. Introduction

Currently leading organizations in international standardization and development of telecommunication technologies such as: ITU, 3GPP, IEEE and ETSI have not formulated a strict definition of "trusted network". However, the trust to communication network significantly affects consumers' choice of communication operator, regulation of operators' activities by state bodies, as well as the market demand on communication services and equipment.

Trust to network or communication technology has market and regulatory aspects that can contribute to the development of the network and technology and increase attractiveness of the services. Therefore, networks and communication technologies should correspond to both market and regulatory requirements of trust.

Given the many factors affecting the trust to 5G networks, in this article authors will briefly review the major factors and examine in details the impact of service quality on the trust to 5G networks.

2. Factors Affecting the Trust to 5G Networks

The existing understanding of "trusted network" is based on the concepts taken by the developers of computer networks, which traditionally include [1]:

- secure guest access – guests obtain restricted network access without threatening the host network;
- user authentication – trusted network integrates user authentication with network access to better manage who can use the network and what they are allowed to do;
- endpoint integrity – trusted network performs a health check for devices connecting to the network. Devices out of compliance can be restricted or repaired;
- clientless endpoint management – trusted network offers a framework to assess, manage and secure clientless end points connected to the network, such as IP phones, cameras and printers;
- coordinated security – security systems coordinate and share information via the Interface for Metadata Access Points (IF-MAP) standard improving accuracy and enabling intelligent response.

According to Kaspersky Internet Security company definition [2], trusted network is a network that can be considered absolutely safe within which your computer or device will not be subjected to attacks or unauthorized attempts to gain access to your data.

Proposed comprehensive look on the issue of trusted communication networks complements the concepts of computer networks developers by the views of consumers, which also comprise quality of services provided by trusted network. The view on the trusted network from the quality aspects is not always taken into account when creating the new mobile technology that reduces trust to the network, both on the part of subscribers and the regulator.

To implement a systematic approach to the trusted communication network the trust of two major players in the telecommunications market should be considered: consumers and regulators that provide both market demand on the communication services and the effectiveness of operators' network infrastructure. As can be seen from Fig. 1, consumers' and regulator's requirements to trusted mobile communication network may either coincide or differ. The main factors affecting the trust of the subscriber and the regulator are shown in Table 1 taking into account their importance in descending order.

Most of consumers' and regulator's factors are the same but factors determining consumer trust, according to the

author's evaluation, have the dominant influence on the mobile network.

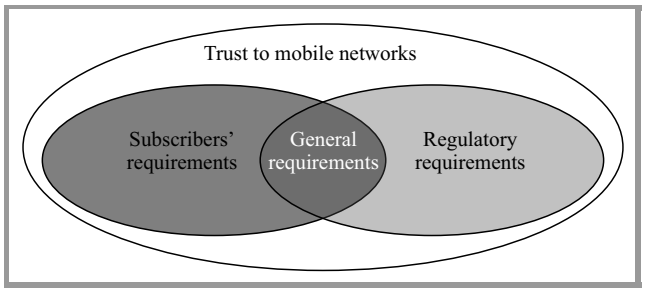


Fig. 1. Domains of trust to mobile networks.

Traditional factors of consumers' and regulator's trust to 5G networks are information security of confidential user data, security of subscriber's devices and network infrastructure. The basis for such security is the resistance to physical attacks on subscriber devices, such as illegal substitution of Subscriber Identification Modules (SIM card), installation of the malicious software on the user device and the impact on the user device configuration, resistance to network attacks on user devices and network infrastructure, such as DoS-attacks and Man-in-the-middle attacks, and resistance to attacks on confidential user data.

Table 1
The main factors affecting the trust of the subscriber and the regulator to network

Importance	Consumer	Regulator
1	Quality of Service	Network security
2	Quality of Experience	Information security
3	Information security	Network performance
4	Network performance	Network reliability
5	Network reliability	Quality of Service
6	Convenience and security of subscriber's equipment	

Ensuring the safety functioning of 5G networks, devices and applications, including the security of transmission and storage of user data, is a major priority for future 5G technologies and networks developers.

In addition to security performance, the trust of users and regulators to 5G networks will depend on quality performance since security of the mobile network itself does not guarantee that the communication service will be provided without interruption and with the stated quality. Reduced quality of 5G networks will lead to a decrease of trust to them, and as a result in an subscribers outflow. Also, given that the 5G network will be used in a variety of financial systems, public safety systems, traffic and energy management systems, the deterioration of their quality could lead to the human life loss, environmental disasters and financial frauds.

Quality parameters of 5G networks can be divided into three levels: Network Performance (NP), Quality of Service (QoS) and Quality of Experience (QoE), as shown in Fig. 2. NP and QoS are objective indicators that can be

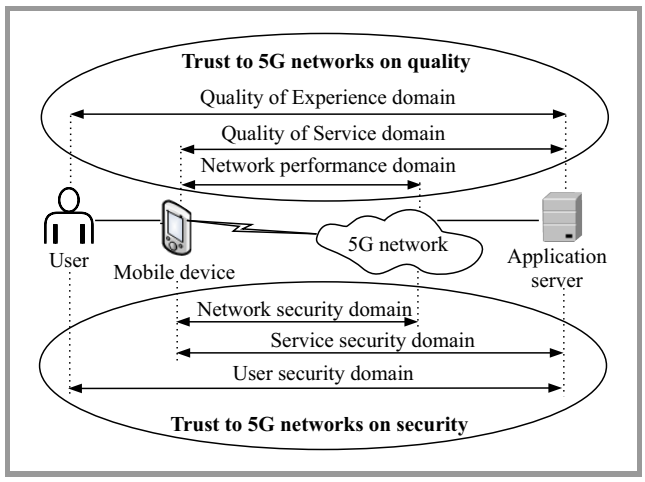


Fig. 2. Quality and security levels of trust to mobile networks.

measured using specialized analyzers while QoE indicators are subjective, estimated by users based on their personal experience. The deterioration of QoS and NP will primarily lead to lower trust to 5G networks of regulators and Business-to-Business (B2B), Business-to-Government (B2G) customers, while the QoE deterioration will lead to lower trust of mass market.

3. Services in 5G Networks

METIS and 5GIS projects consider three basic business models of 5G services: extreme mobile broadband (xMBB), massive machine type communications (M-MTC) and ultra reliable machine type communications (U-MTC) [3].

Forecasts of the leading specialists working in international 5G projects [4], [5] show that video services, such as HD and UHD, with high quality resolution will have a dominant position among services rendered in 5G networks. According to reports of leading 4G networks operators, video services dominate in the subscribers' traffic and will continue to dominate in 5G networks content.

For instance now the traffic volume of video services is estimated by different operators [4] from 66 to 75% of the total traffic in 4G network, including 33% for YouTube services and 34% for clear video as well as CCTV (Closed Circuit TV) video surveillance monitoring in M2M networks. In addition, by 2020 the volume of mobile M2M connections will grow with CAGR index of 45% [6] up to 2.1 billion connections. Given the growing mass scale of M2M services in all industries, they will dominate over basic services (voice & data) in 4G and 5G networks.

5G European development strategy also aims to enable subscribers by 2025 to choose how to connect to TV broadcast: via 5G modem or antenna with DVB-T, so this will require appropriate quality management mechanisms.

Therefore, the efforts of developers to improve the quality management mechanisms will focus on video and M2M services traffic, improvement of quality checking algorithms and creation of new quality assessment methods.

4. Traffic in 5G Networks

When forming requirements to QoS in 5G networks two key traffic models should be firstly considered: high-speed video flow server-subscriber and massive M2M.

Video transmission services will be an important stimulus to development and a rapidly growing segment of 5G networks traffic. Video services accounted for around 45% of mobile data traffic in 2014, and 60% of all mobile data traffic will be from video by 2020 [7]. Mobile video traffic will grow by 55% annually from 2014 to 2020 [8]. Thus, we can already observe the first wave of oncoming "tsunami" of subscribers' traffic in 4G networks. Monthly consumption of data transmission traffic in 4G networks has already reached 2.6 GB and monthly consumption of traffic in 5G networks will exceed 500 GB per user.

The growth of video services traffic volume will be associated with the implementation of various technologies of video services image quality from standard SD TV to UHD TV (8K), which in its turn requires a data transmission speed of up to 10 Gb/s in the network. Technological capabilities of mobile networks of various generations to broadcast video for various video image qualities are shown in Fig. 3 [9], [10]. Capability of video broadcasting depends on data transmission speed in the radio access network.

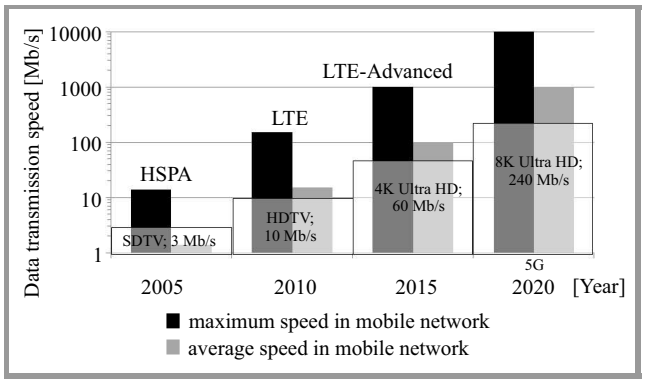


Fig. 3. Technological capabilities of video transfer for mobile networks of various generations.

According to forecasts shown in Fig. 4, in year 2019 the number of M2M connections in the networks of mobile operators (2G, 3G, 4G) will exceed 2.2 billion [11], which is 4 times more than in 2014. The share of M2M connections of the total number of connections in the mobile operators' networks will increase from the current 7% to 22% in 2019.

Strategies of M2M operators are aimed at creating universal M2M platforms capable of operating in multiple vertical economic sectors. This will lead to the possibility to implement approaches, tools, and processing methods for structured and unstructured Big Data derived from M2M networks.

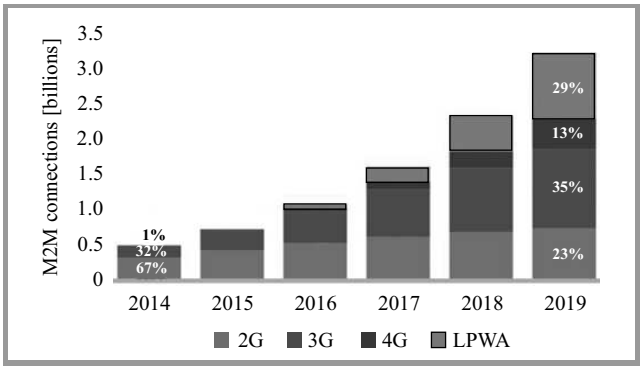


Fig. 4. Number of M2M connections in mobile networks.

According to ABI Research forecasts, the M2M Big Data and analytics industry will grow a robust 53.1% over the next 5 years from 1.9 billion USD in 2013 to 14.3 billion USD in 2018. This forecast includes revenue segmentation for the five components that together enable analytics to be used in M2M services: data integration, data storage, core analytics, data presentation, and associated professional services.

5. Quality Requirements in 5G Networks

METIS project has identified 12 use cases for 5G networks and formed QoE requirements for them [12]. QoE performance requirements that provide trust to network 5G are presented in Table 2. The highest requirements for Experienced user throughput are formed for "Virtual reality office" use case. End-users should be able to experience data rates of at least 1 Gb/s in 95% of office locations and at 99% of the busy period. Additionally, end-users should be able to experience data rates of at least 5 Gb/s in 20% of the office locations, e.g. at the actual desks, at 99% of the busy period. The highest requirements for network latency are formed for "Dense urban information society" use case, device-to-device (D2D) latency is less than 1 ms. The highest requirements for availability and reliability of 5G network are identified for "Traffic safety and efficiency" use case, 100% availability with transmission reliability

Table 2
The main factors affecting the trust of the subscriber and the regulator to network

QoE indicators	Requirements
Experienced user throughput	5 Gb/s in downlink and uplink
Latency	D2D latency less than 1 ms
Availability	$\approx 100\%$
Reliability	99.999%

of 99.999% are required to provide services at every point on the road.

During the evolution of QoS management mechanism in 3GPP (GSM/UMTS/LTE) networks there was a migration from QoS management at the user equipment level to the QoS management at the network level. This approach to QoS management will be maintained in 5G networks as well.

QoS management mechanisms in 5G networks should provide video and VoIP traffic prioritization towards web-search traffic and other applications tolerant to quality.

The service of streaming video transfer without buffering is very sensitive to network delay, so one of the most important parameters that determine QoS requirements is the total packet delay budget (PDB), which is formed on the RAN air interface and is treated as the maximum packet delay with a confidence level of 98%.

Table 3 lists the requirements for delay in 3G/4G/5G networks formed in 3GPP [13] and METIS project [14]. These data demonstrate that with the increase in mobile network's generation the requirements for the lower boundary of the total data delay across the network decline. Also the analysis of the requirements for the overall 5G network delay revealed that given the accumulation effect the delay in 5G Radio Access Network (RAN) should be less than 1 ms.

Table 3
Requirements for delay in 3G/4G/5G networks

QoS terms	Packet Delay Budget [ms]		
	3G	4G	5G
Without quality assurance	Not determined	100-300	Not determined
With guaranteed quality	100-280	50-300	1

On air interface level, the need to transmit control and user data quickly in time domain leads to the demand of fast link direction switching and to short transmission time interval (TTI) length [5]. New 5G frame structure for low latency has been proposed in [15]. Requirements to 5G Radio Access Technology (RAT) delay components in comparison with LTE-Advanced TDD and FDD technologies presented in Table 4. The maximum possible TTI must be less than 0.25 ms for 1 ms radio latency.

Another parameter is the proportion of packets lost due to errors when receiving data packets – IP Packet Error Loss Rate (PELR). Values for this parameter that determines requirements for the largest number of IP packets lost for video broadcasting through 3G/4G/5G mobile networks are shown in Table 5 [16].

For M2M services, the quality also will be determined by the proportion of packets lost when receiving in 3G/4G/5G networks. Given service conditions of M2M subscriber devices determined for both cases: with a guaranteed quality of service and without guarantees, requirements to the share of lost packets differ by three orders. Requirements to the PELR for M2M services are shown in Table 6.

Table 4
Requirements to 5G RAT delay components

Delay component [ms]	5G requirements	LTE-Advanced TDD	LTE-Advanced FDD
User equipment processing	0.2	1	1.5
Frame alignment	0.125	1.1-5	
TTI duration	0.25	1	1
eNB processing	0.3	1.5	1.5
HARQ Re-Tx (10% x HARQ RTT)	0.1	1.16	0.8
Total delay	0.975	5.8-9.7	4.8

Table 5
Requirements to the Packet Error Loss Rate for video broadcasting

QoS terms	Packet Error Loss Rate			
	SDTV	HDTV	4K UHD	8K UHD
Possibilities of mobile communication generation	3G/4G	4G	4G	5G
Video broadcasting with guaranteed quality	10^{-6}	10^{-7}	10^{-8}	10^{-9}

Table 6
Requirements to the Packet Error Loss Rate for M2M services

QoS terms	Packet Error Loss Rate		
	3G	4G	5G
Without guaranteed quality (non-GBR)	10^{-2}	10^{-3}	10^{-4}
With guaranteed quality (GBR)	10^{-2}	10^{-3}	10^{-7}

The development of NFV concept will lead to virtualization of quality management function that could be introduced in the form of two main functions: Cloud QoS management function (CQMF) and Cloud QoS control function (CQCF) [8] shown in Fig. 5.

CQCF function of QoS control provides real-time control of traffic flows in 5G network based on QoS levels established during the connection. Basic QoS control mechanisms include traffic profiling, planning and management of data flows.

CQMF function of QoS management provides QoS support in 5G network in accordance with SLA service contracts, as well as provides monitoring, maintenance, review and scaling of QoS.

Implementation of algorithms for traffic prioritization in 5G networks will be based on traffic classification procedures with a focus on video traffic priorities and M2M traffic. Traffic classification procedure should be done taking into

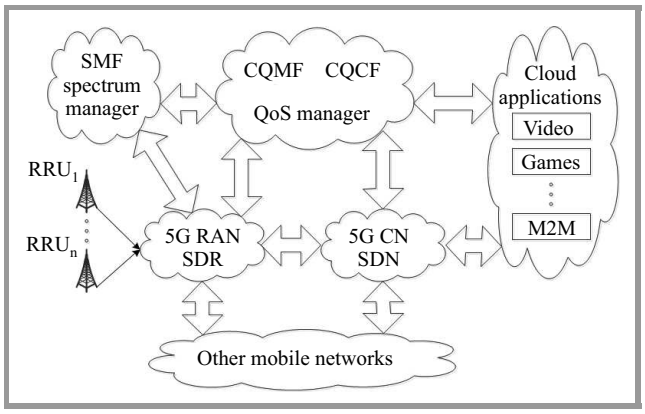


Fig. 5. Virtualization of control and management functions in 5G networks.

consideration the adaptation possibility as the traffic characteristics will dynamically change with the emergence of new applications, both in M2M area and in the field of video services.

In addition to QoS management functions in 5G network, related to traffic management and prioritization, the scope of service quality management also includes management of radio frequency resources used by mobile network (spectrum toolbox). Capabilities of access to the radio spectrum on the principles of Licensed Share Access (LSA) in 5G networks require QoS guarantees to operators who granted access to their spectrum for other operators [17], [18].

The Spectrum Management Function (SMF) in the 5G network is designed to a Spectrum Manager entity. In case of shortage of frequency resources to provide service with required QoS, 5G network must decide to use additional frequency channels for aggregation and select the channel from the frequency ranges which use spectrum based on LSA or Licensed Exempt (LE) principles [17].

Therefore, QoS manager must have information exchange with Spectrum Manager to effectively manage the spectrum resources in the interest of 5G network QoS and trust.

6. Conclusion

The emergence of 5G networks on the market in 2020 will be focused on a significant improvement of characteristics of mobile networks including quality of service that will provide a high level of trust to these networks.

One-sided view on trusted 5G network from security position will limit the growth of trust of customers and regulators. Forming of high level requirements in QoS field will allow 5G developers obtain the trust to 5G on early stage.

Given that the principles of QoS control will be preserved during the transition from 4G to 5G, main effort of 5G developers should be focused on the virtualization of network functions, responsible for the management and control of QoS in the network. Also QoS architecture of 5G should provide information exchange between QoS manager and Spectrum Manager for effective management of spectrum resources for the benefit of ensuring QoS and trust to 5G networks.

References

[1] Trusted Computing Group, Network Access & Identity [Online]. Available: http://www.trustedcomputinggroup.org/

[2] Kaspersky Internet Security, Trusted network [Online]. Available: http://support.kaspersky.com/6423 (accessed 19.06.2015).

[3] ICT-317669-METIS/D6.6, "Final report on the METIS 5G system concept and technology roadmap", Project METIS Deliverable D6.6, 30/04/2015.

[4] Y. Weimin, "No-Edge LTE, Now and the Future", Huawei Pesentation, 5G World Summit 2014 [Online]. Available: http://ws.lteconference.com/

[5] E. Lähetkangas *et al.*, "Achieving low latency and energy consumption by 5G TDD mode optimization", in *Proc. IEEE Int. Conf. Commun. ICC 2014*, Sydney, Australia, 2014.

[6] V. O. Tikhvinskiy, G. S. Bochechka, and A. V. Minov, "LTE network monetization based on M2M services", *Electrosvyaz*, no. 6, pp. 12–17, 2014.

[7] Ericsson Mobility Report, On The Pulse Of The Networked Society, June 2015.

[8] V. Tikhvinskiy and G. Bochechka, "Perspectives and Quality of Service requirements in 5G Networks", *J. Telecommun. Informa. Technol.*, no. 1, pp. 23–26, 2015.

[9] "Series H: Audiovisual and multimedia systems. Infrastructure of audiovisual services – Coding of moving video. High efficiency video coding". Recommendation H.265, ITU-T.

[10] E. Puigrefagut, "HDTV and beyond", in *Proc. ITU Regional Seminar on Transition to Digital Terrestrial Television Broadcasting and Digital Dividend*, Budapest, Hungary, 2012.

[11] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2014–2019, White Paper, 3 Feb., 2015.

[12] ICT-317669-METIS/D1.1, "Scenarios, requirements and KPIs for 5G mobile and wireless system", Project METIS Deliverable D1.1, 29/04/2013.

[13] A. Scrase, "Hot topic: 5G", *BSI/ETSI Telecoms Standards Workshop. The future of telecoms standards*, London, Jun. 2015.

[14] Project METIS Deliverable D2.1 Requirements and general design principles for new air interface, 31.08.2013.

[15] P. Fleming, "Research in 5G Technologies at Nokia Networks", *MIT Wireless Center 5G Day*, May 8, 2015.

[16] ETSI Technical Specification. Digital Video Broadcasting (DVB); Transport of MPEG-2 TS Based DVB Services over IP Based Networks. ETSI TS 102 034 V1.4.1, Aug. 2009.

[17] ICT-317669-METIS/D5.4, "Future spectrum system concept", Project METIS Deliverable D5.4, 30/04/2015.

[18] G. Bochechka and V. Tikhvinskiy, "Spectrum occupation and perspectives millimeter band utilization for 5G networks", in *Proc. ITU-T Conf. "Kaleydoscope 2014"*, St. Petersburg, Russia, 2014.

Valery O. Tikhvinskiy works as Deputy General Director of LLC Icominvest on innovation technologies – the finance investment company in telecommunication sector, and as Chairman of Information and Telecommunication Technologies branch of Russian Academy of Natural Sciences. He is a Doctor Economical Science (2003), and received the Ph.D. degree in Radio engineering (1988), the Government Prize laureate (2002).

He is a Member of State Duma Committee Expert Council (since 2002), Editorial Board Member of Mobile Telecommunications (since 2002), and T-Com Journals (since 2007). He is a Professor of Moscow Technical University of Communications and Informatics (MTUCI, since 2001) and Visit-Professor of Tunisian Telecommunication Institute IsetCom (since 2005).
E-mail: v.tikhvinskiy@icominvest.ru
LLC Icominvest
Ostozhenka st 28
119034 Moscow, Russia

Grigory Bochechka is a Head of Innovation center department of LLC Icominvest and Chairman of WG14 Innovation Management of Telecommunications branch of Russian Academy of Natural Sciences Information and Telecommunication Technologies. He received his Ph.D. degree in specialty Systems, Networks and Telecommunication Devices.
E-mail: g.bochechka@icominvest.ru
LLC Icominvest
Ostozhenka st 28
119034 Moscow, Russia

Andrey Gryazev received his Ph.D. degree in 2015 in specialty of Management and Communication Systems. He is now acting General Director of Russian Federal State Unitary Enterprise Central Science Research Telecommunication Institute. His scientific research interests are in the fields of technologies of modern telecommunications, economical and regulation issues of radio communications, quality of service for fixed and mobile communications.
E-mail: agryazev@zniis.ru
Federal State Unitary Enterprise Central Science
Research Telecommunication Institute
First Passage of Perovo Pole 8
111141 Moscow, Russia

Paper

Digital Audio Broadcasting or Webcasting: A Network Quality Perspective

Przemysław Gilski and Jacek Stefański

Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, Gdańsk, Poland

Abstract—In recent years, many alternative technologies of delivering audio content have emerged, with different advantages and disadvantages. In this paper pros and cons of digital audio broadcasting and webcasting transmission techniques in a network quality perspective are described. A case study of user expectations with respect to currently available services is analyzed, and the perceived quality of real digital broadcasted and webcasted radio stations is examined.

Keywords—broadcast technology, mobile communications, quality of experience, quality of service, wireless communication.

1. Introduction

The current market condition in audio broadcasting and webcasting, also referred to as streaming, is characterized by the convergence of computer, telecommunication, and broadcasting technologies. It also relies on the divergence of different delivery and storage media, which use advanced digital signal processing techniques. The consumers are overwhelmed by new electronic gadgets, which appear each year on the market. They are astonished by new technical innovations that are being designed to change their life habits. The broadcasting sector is facing profound changes, particularly in a growing competition between the public and private sector, especially when it comes to providing high quality content.
With the development of storage media such as hard and flash drives, DVDs, or cloud-based online storage platforms, there is more demand for high quality broadcasted, streamed and downloaded material. Therefore, there is a growing demand for efficient ways of delivering high quality audio material at low bitrates, especially under bandwidth restrictions. Nevertheless, these standards and services sometimes fail to provide many users with the quality they expect in the digital era.

2. Broadcasting Services

The broadcasters are not all the same. They consist of public and private service broadcasters with a variety of national and regional stations. The conventional terrestrial radio transmission is faced with an increasingly strong competition from numerous streaming platforms and non-broadcast media, which use digital multimedia techniques to produce the optimum performance.

2.1. Terrestrial Broadcasting

The terrestrial broadcast delivery is the only free-to-air and cost-effective method for a truly mobile reception. However, in all developed markets, conventional analog and digital radio transmission is constrained by a lack of available spectrum. According to the European Broadcast Union (EBU) [1] the radio is:

- the vital cultural importance throughout Europe,
- consumed by a vast majority of Europeans every week,
- consumed at home, at work and on the move.

The frequency bands available for speech and sound broadcasting are becoming saturated. As a result, the reception quality is suffering more and more from mutual interference between transmissions. In many countries, there are very little or no prospects of additional radio services being provided by means of the existing analogue techniques [2].
Today, one of the main objectives of international broadcasters and content providers is to design and implement viable services, which are based on new universal digital delivery systems.

2.2. Webcasting

The Internet is an increasingly popular means of conveying audio, in particular music, to members of the general public. An audio streaming services are gaining more and more popularity. There are currently thousands of Internet radio stations offering audio streaming on-demand. Broadcasters are investing heavily in the Internet since nearly all of them have their own streaming website. This is also clearly visible in the number of available applications for popular mobile operating systems.
In some cases, the major drawback of streaming platforms is their relatively poor and insufficient sound quality. In order to listen to high quality audio one must purchase a premium account.

2.3. De ning Quality

When it comes to defining quality of a broadcasted or webcasted audio signal one question arises – how much infor-

mation could be lost or changed without seriously affecting the subjective quality of the material? Every lossy compression of audio content transmitted by the telecommunication channel causes degradation in quality. This degradation depends mainly on the transmission bitrate and coding algorithm [3].
The main factors that attract users to a particular service are:

- superior quality,
- stable reception, particularly in mobile environments,
- simple program selection tools,
- various services available at different data rates.

The quality of digital audio signals is defined by Quality of Service (QoS) parameters such as delay, frequency response, linear distortion, quantization noise, Signal-to-Noise Ratio (SNR), frequency bandwidth limitations. Whereas in Internet transmission, smaller or higher number of packets can be lost.
Subscribers expect their mobile devices provide high quality connectivity and performance at all time. Any interruption in data services is as critical as an interruption in voice. Depending on the service being used, subscribers have varying quality expectations for performance and usability. When subscribers consume content, their Quality of Experience (QoE) is not determined strictly by the speed achieved via wireless or wired technologies. They make subjective assessment based on a combination of factors as: speed, smoothness, latency. Service providers know, the better the experience, the longer and more frequently subscribers will consume content. Additional information may be found in [4].

3. Quality Perspective Survey

There are publications concerning popularity of different electronic media, including radio, television and the Internet [5]–[7]. They consist of scientific reports and analysis performed by public and private institutions, including universities. However, they analyze basic user activities and the impact of electronic media on society. These papers focus on, e.g. popular radio or TV channels, net browsing, e-commerce and shopping, as well as writing and receiving e-mails or using social media platforms. Most often, these studies were performed on a population of the so-called typical users, including students of humanities. The authors do not specify whether the surveyed population had a technical background or not. As we know, terms such as bandwidth, bitrate or spectrum may be an abstract concept for some of them.
Hence, authors have decided to carry out a survey on a group of 100 students of the Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology. The research population resembles a group of young people between 18–25 years old, with a particular interest in new technologies. The study was conducted between the 13th and 24th of April 2015 in the form of a questionnaire. The questionnaire consisted of open and closed questions with single and multiple choices. The main aim was to determine what are their particular needs and expectations when it comes to delivering high quality audio content.

3.1. Mobile vs Stationary Devices

According to the study, almost three quarters of students prefer using mobile rather than stationary devices (Fig. 1). When it comes to listening to music or consuming other multimedia content, 39% of them uses a smartphone, whereas only 8% a tablet (Fig. 2).

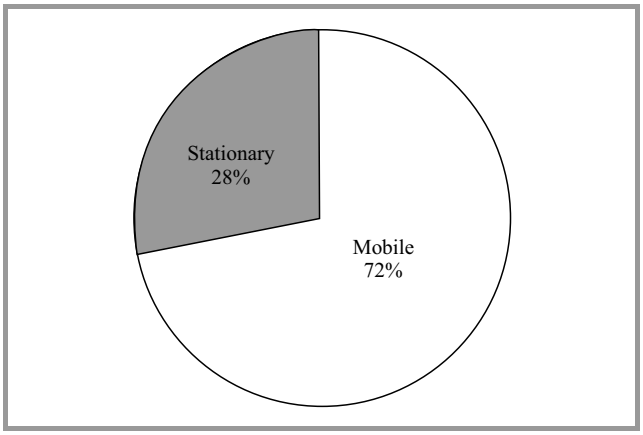


Fig. 1. Preferred type of consumer device.

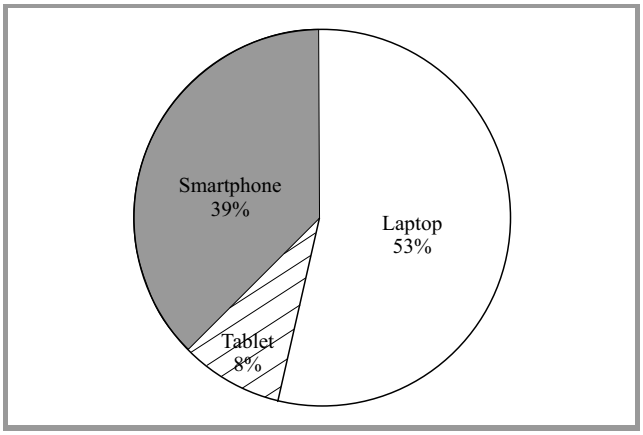


Fig. 2. Popularity of different kinds of mobile devices.

Surprisingly, considering the availability, size and weight of mobile devices such as smartphones and tablets, the laptop still remains the most popular device, with over 50%.

3.2. Streaming Platforms

The streaming platforms are very popular amongst students, 80% of the queried frequently use this type of service (Fig. 3), with over 90% of them being free services (Fig. 4).

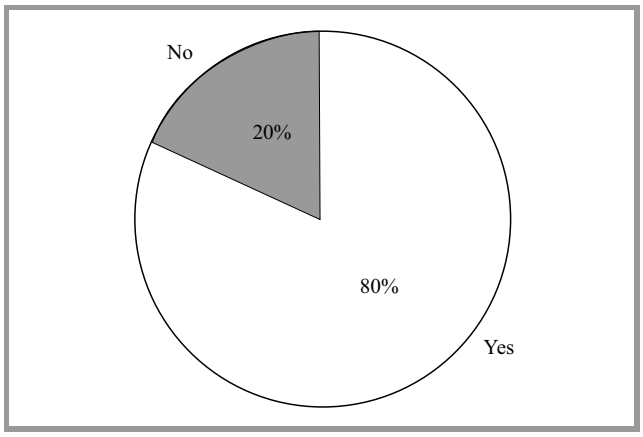


Fig. 3. Frequent use of streaming platforms.

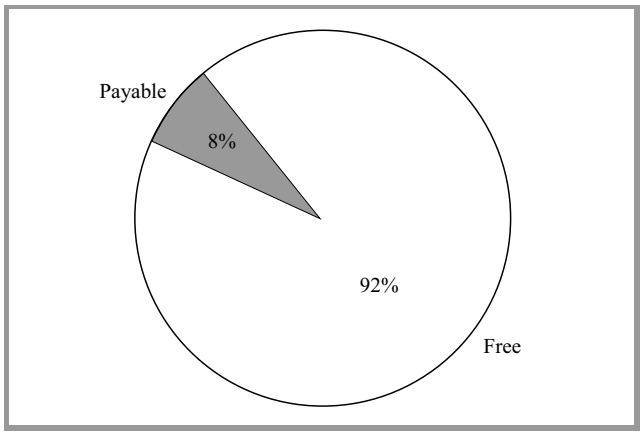


Fig. 4. Types of streaming platforms.

The most popular platforms are Spotify and Open.fm, with 23% and 22% shares respectively. Surprisingly, the majority, being 26%, listens to radio streamed live on the website of a particular radio stations. Streaming platforms such as Twitch.tv or TuneIn gained 9% and 4% respectively, whereas other received 16% (Fig. 5).

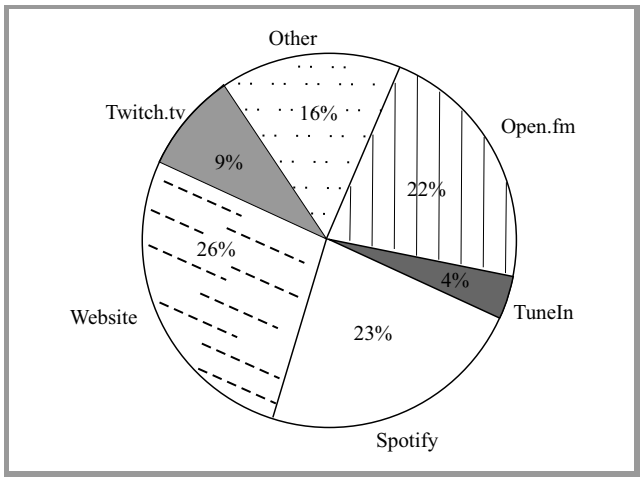


Fig. 5. Popularity of different streaming platforms.

In terms of energy and bandwidth efficiency, these results can be quite intriguing. Immediately, one question arises – is it really necessary to simulcast the same audio material terrestrially and online. The number of active streaming users has a significant impact on network load. As we know, a high number of simultaneous users can lead to higher delay. Furthermore, higher number of simultaneous users leads to less bandwidth allocated per capita. As a result, the user experience related with latency and limited bitrate of the audio stream may be disappointing. On the other hand, when users consume audio content using either analog or digital terrestrial radio transmission, they occupy the same share of bandwidth. The quality of the audio material is nearly the same for all, regardless of the number of active users.

The students responded that the main reason of using these type of services, instead of classical terrestrial radio transmission, is the availability and ease of use. According to them, Internet streaming provides an on-demand richer program offer and since they frequently use mobile devices, it is not any problem to choose a station from available programs. Another issue is, obviously, the lack of analogous or similar offer in terrestrial broadcasting. In their opinion, when it comes to streaming, commercial advertisements are less common.

3.3. Internet Connection

According to obtained data, over 70% of the surveyed group has a mobile data plan (Fig. 6). However, nearly 80% of them prefers fixed, either wired or wireless, over cellular connection (Fig. 7).

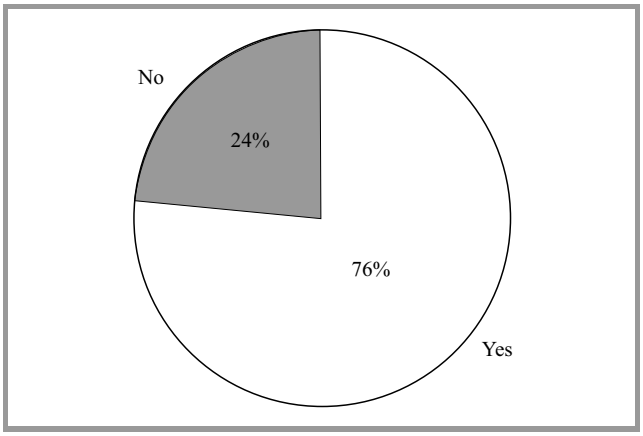


Fig. 6. Users with a mobile data plan.

But do we, as users, really have an option? If we carefully examine the situation in the developing countries, one can be easily noticed – the digital division. An individual that lives in the city center or close to it, has it all – a stable telecom infrastructure, even with Fiber To The Home (FFTH), and a high quality cellular coverage including Long Term Evolution (LTE). However, if a user lives in the suburbs or in a rural area, he or she seldom has any wired infrastructure. The only possible option is either satellite or cellular connectivity.

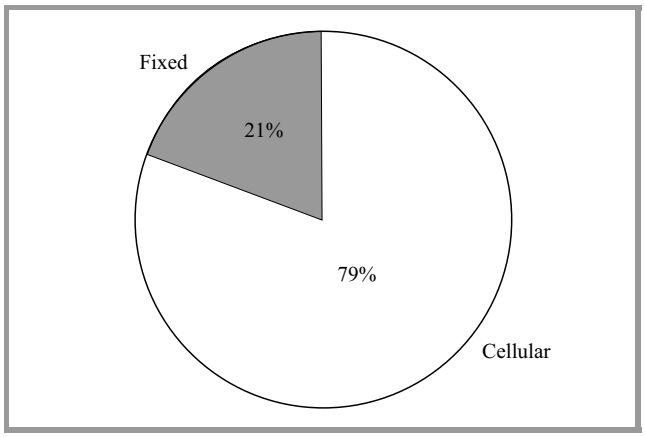


Fig. 7. Preferred type of Internet connection.

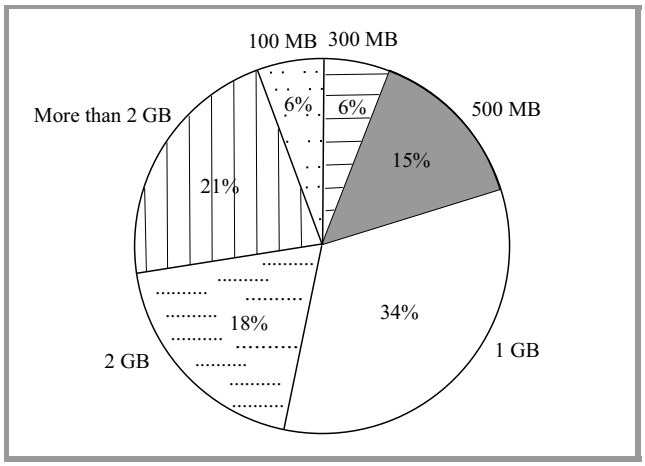


Fig. 8. Preferred type of Internet connection.

It is worth mentioning, that most of the surveyed students have a data limit of a couple of GB and higher (Fig. 8), which has a significant impact on network load.

3.4. Quality vs Network Load

Considering the most frequently chosen bitrate of audio content for either streaming or storing purposes, it is clearly visible that users prefer higher bitrates (Fig. 9). Among them, more than a half selects rates of 256 kb/s and higher, whereas less than 10% rates of 64 kb/s and less. Not surprisingly, users desire to have the best quality available, putting issues such as network load, stress of the mobile device or battery life aside.

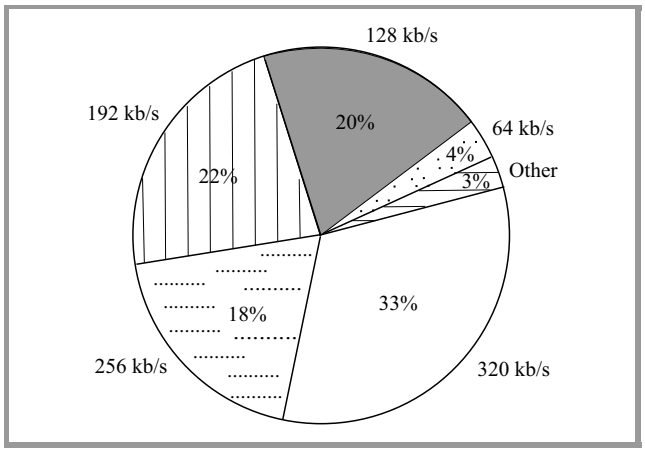


Fig. 9. Most frequently chosen bitrate.

Audio coding systems are used to reduce the amount of data required to represent an audio signal. There may be many reasons to do so, i.e. reduce storage requirements, transfer time or bandwidth requirements. However, there are applications where lower quality audio is acceptable, even unavoidable. The rapid development of the Internet, as a way of distributing audio material where data rates are limited, has led to a compromise in audio quality. Many delivery services, such as Internet streaming, digital satellite services or mobile multimedia applications, may operate at intermediate audio quality.

Considering the user's mobile data plans and selected bitrates, authors have prepared a chart describing how it can affect the network within a time interval (Fig. 10). Users with a data limit of 300 MB and lower can only affect the network under 10 hours per month, regardless of chosen bitrate. If we consider, that about three quarters of them have a mobile data plan of 1 GB and more, their activity will affect the network for tens of hours.

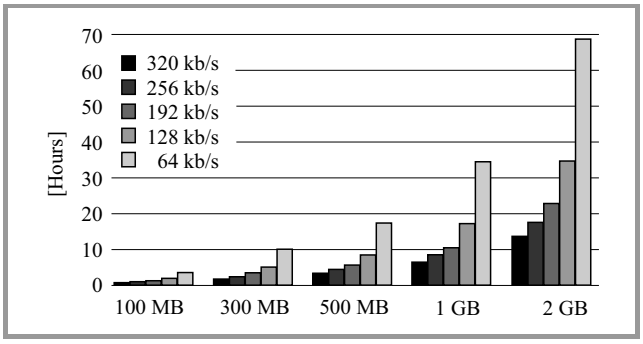


Fig. 10. Time period of user activity.

Nevertheless, mobile contracts, focused mainly on providing unlimited speech signal transmission, prove to be insufficient for the evaluation of long-term streaming of high quality audio content.

4. Perceived Audio Quality Study

The Digital Audio Broadcasting (DAB) [8] standard and its successor Digital Audio Broadcasting plus (DAB+) [9] are the most popular terrestrial broadcasting standards. There are publications concerning both subjective and objective quality assessments of speech and music signals, including [10]–[12]. However, they examine the quality of a predefined set of audio samples that had been processed using different codecs and bitrates. The authors did not encounter any publication on the assessment of an actual real-time live radio transmission.

Considering that the DAB+ platform has been launched in Gdańsk recently, a study was carried out concerning the

quality of the transmitted radio signal. Currently, 10 radio programs are available, with 5 of them being simulcasted in both analogue and digital terrestrial standards. The remaining 5 are new radio stations that are available only on the digital multiplex and online webcasting platforms. The profile and bitrate of new radio programs available on the digital multiplex and streaming platforms is shown in Table 1 and in Fig. 11. Each speech or audio signal was coded using the Advanced Audio Coding (AAC) algorithm.

Table 1
New radio programs available on the digital multiplex and streaming platforms in Gdańsk area

Profile	DAB+ bitrate [kb/s]	Streaming bitrate [kb/s]
Children	72	48
Information EN	64	48
Information PL	64	48
Pop music	96	48
Arts	128	48

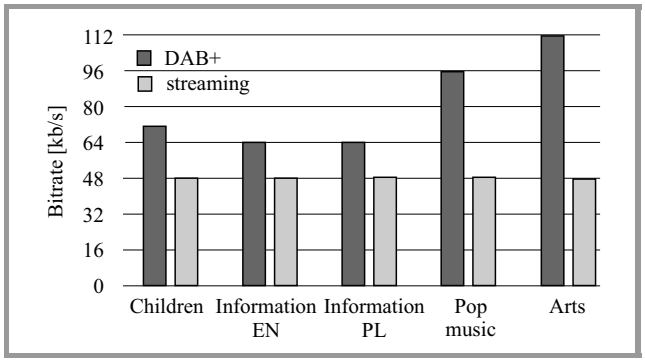


Fig. 11. New radio programs available on the digital multiplex and streaming platforms.

These 5 new stations are dedicated to different audiences. One of them for the youngest listeners, 2 for adults interested in current affairs, both in Polish and English. The remaining 2 are programs playing popular and classical music. It should be understood that the nature of the broadcast material might change in time with future changes in musical styles and preferences.

The study was performed between the 3rd and 21st of October 2015 on a group of 15 students according to recommendation [13], none of them had hearing disorders. Tests were carried out in turns, one participant after another, wearing headphones. Each participant was first instructed about the aim of the study, including the listening environment and equipment, and then asked to assess the quality of the transmitted radio signal.

The study consisted of two parts: Test 1 and Test 2. In Test 1 students were asked to rate the overall quality of each radio program transmitted terrestrially in Absolute Category Rating (ACR) scale, as shown in Fig. 12. In Test 2 they were asked to rate the impairments between "A" and "B", representing the same radio program transmitted terrestrial and online respectively in Degradation Category Rating (DCR) scale, as shown in Fig. 13. The confidence intervals were equal to 95%.

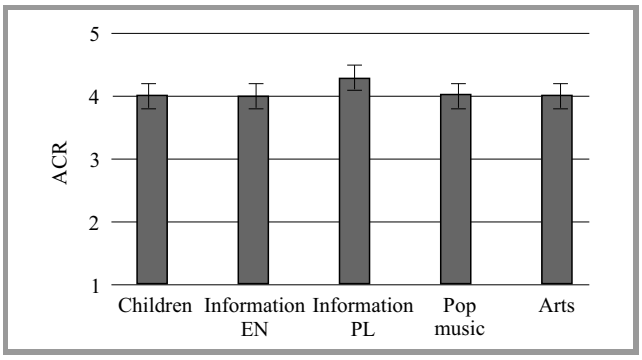


Fig. 12. Perceived audio quality of broadcasted radio programs.

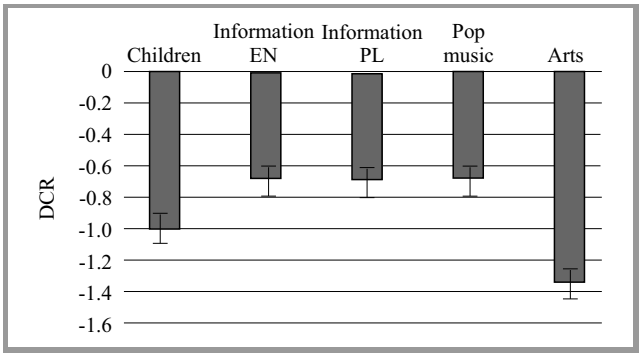


Fig. 13. Audio quality impairments between broadcasted and webcasted radio programs.

In both tests, the quality was assessed by the same group of subjects. Each individual had its own sheet of paper in order to write the score and comments. None of them was informed about the actual bitrate of the transmitted radio program.

According to reports from subjects in earlier listening tests, a fixed listening level was often perceived as annoying, being too low or too high for an individual. In order to overcome such possible problems, listeners were free to adjust the listening level before starting the experiment.

According to the listeners, the overall quality of terrestrial digital radio programs was ranked as good. This proves that the bitrate of each broadcasted radio stations was chosen properly. However, the streamed material was very limited in terms of bandwidth, with a clear cutoff of higher and lower frequencies. The voice of a radio presenter felt unnatural, whereas higher ratings were only observed in case of electronic music.

Quality assessment of speech and sound signals is a complex psychoacoustic phenomena related with human perception. It should be noted that each person interprets quality in a different way. The end perceived quality is sometimes less influenced by the consumer device than it is by the coding algorithm or chosen bitrate.

It can be noticed, that excellent audio quality, generally required from content providers, cannot always be achieved. This is caused either because of too low bitrates used, due to a narrowband transmission channel, or the type of audio material. If there is a serious constraint in terms of bandwidth, so that a broadcaster or webcaster is advised to use lower bitrates, it is often a better strategy to deliver a good stereo audio material than a poor or even bad multichannel audio signal.

One must keep in mind that in most cases, the bitrate of a free audio streaming service is limited. Better quality is reserved only for premium users who decide to switch to a payable service. Every broadcaster wishes to deliver near-studio-quality to the intended audience. Too high compression ratio may severely degrade the user experience. As a result, it will not meet the high expectations associated with new-generation digital broadcasting or webcasting services.

5. Conclusions

According to the study, the users prefer to consume audio content using mobile devices with a fixed Internet connection. However, providing high quality services is not always possible. Terrestrial broadcasting is facing many challenges and competition from webcasting services. It is very important that each service provider knows exactly the advantages and limitations related with different transmission techniques.

Broadcasting systems are capable of providing reliable digital services in real-time to all users located in a predefined covered zone. One of the main factors is clearly the cost of an infrastructure and transmission power required to cover a given area. Delivering high quality content to consumers is one of the most challenging tasks in the world of electronic media. Another crucial aspects is the efficient use of available bandwidth resources.

Broadcasters, telcos and content providers see the opportunity to offer more services, manufacturers look forward to selling larger quantities of devices and associated equipment, network operators are keen to build new telecom infrastructure. It is important to understand the pros and cons of different technologies and their commercial, economic, and operational implications. Broadcasters will always aim to use the best possible means to reach the user in the most effective way. Listeners will welcome every new technology that offers more features and higher audio quality. However, users do not mind about the technology used, they are only interested in the quality and the cost of a particular service.

References

[1] EBU website [Online]. Available: http://www3.ebu.ch/home (accessed 20.10.2015).

[2] F. Kozamernik, "Digital Audio Broadcasting – radio now and for the future", *EBU Tech. Rev.*, pp. 2–27, Autumn 1995.

[3] M. Kin, "Subjective evaluation of sound quality of musical recordings transmitted via DAB+ system", in *Proc. 134th Audio Engineering Society Convention*, Rome, Italy, 2013, pp. 1231–2366.

[4] P. Gilski and J. Stefański, "Quality expectations of mobile subscribers", *J. Telecom. Inform. Technol.*, no. 1, pp. 15–19, 2015.

[5] M. Kowalska, "Jakakolwiek, dla kogokolwiek, gdziekolwiek – obecność książki w polskich mediach elektronicznych" [Online]. Available: https://repozytorium.umk.pl/handle/item/413 (accessed 20.10.2015) (in Polish).

[6] L. Dyczewski, "Więź rodzinna a media elektroniczne (Family bonds vs electronic media)", *Ruch Prawniczy, Ekonomiczny i Socjologiczny*, no. 1, pp. 225–242, 2005 (in Polish).

[7] M. Brzozowska-Woś, "Social commerce – nowy trend w handlu elektronicznym (Social commerce – new trend in electronic commerce)", *Zeszyty Naukowe*, Uniwersytet Ekonomiczny w Poznaniu, pp. 221–231, 2011 (in Polish).

[8] ETSI EN 300 401 V1.4.1 (2006-01) Radio Broadcasting Systems; Digital Audio Broadcasting (DAB) to mobile, portable and fixed receivers [Online]. Available: http://www.etsi.org

[9] ETSI TS 102 563 V1.2.1 (2010-05) Digital Audio Broadcasting (DAB); Transport of Advanced Audio Coding (AAC) audio [Online]. Available: http://www.etsi.org

[10] S. Meltzer and G. Moser, "MPEG-4 HE-AAC v2 – audio coding for today's digital media world", *EBU Tech. Rev.*, pp. 1–12, Jan. 2006.

[11] K. Błasiak, A. Dobrucki, M. Kin, and M. Ostrowski, "Badanie jakości dźwięku programów muzycznych przesyłanych w systemie DAB+ (Sound quality evaluation of DAB+ musical programs)", in *Proc. 14th Int. Symp. Sound Engin. Tonmeistering ISSET 2011*, Wrocław, Poland, 2011.

[12] P. Pocta and J. G. Beerends, "Subjective and objective assessment of perceived audio quality of current digital audio broadcasting systems and web-casting applications", *IEEE Trans. Broadcast.*, vol. 61, no. 3, pp. 407–415.

[13] ITU-R BS.1284-1 (1997) Methods for the subjective assessment of sound quality – general requirements.

Przemysław Gilski received his B.Sc. and M.Sc. degrees in Telecommunications Engineering from Gdańsk University of Technology (GUT), Poland, in 2012 and 2013, respectively. Currently he is a Ph.D. student at the Department of Radio Communication Systems and Networks (DRCSN), GUT. His research and development interests include digital video and audio broadcasting systems, software-defined radio technology, location services and radio navigation systems, as well as quality measurements in mobile networks.

E-mail: pgilski@eti.pg.gda.pl
Faculty of Electronics, Telecommunications and Informatics
Department of Radio Communication Systems and Networks
Gdańsk University of Technology
Gabriela Narutowicza st 11/12
80-233 Gdańsk, Poland

Jacek Stefański received his M.Sc., Ph.D. and D.Sc. degrees in Telecommunications Engineering from Gdańsk University of Technology (GUT), Poland, in 1993, 2000 and 2012, respectively. From 1993 to 2000 he worked as an assistant professor at the Department of Radio Communication Systems and Networks (DRCSN), GUT. Since 2001 he has been working as an associate professor at the DRCSN. His research and development interests include analysis, simulation, design and measurements of cellular, wireless and trunked radio systems, techniques of digital modulation, channel coding, signal spreading, radio signal reception, measurement of radio wave propagation, field strength prediction, software radio design, location services, ad-hoc sensor networks, radio monitoring systems and radio navigation systems. He is the author and co-author of more than 250 papers. He is a member of the Electromagnetic Compatibility Section of the Electronics and Telecommunications Committee, Polish Academy of Science and the Institute of Electrical and Electronics Engineers organization.

E-mail: jstef@eti.pg.gda.pl
Faculty of Electronics, Telecommunications
and Informatics
Department of Radio Communication Systems
and Networks
Gdańsk University of Technology
Gabriela Narutowicza st 11/12
80-233 Gdańsk, Poland

Paper

Monitoring of a Cloud-Based Environment for Resilient Telecommunication Services

Grzegorz Wilczewski

Faculty of Electronics and Information Technology, Warsaw University of Technology, Warsaw, Poland

Abstract—This article depicts insights and in-depth presentation of a new tool, specifically designed for Data Center resources monitoring purpose. It enables physical and virtual resources monitoring and is capable of performing advanced analysis on the resulting, measured data. Here in presented are exemplary scenarios conducted over the proprietary Data Center unit, delivering specific information on the behavior of the analyzed environment. Presented results create a base layer for a high level resiliency analysis of telecommunication services.

Keywords—cloud computing, Data Center, resiliency, resources monitoring.

1. Introduction

Nowadays, most telecommunication services are stated upon a concept of a virtualized, cloud-based functionalities serving contemporary telco products, for instance video streaming capabilities, storage services or big data processing functionalities [1]. Majority of workload volume is digested within Data Center units (DC), being it a vast computational power and massive storage spaces. Services being deployed utilizing such solutions require emerging but reliable systems overcoming possible flaws in the DCs design [2], [3]. Reliability of a cloud environment is the key feature that determines the success of a service being created within such environment [4]. Popular cloud-based services contribute to the XaaS model – *X as a Service*, where *X* defines what users can choose from the Cloud Service Provider (CSP) offer. Is it infrastructure, platform or application or many others, the common point is to deliver stable fundament for the overlaid services.

Creating reliable Cloud or DC environment requires multiplicity of features being supported by the discussed unit management system. One of the upmost importance characteristic of that system is to deliver a crude monitoring functionality. However the concept is not only limited to a basic parameters or resources observation but also requires appropriate examination of those gathered results. Active monitoring function has to cope with both hardware and software domains of the Data Center environment, moreover in both classification of the resources, namely, virtual and physical [5], [6]. The analysis of the current state of the Cloud unit delivers essential data for the CSP to act with. Whenever troubleshooting of a current problem is required or appropriate adjustment of a Service Level Agreement (SLA) is necessary, the monitoring application clarifies what element or node of the environment is responsible for such a state of a matter [7], [8]. Moreover, monitoring variety of parameters can be a successful approach towards isolation of specific virtual units whenever the assumption is that user's activities led to unstable or overloaded working conditions.

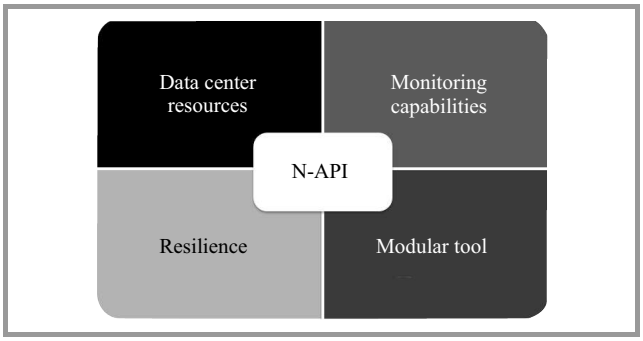


Fig. 1. Functional stages of N-API project.

To overcome the aforementioned scenarios and to fulfill previously assessed functionalities the concept of an N-API project was defined. The following paper covers the insights towards design, creation and initial test-bed investigation of a tool in its native, cloud environment. The common goals of the project, presented within Fig. 1, define the following features of the N-API tool:

- an unified structure of N-API supporting networking Application Programming Interface (API) for the required functionalities,
- a definition and description of a set of resources and parameters of a DC (concerning processing capabilities, storage information, networking means and topology),
- a monitoring capability in all of the DCs structural layers (both virtual and physical),
- a modular structure of itself, supporting upcoming analysis expansion features and additional vendor-specific development,
- a resiliency evaluation module.

Table 1
N-API S2 REST functionalities – available methods

Method	Parameters	Exemplary call	Returned	Comments
getVMS	None	http://localhost:8084/napi/analysis/getVMS	{"vmIds":[id1, id2, ..., idn]}	• Check the list of all available VMs of a DC
getCPU Analysis	vmId timePeriod	http://localhost:8084/napi/analysis/getCPUAnalysis?vmId=488&timePeriod=hourly	AnalysisResult Combined	• vmId has to be one of the vmIds elements set • timePeriod has to be one of the following set: hourly, daily, weekly, monthly
getMemory Analysis	vmId timePeriod	http://localhost:8084/napi/analysis/getMemory Analysis?vmId=488&timePeriod=hourly	AnalysisResult Combined	• vmId has to be one of the vmIds elements set • timePeriod has to be one of the following set: hourly, daily, weekly, monthly
getDisk Analysis	vmId TimePeriod	http://localhost:8084/napi/analysis/getDiskAnalysis ?vmId=488 timePeriod =hourly	AnalysisResult Combined	• vmId has to be one of the vmIds elements set • timePeriod has to be one of the following set: hourly, daily, weekly, monthly

Considering listed attributes of a tool being designed one has to divide the overall project into a three, complementary steps being realized:

- design and deployment of a functional API for DC resource monitoring,
- positioning of parameters being taken into consideration while evaluating measured unit characteristic (i.e., storage utilization, computational resources charge, hierarchy and topology of a Data Center),
- design and development of an application utilizing aforementioned modules.

2. Application Programming Interface

Approaching realization of the predefined project goals required access to the deployed Data Center environment. As a model one, the unit serving telecommunication services, manufactured by Cisco was selected. The management system being utilized within this unit was of the UCS Director software package. Thus, the designed N-API API is compatible with the mechanisms, functions and data model of the mentioned Cisco product line. Architecture of the designed API (presented on the Fig. 2) delivers essential information of how the tool is constructed and what are the possible utilization schemes. The programming interface of the N-API enables to retrieve essential information about the monitored DC unit, while API functions are used directly by the Application (APP) module to deliver raw and processed information about the environment (from the Analysis Extension section).

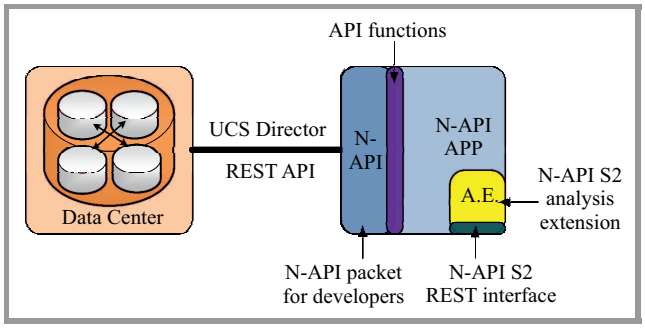


Fig. 2. Insights of the N-API architecture.

The N-API tool is purely designed as a networking, mobile software package, thus it is delivered in Java environment supporting data representation and manipulation by JSON (JavaScript Object Notation) and REST (Representational State Transfer) functionalities. In case of available connectivity schemes (depicted on the graph in Fig. 3) the interaction may be led in two approaches: either local regime or remote access. To support efficient security while accessing DC unit, appropriate secure API key is being utilized. Latter, a user can define specific connectivity socket (i.e., protocol, port, address) as well as establish a Joint Singleton interaction, what improves performance of an access towards monitored Data Center unit, deployed with use of the Cisco UCS package.

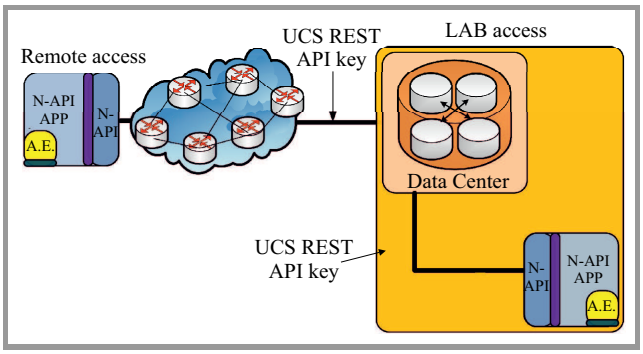


Fig. 3. Available connectivity schemes.

Monitored, raw data of the specified parameters is being delivered by a detailed description report files. Eligible entities possessing such a functionality are contained is the following context units (i.e., reflecting the hierarchy of the DC):

- *GlobalClient* – reveals the overall DC's structural outline, especially topology; is the highest order element in the hierarchical layout;

- *Cloud* – presents information concerning all of the Cloud units being realized in the Data Center; supports user with the intrinsic data about addressing, structure, contents and topology of the Cloud unit;
- *VirtualDataCenter* (vDC) – intermediate partition of a Cloud unit; enables layer isolation on the level required for IaaS and PaaS functionality support;
- *VirtualMachine* (VM) – is the lowest order entity in the hierarchy of the monitored Data Center unit; contributes to the basic SLA and billing handling; guarantees the lowest level of separation and isolation within the DC resources; initiates foundation for *SaaS* services;
- *HostNode* and *DataStore units* – deliver essential information about physical assets of the Data Center; presents configuration of selected peripherals and delivers network topology of the analyzed environment.

Table 2
N-API analysis results – returned response structures

Analysis Result Combined	{“xyResults”:[AnalysisResultXYChart], “barResults”:[AnalysisResultBarChart], “pieResults”:[AnalysisResultPieChart], “singleResults”:[AnalysisResultSingle], “comment”:“comment”}
Analysis Result XYChart	{“series”:[{ “xValues”:[1.1,2.2,3.3 ...], “yValues”:[1.1,2.2,3.3 ...], “comment”:“comment”},...], “chartTitle”:“Title”,“units”:“range label”, “domainLabel”:“label”, “comment”:“comment”}
Analysis Result BarChart	{“series”:[{“values”:[{“xValue”: 1.1, “comment”:“comment”},...], “comment”:“seriesName”},...], “chartTitle”:“Title”,“units”:“range label”, “domainAxisLabel”:“label”, “comment”:“comment”}
Analysis Result PieChart	{“values”:[{“xValue”:1.1,“comment”: “bar label”}, ...], “chartTitle”:“Title”,“units”: “range label”,“comment”:“comment”}
Analysis Result Single	{“xValue”:1.1,“comment”:“comment”}

Nonetheless, architecture depicted on the Fig. 2 requires clarification on the REST functionalities presented in the Analysis Extension part of the N-API S2 module. Supported monitoring activities are expanded by the analysis module that supplies user or third party Cloud Manager with essential data elements/structures. The designed interface utilizes simple GET functionality and enables methods for specific inquiry. Parameters of those calls (i.e. VM identifiers) as far as an exact actions are presented in the positioning across Table 1. Resulting responses, given by means of JSON data modules are presented in the Table 2, accordingly. Whenever called action cannot be executed, i.e. due to user’s inappropriate parameter input, the server returns http 500 status.

3. Monitoring Application and Results Analysis

Previous sections presented the insights towards N-API’s utilization and its capabilities, thus herein are discussed functionalities of the application, being a complementary part of the N-API package (denoted across Figs. 2 and 3 by N-API APP phrase). To start with, it is a standalone ap-

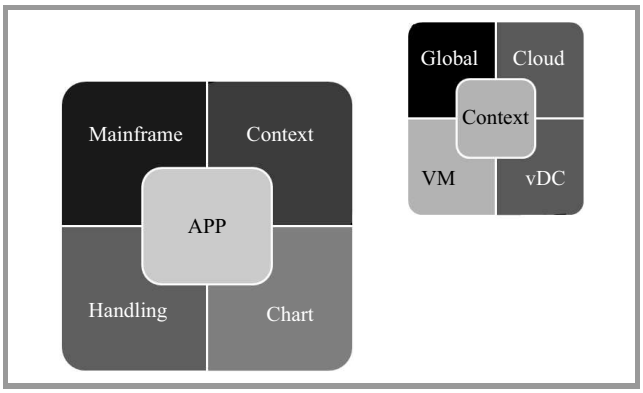


Fig. 4. Functional modules of N-API application.

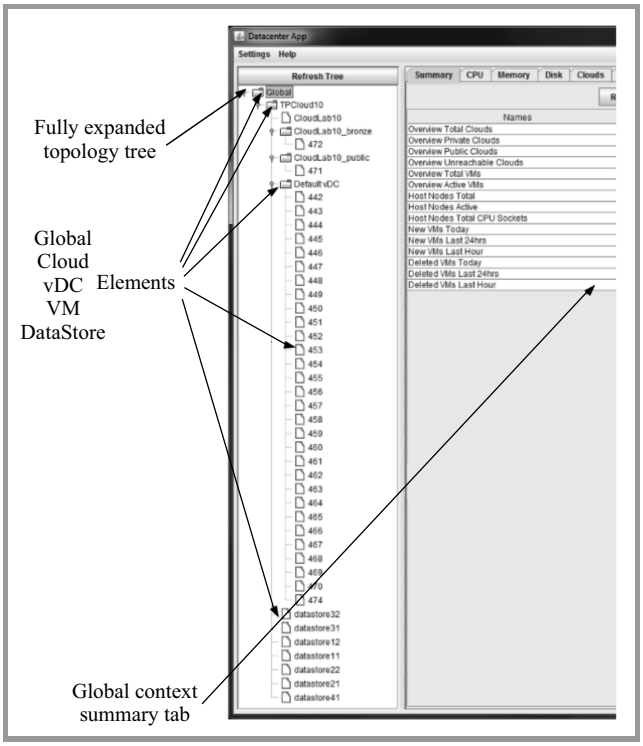


Fig. 5. Topology tree – discovered hierarchy.

Table 3
Functionalities of the N-API Analysis Extension toolkit

Tab	Tool/plot	Comment
Interpolation	Input values	Plot of the original values of the performance indicator.
	Input values; Chunked grouped by counting; Operation on chunk: median	Plot of the median value of the chunked samples from the original set of values.
	Input values; loess interpolation; Cut to max 100.0, min 0.0.	Plot of the interpolated values of the performance indicator. Interpolation method: LOcal regrESSion algorithm with the default configuration (please refer to the Apache Commons Math 3 v3.3 toolbox). Whenever interpolated values overshoot the boundary value, the cutoff below 0 and above 100 is applied.
Pie chart	Percentile containers	Pie chart represents the total time the VM's performance indicator spent in a specific percentile range. Containers represent intervals: 0–50%; 50–75%; 75–90%; 90–100%.
Resiliency monitoring	Color indicators	Visualization of the resiliency monitoring by means of distinctive color markers, representing total time spent in the appropriate percentile container. In case of a RED marker there is a Note informing about suggested activation of a chosen resiliency mechanism.
Bar chart	Minimum	Displays the minimum value from the analyzed set of values.
	Maximum	Presents the maximum value out of the analyzed sequence.
	Arithmetic mean	Calculates an arithmetic mean from the analyzed sequence
	Median	Calculates a median value from the analyzed set of values.
	Standard deviation	Calculates standard deviation value out of the considered set.

plication what means it only requires a compatible runtime environment to be able to function properly and deliver support package for developing purposes. It has specific modules and plug-ins integrated within programming environment, alongside with the JFreeChart and Apache Commons Math 3 v3.3 toolboxes. The general structure of the N-API APP is presented on the Fig. 4. What is noticeable is the modular built that supports basic activities delivered by means of the API (Context Block) and enables handling

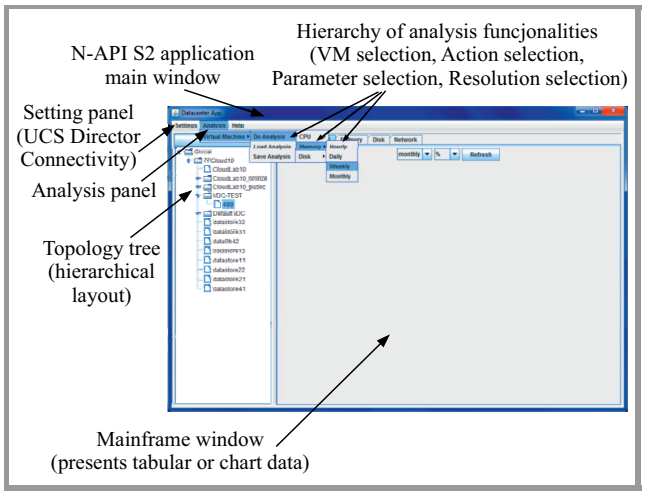


Fig. 6. Analysis Extension GUI.

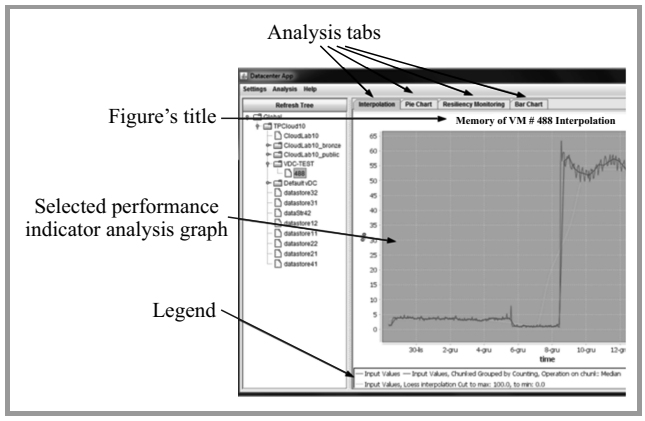


Fig. 7. Analysis of the selected VM's resources.

of the monitoring data (both raw and analyzed) in a form of an intuitive Graphic User Interface (GUI). Designed N-API application enables user to discover following information concerning monitored DC unit:

- topology and structure in a form of an expandable, active tree,
- status of the peripherals and inventory of the structural unit, i.e. configuration of the VM entity,
- monitoring data of the essential resources: CPU, RAM memory, storage space.

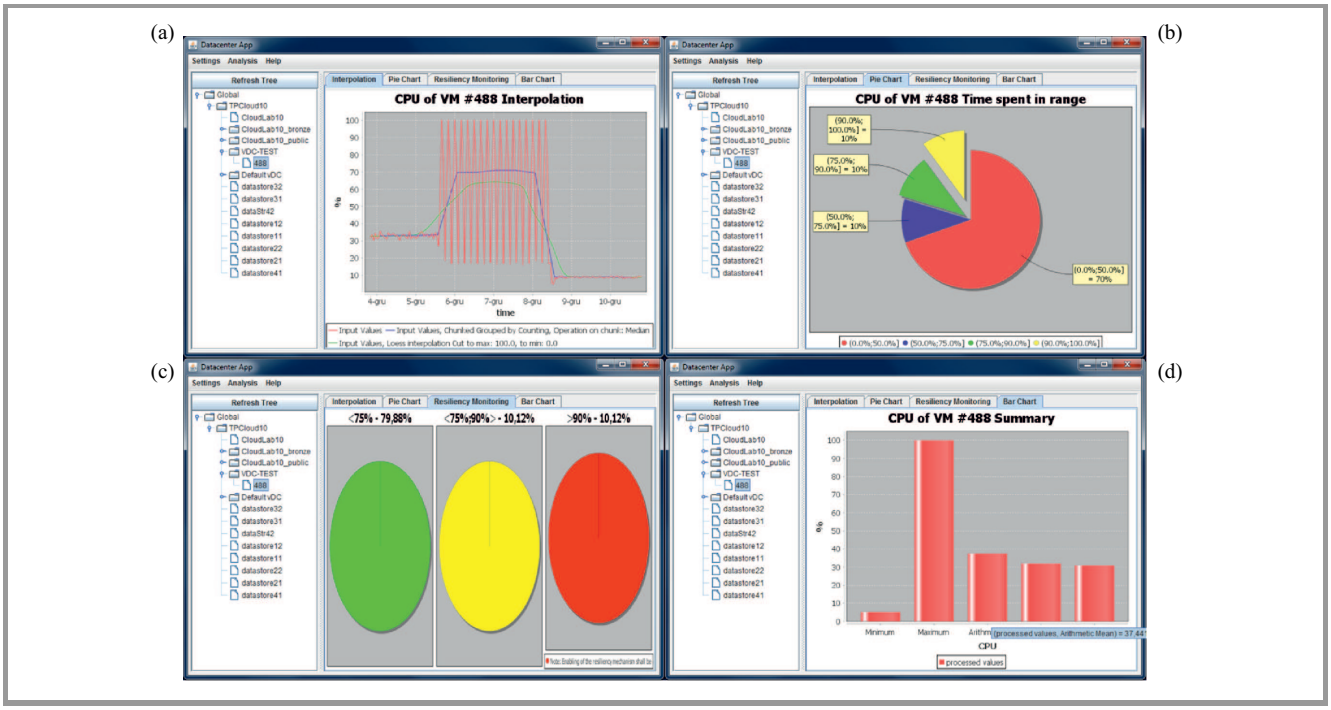


Fig. 8. Resiliency evaluation tools of the N-API: Interpolation method (a), Pie chart analysis (b), Resiliency monitoring (c), Bar chart samples analysis (d).

Functionality-wise, the application delivers reporting capabilities for the export/process data in a range of selectable units (relative to the actual resource metric): CPU (MHz, in percent), RAM and Storage (GB, in percent). In case of defining temporal resolution of the report, following options are exposed: instant report (average value of parameter over the period of 5 min.), historical data (resolution: hourly, daily, weekly, monthly).

Concerning working with the executed application, the graphical representations of the selected states are delivered over the sequence from Fig. 5 to Fig. 7. Describing operation scheme of the application, one has to start with the activation of the connectivity module. After selecting appropriate socket and its configuration a secure link with the DC is being established. Whenever a successful connection is present the Context section is being activated and specific data obtained. Scrolling through the topology tree (Fig. 5) one can select a VM of choice and afterwards perform a set of analysis methods. In order to achieve that goal selection of the Analysis List from the top bar of the window is necessary. Mechanisms incorporated within Analysis Tab enable users to perform specific actions over the Virtual Machine unit, and those are: Do Analysis, Load Analysis, Save Analysis. The intermediate storage format of the data is coherent across the whole project and contributes to the .json format files. Furthermore, the application is capable of presenting the analysis of performance data for the following group of resources: CPU, RAM and Disk (Storage). Time resolutions are in coherence with instant and historical reports. Worth noticing is the fact of an abundant data being transferred whenever high resolution of samples across longest period is required. Positioning in the Table 3 covers the information about Analysis Extension module, while composed graphics in the Fig. 8 depict outcomes of all available and applied analysis tools.

4. Conclusions

Testing the designed API and application within its native, Data Center environment delivered a vast range of results. Selected scenarios covered various states of workload being applied to the controlled unit. Performed analyses of the overloaded resources delivered unique characteristics of monitored Cloud structure (depicted by means of Fig. 8). Deployed N-API tool enables CSPs to call the appropriate actions whenever troubleshooting or SLAs improvement are to be done. Advanced statistical analysis might improve resource allocation or introduce dynamically adjustable pools of resources for VMs. Finally, one ought to perceive the upmost importance of the resiliency analysis within a DC unit, in order to create a successful telecommunication service.

References

[1] G. Wilczewski, "Analysis of content quality evaluation within 3DTV service distribution systems", *J. Telecommun. Inform. Technol. (JTIT)*, no. 1, pp. 16–20, 2014.

[2] G. Wilczewski, "Utilization of the software-defined networking approach in a model of a 3DTV service", *J. Telecommun. Inform. Technol. (JTIT)*, no. 1, pp. 32–36, 2015.

[3] S. Musman and S. Agbolosu-Amison, "A measurable definition of resiliency using mission risk as a metric", Mitre Tech. Rep., McLean, VA, 2014.

[4] Y. H. Khalil, A. Elmaghraby, and A. Kumar, "Evaluation of resilience for data center systems", in *Proc. IEEE Symp. Comp. Commun. ISCC'08*, Marrakech, Morocco, 2008, pp. 340–345.

[5] Riverbed Migration Mitigation, The Complete Data Center Consolidation Guide, Corporate Whitepaper, 2012.

[6] Cisco Data Center Solution Brief, Mitigate Risk for Data Center Network Migration, Corporate Whitepaper, 2014.

[7] R. Nasiri and S. Hosseini, "A case study for a novel framework for cloud testing", in *Proc. 11th Int. Conf. Electron., Comp. Comput. ICECCO 2014*, Abuja, Nigeria, 2014, pp. 52–56 (doi: 10.1109/ICECCO.2014.6997566)

[8] D. Agarwal and S. K. Prasad, "AzureBench: Benchmarking the storage services of the azure cloud platform", in *Proc. IEEE 26th Int. Parallel and Distrib. Process. Symp. Worksh. & PhD Forum IPDPSW 2012*, Shanghai, China, 2012, pp. 1048–1057.

Grzegorz Wilczewski received his B.Sc. and M.Sc. degrees in Electrical and Computer Engineering from Warsaw University of Technology, Poland, in 2009 and 2011, respectively. He is currently a Ph.D. candidate at Warsaw University of Technology. His research interests include 3DTV service quality monitoring, 3D imagery and digital signal processing.
E-mail: g.wilczewski@tele.pw.edu.pl
Faculty of Electronics and Information Technology
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw

Paper

Study of No-Reference Video Quality Metrics for HEVC Compression

Kais Rouis[1], Mikołaj Leszczuk[2], Lucjan Janowski[2], Zdzisław Papir[2], and Jamal Bel Hadj Tahar[1]

[1] *Innov'COM Lab, Sup'Com, University of Carthage, Ecole Nationale d'Ingénieurs de Tunis, Université de Tunis El Manar, Tunisia*

[2] *AGH University of Science and Technology, Department of Telecommunications, Krakow, Poland*

Abstract—The paper proposes a No-Reference (NR) quality assessment measurement originally developed for H.264, used for High Efficiency Video Coding (HEVC). In particular, authors present an investigation of NR metrics to objectively estimate the perceptual quality of a set of processed video sequences. The authors take into account typical distortions introduced by the block-based coding approaches like HEVC codec. The underlying processing used for the quality assessment considers the blockiness caused by the boundaries of each coded block and the blurring as a lack of spatial details. The correlation between the NR quality metrics and the well-known and most widely used objective metric, the Video Quality Model (VQM), is performed to validate the quality prediction accuracy based on the provided scores. The Pearson correlation coefficients obtained stand for promising results for different types of videos.

Keywords—High Efficiency Video Coding, No-Reference metrics, Quality of Experience, Video Quality Assessment.

1. Introduction

In addition to traditional Quality of Service (QoS), Quality of Experience (QoE) poses a real challenge for Internet service providers, audiovisual services, broadcasters, and new Over-The-Top (OTT) services. The leading operators have to solve the problem of accurate QoE prediction since the end-user satisfaction is a real added value in the market competition. QoE tools should be proactive and provide innovative solutions that are well adapted for new audiovisual technologies. Therefore, objective audiovisual metrics are frequently dedicated to monitoring, troubleshooting, investigating, and setting benchmarks of content applications working in real-time or off-line.

To advance the field of video quality assessment, Video Quality Experts Group (VQEG) performs subjective video quality experiments, validates objective video quality models, and collaboratively develops new techniques. VQEG proposed to monitor audio visual quality by Key Performance Indicators (KPI), which are able to isolate and focus investigation, set-up algorithms, increase the monitoring period, and guarantee good prediction of video quality. It is known that, depending on the technologies used in audiovisual services, the impact of QoE can change completely. So, based on that proposed concept, it is possible to select the best algorithms and activate or switch off features in a default audiovisual perceived list. The scores are separated for each algorithm and preselected before the testing phase. Then, each artifact KPI can be analyzed by working on the spatially and/or temporally perceived axes [1].

The proposed concept is an interesting approach because it can detect the artifacts present in videos, as well as predict the quality as described by consumers. In realistic situations, when video quality decreases in audiovisual services, customers can call a helpline to describe the annoyance and visibility of the defects or degradations in order to describe the outage. In general, they are not required to provide a Mean Opinion Score (MOS). As such, the concept is completely in phase with user experience. There are many possible reasons for video disturbance, and they can arise at any point along the video chain transmission (filming stage to end-user stage).

VQEG experiments were carried out over several steps with experimental set-ups for concept verification. The impairments included in the experiments were limited to MPEG-2 and H.264. Nevertheless, in year 2013, the first version of the High Efficiency Video Coding (HEVC) standard was completed, approved, and published. HEVC is a video compression standard, a successor to H.264/MPEG-4 AVC (Advanced Video Coding), which was jointly developed by the ISO/IEC JTC 1/SC 29/WG 11 Moving Picture Experts Group (MPEG) and ITU-T SG16/Q.6 Video Coding Experts Group (VCEG) as ISO/IEC 23008-2 MPEG-H Part 2 and ITU-T H.265 [2], [3].

In this paper, the experiments carried out over several steps with an HEVC experimental set-up for the proposed concept verification are presented.

The remainder of this paper is structured as follows. Section 2 is devoted to the state-of-the-art background. Section 3 discusses NR video quality assessment. Section 4 presents objective video quality methods. Section 5 analyses results on KPI. Section 6 discusses further work and summarizes the paper.

2. Related Works

This section presents brief survey of current NR approaches for standardized models together with their limitations. Most of the models in ITU-T recommendations were val-

idated on video databases that used one of the following hypotheses:

- frame freezes lasting up to 2 s,
- no degradation at the beginning or at the end of the video sequence; no skipped frames,
- clean video reference (no spatial or temporal distortions),
- minimum delay supported between video reference and video (sometimes with constant delay),
- up or down-scaling operations not always taken into account [4].

As mentioned earlier, most quality models are based on measuring common artifacts/KPI, such as blur, blocking, and jerkiness, for producing a prediction of the MOS. Consequently, the majority of the algorithms generating a predicted MOS show a mix of blur, blocking, and jerkiness metrics. The weighting between each KPI could be a simple mathematical function. If one of the KPIs is not correct, the global predictive score is completely wrong. Other KPIs mentioned by VQEG are usually not taken into account (exposure time distortion, noise, block loss, freezing, slicing, etc.) in predicting MOS [4]. ITU-T has been working on similar distortions for many years [5]. However, only for Full-Reference (FR) and Reduced-Reference (RR) approaches. The history of the ITU-T Recommendations for video quality metrics is shown in Table 1. Table 2 shows a synthesis of the set of standardized metrics that are based on video signals [4]. As can be noticed from both tables, there is a lack of developments for the NR approach.

Table 1
The history regarding ITU-T Recommendations

Model type	Format	Recommendation	Year
FR	SD	J.144 [6]	2004
FR	QCIF–VGA	J.247 [7]	2008
RR	QCIF–VGA	J.246 [8]	2008
FR	SD	J.144 [6]	2004
RR	SD	J.249 [9]	2010
FR	HD	J.341 [10]	2011
RR	HD	J.342 [11]	2011
Bitstream	VGA–HD	P.1202 [12]	2013
Hybrid	VGA–HD	J.343 [13]	2014

In a related research, Gustafsson *et al.* [14] addressed the problem of measuring multimedia quality in mobile networks with an objective parametric model [4]. Closely related work are ongoing standardization activities at ITU-T SG12 on models for multimedia and Internet Protocol Television (IPTV) based on bit-stream information. SG12 is currently working on models for IPTV. Q.14/12 is responsible for these projects, provisionally known as non-intrusive parametric model for assessment of performance of multimedia streaming (P.NAMS) and non-intrusive bit-stream model for assessment of performance of multimedia streaming (P.NBAMS) [4]. P.NAMS uses packet-header information (e.g., from IP through MPEG2-TS), while P.NBAMS also uses payload information, i.e., coded bit-stream [15]. However, this work focuses on the overall quality (in MOS units), while the proposed concept is focused on KPIs [4].

Table 2
Synthesis of FR, RR and NR MOS models

Resolution	Type of ITU-T model		
	FR	RR	NR
HDTV	J.341 [10]	n/a	n/a
SDTV	J.144 [6]	n/a	n/a
VGA	J.247 [7]	J.246 [8]	n/a
CIF	J.247 [7]	J.246 [8]	n/a
QCIF	J.247 [7]	J.246 [8]	n/a

Most of the recommended models are based on global quality evaluation of video sequences, as in the P.NAMS and P.NBAMS projects. The predictive score is correlated to subjective scores obtained with global evaluation methodologies (SAMVIQ, DSCQS, ACR, etc.). Generally, the duration of video sequences is limited to 10 or 15 s in order to avoid a forgiveness effect (the observer is un-enable to score the video properly after 30 s and may give more weight to artifacts occurring at the end of the sequence). When one model is deployed for monitoring video services, the global scores are provided for fixed temporal windows and without any acknowledgement of the previous scores [4].

Recently, the interest is oriented toward the HEVC standard, which has proved high efficiency compared to its predecessors. Several tools are introduced in the coding process, such as the increasing number of intra prediction modes and the frequent use of inter coded pictures within a closed Group Of Pictures (GOP). These characteristics ensure an important coding gain relative to the encoding parameters but in the other hand, the complex structure of picture division and the new configurations' models can be the source of certain artifacts. However, very limited works concern the quality assessment approaches for HEVC compression. In particular, the coding parameters and the impact of network losses on the decoder side were investigated [16]. The distortions of HEVC videos are more significant than H.264 videos. The proposed NR distortion measure exploits the spectral densities between the frames and precisely, the energy variation in the temporal domain for each coding unit.

One can bear in mind that FR measures are in general not applicable as the reference content might be not available. In the same vein, the bitstream features were selected to estimate the perceptual quality, including the different prediction modes and statistics of the motion vector [17]. In this method the measures are predicted in a NR manner.

The quality monitoring becomes primordial in communication and broadcasting environments for improving the end user's QoE [18]. A NR Peak Signal-to-Noise Ratio (PSNR) estimation was proposed for such a model [19]. Distributions of transform coefficients are considered based on the quad-tree coding structure and the distortion model was derived according to the coding unit depth level.
The concept of QoE in [20] is used for a practical recognition problem for video transmitted over a network link, where subjective satisfaction of the user is imperative. This latter requires achieving specific functionalities such as even detection and object recognition. The proposed methods measure the usefulness of degraded quality video and the solutions have been proposed to optimize the network QoS parameters.
Designing algorithms for video quality assessment requires a consistent dataset of coded video sequences. For the case of HEVC it is a key factor for an effective performance evaluation of developed metrics, to take advantage of a publicly available database, which includes several compressed versions of different sequences. In [21] Full-Reference measurements are provided with a large database of FULL-HD HEVC encoded videos based on a variety of HEVC compression characteristics.
A variety of NR quality estimation methods exist for the AVC videos but on the other side, widely used examples such pixel-based approaches are still not applied or tested for the HEVC compressed videos.

3. No-Reference Video Quality Assessment

In this section, NR measurement techniques in the spatial domain for two KPI are proposed: blur and blockiness. Assuming that we do not own a knowledge and assumptions of the original content or the distortion process of the HEVC compression. In fact, the NR pixel-based approach for measuring artifacts of the visual quality is proposed by considering a given model of degradation to investigate the performance of the mentioned metrics.

3.1. No-Reference Blockiness Metric

The same approach is used for calculating the blockiness artifact published in [22]. It is calculated locally for each coding block. Absolute differences in pixel luminance were calculated separately for intra-pairs, represented by neighboring pixels from a single coding block, and inter-pairs, represented by pixels from neighboring blocks. A ratio between the total values of intra- and inter-differences is calculated over the entire video frame. For a real time application the metric should be calculated over a time window (the number of video frames). Mean value for the window represents a blockiness level. For the purposes of the experiment the window size was equal to the sequence length (10 s). It was verified that the level of the blockiness artifact does not change significantly over time within the same video scene. Thus, any other window size or different method for temporal pooling would yield similar results.

3.2. No-Reference Blur Metric

The blurred image in compression techniques appears when high spatial frequency components of the image spectrum are truncated. For instance, possible reasons of blurring can be out-of-focus capturing or relative motion between the camera and the captured object. Besides, high compression performance can introduce blur when processing the data of images' sequence. Perceptually, the blur artifact appears along edges and textured regions. In this work, the width of the edges is measured in order to characterize smoothing blur effect [23]. First, the Sobel filter as an edge detector is applied to find the gradient of the image. It is obvious that below a certain threshold, blur remains as just noticeable and visually unperceived. According to that threshold, the pixels being the part of the edges are differentiated. Then the width of an edge is measured, depending on its growth direction (left or right). Finally, the global blur value is obtained by averaging over all edges of the whole image.

4. Objective Video Quality Methods

Huge variety of proposed works concerning the video quality measurement use the objective metrics such as the simplest and commonly used ones: the PSNR and Mean-Squared Error (MSE). But in general, it is not ensured that error visibility would always the appearance of quality artifacts for most of distortions. Assuming that the structural information is highly captured from the viewing field by the human visual system, extracting this kind of information provides a good estimation of the perceived distortion. Therefore, the Structural Similarity (SSIM) has been used recently to characterize complex structured signals [24].
However, the different types of video coding and transmission systems require a more general model that covers a wide range of quality degradations. In fact an extensive objective and subjective tests should be performed to provide an effective perceptual measurement. The Video Quality Model (VQM) was indeed proposed by the Institute for Telecommunication Science (ITS) [25] and standardized by the American National Standards Institute (ANSI). It was further included in Draft Recommendations from ITU-T SG9. The VQM has proved a good performance for measuring perceptual effects of different types of video impairments such as blurring, jerkiness and block distortion. The calculation of VQM taking as input the original and processed videos follows these main steps:

- calibrate the processed video with respect to the original sequence by estimating and correcting the spatial-temporal shifts, as well as adjusting the contrast and brightness,

- extract a set of quality features to characterize perceptual changes from particular spatial-temporal regions in the video stream; for instance, in the chrominance, temporal and spatial properties,
- compare the extracted features form the processed video with those of the original sources,
- conclude the VQM value using a linear combination of the obtained parameters.

From the described functions, it makes sense that the VQM has a high correlation with subjective scores, which makes us believe that using it as a reference metric would provide accurate testing results.

5. Experiments and Results

In order to effectively evaluate the video quality based on HEVC compressions, the dataset of the project developed by the Joint Effort Group (JEG) of Video Quality Experts Group (VQEG) is used [21]. It presents a large-scale database of HEVC coded videos for researchers involved in designing hybrid quality metrics. Different encoding parameters were performed on ten sequences representing different characteristics. Among interesting benefits of the mentioned dataset, objective quality measurements are provided at frame-level granularity. This database is exploited by applying the NR metrics of blur and blockiness. It is primordial to investigate the accuracy of these metrics for the HEVC distortions and make useful interpretations about the specificities of the target approach.

5.1. Selected Compression Parameters

The performance of the quality metrics is investigated based on a diverse set of encoding parameters. Table 3 presents the retained HEVC configurations in order to carry out tests over an increasing data compression. The distortion is intrinsically related to the following values selected from the adopted database [21].

Table 3
Encoding parameters

Parameter	Value
WIDTH	1280
GOPTYPESIZE	GOP8
RATECONTROL_QP	26, 32, 38, 46
RATECONTROL_FRAME_mbit/s	1, 2, 4, 8, 16
REFRESH	1
INTRAPERIOD	16
SLICEARGUMENT	0

The resolution of the ten original sources is 1280×720 pixels. The authors take into account all available fixed QP values as it represents a basic distortion source along with the frame rate control. The refresh number corresponds to the decoding refresh type, to apply a non-IDR clean random access point. This encoder option allows the use of an open GOP. The slicing value signifies one slice per frame. As a result, 90 processed video sequences are generated based on the above parameters. The prior-knowledge of these settings is not considered in developing the NR metrics and authors just provide it for a precise description of the compression rate and consequently the distortion strength.

5.2. Results and Analysis

Table 4 displays the results of the applied metrics on the ten processed videos. For each sequence, the Pearson correlation coefficient is used to validate the performance of the blur and blockiness measurements relative to the VQM values, offered by the JEG project for each encoded sequence according to the given parameters. From the shown results, the efficiency of the blur metric is confirmed for each source which means that the distorted edges are well predicted, providing high correlation values. It is further clear that the blockiness metric works well for the majority of the sources.

Table 4
Pearson correlation coefficients with VQM

Source	Blockiness	Blur
*src*01	–0.67	1.00
*src*02	–0.97	0.91
*src*03	–0.97	0.96
*src*04	–0.77	0.99
*src*05	–0.87	0.99
*src*06	–0.57	0.91
*src*07	–0.96	0.96
*src*08	–0.95	0.92
*src*09	–0.36	0.95
*src*10	0.69	0.99

The authors mention here that the origin of the negative scores is caused by the metric's construction, as increasing the compression rate corresponds to lower values of blockiness and vice versa. However, the correlation tends to drop for the case of src09 due to the complex nature of the motion and spatial activity in the video. Src09 consists of several combined shots separated with a black-pixels frame. Besides, the positive correlation of the src10 means that the trends of values are opposite to the expected ones.

The scatter plots in Figs. 1 and 2, representing the blockiness and blur metrics for all sequences, respectively, reveal a partial success even the measures are convincing for each source separately. The global correlation coefficient of blockiness is 0.55 whilst 0.23 for blur, which gives rise to useful interpretations for a more complete evaluation ap-

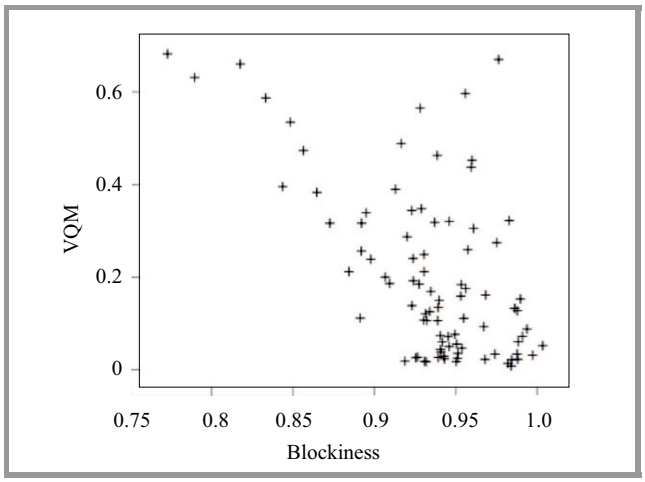


Fig. 1. Correlation with VQM for blockiness.

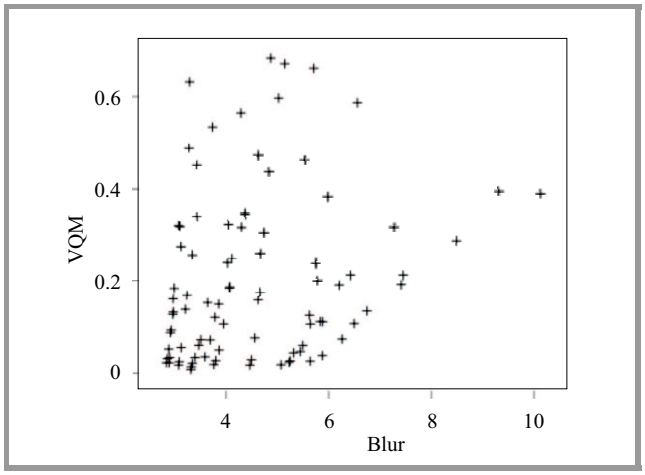


Fig. 2. Correlation with VQM for blur.

proach. The temporal and spatial inconsistency could be incorporated to overcome such problems.

6. Conclusion and Further Work

In case of pixel-based NR methods, an accurate model which combines different kind of artifacts would generate an estimation of the perceptual quality using weighting factors. The strength of weights, which could be determined by a regression analysis, is computed with respect to a particular single metric. This latter is combined to another distortion measure, based on a linear or non-linear model. Furthermore, even the VQM measurement combines several features and represents with a certain precision perceptual characteristics, implicating subjective scores in the assessment process still more effective. For instance, correctness functions such as sigmoid model, can be applied on the predicted measures according to the subjective evaluation as it requires parameters' estimation.

The HEVC specificities as the highly flexible quad-tree structure and effective prediction tools allow an accurate exploitation of the video content in addition to the high compression performance. Assessing quality of HEVC processed videos for different types of distortions require sophisticated techniques for a successful NR approach. In this work, the proposed metrics as a basic step to establish a completing framework of quality assessment are analyzed, taking into consideration particular aspects introduced in this new codec.

Acknowledgments

The work was co-financed by The Polish National Centre for Research and Development (NCBR), as a part of the EUREKA Project no. C 2012/1-5 MITSU and contract number 11.11.230.018. This research was further supported by the Bureau For Academic Recognition and International Exchange.

References

[1] M. Leszczuk, M. Hanusiak, M. Farias, E. Wyckens, and G. Heston, "Recent developments in visual quality monitoring by key performance indicators", *Multimed. Tools & Appl.*, pp. 1–23, 2014 [Online]. Available: http://dx.doi.org/10.1007/s11042-014-2229-2

[2] G. Sullivan, J. Ohm, W. J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard", *IEEE Trans. Circ. Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, 2012.

[3] ITU-T Recommendation H.265, "High efficiency video coding", 2015 [Online]. Available: http://www.itu.int/rec/T-REC-H.265

[4] M. Leszczuk, M. Hanusiak, I. Blanco, A. Dziech, J. Derkacz, E. Wyckens, and S. Borer, "Key indicators for monitoring of audiovisual quality", in *Proc. 22nd Sig. Process. Commun. Appl. Conf. SIU 2014*, Trabzon, Turkey, 2014, pp. 2301–2305.

[5] ITU-T Recommendation P.930, "Principles of a reference impairment system for video", 1996 [Online]. Available: http://www.itu.int/rec/T-REC-P.930-199608-I

[6] ITU-T Recommendation J.144, "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference", 2004 [Online]. Available: http://www.itu.int/rec/T-REC-J.144-200403-I

[7] ITU-T Recommendation J.247, "Objective perceptual multimedia video quality measurement in the presence of a full reference", 2008 [Online]. Available: http://www.itu.int/rec/T-REC-J.247-200808-I

[8] ITU-T Recommendation J.246, "Perceptual visual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference", 2008 [Online]. Available: http://www.itu.int/rec/T-REC-J.246-200808-I

[9] ITU-T Recommendation J.249, "Perceptual video quality measurement techniques for digital cable television in the presence of a reduced reference", 2010 [Online]. Available: http://www.itu.int/rec/T-REC-J.249-201001-I

[10] ITU-T Recommendation J.341, "Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference", 2011 [Online]. Available: http://www.itu.int/rec/T-REC-J.341-201101-I

[11] ITU-T Recommendation J.342, "Objective multimedia video quality measurement of HDTV for digital cable television in the presence of a reduced reference signal", 2011 [Online]. Available: http://www.itu.int/rec/T-REC-J.342-201104-I

[12] ITU-T Recommendation P.1202, "Parametric non-intrusive bitstream assessment of video media streaming quality", 2013 [Online]. Available: http://www.itu.int/rec/T-REC-P.1202

[13] ITU-T Recommendation J.343, "Hybrid perceptual bitstream models for objective video quality measurements", 2014 [Online]. Available: http://www.itu.int/rec/T-REC-J.343

[14] J. Gustafsson, G. Heikkila, and M. Pettersson, "Measuring multimedia quality in mobile networks with an objective parametric model" in *Proc. 15th IEEE Int. Conf. Image Process. ICIP 2008*, San Diego, CA, USA, 2008, pp. 405–408.

[15] A. Takahashi, K. Yamagishi, and G. Kawaguti, "Global standardization activities recent activities of QoS/QoE standardization in ITU-T SG12", *NTT Tech. Rev.*, vol. 6, no. 9, pp. 1—5, 2008.

[16] M. Aabed *et al.*, "No-reference quality assessment of hevc videos in loss-prone networks", in *Proc. IEEE Int. Conf. Acoust., Speech Sig. Process. ICASSP 2014*, Florence, Italy, 2014, pp. 2015–2019.

[17] M. Shahid, J. Panasiuk, G. Van Wallendael, M. Barkowsky, and B. Lövström, "Predicting full-reference video quality measures using hevc bitstream-based no-reference features", in *Proc. 7th Int. Worksh. Quality of Multimed. Exper. QoMEX 2015*, Costa Navarino, Messinia, Greece, 2015.

[18] N. Staelens, S. Moens, W. Van den Broeck, I. Marien, B. Vermeulen, P. Lambert, R. Van de Walle, and P. Demeester, "Assessing quality of experience of IPTV and video on demand services in real-life environments", *IEEE Trans. Broadcast.*, vol. 56, no. 4, pp. 458–466, 2010.

[19] B. Lee and M. G. Kim, "No-reference psnr estimation for hevc encoded video", *IEEE Trans. Broadcast.*, vol. 59, no. 1, pp. 20–27, 2013.

[20] L. Janowski, P. Kozłowski, R. Baran, P. Romaniak, A. Głowacz, and T. Rusc, "Quality assessment for a visual and automatic license plate recognition", *Multimed. Tools & Appl.*, vol. 68, no. 1, pp. 23–40, 2014.

[21] G. Van Wallendael, N. Staelens, E. Masala, and M. Barkowsky, Full-HD HEVC-encoded video quality assessment database", in *Proc. 9th Int. Worksh. Video Process. Quality Metrics VPQM 2015*, Chandler, AZ, USA, 2015.

[22] M. Mu, P. Romaniak, A. Mauthe, M. Leszczuk, L. Janowski, and E. Cerqueira, "Framework for the integrated video quality assessment", *Multimed. Tools & Appl.*, vol. 61, no. 3, pp. 787–817, 2012.

[23] E. Cerqueira, L. Janowski, M. Leszczuk, Z. Papir, and P. Romaniak, "Video artifacts assessment for live mobile streaming applications", in *Future Multimedia Networking*, A. Mauthe, S. Zeadally, E. Cerqueira, M. Curado, Eds. *LNCS* 5630, pp. 242–247. Springer, 2009.

[24] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity", *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[25] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality", *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, 2004.

Kais Rouis received his B.Sc. in Computer Sciences in 2009 and his M.Sc. in Intelligent and Communicating Systems in 2011. He is currently a Ph.D. student at the National Engineering School of Tunis at the University of Tunis El Manar. His work revolves around image and video coding and wireless communication techniques.

E-mail: kais.rouis.phd@ieee.org
Innov'COM Lab, Sup'Com
University of Carthage

Ecole Nationale d'Ingénieurs de Tunis
Université de Tunis El Manar
Tunisia

Mikołaj Leszczuk started his professional career in 1996 at Comarch SA as manager of the Multimedia Technology Department, and then at Comarch Multimedia as the CEO. Since 1999 has been employed at the AGH Department of Telecommunications. In 2000 he moved to Spain for a four-month scholarship at the Universidad Carlos III de Madrid. After returning to Poland, he was employed at the Department of Telecommunications as a research and teaching assistant, and in 2006, he successfully defended his doctoral dissertation as an assistant professor. His current research interests are focused on multimedia data analysis and processing systems, with particular emphasis on QoE. He has participated more than 20 major research projects, including FP4, FP5, FP6, FP7, Horizon 2020, OPIE, Culture 2000, PHARE, eContent+, and Eureka!. Between 2009 and 2014, he was the administrator of the major international INDECT research project, dealing with solutions for intelligent surveillance and automatic detection of suspicious behavior and violence in urban environments. He is a member of VQEG, IEEE, and GAMA. He (co-)authored approximately 130 scientific publications.

E-mail: leszczuk@agh.edu.pl
AGH University of Science and Technology,
Department of Telecommunications
Mickiewicza av. 30
30-059 Krakow, Poland

Lucjan Janowski is an assistant professor with the Department of Telecommunications, AGH University of Science and Technology. He received his Ph.D. degree in telecommunications in 2006 from the AGH. In 2007, he worked in a postdoctoral position at the Centre National de la Recherche Scientifique (CNRS), LAAS (Laboratory for Analysis and Architecture of Systems of CNRS) in France, where he prepared both malicious traffic analysis and anomaly detection algorithms. In 2010–2011, he spent half a year in a postdoctoral position at the University of Geneva, working on quality of experience (QoE) for health applications. In 2014–2015, he spent half a year in a postdoctoral position at The Telecommunications Research Center Vienna (FTW), working on quality of experience for IPTV customers. His main interests are statistics and probabilistic modeling of subjects and subjective rates used in QoE evaluation.

E-mail: janowski@kt.agh.edu.pl
AGH University of Science and Technology,
Department of Telecommunications
Mickiewicza av. 30
30-059 Krakow, Poland

Zdzisław Papir is professor and a deputy chair at Department of Telecommunications, AGH University of Science and Technology. Between 1999–2006 he was a guest co-editor for IEEE Communications Magazine responsible for the Broadband Access Series. He has been participating in several R&D IST European projects being responsible for network performance evaluation and quality assessment of communication services. He has also been appointed as an ICT expert by the European Commission. His current research interests include modeling of telecommunication networks/services and measuring quality of experience.

E-mail: papir@kt.agh.edu.pl
AGH University of Science and Technology
Department of Telecommunications
Mickiewicza av. 30
30-059 Krakow, Poland

Jamal Bel Hadj Tahar received his Ph.D. in 1993 at the Polytechnic Institute of Grenoble, INPG, and became master conference in 2011 at the Higher School of Communications of Tunis, on Information and communication technologies TIC. Currently, he is professor at the National Engineering School of Sousse and responsible for the training of masters research in telecommunications engineering. His research focuses on two areas one is one wireless communication techniques and systems while the other is optical systems.

E-mail: belhadj.tahar@supcom.rnu.tn
Innov'COM Lab, Sup'Com
University of Carthage

Ecole Nationale d'Ingénieurs de Tunis
Université de Tunis El Manar
Tunisia

Paper

Quantifying the Suitability of Reference Signals for the Video Streaming Analysis for IPTV

Christian Hoppe[1], Robert Manzke[1], Marcus Rompf[1], and Tadeus Uhl[2]

[1] *Kiel University of Applied Sciences, Kiel, Germany*
[2] *Maritime University of Szczecin, Szczecin, Poland*

Abstract—IP networks are indispensable nowadays and they are some of the most efficient platforms. The constantly growing number of users and new services in these networks – the largest being the Internet – require a satisfactory quality of service from any application they use. So, determining the QoS in real-time services is particularly important. This work shows how to quantify the suitability of reference signals for analyzing the quality of video streaming in IPTV. The assessment relies on two different algorithms: PEVQ and VQuad-HD. Three different reference signals – two real ones and an artificial one – are used in this study, and a numerical measurement system is used, which simulates mean network impairments. These measurements provide valuable information for determining the QoS of actual IPTV services in practice.

Keywords—communication network, IPVT service, ITU-T J.247, measurement tool, PEVQ, QoS/QoE determination, reference signals, Triple Play Services, VQuad-HD, ITU-T J.341.

1. Introduction

3G and Triple play networks are expanding day by day. Their new applications and services include video telephony, video conferencing, video streaming and video podcasts. Although networks have never been as powerful and reliable as they are today, IPTV, mobile TV and others call for new fixed and mobile applications. A major factor for their increasing success will be their ability to satisfy their customers' high expectations while keeping down the costs. Operators and service providers achieve this by employing new powerful technology for their setups as well as new measurement tools that help to maintain a satisfactory level of Quality of Experience (QoE).

One of the major uses of next-generation networks is simulcast streaming (or broadcasting) of identical contents in various formats for different applications. Also referred to as the "Triple Screen" scenario, video content will typically be transmitted in high quality over cable or satellite HDTV networks. Medium quality will be available over the Internet for streaming to clients on PCs and laptops while the lowest quality will be offered on mobile multimedia devices such as mobile phones, smartphones and tablets. Triple Screen scenarios involve many steps of signal post-processing, including reformatting (e.g. 16:9 to 4:3), rescaling (e.g. from HD to VGA or CIF), reframing (e.g. from 50 f/s to 25 f/s), transcoding, and retransmission over IP-based networks. The issue for the test engineer is to maintain the best QoE possible across the various formats, given the system-bound limitations of each.

Three important measurement techniques [1] are used to assess QoE and Quality of Service (QoS). The "Full Reference Model", the "Reduced Reference Model" and the "No Reference Model" are shown in Fig. 1. These measurement techniques are also to be found in standard QoE/QoS measurement models, as described in [2] and [3]. The short texts contained in Fig. 1 explain briefly the procedure used in each of the measurement techniques and list the typical application scenarios of each.

The Full Reference Model technique was used in conjunction with two algorithms, PEVQ (ITU-T J.247) [4] and VQuad-HD (ITU-T J.341) [5], for the bulk of this study. Using these algorithms means using suitable reference signals that satisfy a number of requirements not the least of which are format and length. However, selecting suitable reference signals is not as easy as it might at first seem, as this paper will show.

First of all, PEVQ, the algorithm primarily used for this study, will be explained briefly. The chief requirements on reference signals according to international recommendations are then presented. The paper goes on to describe the selection of suitable reference signals. The investigation's goal is to find reference signals which, on the one hand, fulfill the main requirements laid out in the international recommendations and, on the other hand, deliver the best QoS/QoE values on the MOS scale. The following Section 2 will then present the analysis architecture and the scenarios chosen. A further Section 3 presents graphically several examples, the analysis results, and their interpretation. The concluding series of analyses in Section 4 presents a comparison of the algorithms PEVQ and VQuad-HD. The paper closes with a summary and an outlook on future areas of study in Section 5.

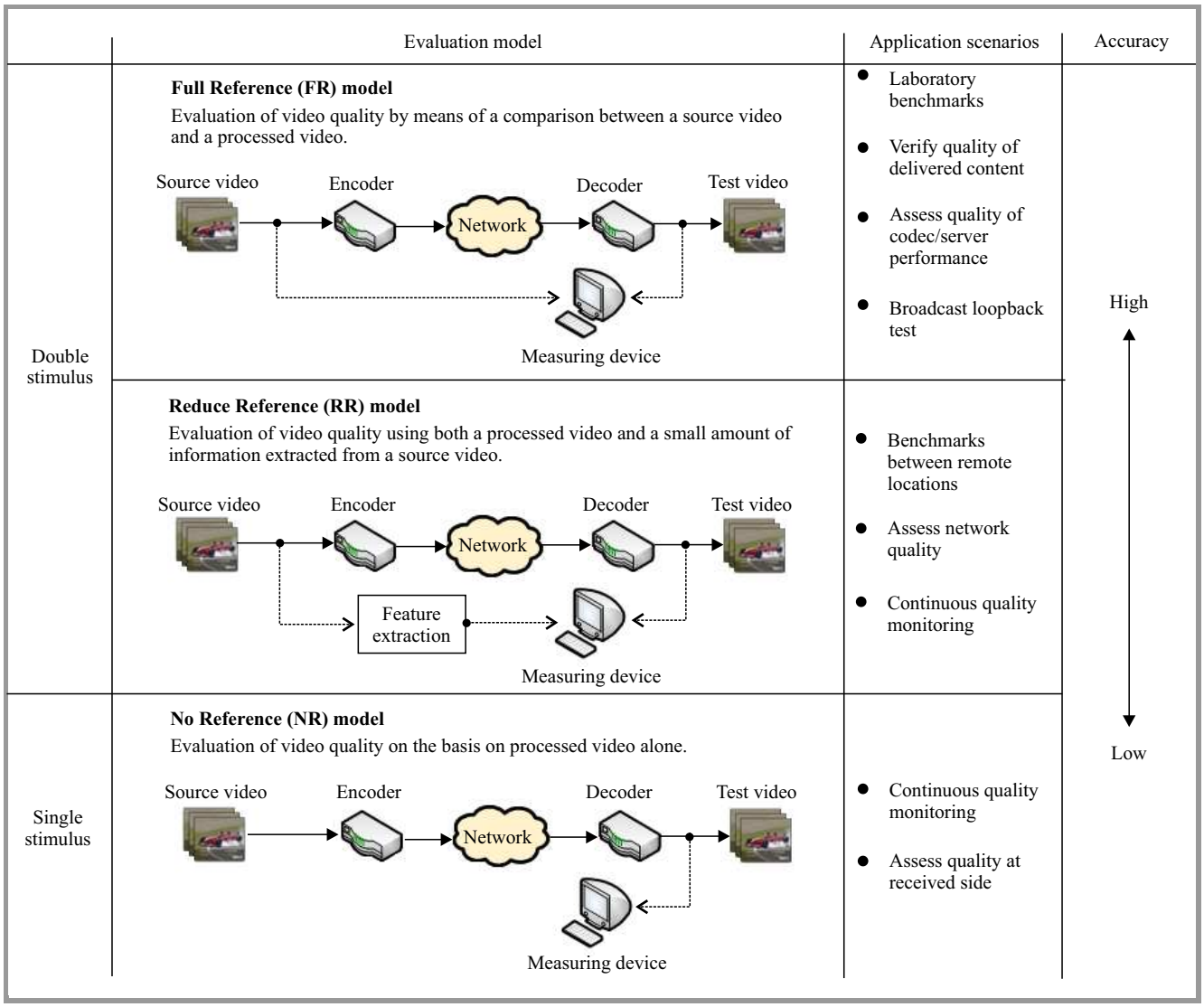


Fig. 1. Overview of QoE/QoS measurement techniques. (See color pictures online at www.nit.eu/publications/journal-jtit)

2. The PEVQ Algorithm

PEVQ is designed to predict the effects of transmission impairments on the video quality as perceived by a test person. Its main application areas are mobile multimedia applications and IPTV. It fulfills the ITU-T Recommendation J.247 [4] for full reference quality measurements. The key features of PEVQ (Fig. 2):

- temporal alignment of the input sequences based on multi-dimensional feature correlation analysis with limits that reach far beyond those tested by the Video Quality Experts Group (VQEG), especially with regard to the amount of time clipping, frame freezing and frame skipping which can be handled;
- full frame spatial alignment;
- color alignment algorithm based on cumulative histograms;
- enhanced frame rate estimation and rating;
- detection and perceptually compliant weighting of frame freezes and frame skips;
- only four indicators are used to detect the video quality. Those indicators operate in different domains (temporal, spatial, chrominance) and are motivated by the Human Visual System (HVS). Perceptual masking properties of the HVS are modeled at several stages of the algorithm. These indicators are integrated using a sophisticated spatial and temporal integration algorithm.

Apart from the MOS value, which is the ultimate yardstick of quality, PEVQ offers several other indicators that are used to analyse the reasons for quality impairments such as:

- distortion,
- delay,

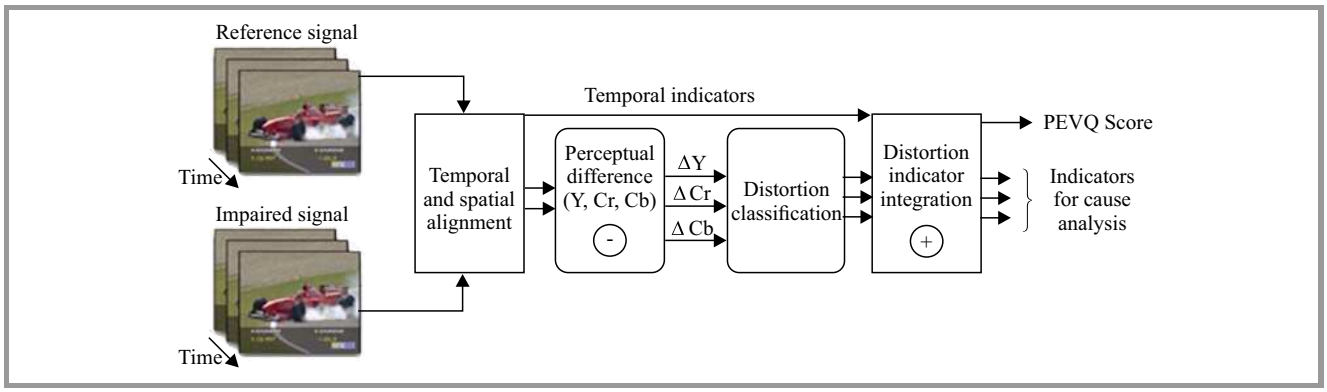


Fig. 2. Sequence diagram of PEVQ score calculation.

- luminance,
- contrast,
- peak signal to noise ratio,
- jerking,
- blurring,
- block constriction,
- frame freezing and frame skipping,
- effective picture rate,
- time and areal activity.

The PEVQ algorithm is the tool used for the bulk of the analyses described in this paper. For the sake of comparison a second algorithm, VQuad-HD, is introduced. The two algorithms necessitate the use of specifically formatted reference signals. That is the theme of the next chapter.

3. Requirements on Reference Signals

Many factors need to be taken into consideration when selecting reference signals. These factors can be found in the ITU-T Tutorial [6] and in publications of the VQEG [7]. The video format requirements and recommendations of the algorithms and tools used state that the best results will be obtained if the reference file is an uncompressed AVI (Audio Video Interleaving) file in the YUV 4:2:0 color space. A short video sequence of around 10 s is ideal since the algorithms would take far too long to process longer sequences whilst the influence of network impairments in shorter sequences would hardly coax sufficiently meaningful responses from the algorithms. In Europe full HD videos are usually in 1080i50, which means a resolution of 1920×1080 pixels at 25 full frames per second are ideal parameters for the reference signals. The reference signals should of course be free from distortions, errors and coding artifacts to preclude influences above and beyond network impairments.
The sequences selected for the comparison described here differ with regard to spatial details, motion and color intensity. A selection of reference files can be found in the consumer digital video library [8] or from the license holders of the two measurement algorithms (Opticom [9] and SwissQual [10]).

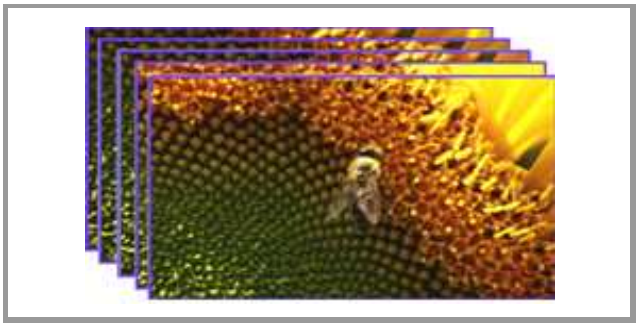


Fig. 3. Sunflower images (name: Sunflower) [9].

Fig. 4. Tractor images (name: Tractor) [9].

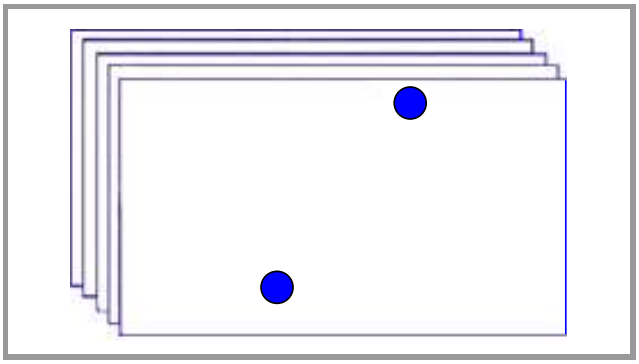


Fig. 5. Videotoms images (name: Artificial) [11].

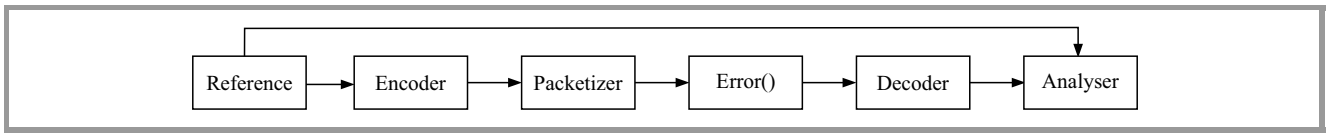


Fig. 6. Schematic representation of quality measurement by QoSCalc(IPTV).

The following reference signals were chosen for this analysis:

- little movement and slow, small changes in the images with relatively high color intensity (Fig. 3),
- greater movement and changes in the image background due to zooming with relatively low color intensity (Fig. 4),
- simple geometric shapes (circles) with rapid and random movement with minimal changes in color intensity (Fig. 5).

QoE/QoS measurements can be conducted in two different ways: in a real environment, or in an emulated environment. One evaluation of QoE/QoS on the basis of a reference signal takes several minutes. A number of measurements for each parameter setting are needed to arrive at any meaningful results. In a real environment, this ties up both network and measurement resources. That is why it is better to conduct analysis like this in an emulated measurement environment. This also allows a range of parameter settings to be used automatically and yields results that are duplicable in similar measurement scenarios. That explains why a numerical tool has been used for the analysis described in this paper. The next chapter is a brief description of the tool.

4. The Analysis Environment

For the reasons given above a numeric software tool QoSCalc (IPTV) [12] was used to analyze the quality of a video stream. The tool automates the entire measurement process.
The following explains each block in the sequence shown in Fig. 6. in order to compare the real environment with the measurement tool:

- a reference video file is loaded;
- the video is encoded in accordance with the selected codec by FFmpeg [13];
- the coded data is encapsulated according to the selected transport protocol (e.g. native RTP [14], MPEG2-TS [15], etc.) by FFmpeg;
- the block "Error" represents the generation of a selected level of network impairments;
- the packeted video is decoded to the same format as the reference (raw video, same resolution and bitrate) by FFmpeg;
- finally, the decoded data and the video reference file loaded at the start are passed on to the evaluation algorithm (here PEVQ or VQuad-HD). This computes the quality score on the MOS [16] scale and then saves it.

The "Error" block has been designed for non-deterministically distributed packet loss (binominal distribution with probability P) and non-deterministically distributed burst size (exponential distribution) with a selectable mean value.
Two different versions of the tool QoSCalc(IPTV) were used, utilizing different versions of FFmpeg. The first version is the default FFmpeg with its main error concealment techniques enabled. In the second version the error concealment methods are disabled. This is done specifically to analyze the influence of the error concealment methods.
Different error concealment algorithms for video streaming exist [17]. The FFmpeg uses the techniques "Macro Block Detection" [18], and "Motion Vector Search" [19], which are designed to detect and predict movement of macro blocks in the pictures and substitute missing information. FFmpeg first counts how many macro blocks are intact (not lossy). If that number is above a set threshold then intra concealment is used. Otherwise, an inter error concealment is used.
Intra error concealment involves averaging the pixels of the macro blocks surrounding the damaged one. The result of weighting and averaging the uncorrupted blocks is the block used for concealment. Inter error concealment works differently for I, P and B frames. In I and P frames the surrounding blocks are analyzed using the motion vectors, and several replacement block candidates are calculated using different methods including median and means. The block which produces the smoothest transitions is then chosen. In B frames the decoder uses the nearest P reference frame and creates a forwardly and backwardly weighted version of the motion vector.
The following configuration has been chosen for the measurement scenario in the testing environment:

- Reference files: Sunflower 1080p25 (similar 1080i50), Tractor 1080p25, Artificial: 1080p25,
- Packet loss: 0–12 (in steps of 1), 14, 16, 20%,
- Burst size: 1–5 (in steps of 1),
- Packaging: MPEG2-TS,
- Encoding: H.264 (medium),
- Bitrates: 1,000, 3,875, 6,750, 8,625 and 10,500 kb/s.

Using the numerical tool described above several analysis were conducted over several days for each scenario. The most significant results of the measurement scenario are presented and interpreted in the following section.

5. Quantitative Comparison of the Reference Signals

First of all it is necessary to describe the expectations which might be had. Due to network impairments, in this case packet loss, the expectation is that higher packet loss would result in a lower MOS value. Regarding the video content at one test point, e.g. 1% packet loss, and assuming that at this test point information is missing in scenes with a large degree of motion or rapid changes in color intensity the expectation would be a lower MOS value.
Given the nature of the test results from the configuration described in Section 4 a representations of the following configurations have been selected:

- Reference files: Sunflower.avi, Tractor.avi and Artificial.avi video content,
- Packet loss: 0–12 (in steps of 1), 14, 16, 20%,
- Burst size: 1 and 2,
- Packaging: MPEG2-TS,
- Encoding: H.264 (medium),
- Bitrates: 3,875 kb/s and 10,500 kb/s,
- Evaluation algorithm: PEVQ.

Figures 7–10 represent the results, starting with 3,875 kb/s and burst size 1.

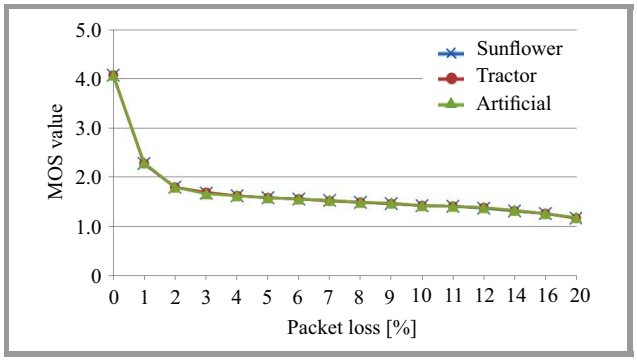


Fig. 7. Comparison of all three reference signals at 3,875 kb/s and burst size 1.

From these results, it is obvious that the MOS values for all bit rates and bursts are very close to each other. These results differ widely from the expectations, which led to two assumptions: either the video content does not affect the MOS value at all, or the functionality of FFmpeg decoding techniques is fully able to cope, or both. So the FFmpeg decoding techniques had to be examined in greater depth.
Two techniques, called “Macro Block Detection” and “Motion Vector Search”, are used to conceal errors. They

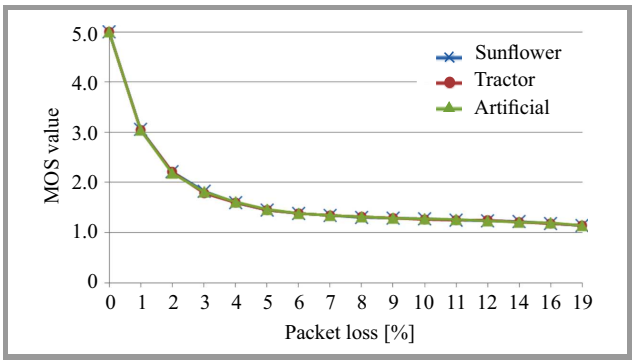


Fig. 8. Comparison of all three reference signals at 10,500 kb/s and burst size 1.

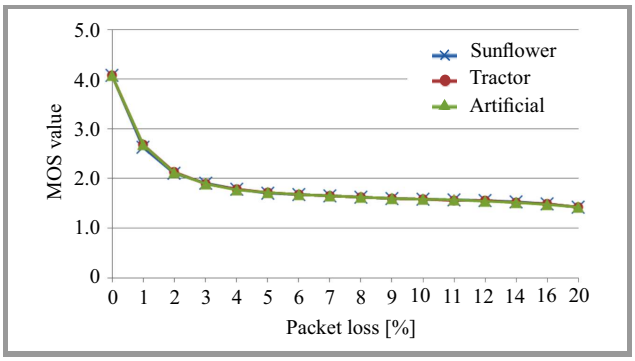


Fig. 9. Comparison of all three reference signals at 3,875 kb/s and burst size 2.

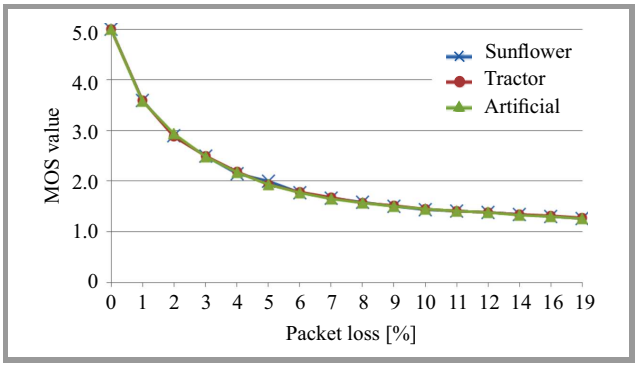


Fig. 10. Comparison of all three reference signals at 10,500 kb/s and burst size 2.

obviously do a good job. They were the subject of the next series of tests with the expectation being a lower MOS value when error concealment techniques are disabled. Figures 11–14 represent the results; they include the representation to allow a comparison of the Tractor.avi reference signal with and without error concealment.

In conclusion, it can be said that the expectation was justified, at least as far as lower packet losses as the network impairment are concerned. When error concealment is enabled, the MOS value is indeed higher. With regard to the second point of intersection of all diagrams and curves the following observations can be made:

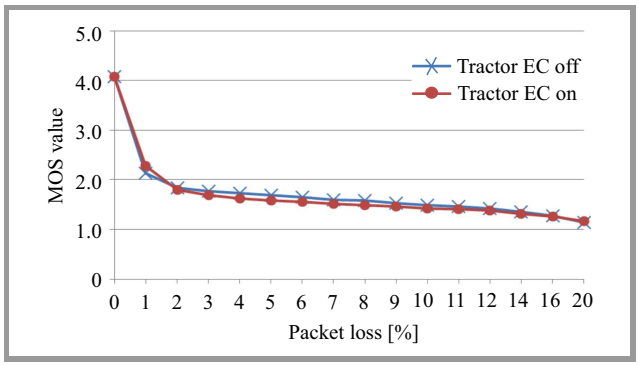


Fig. 11. Comparison Tractor with and without EC at 3,875 kb/s and burst size 1.

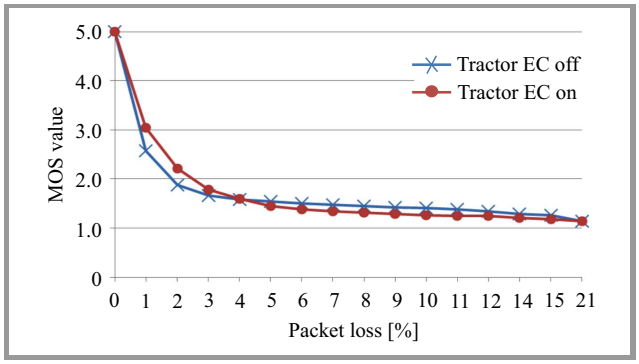


Fig. 12. Comparison Tractor with and without EC at 10,500 kb/s and burst 1.

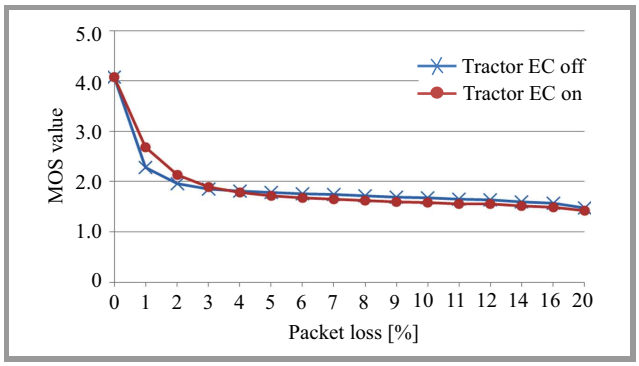


Fig. 13. Comparison Tractor with and without EC at 3,875 kb/s and burst size 2.

First, at some points with high packet losses, the MOS value obtained when error concealment is disabled is actually higher than that obtained when it is enabled. The video quality, with a MOS value of less than 2, is really poor nonetheless. The reason for that could be that these techniques substitute wrong video content. In severely lossy networks it might be better to disable error concealment techniques.

Second, the second point of intersection of both curves can be shifted in the direction of higher packet loss by increasing either the bit rate or the burstiness, so that the resulting higher MOS value, with error concealment enabled, would lead to improved video quality. These observations could

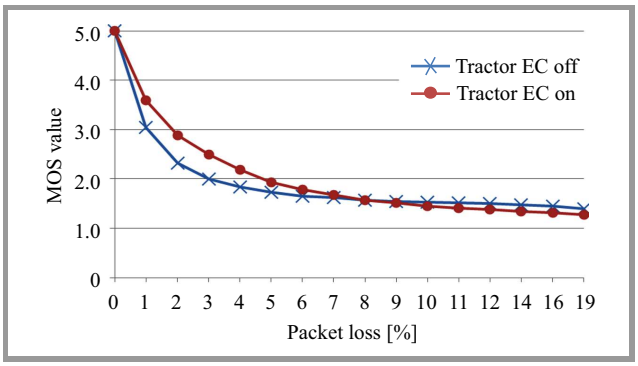


Fig. 14. Comparison Tractor with and without EC at 10,500 kb/s and burst size 2.

lead to useful implementations which improve video quality by artificially increasing network burstiness, which is already present anyway, whenever packet losses occur.

As far as reliability is concerned Figs. 15–16 represent results using both PEVQ (ITU-T J.247 [4]) and VQuad-HD (ITU-T J.341 [5]) prediction algorithms for video content of the reference signals Tractor.avi and Artificial.avi, the expectation being that these algorithms should yield only slightly differing respective MOS values.

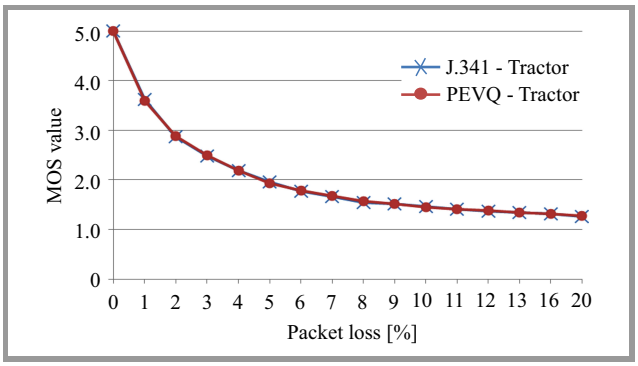


Fig. 15. Comparison of J.341 and PEVQ for Tractor at 10,500 kb/s and burst size 2, with EC.

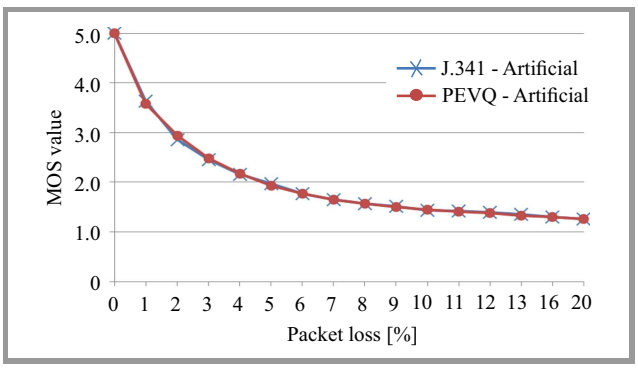


Fig. 16. Comparison of J.341 and PEVQ for Artificial at 10,500 kb/s and burst size 2, with EC.

Again, it can be said that the expectation was justified, which leads to the following two final conclusions. First, both algorithms are suitable for the perceptual evaluation

and measurement of HD video quality and second, artificial video content is suitable as a reference signal: its use leads to simpler realization of initial scenario criteria.

6. Summary and Outlook

This paper has assessed the suitability of video reference signals for the PEVQ (ITU-T J.247) and the VQuad-HD (ITU-T J.341) algorithms for evaluating the video quality in IPTV. To that end numerical software tool was used that had been developed previously on the basis of FFmpeg to provide encoding, packaging, degrading (packet loss, burst) and decoding techniques. Both algorithms are full reference models, that is: they necessitate the use of two signals – the original signal on the one hand, and a degraded signal on the other. Research on the topic of suitability has shown that there are recommendations regarding the composition of reference files with regard to, for example, changes in movement, color and luminescence. Accordingly, two reference files provided by "Opticom" and one provided by [11] were selected and the analysis environment was set up to implement the files and initiate the measurements. The results obtained for the video quality under evaluation differed from the expectations, one of them having been, for example, better video quality for the reference file that contained less movement when video content information loss occurs. There were, however, hardly any differences. That led directly to an investigation of FFmpeg's "Decoder", which showed that the existing error concealment techniques provide very good functionality in repairing and concealing issues. Nevertheless, as was expected, increasing packet loss caused an exponential decrease in the resulting MOS value for the video quality of a reference file examined in isolation. One surprising result must be spotlighted: the "artificial" signal Artificial.avi proved to be just as suitable for use in QoE/QoS measurements as the very complex reference signals recommended by the license holder Opticom [9]. Further analysis, which cannot be described here owing to lack of space, confirm the good functionally of the EC techniques implemented in FFmpeg.

The results obtained in the course of the work described here could serve as a basis on which to develop parameterized QoE/QoS models, that are widely known to be simple and easy to implement in practice, yet provide reliable meaning results. It is therefore very worthwhile developing such QoE/QoS models. Further work in this direction is already being planned.

References

[1] Network Technology Laboratories, Objective Video Quality Assessment Methods [Online]. Available: http://www.ntt.co.jp/qos/qoe/eng/technology/visual/02_3.html (accessed Nov. 2015).

[2] A. Raake, *Speech Quality of VoIP*. Chichester: Wiley, 2006.

[3] T. Uhl, "E-model and PESQ in the VoIP environment: A comparison study", in *Proc. 5th Polish-German Teletraffic Symp.*, Berlin, Germany, 2008, pp. 207–216.

[4] ITU-T Recommendation J.247: Objective perceptual multimedia video quality measurement in the presence of a full reference [Online]. Available: http://www.itu.int/rec/T-REC-J.247-200808-I (accessed Nov. 2015)

[5] ITU-T Recommendation J.341: Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference [Online]. Available: http://www.itu.int/rec/T-REC-J.341-201101-I/en (accesed Nov. 2015).

[6] ITU-T Tutorial, Objective perceptual assessment of video quality: Full reference television [Online]. Available: http://www.itu.int/ITU-T/studygroups/com09/docs/tutorial_opavc.pdf (accesed Nov. 2015).

[7] Video Quality Experts Group (VQEG), HDTV quality determination [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/hdtv/hdtv.aspx (accessed Jul. 2015)

[8] The consumer digital video library, reference file collection [Online]. Available: http://www.cdvl.org/login.php (accesed Nov. 2015).

[9] Company Opticom website [Online]. Available: http://www.opticom.de (accesed Nov. 2015).

[10] Company SwissQual AG – A Rohde & Schwarz Company [Online]. Available: http://www.swissqual.com (accesed Nov. 2015).

[11] Videotoms samples [Online]. Available: https://drive.google.com/folderview?id=0B4qqvf83yCxtTHBWZGtZd3R6Y2M&usp=sharing (accesed Nov. 2015).

[12] T. Uhl and H. Jürgensen, "New tool for examining QoS in the IPTV service", in *Proc. World Telecommun. Congr. WTC 2014*, Berlin, Germany, 2014.

[13] Zeranoe FFmpeg software (current Windows builds) [Online]. Available: http://FFmpeg.zeranoe.com/builds (accesed Nov. 2015).

[14] S. Wenger, M. Hannuksela, T. Stockhammer, M. Westerlund, and D. Singer, "RTP Payload Format for H.264 Video", RFC 3984, Feb. 2005.

[15] MPEG-2 Transport Stream, Electronics Research Group (ERG) [Online]. Available: http://erg.abdn.ac.uk/research/future-net/digital-video/mpeg2-trans.html (accesed Nov. 2015).

[16] ITU-T Recommendation P.800: MOS-Scale [Online]. Available: http://www.itu.int/T-REC-P.800/en (accesed Nov. 2015).

[17] M. J. Bustamante, "Comparison of algorithms for concealing packet losses in the transmission of compressed video", Master Thesis, University of California, San Diego, CA, USA, 2010.

[18] FFmpeg, Macroblocks and Motion Vectors [Online]. Available: https://trac.FFmpeg.org/wiki/Debug/MacroblocksAndMotionVectors (accesed Nov. 2015).

[19] D. Marshall, "Motion vector search" [Online]. Available: https://www.cs.cf.ac.uk/Dave/Multimedia/node252.html (accesed Nov. 2015).

Christian Hoppe received his B.Eng. in Communications Technology from the Flensburg University of Applied Sciences (Germany) in 2010. Today he is student for Master's degree in Information Technology at Kiel University of Applied Sciences (Germany). His main activities cover the following areas: quality assurance for Triple Play Services and medical imaging solutions.

E-mail: christian.hoppe@fh-kiel.de
Kiel University of Applied Sciences
Grenzstraße 5
D 24149 Kiel, Germany

Robert Manzke received his M.Sc. in Electrical Engineering from Kiel University of Applied Sciences, Germany in 2001. He finished his Ph.D. at King's College London, University of London, UK in 2004 working on image reconstruction algorithms. Subsequently, he gathered multiple years of industrial experience with Philips Research in the field of interventional guidance techniques in medicine with focus on real-time data visualization. He authored about 80 publications, 2 book chapters and over 60 patents. In 2012 he joined the Faculty of Computer Science and Electrical Engineering at Kiel University of Applied Sciences as tenured Professor with focus on ubiquitous and mobile computing and is currently the managing director of the Institute of Applied Computer Science.

E-mail: robert.manzke@fh-kiel.de
Kiel University of Applied Sciences
Grenzstraße 5
D 24149 Kiel, Germany

Marcus Rompf received his B.Eng. in Computer Engineering from the Flensburg University of Applied Sciences (Germany) in 2014. Today he is student for Master's degree in Information Technology at Kiel University of Applied Sciences, (Germany). His main activities cover the following areas: quality assurance for Triple Play Services and medical imaging solutions.

E-mail: marcus.rompf@fh-kiel.de
Kiel University of Applied Sciences
Grenzstraße 5
D 24149 Kiel, Germany

Tadeus Uhl received his M.Sc. in Telecommunications from Academy of Technology and Agriculture in Bydgoszcz in 1975, Ph.D. from Gdańsk University of Technology in 1982 and D.Sc. from University at Dortmund (Germany) in 1990. Since 1992 he works as Professor at the Institute of Communications Technology, Flensburg University of Applied Sciences (Germany) and additionally since 2013 as Professor at the Institute of Transport Engineering, Maritime University of Szczecin, Poland. His main activities cover the following areas: traffic engineering, performance analysis of communications systems, measurement and evaluation of communication protocols, QoS and QoE by Triple Play Services, Ethernet and IP technology. He is author or co-author of three books and about 130 papers on the subjects LAN, WAN and NGN.

E-mail: t.uhl@am.szczecin.pl
Maritime University of Szczecin
Henryka Pobożnego st 11
PL 70-507 Szczecin, Poland

Paper

Properties of the Multiservice Erlang's Ideal Gradings

Sławomir Hanczewski and Damian Kmiecik

Faculty of Electronics and Telecommunications, Poznan University of Technology, Poznan, Poland

Abstract—The design and optimization process of modern telecommunications networks is supported by a range of appropriate analytical models. A number of these models are based on the Erlang's Ideal Grading (EIG) model, which is a particular case of non-full-availability groups. A possibility of the application of the EIG model results from the fact that telecommunications systems show properties and features distinctive to non-full-availability systems. No detailed studies that would decisively help determine appropriate conditions for the application of the EIG model for modeling of other non-full-availability groups, that would be models corresponding to real telecommunications systems, have been performed. Therefore, this article attempts to find an answer to the following question: what are the prerequisite conditions for the application of the EIG model and when the model can be reliably used?

Keywords—Erlang's Ideal Grading, multiservice systems, traffic engineering.

1. Introduction

For the past number of years, we have been witnessing an exponential growth in the development of wired and wireless telecommunications networks [1], [2]. The constantly decreasing access services prices have made the number of network users (devices that make use of data transmission) growing rapidly. This, in turn, have effected in the increase in the amount of data sent over networks, particularly in wireless networks. Transmitting such a data poses an enormous challenge to telecommunications and computer networks and telcos. In order to use network resources in the best possible way operators are forced to implement advanced traffic management mechanisms, such as reservation [3]–[5], compression [6]–[12], priorities [13]–[15] or traffic overflow [16]–[19]. Those mechanisms influence advantageously the parameters of sent data streams and, in this way, make all resources of a network available in optimal way. The resource optimization process and network design are supported by and benefited from analytical modeling that allows characteristics of telecommunications systems to be determined on the basis of appropriate mathematical dependencies. The bulk of the models of telecommunications systems addressed in the literature of the subject uses either multiservice models of the full-availability group [20], [21] or limited-availability group [22]. An alternative solution for these groups of models, however, are models that make use of non-full-availability group models, i.e. Erlang's Ideal Grading (EIG). EIG is a particular (ideal) case of a non-full-availability group, since in this group a uniform load of group resources is assumed (which results from an appropriate number of load groups), despite the fact that individual traffic sources have no access to all resources of the group, but only to a part of it. The adoption of such assumption has made it possible to develop a simple analytical model of this group [23]. A. K. Erlang, who developed the structure of the EIG group and its analytical model for single-service traffic, noticed that this particular model could also be applied to approximate other non-full-availability systems (those with non-uniform loads). It is worthwhile to add that the EIG group model has been successfully used for modeling single-service switching networks [24], [25]. Regrettably, the developments in technology and the subsequent appearance of multiservice systems caused the EIG model to be abandoned and left out in the early 1980s, since a multiservice model of EIG was nonexistent at the time. This unfavorable situation for the non-full-availability group was changed, however, when the model presented in [26], and derivations thereof, were proposed for multiservice traffic with differentiated availabilities, including non-integer availabilities [5], [27].

Present-day telecommunications systems can be viewed as non-full-availability systems. This assumption is confirmed by models of systems that use the EIG group model described in [18], [28], [29]. However, no available publications provide key information that would, in an unambiguous way, determine the range of versatile application possibilities for EIG in modeling other non-full-availability systems. The present article is an attempt to provide an answer to these questions.

The remaining part of the article has been structured as follows. Section 2 presents the issues related to non-full-availability systems and an analytical model of the EIG group. Exemplary results are provided in Section 3, whereas Section 4 sums up the article.

2. Non-full-availability Systems

Non-full-availability systems are characterized by the fact that individual traffic sources do not have access to all resources of the system (expressed in BBU[1]), but only to a part of them. A good example of these systems are

[1]The BBU is defined as the greatest common divisor of equivalent bandwidths of all call streams offered to the system [30].

switching networks in which, due to the connecting paths set up in a given state of the network, a connecting path between a given input and output is not possible [24], [26].

Another example of modern telecommunications systems that can be treated as non-full-availability systems is the radio interface in a 3G mobile network. In this particular case, this non-full-availability stems from limitations in available resources of the interface imposed by noise and signal characteristics, e.g. interference from neighboring cells [31]. Yet another examples are the traffic overflow system in which non-full-availability results from limited access to resources to which connections are transferred (overflow) [29] and the VoD system [28].

In traffic engineering non-full-availability systems are modeled by non-full-availability groups. Each group of this type is described by three parameters: capacity V, availability d and the number of load groups g. Availability d defines the amount of resources of the group to which a traffic source has access. Traffic sources that have access to the same BBUs in the system create the so-called load group (component group). Conventionally, non-full-availability groups are divided into: graded and uniform (homogenous) groups [32]. In graded groups, with an increase in the number of BBUs, the number of load groups that have access to this BBU increases (or remains unchanged). In uniform groups, each BBU is always available to the same number of load groups. Figure 1 shows both examples. A particular case of uniform groups is the Erlang's Ideal Grading – ideally symmetrical non-full-availability groups. The latter group assumes all resources of the group to be uniformly loaded, while the number of load groups is equal to the number of possible choices d of resources, from among V:

$$g = \binom{V}{d}. \tag{1}$$

In Fig. 2a the example of EIG with single service traffic is presented. The capacity (V) of this grading is equal to 3 BBUs. The availability is equal to 2 BBUs. Figure 2b presents the idea of availability.

2.1. Model of Erlang's Ideal Grading with Various Availabilities

Let us consider Erlang's Ideal Grading [33] with various availabilities that is offered m independent Poisson call streams with the intensities $\lambda_1, ..., \lambda_i, ..., \lambda_m$. The service time of calls of particular classes has an exponential distribution with the parameters $\mu_1, ..., \mu_i, ..., \mu_m$. Therefore, traffic offered A_i by individual call streams can be determined on the basis following formula:

$$A_i = \frac{\lambda_i}{\mu_i}. \tag{2}$$

The calls offered to grading are characterized by different values of demanded BBUs to set up a connection $t_1, ..., t_i, ..., t_m$ and different availability $d_1, ..., d_i, ..., d_m$. This means that each class of calls is related to a different number of load groups:

$$g_i = \binom{V}{d_i}. \tag{3}$$

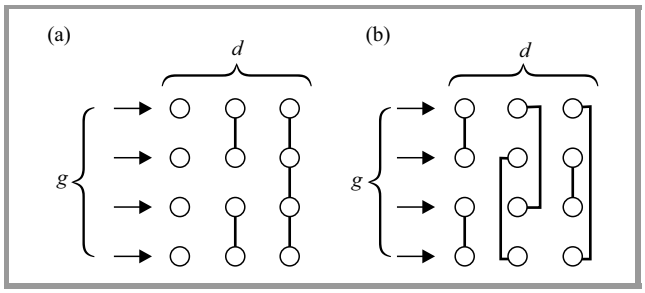


Fig. 1. Non-full-availability group for $V = 7$ BBUs, $g = 4$ and $d = 3$ BBUs: (a) graded group, (b) uniform group.

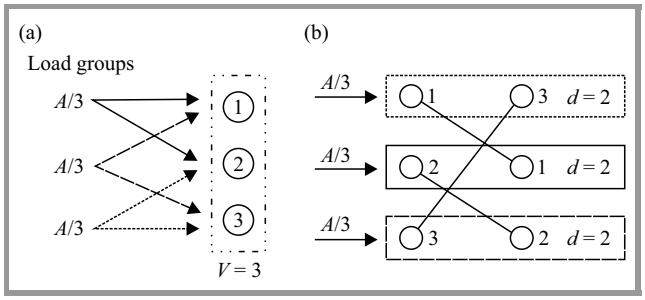


Fig. 2. Erlang's Ideal Grading with single-service traffic for $V = 3$ BBUs, $g = 3$ and $d = 2$ BBUs: (a) offered traffic distribution, (b) idea of availability.

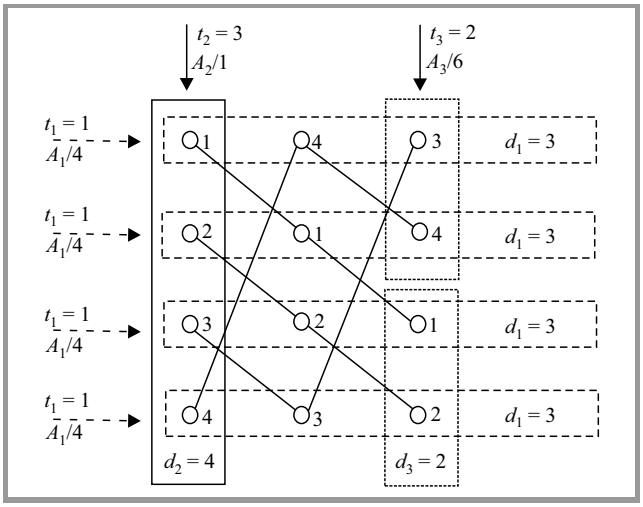


Fig. 3. Erlang's Ideal Grading: $V = 4$, m = 3, t_1=1, $d_1 = 3$, $g_1 = 4$, t_2=3, $d_2 = 4$, $g_2 = 1$, t_3=2, $d_3 = 2$, $g_3 = 6$.

Figure 3 presents an example of such grading. This EIG is composed of 4 BBUs ($V = 4$). The grading services $m = 3$ class of calls: $t_1 = 1$, $d_1 = 3$, $t_2 = 3$, $d_2 = 4$, $t_3 = 2$, $d_3 = 2$. The number of load groups for particular class of call is equal: $g_1 = 4$, $g_2 = 1$, $g_3 = 6$.

According to the model [27], [33], the occupancy distribution $P(n)$ is expressed by the formula:

$$nP(n) = \sum_{i=1}^{m} A_i t_i \left[1 - \sigma_i(n - t_i)\right] P(n - t_i)\,, \tag{4}$$

where A_i is traffic offered to the group by a call of class i – Eq. (2) – and $\sigma_i(n)$ is the conditional probability of transition for a traffic stream of class i in occupancy state n in the group

$$\sigma_i(n) = \frac{1 - \sum_{d-t_i+1}^{k} \binom{d_i}{x}\binom{V-d_i}{n-x}}{\binom{V}{n}}, \tag{5}$$

where:

- $k = n - t_i$, if $(d_i - t_i + 1) \leq (n - t_i) < d_i$,
- $k = d_i$, if $(n - t_i) \geq d_i$.

It should be stressed that the conditional probability of transition ($\sigma_i(n)$) is combinatorial function of availability and it is independent of offered traffic.
The blocking probability for calls of class i can be determined on the basis of the following formula:

$$E_i = \sum_{n=d_i-t_i+1}^{V} \left[1 - \sigma_i(n)\right] P(n)\,. \tag{6}$$

2.2. Non-integer Availability

Presented model in Subsection 2.1 enables authors to determine the values of blocking probabilities in EIG only for integer values of availability parameter. In [27] the model for non-integer value of this parameter was proposed. According to this model the blocking probability is calculated as follows.
Let us assume that for class i the availability parameter takes on non-integer values. This class of calls is replaced by two fictitious classes with integer values of availability (d_{i1}, d_{i2}) and offered traffic (A_{i1}, A_{i2}). Values of these parameters are defined in the following way:

$$d_{i1} = \lfloor d_i \rfloor\,, \tag{7}$$

$$d_{i2} = \lceil d_i \rceil\,. \tag{8}$$

The traffic offered by the new fictitious call classes is respectively equal to:

$$A_{i1} = A_i[1 - (d_i - d_{i1})] = A_i(d_{i2} - d_i), \tag{9}$$

$$A_{i2} = A_i(d_i - d_{i1}), \tag{10}$$

where the difference $d_i - d_{i1}$ determines the fractional part of the parameter d_i. Such a definition of the parameters A_{i1}, A_{i2}, d_{i1}, d_{i2} means that the values of the fictitious traffic A_{i2} is directly proportional to the fractional part of the availability parameter, i.e. to $\Delta_i = d_i - d_{i1}$, while the value of the fictitious traffic A_{i1} is directly proportional to the complement Δ_i , i.e. to the value $1 - \Delta_i = 1 - (d_i - d_{i1}) = d_{i2} - d_i$ [27].
After replacing class with two fictitious classes: $i1$, and $i2$, with assigned values of availability and traffic intensity, it is possible to determine, on the basis of Eqs. (4)–(6), the blocking probabilities of all classes of calls, including the blocking probability of new classes of calls. The blocking probability of calls of class i for non-integer availability d_i can be determined in the following way:

$$E_i = \frac{A_{i1}E_{i1} + A_{i2}E_{i2}}{A_i}\,. \tag{11}$$

In the case of a higher number of classes with non-integer availabilities, each class of calls is replaced by two fictitious classes with the parameters determined by Eqs. (7)–(10). The maximum number of fictitious classes is equal to $2m$.

3. The Results

In order to properly define the scope of the applicability of the EIG model for modeling of non-full-availability groups with a different number of load groups and different load in a single BBU as well as imprecisely estimated availability values, appropriate simulation experiments were carried out. For this purpose, a dedicated simulator was devised and successfully implemented. The simulator makes it possible to perform simulations for EIG groups, non-full-availability groups as well as other telecommunications systems. The simulator was implemented in the C++ language according to the event scheduling method [34].
The input data for the simulator were the parameters that described the system, i.e. its structure, capacity and the parameters that describe the call stream offered to the system (the number of classes m, demands of individual classes t_i and availability d_i). Additionally, it is also possible to determine the parameters of the simulation experiment itself, such as the total number of simulation series and the length of a single simulation series (expressed in the number of defined events). Results obtained in this way make a determination of 95% confidence intervals possible.

3.1. Erlang's Ideal Grading vs. Full Availability Group

A full-availability group with multiservice traffic is the most frequently used model of telecommunications systems. The occupancy distribution in this group can be determined on the basis of the recurrent dependence known as the Kaufman-Roberts [20], [21] formula:

$$nP(n) = \sum_{i=1}^{m} A_i t_i P(n - t_i)\,. \tag{12}$$

It should be noticed that this group is in fact a particular case of the EIG group, the fact that seems to be notoriously overlooked by researchers studying telecommunications traffic engineering. Observe that in the case where

availability of all classes' is equal to the capacity of a considered system, i.e. $d_i = V$, $(1 \leq i \leq m)$, Equation (4) will be simplified to Eq. (12) (parameter $\sigma_i(n) = 1$). The multiservice EIG model, because of its general nature, is thus even a more universal and versatile tool supporting any analysis of modern telecommunication systems.

3.2. The In uence of the Evaluation on the Results

In order to use the EIG model to model present-day telecommunications systems it is necessary to determine availability values for all serviced traffic classes. Availability parameters are generally defined by the structure of a modeled system and offered traffic. In most cases, this availability can be determined on the basis of a relatively simple mathematical dependence [22], [35], [29], because the accuracy of obtained results directly derives from and depends on the precision of the evaluation of the value of individual availability parameters. To illustrate this problem, an experiment for an EIG group with the capacity of 30 BBUs servicing $m = 3$ classes of calls that demanded respectively $t_1 = 1$, $t_1 = 3$, $t_3 = 5$ BBUs was carried out. The assumption was that the accurate availability values for the system were equal to: $d_1 = 10$, $d_2 = 15$, $d_3 = 20$ BBUs.

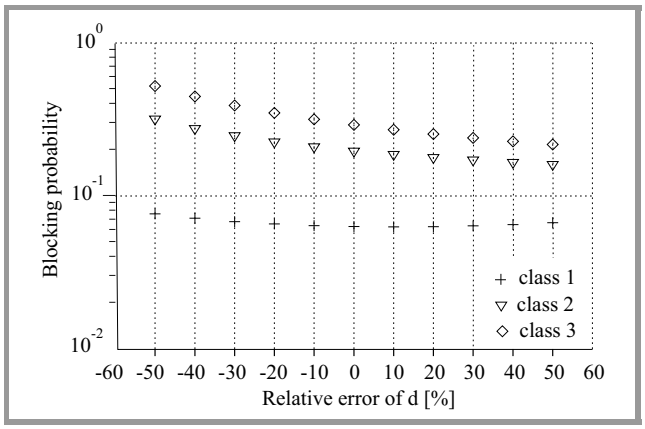


Fig. 4. Blocking probability as a function of relative error of availability.

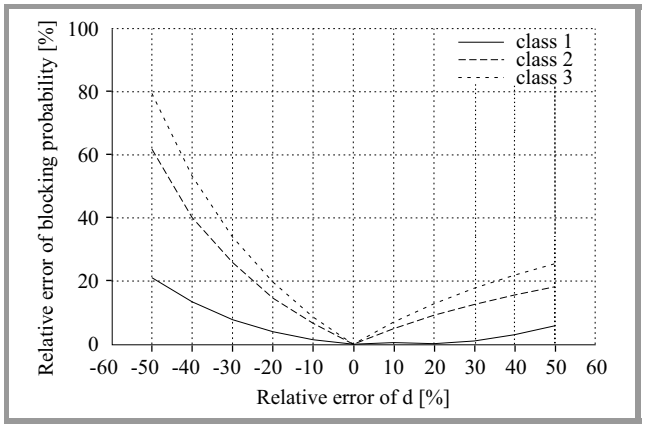


Fig. 5. Relative error of blocking probability as a function of relative error of availability.

Figure 4 shows the blocking probability as a function of relative error of availability. If the availability parameter is underestimated (the determined values are lower than the precise values), the blocking probability is higher than in the reference EIG group. If, on the other hand, the values of availability parameters are overestimated, the values of probability are lower than in the reference EIG group (the relationship is least evident in the class demanding the lowest number of BBUs to be served). This occurs regardless of the offered traffic value. In turn, Fig. 5 shows the relative error determined on the basis of the EIG model with the assumption that the availabilities of all classes were not accurately estimated. The identical nature of underestimation was adopted for all classes. In the second case (Figs. 6 and 7) presented here, erroneous estimation of the

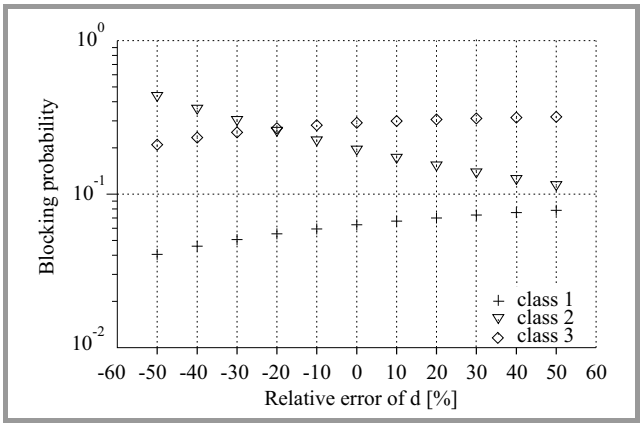


Fig. 6. Blocking probability as a function of relative error of availability.

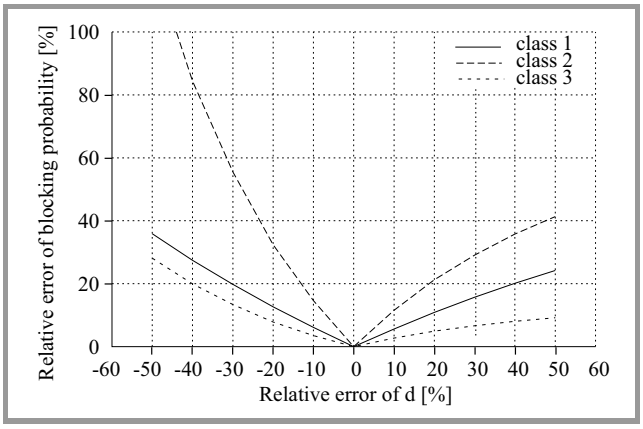


Fig. 7. Relative error of blocking probability as a function of relative error of availability.

value of the availability parameter was to be found for only one class (class 2). As it is easy to observe, an erroneous estimation of availability parameters has a detrimental and negative influence on the correctness of results to be obtained. The results of blocking probability are better when the values of availability parameters are overestimated. For the group under investigation, acceptable results are obtained when it does not exceed about 20%. Presented re-

sult were calculated for offered traffic by one BBU equal to 0.8 Erl and offered traffic by all serviced classes is in relation $A_1t_1 : A_2t_2 : A_3t_3 = 1 : 1 : 1$.

3.3. Other No-full-availability Groups

When considering real systems as non-full-availability systems, the fact that the number of load groups in such a system is lower than the number of groups in the EIG group should be taken into consideration. The next step then is to examine what influence the structure of the approximated system has upon the accuracy of obtained results. Figures 8–11 show the results for a non-full-availability

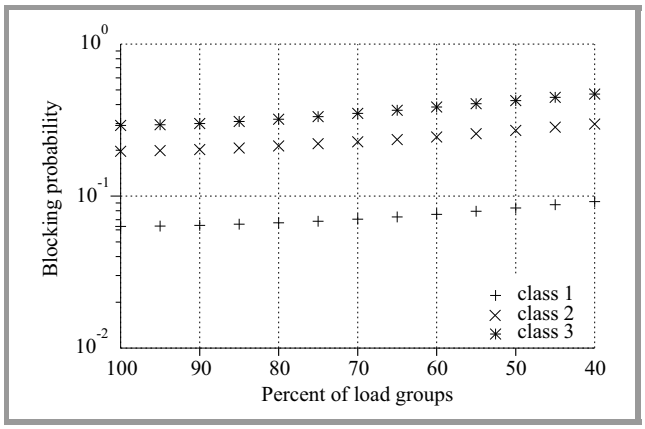


Fig. 8. Blocking probability as a function of number of load groups.

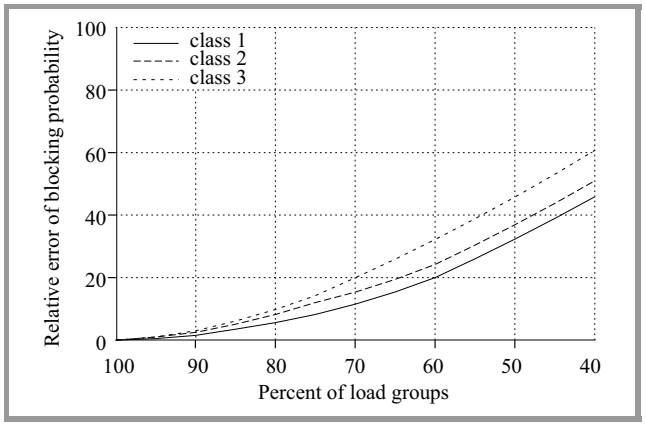


Fig. 9. Relative error of blocking probability as a function of number of load groups.

group with the capacity $V = 20$ BBUs that services two classes of calls demanding $t_1 = 1$ and $t_2 = 3$ BBUs, respectively. The availability is equal to $d_1 = 10$ BBUs and $d_2 = 15$ BBUs. The assumption is that presented real system has a structure of a homogenous group (Fig. 1a). The adoption of this assumption introduces the possibility that, despite a decreasing number of load groups, the load in each BBU is uniform. Hence, even when this decrease in the number of load groups is significantly high (in the considered case, acceptable results are still obtained

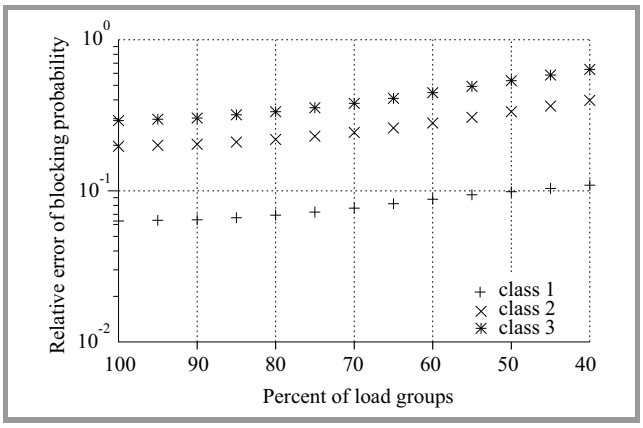


Fig. 10. Blocking probability as a function of relative error of availability.

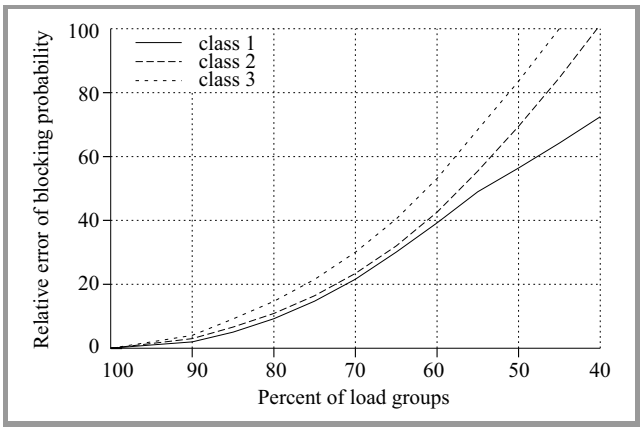


Fig. 11. Relative error of blocking probability as a function of relative error of availability.

with the number of groups lower by even 40%), Figs. 8 and 9, the load in individual groups remains equal. A different situation, however, is to be observed with the case of a system that has a structure of a grading group (Fig. 1b). In this case, even a 30% change in the number of load groups results in a significant impact on the obtained results (Figs. 10 and 11). This phenomenon results from the occurrence of the uneven load of BBU in a group.

4. Summary

This article presents the results of an investigation into a broad range of potential applications of the EIG group model for modeling of telecommunications systems. Even though only a small excerpt of the case study is presented here, the results are robust enough to make a conclusion that the EIG group and its model are indeed ideal tools for modeling telecommunications systems, provided a proper evaluation (with a certain degree of accuracy) of the value of availability parameters can be executed. It has to be stressed that the number of load groups in a system has a lower influence on obtained results than an error in the estimation of availability parameters. As yet the authors

have managed to find simple dependencies between the structure of a real system and the availability that characterizes particular classes of calls in the system. The only exception is the system with reservation. For this particular case, however, an algorithm has been developed that makes a precise evaluation of values of these parameters possible [27].

Acknowledgements

The presented work has been funded by the Polish Ministry of Science and Higher Education within the status activity task "Structure, analysis and design of modern switching system and communication networks" in 2016.

References

[1] "Ericsson Mobility Report", Tech. Rep., Ericsson, 2014.

[2] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2014-2019 – white paper", Tech. Rep., Cisco, 2015.

[3] J. Roberts, "Teletraffic models for the Telcom 1 integrated services network", in *Proc. 10th Int. Teletraff. Congr. ITC 83*, Montreal, Canada, 1983, p. 1.1.2.

[4] P. Tran-Gia and F. Hübner, "An analysis of trunk reservation and grade of service balancing mechanisms in multiservice broadband networks", in *Modelling and Evaluation of ATM Networks*, H. G. Perros, G. Pujolle, and Y. Takahashi, Eds., *IFIP Trans*., vol. C-15, pp. 83–97. Amsterdam: North-Holland, 1993.

[5] M. Stasiak and S. Hanczewski, "Approximation for multi-service systems with reservation by systems with limited-availability", in *5th European Performance Engineering Workshop*, N. Thomas and C. Juiz, Eds., *LNCS*, vol. 5261, pp. 257–267. Palma de Mallorca, Spain: Springer, 2008.

[6] M. Głąbowski, A. Kaliszan, and M. Stasiak, "Modeling product-form state-dependent systems with BPP traffic", *J. Perform. Eval.*, vol. 67, no. 3, pp. 174–197, 2010.

[7] I. D. Moscholios, J. S. Vardakas, M. D. Logothetis, and A. C. Boucouvalas, "Congestion probabilities in a batched poisson multirate loss model supporting elastic and adaptive traffic", *Annales des Télécommun.*, vol. 68, no. 5-6, pp. 327–344, 2013.

[8] J. Kaufman, "Blocking with retrials in a completly shared recource environment", *J. Perform. Eval.*, vol. 15, no. 2, pp. 99–113, 1992.

[9] S. Rácz, B. P. Gerö, and G. Fodor, "Flow level performance analysis of a multi-service system supporting elastic and adaptive services", *Perform. Eval.*, vol. 49, no. 1-4, pp. 451–469, 2002.

[10] M. Sobieraj, M. Stasiak, J. Weissenberg, and P. Zwierzykowski, "Analytical model of the single threshold mechanism with hysteresis for multi-service networks", *IEICE Trans. Commun.*, vol. E95-B, no. 1, pp. 120–132, 2012.

[11] I. Moscholios, M. Logothetis, and G. Kokkinakis, "Connection-dependent threshold model: a generalization of the Erlang multiple rate loss model", *Perform. Eval.*, vol. 48, no. 1-4, pp. 177–200, 2002.

[12] M. Stasiak, M. Głąbowski, A. Wiśniewski, and P. Zwierzykowski, *Modeling and Dimensioning of Mobile Networks*. Wiley, 2011.

[13] M. Stasiak, P. Zwierzykowski, J. Wiewióra, and D. Parniewicz, "An approximate model of the WCDMA interface servicing a mixture of multi-rate traffic streams with priorities", in *Computer Performance Engineering*, D. Parniewicz, M. Stasiak, J. Wiewióra, and P. Zwierzykowski, Eds. *LNCS*, vol. 5261, pp. 168–180. Springer, 2008 (doi: 10.1007/978-3-540-87412).

[14] L. Katzschner and R. Scheller, "Probability of loss of data traffics with different bit rates hunting one common PCM-channel", in *Proc. 8th Int. Teletraff. Congr.*, Melbourne, Australia, 1976, pp. 525/1–8.

[15] K. Subramaniam and A. A. Nilsson, "An analytical model for adaptive call admission control scheme in a heterogeneous UMTS-WCDMA system", in *Proc. IEEE Int. Conf. Commun. ICC 2005*, Seoul, South Korea, 2005, vol. 5, pp. 3334–3338.

[16] A. Fredericks, "Congestion in blocking systems – a simple approximation technique", *Bell System Tech. J.*, vol. 59, no. 6, pp. 805–827, 1980.

[17] M. Głąbowski, K. Kubasik, and M. Stasiak, "Modeling of systems with overflow multi-rate traffic", *Telecommun. Syst.*, vol. 37, no. 1–3, pp. 85–96, 2008.

[18] M. Głąbowski, S. Hanczewski, and M. Stasiak, "Erlang's ideal grading in diffserv modelling", in *Proc. IEEE Africon 2011*, Livingstone, Zambia, 2011, pp. 1–6.

[19] M. Głąbowski, A. Kaliszan, and M. Stasiak, "Two-dimensional convolution algorithm for modelling multiservice networks with overflow traffic", *Mathem. Problems in Engin.*, vol. 2013, p. 18, 2013 (article ID 852082).

[20] J. Kaufman, "Blocking in a shared resource environment", *IEEE Trans. Commun.*, vol. 29, no. 10, pp. 1474–1481, 1981.

[21] J. Roberts, "A service system with heterogeneous user requirements – application to multi-service telecommunications systems", in *Performance of Data Communications Systems and their Applications*, G. Pujolle, Ed. Amsterdam: North Holland, 1981, pp. 423–431.

[22] M. Stasiak, "Blocking probability in a limited-availability group carrying mixture of different multichannel traffic streams", *Annales des Télécommun.*, vol. 48, no. 1–2, pp. 71–76, 1993.

[23] E. Brockmeyer, H. Halstrom, and A. Jensen, "The life and works of A. K. Erlang", *Acta Polytechnika Scandinavia*, vol. 6, no. 287, 1960.

[24] V. Eršov, "Rasčët komutacionnych sistem metodom rasdel'nyh poter'", in *Sistemy massovogo obsluživaniâ i kommutacii*, Moscow: Nauka, 1974, pp. 54–66 (in Russian).

[25] V. Eršov, "Srednââ dostupnost' i rekurrentnyj rasčët poter' v kommutacionnych sistemah", in *Sistemy upravleniâ setâmi*, Moscow: Nauka, 1980, pp. 121–126 (in Russian).

[26] M. Stasiak, "An approximate model of a switching network carrying mixture of different multichannel traffic streams", *IEEE Trans. Commun*., vol. 41, no. 6, pp. 836–840, 1993.

[27] M. Stasiak and S. Hanczewski, "A model of WCDMA radio interface with reservation", in *Proc. Int. Symp. Inform. Theory and Its Appl. ISITA 2008*, Auckland, New Zeland, 2008.

[28] S. Hanczewski and M. Stasiak, "Performance modelling of video-on-demand systems", in *Proc. 17th Asia-Pacific Conf. Commun. APCC 2011*, Kuala Lumpur, Malaysia, 2011, pp. 784 –788.

[29] M. Głąbowski, S. Hanczewski, and M. Stasiak, "Modelling of cellular networks with traffic overflow", *Mathem. Problems in Engin.*, vol. 2015, p. 158, 2015 (article ID 286490).

[30] *Broadband Network Teletraffic, Final Report of Action COST 242*, J. Roberts, V. Mocci, and I. Virtamo, Eds. Berlin: Springer, 1996.

[31] M. Stasiak, M. Głąbowski, and S. Hanczewski, "The application of the Erlang's Ideal Grading for modelling of UMTS cells", in *8th Int. Symp. Commun. Syst., Netw. Digit. Sig. Process. CSNDSP 2012*, Poznan, Poland, 2012, pp. 1–6.

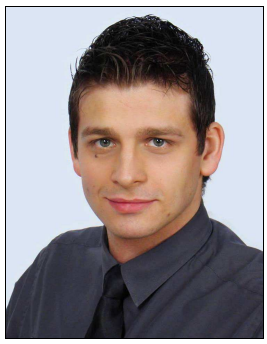
[32] A. Jajszczyk, *Wstęp do Telekomutacji*. Wydawnictwa Naukowo-Techniczne, 1998 (in Polish).

[33] M. Głąbowski, S. Hanczewski, M. Stasiak, and J. Weissenberg, "Modeling Erlang's Ideal Grading with multirate BPP traffic", *Mathem. Problems in Engin.*, vol. 2012, p. 35, 2012 (article ID 456910).

[34] J. Tyszer, *Object-Oriented Computer Simulation of Discrete-Event Systems*. Kluwer, 1999.

[35] S. Hanczewski and M. Stasiak, "Point-to-group blocking in 3-stage switching networks with multicast traffic streams", in *Proceedings of First International Workshop (SAPIR 2004)*, P. Dini, P. Lorenz, and J. N. de Souza, Eds. LNCS, vol. 3126, pp. 219–230. Springer, 2004.

Sławomir Hanczewski received M.Sc. and Ph.D. degrees in Telecommunications from Poznan University of Technology, Poland, in 2001 and 2006, respectively. Since 2007 he has been working in the Faculty of Electronics and Telecommunications, Poznan University of Technology. He is an Assistant Professor in the Chair of Communications and Computer Networks. He is the author, and co-author, of more than 50 scientific papers. He is engaged in research in the area of performance analysis and modeling of queuing systems, multiservice networks and switching systems.
E-mail: slawomir.hanczewski@et.put.poznan.pl
Faculty of Electronics and Telecommunication
Poznan University of Technology
Polanka st 3
60-965 Poznan, Poland



Damian Kmiecik received his M.Sc. degree in Telecommunications from Poznan University of Technology, Poland, in 2014. Since 2015 he is a Ph.D. student at the Chair of Communications and Computer Networks at Poznan University of Technology. Damian Kmiecik is engaged in research and teaching in the area of performance analysis and modeling of queuing systems.
E-mail: damian.kmiecik@et.put.poznan.pl
Faculty of Electronics and Telecommunication
Poznan University of Technology
Polanka st 3
60-965 Poznan, Poland

Paper

Call Blocking Probabilities of Multirate Elastic and Adaptive Traffic under the Threshold and Bandwidth Reservation Policies

Ioannis D. Moscholios[1], Michael D. Logothetis[2], Anthony C. Boucouvalas[1], and Vassilios G. Vassilakis[3]

[1] *Dept. of Informatics and Telecommunications, University of Peloponnese, Tripolis, Greece*
[2] *WCL, Dept. of Electrical and Computer Engineering, University of Patras, Patras, Greece*
[3] *Computer Laboratory, University of Cambridge, Cambridge, United Kingdom*

Abstract—This paper proposes multirate teletraffic loss models of a link that accommodates different service-classes of elastic and adaptive calls. Calls follow a Poisson process, can tolerate bandwidth compression and have an exponentially distributed service time. When bandwidth compression occurs, the service time of new and in-service elastic calls increases. Adaptive calls do not alter their service time. All calls compete for the available link bandwidth under the combination of the Threshold (TH) and the Bandwidth Reservation (BR) policies. The TH policy can provide different QoS among service-classes by limiting the number of calls of a service-class up to a predefined threshold, which can be different for each service-class. The BR policy reserves part of the available link bandwidth to benefit calls of high bandwidth requirements. The analysis of the proposed models is based on approximate but recursive formulas, whereby authors determine call blocking probabilities and link utilization. The accuracy of the proposed formulas is verified through simulation and found to be very satisfactory.

Keywords—adaptive traffic policy, Call Blocking Probabilities, Multirate Loss Model, threshold and bandwidth reservation policies.

1. Introduction

Multirate elastic traffic refers to in-service calls of different service-classes which have the ability to compress/expand their bandwidth and simultaneously increase/decrease their service time, during their lifetime in a system. A variation of elastic traffic is the so-called adaptive traffic. The service time of adaptive in-service calls is not affected by their bandwidth compression/expansion. Assuming that the system behaves as a loss system (i.e. calls are not allowed to wait in order to be serviced) and that the call arrival process is Poisson then the calculation of various performance measures such as Call Blocking Probabilities (CBP), and system's utilization can be based on the classical Erlang Multirate Loss Model (EMLM) [1], [2]. The latter has led to numerous loss models proposed for the call-level analysis of wired (e.g. [3]–[19]), wireless (e.g. [20]–[32]) and optical networks (e.g. [33]–[37]).

In the EMLM, a link accommodates calls of different service-classes. New calls compete for the available link bandwidth according to the Complete Sharing (CS) policy (i.e., calls compete for all bandwidth resources) and have fixed bandwidth requirements and generally distributed service time [1]. The term "fixed" means that in-service calls do not compress their bandwidth during their lifetime in the system. A new call is blocked and lost if its required bandwidth is not available. The steady state probabilities in the EMLM have a Product Form Solution (PFS), which leads to an accurate CBP calculation [1], [2]. In [5], the EMLM is extended to include the case of elastic traffic. The authors name this model Elastic EMLM (E-EMLM). In the E-EMLM, instead of rejecting immediately a blocked call, the link may accept this call by compressing its bandwidth and the bandwidth of all in-service calls. Bandwidth compression is permitted down to a minimum bandwidth, which can be different for each service-class. Elastic calls increase their service time so that the product *bandwidth by service time* remains constant. When a call with compressed bandwidth leaves the system, then the remaining calls expand their bandwidth in proportion to their initial bandwidth requirement. Call blocking occurs when the value of the minimum bandwidth requirement is still higher than the available bandwidth. The model of [5] has been extended in [9] in order to include the case of adaptive traffic. The authors name the model of [9], Elastic-Adaptive EMLM (EA-EMLM).

In this paper, authors initially consider a link that accommodates Poisson arriving calls of elastic service-classes and modify the admission mechanism to include the Threshold (TH) and the Bandwidth Reservation (BR) policies. The proposed model is named E-EMLM/TH-BR. In addition, authors propose the EA-EMLM/TH-BR whereby a link accommodates elastic and adaptive service-classes. In the TH policy, the number n_k of in-service calls k of service-class should not exceed a pre-defined threshold, after the acceptance of a new service-class k call. Otherwise, the

call is blocked and lost. The TH policy is significant since it analyzes a multirate access tree network which accommodates different service-classes [38] and may differentiate service-classes in terms of CBP or revenue rates by a proper threshold selection (see e.g. [39], [40]). The BR policy is used to reserve bandwidth to benefit calls of high bandwidth requirements and is mainly used when CBP equalization is required among calls of different service-classes. The fact that the BR policy has been extensively applied in the literature (e.g. [6], [8], [18], [28], [42]–[47]) evinces its importance in call admission control.

To model the proposed E-EMLM/TH-BR and EA-EMLM/TH-BR, the Markov chain method is used. However, due to the existence of the compression/expansion mechanism and the BR policy, the reversibility of the Markov chains is destroyed, and the steady state probabilities in the proposed models cannot be determined via a PFS. Therefore, authors resort to approximate Markov chains which provide recursive formulas for the efficient determination of the link occupancy distribution and, consequently, CBP and link utilization. The accuracy of the proposed formulas is verified through simulation and found to be very satisfactory. On the other hand, the comparison of the proposed models with existing models shows the necessity of the new models, as well as their consistency over changes of their parameters (e.g. compression factor and offered traffic-load). The term "consistency" is referred to the anticipated behavior of the proposed models over changes, such as the increase of offered traffic-load or the increase of the compression factor.

This paper is organized as follows. In Section 2, the E-EMLM/TH is presented and formulas for the calculation of the various performance measures are proved. In Section 3, the E-EMLM/TH is extended to include the BR policy. In Section 4, the E-EMLM/TH-BR is extended to include the case of adaptive traffic. In Section 5, authors provide numerical results whereby the E-EMLM/TH and the E-EMLM/TH-BR are compared to existing models and evaluated via simulation. The paper is concluded in Section 6.

2. The Elastic EMLM/TH (E-EMLM/TH)

2.1. The System Model

Consider a link of capacity C bandwidth units (b.u.) that accommodates K elastic service-classes. Calls of service-class k ($k = 1, \ldots, K$) follow a Poisson process with arrival rate λ_k and request b_k b.u. Bandwidth compression is introduced in the system by allowing the occupied link bandwidth to virtually exceed C up to T b.u., i.e. $j = 0, 1, \ldots, T$. Let $\mathbf{n} = (n_1, \ldots, n_K)$ be the vector of all in-service calls and $\mathbf{b} = (b_1, \ldots, b_K)$ the vector of peak-bandwidth requirements, then $j = \mathbf{nb}$.

The decision to accept a new service-class k call in the system is based on the following constraints:

(a) the number of in-service calls of service-class k, n_k, together with the new call, should not exceed a maximum threshold n_k^*, i.e. $n_k + 1 \le n_k^*$. Otherwise the call is blocked. This constraint expresses the TH policy;

(b) if constraint (a) is met then:

(b1) if $j + b_k \le C$, the call is accepted in the system with b_k b.u. and remains in the system for an exponentially distributed service time with mean μ_k^{-1};

(b2) if $T \ge j + b_k > C$ the call is accepted by compressing its b_k together with the bandwidth of all in-service calls of all service-classes.

The compressed bandwidth of the new service-class k call is:

$$b_k' = r b_k = \frac{C b_k}{j + b_k}, \tag{1}$$

where $r \equiv r(\mathbf{n}) = \frac{C}{\mathbf{nb} + b_k} = \frac{C}{j + b_k}$.

The product *service time by bandwidth per call* is kept constant by changing the mean value of the service time of the new service-class k call to $\frac{1}{\mu_k'} = \frac{j + b_k}{C \mu_k}$. The compressed bandwidth of all in-service calls becomes $\frac{C b_i}{j + b_k}$ for $i = 1, \ldots, K$. When all calls have compressed their bandwidth, then $j = C$. Note that the minimum bandwidth that a call tolerates is:

$$b_{k,\min}' = r_{\min} b_k = \frac{C b_k}{T}, \tag{2}$$

where $r_{\min} = \frac{C}{T}$ is the min. proportion of the required peak-bandwidth and is common for all service-classes.

A new service-class k call is blocked if $j + b_k > T$.

When an in-service call, with compressed bandwidth b_i' departs from the system then the remaining calls expand their bandwidth to b_i^* in proportion to their b_i, as follows:

$$b_i'' = \min \left(b_i,\ b_i' + \frac{b_i b_k'}{\sum_{k=1}^{K} n_k b_k} \right). \tag{3}$$

2.2. The Analytical Model

The existence of the bandwidth compression mechanism destroys reversibility in the E-EMLM/TH and therefore the steady state probabilities have no PFS. To circumvent this problem, the state-dependent factors $\phi_k(\mathbf{n})$ are used, which lead to a reversible Markov chain:

$$\phi_k(\mathbf{n}) = \begin{cases} 1, & \text{when } \mathbf{nb} \le C \text{ and } \mathbf{n} \in \boldsymbol{\Omega} \\ \dfrac{x(\mathbf{n}_k^-)}{x(\mathbf{n})}, & \text{when } C < \mathbf{nb} \le T \text{ and } \mathbf{n} \in \boldsymbol{\Omega} \end{cases}, \tag{4}$$

where:

$$\boldsymbol{\Omega} = \{\mathbf{n} : 0 \le \mathbf{nb} \le T,\ n_k \le n_k^*,\ k = 1, \ldots, K\},$$

$$\mathbf{n} = (n_1, \ldots, n_k, \ldots, n_K),$$
$$\mathbf{n}_k^- = (n_1, \ldots, n_k - 1, \ldots, n_K)$$

and

$$x(\mathbf{n}) = \frac{1}{C}\sum_{k=1}^{K} n_k b_k x(\mathbf{n}_k^-), \text{ when } C < \mathbf{nb} \le T, \ \mathbf{n} \in \mathbf{\Omega}. \quad (5)$$

To prove a recursive formula for the link occupancy distribution, $G(j)$, initially the global balance equation for state $\mathbf{n}$, expressed as *rate into state* $\mathbf{n}$ = *rate out of state* $\mathbf{n}$ is considered:

$$\sum_{k=1}^{K} \lambda_k P(\mathbf{n}_k^-) + \sum_{k=1}^{K} (n_k+1)\mu_k \phi_k(\mathbf{n}_k^+) P(\mathbf{n}_k^+) = \\ = \sum_{k=1}^{K} \lambda_k P(\mathbf{n}) + \sum_{k=1}^{K} n_k \mu_k \phi_k(\mathbf{n}) P(\mathbf{n}),$$

where $\mathbf{n}_k^+ = (n_1, \ldots, n_k + 1, \ldots, n_K)$ and $P(\mathbf{n})$, $P(\mathbf{n}_k^-)$, $P(\mathbf{n}_k^+)$ are the probability distributions of states $\mathbf{n}$, $\mathbf{n}_k^-$, $\mathbf{n}_k^+$, respectively.

Assume now, the existence of Local Balance (LB) between adjacent states. Then the following LB equations can be extracted, for $k = 1, \ldots, K$ and $\mathbf{n} \in \mathbf{\Omega}$:

$$\lambda_k P(\mathbf{n}_k^-) = n_k \mu_k \phi_k(\mathbf{n}) P(\mathbf{n}), \quad (6)$$

$$\lambda_k P(\mathbf{n}) = (n_k+1)\mu_k \phi_k(\mathbf{n}_k^+) P(\mathbf{n}_k^+). \quad (7)$$

Based on the assumption of LB, $P(\mathbf{n})$ can be determined by

$$P(\mathbf{n}) = G^{-1}\left(x(\mathbf{n}) \prod_{k=1}^{K} \frac{a_k^{n_k}}{n_k!}\right), \quad (8)$$

where $a_k = \frac{\lambda_k}{\mu_k}$ is the offered traffic-load (in Erlangs) of service-class k and $G \equiv G(\Omega) = \sum_{n \in \mathbf{\Omega}} \left(x(\mathbf{n}) \prod_{k=1}^{K} \frac{a_k^{n_k}}{n_k!}\right)$.

Since j is the occupied link bandwidth, $G(j)$ is defined as:

$$G(j) = \sum_{\mathbf{n} \in \mathbf{\Omega}_j} P(n), \quad \mathbf{\Omega}_j = \{\mathbf{n} \in \mathbf{\Omega} : \mathbf{nb} = j\}, \quad (9)$$

Consider now two sets: 1) $0 \le j \le C$ and 2) $C < j \le T$. For set 1), we have the EMLM/TH and $G(j)$'s are given by the following formula [41]:

$$G(j) = \frac{1}{j}\sum_{k=1}^{K} a_k b_k \big[G(j-b_k) - T_k(j-b_k)\big], \text{ for } j = 1, \ldots, C, \quad (10)$$

where

$$T_k(x) := Pr\big[j = x, \ n_k = n_k^*\big]. \quad (11)$$

In Eq. (11) the fact that $n_k = n_k^*$ implies that $j \ge n_k^* b_k$.

When $C < j \le T$, Eq. (4) is substituted in Eq. (6) to have:

$$a_k x(\mathbf{n}) P(\mathbf{n}_k^-) = n_k x(\mathbf{n}_k^-) P(\mathbf{n}). \quad (12)$$

Multiplying both sides of Eq. (12) by b_k and summing over k we obtain:

$$x(\mathbf{n}) \sum_{k=1}^{K} a_k b_k P(\mathbf{n}_k^-) = P(\mathbf{n}) \sum_{k=1}^{K} n_k b_k x(\mathbf{n}_k^-). \quad (13)$$

Equation (13), due to Eq. (5) is written as:

$$P(\mathbf{n}) = \frac{1}{C}\sum_{k=1}^{K} a_k b_k P(\mathbf{n}_k^-). \quad (14)$$

Summing both sides of Eq. (14) over $\mathbf{\Omega}_j = \{n \in \mathbf{\Omega} : \mathbf{nb} = j\}$ and based on Eq. (9), we obtain:

$$G(j) = \frac{1}{C}\sum_{k=1}^{K} a_k b_k \sum_{\mathbf{n} \in \mathbf{\Omega}_j} P(\mathbf{n}_k^-). \quad (15)$$

Since $n_k \le n_k^*$ then

$$\sum_{\mathbf{n} \in \mathbf{\Omega}_j} P(\mathbf{n}_k^-) = G(j - b_k) - Pr\big[x = j - b_k, \ n_k = n_k^*\big].$$

Thus, Eq. (15) can be written as:

$$G(j) = \frac{1}{C}\sum_{k=1}^{K} a_k b_k \big[G(j - b_k) - T_k(j - b_k)\big], \text{ for } j = C+1, \ldots, T, \quad (16)$$

where $T_k(x)$ is given by Eq. (11).

Equations (10) and (16) result in the following approximate but recursive formula for the calculation of $G(j)$'s in the E-EMLM/TH:

$$G(j) = \frac{1}{\min(C, j)}\sum_{k=1}^{K} a_k b_k \big[G(j - b_k) - T_k(j - b_k)\big], \\ \text{for } \ j = 1, \ldots, T. \quad (17)$$

Having determined $G(j)$'s the CBP of service-class k, B_k, and the link utilization, U, are calculated as:

$$B_k = \sum_{j=T-b_k+1}^{T} G^{-1} G(j) + \sum_{j=n_k^* b_k}^{T-b_k} G^{-1} T_k(j), \quad (18)$$

$$U = \sum_{j=1}^{C} j G^{-1} G(j) + \sum_{j=C+1}^{T} C G^{-1} G(j), \quad (19)$$

where $G = \sum_{j=0}^{T} G(j)$ is the normalization constant.

In Eqs. (17) and (18) the knowledge of $T_k(j)$ is required. Since $T_k > 0$ when $j = n_k^* b_k, \ldots, T - b_k$, two subsets are considered: 1) $n_k^* b_k \le j \le C$ and 2) $C + 1 \le j \le T - b_k$.

For the first subset, let a system of capacity $F_k = T - b_k - n_k^* b_k$ that accommodates all service-classes but service-class k. For this system, $r_k(j)$ is defined as:

$$r_k(j) = \frac{1}{j}\sum_{\substack{i=1 \\ i \ne k}}^{K} a_i b_i \big[r_k(j - b_i) - T_i(j - b_i)\big], \\ \text{for } \ j = 1, \ldots, F_k. \quad (20)$$

Based on $r_k(j)$'s, $T_k(j)$ is computed via the formula

$$T_k(j) = \frac{a_k^{n_k^*}}{n_k^*!} r_k(j - n_k^* b_k). \quad (21)$$

For the second subset, $T_k(j)$ can be determined by

$$T_k(j) = \frac{a_k^{n_k^*}}{n_k^*!} \sum_{\mathbf{n}\in\mathbf{\Omega}} x(\mathbf{n}) \prod_{\substack{i=1 \\ i\neq k}}^{K} \frac{a_i^{n_i}}{n_i!}, \tag{22}$$

where $\mathbf{\Omega} = \left\{\mathbf{n}\in\mathbf{\Omega} : n_k^* b_k + \sum_{i=1, i\neq k}^{K} n_i b_i = j,\ C+1 \le j \le T-b_k\right\}$.

In Eq. (22), $T_k(j)$ is determined only for a subset of $\mathbf{\Omega}$, defined by $C+1 \le j \le T-b_k$ and only under the assumption that $n_k = n_k^*$. This means that enumeration of the subset of Ω is needed for those states $\mathbf{n} = (n_1, n_2, \ldots, n_k = n_k^*, \ldots, n_K)$ where $C+1 \le \mathbf{nb} \le T-b_k$. Based on the fact that the value of T should not be much higher than the corresponding value of C (otherwise the increase of delay for elastic calls may be unacceptable for some applications) the subset of $\mathbf{\Omega}$ will not become large. In general, the computational complexity of Eq. (22) grows exponentially with $K-1$ (since for service-class k we have $n_k = n_k^*$) and can be in the order of $O\left((T-b_k-C)^{(K-1)}\right)$. Assuming the existence of the CS policy and ignoring the bandwidth compression mechanism, then the computational complexity becomes $O(C^K)$ [1].

To further reduce the computational complexity of the proposed model, the application of convolutional algorithms may be considered [48], but this is left for future work.

3. The Elastic EMLM/TH-BR

Consider again a link of capacity C b.u. that accommodates K elastic service-classes of Poisson arriving calls. A new service-class k ($k = 1, \ldots, K$) call has a peak-bandwidth requirement of b_k b.u. and a BR parameter t_k that expresses the reserved b.u. used to benefit calls of all other service-classes except k. If $j + b_k \le T - t_k$ and $n_k^* + 1 \le n_k^*$ then the call is accepted in the link and remains for an exponentially distributed service time with mean μ_k^{-1}. Otherwise the call is blocked and lost.

To determine $G(j)$'s in the E-EMLM/TH-BR the authors propose the following approximate but recursive formula:

$$G(j) = \left\langle \begin{array}{ll} 1, & \text{for } j = 0 \\ \dfrac{1}{\min(C,j)} \displaystyle\sum_{k=1}^{K} a_k D_k(j-b_k) \times & \\ \times \left[G(j-b_k) - T_k(j-b_k)\right] & \text{for } j = 1, \ldots, T \\ 0, & \text{otherwise} \end{array} \right\rangle, \tag{23}$$

where

$$D_k(j-b_k) = \begin{cases} b_k & \text{for} \quad j \le T - t_k \\ 0 & \text{for} \quad j > T - t_k \end{cases}. \tag{24}$$

A characteristic of the BR policy is that it ensures CBP equalization among different service-classes by a proper selection of the BR parameters. If, for example, CBP equalization is required between calls of two service-classes with $b_1 = 1$ and $b_2 = 10$ b.u., respectively, then $t_1 = 9$ b.u. and $t_2 = 0$ b.u. so that $b_1 + t_1 = b_2 + t_2$.

The application of the BR policy in the E-EMLM/TH-BR is based on the assumption that the number of service-class k calls is negligible in states $j > T - t_k$ and is incorporated in Eq. (23) by the variable $D_k(j-b_k)$ given in Eq. (24). The states $j > T - t_k$ belong to the so-called reservation space. Note that the population of calls of service-class k in the reservation space may not be negligible. In [6], [11] a complex procedure is implemented that takes into account this population in the EMLM and Engset multirate state-dependent loss models, respectively. However, this procedure may not always increase the accuracy of the CBP results compared to simulation [11].

Similarly to the E-EMLM/TH, the CBP of service-class k, B_k, is determined based on two groups of states:

- those where the available link bandwidth is less than $b_k + t_k$ b.u. when the new call arrives in the system; this happens when $T - b_k - t_k + 1 \le j \le T$;
- those where the available link bandwidth is enough to accept the new call, i.e. $j \le T - b_k - t_k$ but $n_k = n_k^*$.

The latter implies that $j \ge n_k^* b_k$, or $n_k^* b_k \le j \le T - b_k - t_k$. Thus, the values of B_k are calculated by:

$$B_k = \sum_{j=T-b_k-t_k+1}^{T} G^{-1} G(j) + \sum_{j=n_k^* b_k}^{T-b_k-t_k} G^{-1} T_k(j), \tag{25}$$

where $G = \sum_{j=0}^{C} G(j)$ is the normalization constant.

As far as U is concerned, it can be determined by Eq. (19).

In Eqs. (23) and (25) the knowledge of $T_k(j)$ is required for $n_k^* b_k \le j \le T - b_k - t_k$. The authors consider again two subsets: 1) $n_k^* b_k \le j \le C$ and 2) $C+1 \le j \le T - b_k - t_k$. For the first subset, authors use Eqs. (20), (21) where $F_k = T - b_k - t_k - n_k^* b_k$, while for subset (2) we use Eq. (22) where $x(\mathbf{n})$ is given by Eq. (5) and

$$\mathbf{\Omega'} = \left\{\mathbf{n}\in\mathbf{\Omega'} : n_k^* b_k + \sum_{\substack{i=1 \\ i\neq k}}^{K} n_i b_i = j,\ C+1 \le j \le T - b_k - t_k\right\}.$$

If $C = T$ and both the TH and the BR policies are considered, then calls are not allowed to compress their bandwidth. In this case, the proposed E-EMLM/TH-BR coincides with the EMLM/TH-BR of [45]. The values of $G(j)$'s and CBP are given by Eqs. (26), (27), respectively:

$$G(j) = \left\langle \begin{array}{ll} 1, & \text{for } j = 0 \\ \dfrac{1}{j} \displaystyle\sum_{k=1}^{K} a_k D_k(j-b_k) \times & \\ \times \left[G(j-b_k) - T_k(j-b_k)\right] & \text{for } j = 1, \ldots, C \\ 0, & \text{otherwise} \end{array} \right\rangle, \tag{26}$$

$$B_k = \sum_{j=C-b_k-t_k+1}^{C} G^{-1} G(j) + \sum_{j=n_k^* b_k}^{C-b_k-t_k} G^{-1} T_k(j), \tag{27}$$

where

$$D_k(j-b_k)=\begin{cases} b_k & \text{for} \quad j \le C-t_k \\ 0 & \text{for} \quad j > C-t_k \end{cases}. \tag{28}$$

The link utilization can be calculated by

$$U=\sum_{j=1}^{C} jG^{-1}G(j). \tag{29}$$

Finally, if $C=T$ and we do not consider the TH and the BR policies, then the proposed E-EMLM/TH-BR coincides with the classical EMLM of [1], [2]. In that case, the link occupancy distribution is determined by the well-known Kaufman-Roberts recursion:

$$G(j)=\left\langle \begin{array}{ll} 1, & \text{for } j=0 \\ \frac{1}{j}\sum_{k=1}^{K} a_k b_k G(j-b_k) & \text{for } j=1,\ldots,C \\ 0, & \text{otherwise} \end{array} \right\rangle, \tag{30}$$

The CBP of service-class k is given by [1], [2]:

$$B_k=\sum_{j=C-b_k+1}^{C} G^{-1}G(j), \tag{31}$$

while the link utilization can be determined by Eq. (29).

4. The Elastic-Adaptive EMLM/TH-BR

Adaptive traffic is a variant of elastic traffic since adaptive calls can tolerate bandwidth compression without altering their service time. To include adaptive traffic in the E-EMLM/TH, authors assume that K_e and K_a are the number of elastic and adaptive service-classes, respectively. The single link accommodates K service-classes of Poisson arriving calls, where $K=K_e+K_a$.

The existence of the bandwidth compression mechanism destroys reversibility in the proposed model and therefore the steady state probabilities have no PFS. To circumvent this problem, we use $\phi_k(\mathbf{n})$ based on Eq. (4) while the values of $x(\mathbf{n})$ are given by:

$$x(\mathbf{n})=\begin{cases} 1, & \text{when } \mathbf{nb} \le C,\ \mathbf{n} \in \mathbf{\Omega} \\ \frac{1}{C}\left(\sum_{k\in K_e} n_k b_k x(\mathbf{n}_k^-) + \right. & \\ \left. + r(\mathbf{n}) \sum_{k\in K_a} n_k b_k x(\mathbf{n}_k^-) \right) & \text{when } C < \mathbf{nb} \le T,\ \mathbf{n} \in \mathbf{\Omega} \\ 0, & \text{otherwise} \end{cases}. \tag{32}$$

The derivation of Eq. (32) is based on the assumptions:

- The bandwidth of all in-service calls of service-class $k \in K$ (elastic or adaptive) is compressed by a factor $\phi_k(\mathbf{n})$, in state $C < \mathbf{nb} \le T$, so that:

$$\sum_{k\in K_e} n_k b'_k + \sum_{k\in K_a} n_k b'_k = C \tag{33}$$

- The product *service time by bandwidth per call* of service-class k calls, $k \in K$, remains the same in state $\mathbf{n}$ either of the irreversible or the reversible Markov chain. In other words:

  For elastic service-classes:

$$\frac{b_k r(\mathbf{n})}{\mu_k r(\mathbf{n})} = \frac{b'_k}{\mu_k \phi_k(\mathbf{n})} \Rightarrow b'_k = b_k \phi_k(\mathbf{n}). \tag{34}$$

  For adaptive service-classes:

$$\frac{b_k r(\mathbf{n})}{\mu_k} = \frac{b'_k}{\mu_k \phi_k(\mathbf{n})} \Rightarrow b'_k = b_k \phi_k(\mathbf{n}) r(\mathbf{n}). \tag{35}$$

Under these assumptions, Eq. (32) can be derived by substituting Eqs. (34), (35) and Eq. (4) into Eq. (33).

Based on Eqs. (32)–(35) and following the analysis of Section 2, it can be proved that $G(j)$'s are given by the following formula for the EA-EMLM/TH:

$$G(j)=\left\langle \begin{array}{ll} 1, & \text{for } j=0 \\ \frac{1}{\min(C,j)} \sum_{k=1}^{K_e} a_k b_k \left[(G(j-b_k) - T_k(j-b_k) \right] & \\ + \frac{1}{j} \sum_{k=1}^{K_a} a_k b_k \left[G(j-b_k) - T_k(j-b_k) \right], & \text{for } j=1,\ldots,T \\ 0, & \text{otherwise} \end{array} \right\rangle. \tag{36}$$

Having determined $G(j)$'s, the CBP of service-class k, B_k, and the link utilization, U, can be calculated via Eqs. (18) and (19), respectively.

To determine $G(j)$'s, in the EA-EMLM/TH-BR the following approximate but recursive formula is proposed:

$$G(j)=\left\langle \begin{array}{ll} 1, & \text{for } j=0 \\ \frac{1}{\min(C,j)} \sum_{k=1}^{K_e} a_k D_k(j-b_k) \left[G(j-b_k) - T_k(j-b_k) \right] & \\ + \frac{1}{j} \sum_{k=1}^{K_a} a_k D_k(j-b_k) \left[G(j-b_k) - T_k(j-b_k) \right], & \text{for } j=1,\ldots,T \\ 0, & \text{otherwise} \end{array} \right\rangle, \tag{37}$$

where the values of $D_k(j-b_k)$ are given by Eq. (24).

Similar to the E-EMLM/TH-BR, the CBP of service-class k, B_k, and the link utilization, U are determined, via Eqs. (25) and (19), respectively.

5. Numerical Examples – Evaluation

In this section, an application example of the proposed E-EMLM/TH-BR and the model of [49] (EMLM/TH-BR) is presented. Through the proposed model authors obtain analytical CBP and link utilization results, and compare them with the corresponding simulation results, in order to reveal the accuracy of the proposed model. Similar accuracy appears in the case of the EA-EMLM/TH-BR and therefore these results are not presented herein. The simulation model is based on the bandwidth compression/expansion mechanism described by $r(\mathbf{n})$'s. On the other hand, the proposed analytical models are based

on $\phi_k(\mathbf{n})$'s. In that sense, the comparison of analytical with simulation results shows how satisfactory the approximation of $\phi_k(\mathbf{n})$'s is. Simulation results are mean values of 7 runs. Each run is based on the generation of four million calls. To account for a warm-up period, the blocking events of the first 5% of these generated calls are not considered in the results. Due to the fact that reliability ranges are very small, they are not presented in the figures that follow. The simulation language used is Simscript III [50].
As an application example, a link of capacity $C = 70$ b.u. is considered and three values of T:

1) $T = C = 70$ b.u.,
2) $T = 75$ b.u. with $r_{\max} = 70/75$,
3) $T = 80$ b.u. with $r_{\max} = 70/80$.

The link accommodates three service-classes, with the following characteristics:

- 1st service-class: $a_1 = 5.0$ Erl, $b_1 = 2$, $n_1^* = 25$, $t_1 = 7$,
- 2nd service-class: $a_2 = 1.5$ Erl, $b_2 = 5$, $n_2^* = 11$, $t_2 = 4$,
- 3rd service-class: $a_3 = 1.0$ Erl, $b_3 = 9$, $n_3^* = 6$, $t_2 = 0$.

In the x axis of all figures, traffic loads α_1, α_2 and α_3 increase in steps of 1, 0.5 and 0.25 Erl, respectively. So, Point 1 refers to $(a_1, a_2, a_3) = (5.0, 1.5, 1.0)$ while Point 7 is $(a_1, a_2, a_3) = (11.0, 4.5, 2.5)$.
In Figs. 1–3, authors consider the proposed E-EMLM/TH-BR and present the analytical and simulation CBP results of the three service-classes, respectively, for all values of T. For comparison, the corresponding analytical results of the EMLM/TH-BR (when $T = C = 70$) are presented. According to Figs. 1–3, authors deduce that:

- the results obtained by the proposed formulas are very close to the simulation results;
- the bandwidth compression mechanism reduces CBP as expected (higher reduction is achieved for $T = 80$ b.u.);
- the analytical CBP results obtained by the existing EMLM/TH-BR fail to approximate the simulation CBP results of the E-EMLM/TH-BR;
- the application of the BR policy in the E-EMLM/TH-BR results in the CBP increase of the 1st and 2nd service-classes and the CBP decrease of the 3rd service-class. This behavior is expected since the BR parameters are chosen to favor the 3rd service-class.

In Fig. 4, the link utilization results (in b.u.) are presented. Again, the analytical results are very close to simulation, while the existing EMLM/TH-BR fails to approximate the results obtained by the proposed model.

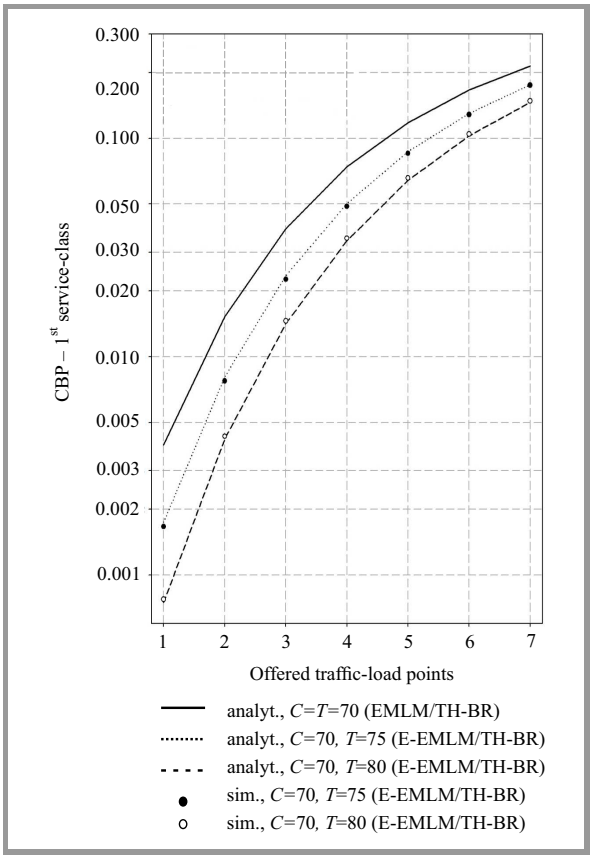


Fig. 1. CBP – 1st service-class.

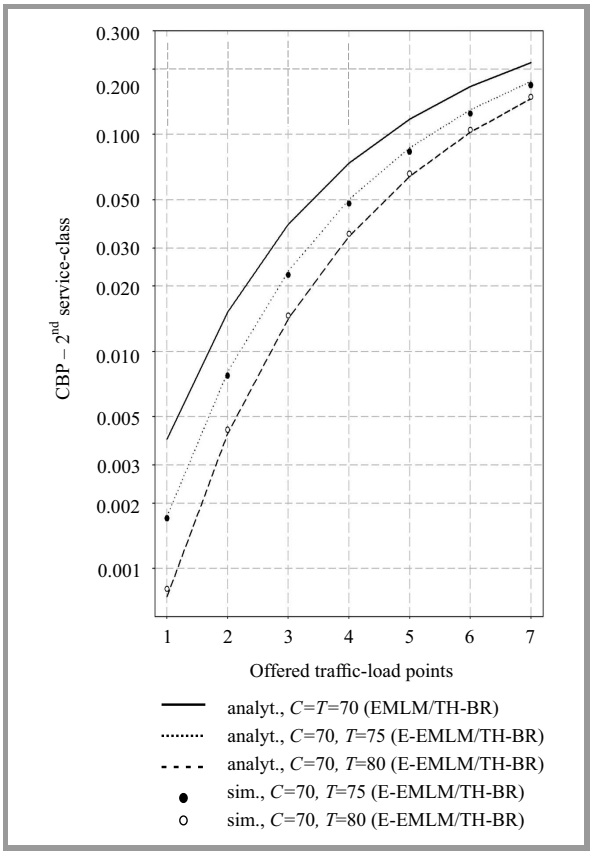


Fig. 2. CBP – 2nd service-class.

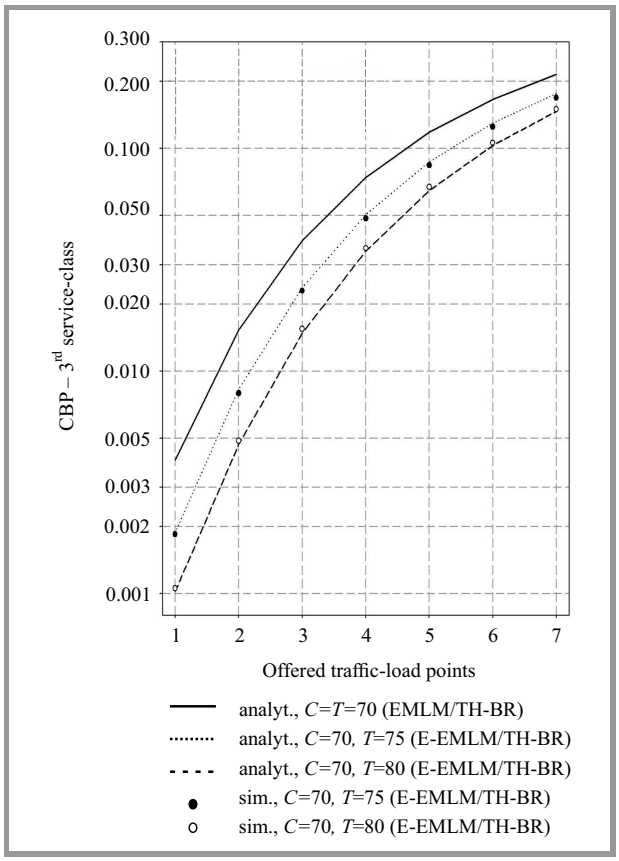


Fig. 3. CBP – 3rd service-class.

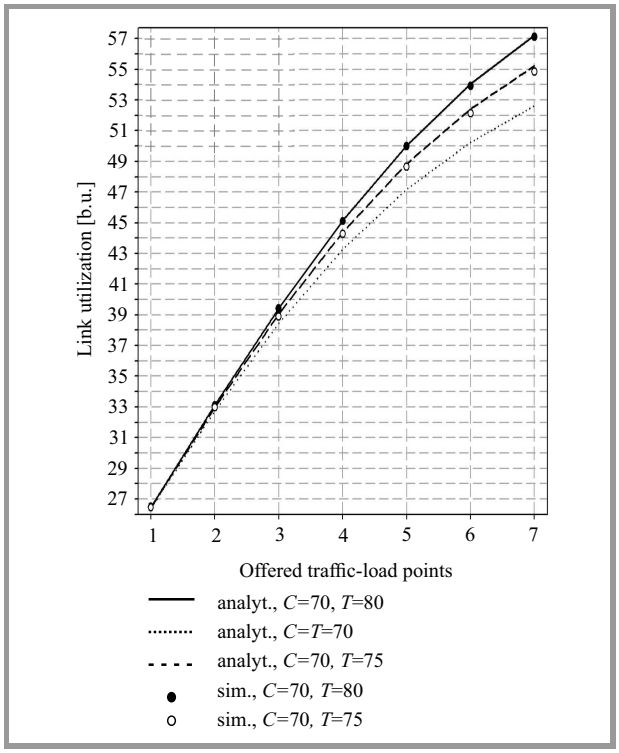


Fig. 4. Link utilization.

As a final comment, the results obtained by the proposed formulas are very close to the simulation results even for quite large values of T compared to C. However, increasing T results in a delay increase of elastic calls, which may be unacceptable for some applications. Thus, T should be chosen so that this delay remains within acceptable levels.

6. Conclusion

In this paper authors propose multirate loss models where Poisson arriving calls compete for the available link bandwidth under the TH and the BR policies. Calls are of elastic or adaptive type, i.e., they can tolerate bandwidth compression while in-service. When bandwidth of in-service elastic calls is compressed then their remaining service time is increased. Adaptive in-service calls do not alter their service time. The analysis of the proposed models leads to approximate but recursive formulas for the calculation of the steady-state probabilities and consequently CBP and link utilization. Simulation results verify the accuracy of the proposed models. In addition, numerical results show the necessity of the proposed models, since existing models fail to approximate the results obtained by the proposed models, and their consistency.

References

[1] J. Kaufman, "Blocking in a shared resource environment", *IEEE Trans. Commun.*, vol. 29, no. 10, pp. 1474–1481, 1981.

[2] J. Roberts, "A service system with heterogeneous user requirements", in *Performance of Data Communications Systems and their Applications*, G. Pujolle, Ed. Amsterdam: North-Holland Pub., 1981, pp. 423–431.

[3] J. Kaufman, "Blocking with retrials in a completely shared resource environment", *Perform. Eval.*, vol. 15, no. 2, pp. 99–113, 1992.

[4] G. Stamatelos and J. Hayes, "Admission control techniques with application to broadband networks", *Comput. Commun.*, vol. 17, no. 9, pp. 663–673, 1994.

[5] G. Stamatelos and V. Koukoulidis, "Reservation-Based Bandwidth Allocation in a Radio ATM Network", *IEEE/ACM Trans. Netw.*, vol. 5, no. 3, pp. 420–428, 1997.

[6] M. Stasiak and M. Głąbowski, "A simple approximation of the link model with reservation by a one-dimensional Markov chain", *Perform. Eval.*, vol. 41, no. 2–3, pp. 195–208, 2000.

[7] I. Moscholios, M. Logothetis, and G. Kokkinakis, "Connection dependent threshold model: a generalization of the erlang multiple rate loss model", *Perform. Eval.*, vol. 48,no. 1–4, pp. 177–200, 2002.

[8] M. Głąbowski and M. Stasiak, "Point-to-point blocking probability in switching networks with reservation", *Annals of Telecommun.*, vol. 57, no. 7–8, pp. 798–831, 2002.

[9] S. Rácz, B. Gerő, and G. Fodor, "Flow level performance analysis of a multi-service system supporting elastic and adaptive services", *Perform. Eval.*, vol. 49, no. 1–4, Sept. 2002, pp. 451–469.

[10] I. Moscholios, P. Nikolaropoulos, and M. Logothetis, "Call level blocking of ON-OFF traffic sources with retrials under the complete sharing policy", in *Proc. 18th Int. Teletraffic Congr. ITC-18*, Berlin, Germany, 2003, vol. 31, pp. 811–820.

[11] I. Moscholios and M. Logothetis, "Engset multirate state-dependent loss models with QoS guarantee", *Int. J. of Commun. Syst.*, vol. 19, no. 1, pp. 67–93, 2006.

[12] M. Głąbowski, "Recurrent calculation of blocking probability in multiservice switching networks", in *Proc. Asia-Pacific Conf. Commun. APCC 2006*, Busan, South Korea, 2006.

[13] I. Moscholios, M. Logothetis, and M. Koukias, "An ON-OFF multi-rate loss model of finite sources", *IEICE Trans. Commun*., vol. E90-B, no. 7, pp. 1608–1619, 2007.

[14] Q. Huang, King-Tim Ko, and V. Iversen, "Approximation of loss calculation for hierarchical networks with multiservice overflows", *IEEE Trans. Commun.*, vol. 56, no. 3, pp. 466–473, 2008.

[15] M. Głąbowski, A. Kaliszan, and M. Stasiak, "Modelling product-form state-dependent systems with BPP traffic", *Perform. Eval.*, vol. 67, no. 3, pp. 174–197, 2010.

[16] I. Moscholios, J. Vardakas, M. Logothetis, and A. Boucouvalas, "QoS guarantee in a batched poisson multirate loss model supporting elastic and adaptive traffic", in *Proc. IEEE Int. Conf. on Commun. IEEE ICC 2012*, Ottawa, Canada, 2012.

[17] I. Moscholios, J. Vardakas, M. Logothetis, and A. Boucouvalas, "Congestion probabilities in a Batched Poisson multirate loss model supporting elastic and adaptive traffic", *Annals of Telecommun*., vol. 68, no. 5, pp. 327–344, 2013.

[18] I. Moscholios, J. Vardakas, M. Logothetis, and M. Koukias, "A quasi-random multirate loss model supporting elastic and adaptive traffic under the bandwidth reservation policy", *Int. J. on Adv. in Netwo. and Serv*., vol. 6, no. 3–4, pp. 163–174, 2013.

[19] S. Hanczewski, M. Stasiak, and J. Weissenberg, "A queueing model of a multi-service system with state-dependent distribution of resources for each class of calls", *IEICE Trans. Commun*., vol. E97-B, no. 8, pp. 1592–1605, 2014.

[20] D. Staehle and A. Mäder, "An analytic approximation of the uplink capacity in a UMTS network with heterogeneous traffic", in *Proc. 18th Int. Teletraffic Congr. ITC-18*, Berlin, Germany, 2003, pp. 81–90, 2003.

[21] V. Iversen, V. Benetis, N. Ha, and S. Stepanov, "Evaluation of multiservice CDMA networks with soft blocking", in *Proc. ITC Specialist Seminar*, Antwerp, Belgium, 2004, pp. 223–227.

[22] G. Fodor and M. Telek, "Bounding the blocking probabilities in multirate CDMA networks supporting elastic services", *IEEE/ACM Trans. Netw.*, vol. 15, no. 4, pp. 944–956, 2007.

[23] V. Vassilakis, G. Kallos, I. Moscholios, and M. Logothetis, "Call-Level Analysis of W-CDMA Networks Supporting Elastic Services of Finite Population", in *Proc. IEEE Int. Conf. on Commun. IEEE ICC 2008*, Beijing, China, 2008.

[24] M. Głąbowski, M. Stasiak, A. Wisniewski, and P. Zwierzykowski, "Blocking probability calculation for cellular systems with WCDMA Radio Interface Servicing PCT1 and PCT2 Multirate Traffic", *IEICE Trans. Commun.*, vol. E92-B, pp. 1156–1165, 2009.

[25] I. Widjaja and H. Roche, "Sizing X2 bandwidth for Inter-connected eNBs", in *Proc. 70th Vehi. Technol. Conf. Fall VTC 2009-Fall*, Anchorage, Alaska, USA, 2009, pp. 420–424.

[26] M. Stasiak, P. Zwierzykowski, and D. Parniewicz, "Modelling of the WCDMA interface in the UMTS network with soft handoff mechanism", in *Proc. IEEE Global Commun. Conf. Globecom 2009*, Honolulu, Hawaii, 2009.

[27] D. Parniewicz, M. Stasiak, and P. Zwierzykowski, "Analytical model of the multi-service cellular network servicing multicast connections", *Telecommun. Syst*., vol. 52, no. 2, pp. 1091–1100, 2013.

[28] I. Moscholios, G. Kallos, M. Katsiva, V. Vassilakis, and M. Logothetis, "Call blocking probabilities in a W-CDMA cell with interference cancellation and bandwidth reservation", in *Proc. IEICE Informa. Commun. Technol. Forum IEICE ICTF 2014*, Poznań, Poland, 2014.

[29] I. Moscholios, G. Kallos, V. Vassilakis, and M. Logothetis, "Congestion probabilities in CDMA-based networks supporting batched Poisson input traffic", *Wirel. Personal Commun.*, vol. 79, no. 2, pp. 1163–1186, 2014.

[30] V. Vassilakis, I. Moscholios, J. Vardakas, and M. Logothetis, "Handoff modeling in cellular CDMA with finite sources and state-dependent bandwidth requirements", in *Proc. 19th Int. Worksh. Comp. Aided Model. & Design Commun. Links Netw. IEEE CAMAD 2014*, Athens, Greece, 2014.

[31] V. Vassilakis, I. Moscholios, and M. Logothetis, "Performance evaluation of priority-based CDMA systems in the presence of multirate poisson traffic", in *Proc. IEICE Inform. Commun. Technol. Forum ICTF 2015*, Manchester, United Kingdom, 2015.

[32] I. Moscholios, V. Vassilakis, G. Kallos, and M. Logothetis, "Performance analysis of CDMA-based networks with interference cancellation, for batched Poisson traffic under the bandwidth reservation policy", in *Proc. 13th Int. Conf. Telecommun. ConTEL 2015*, Graz, Austria, 2015.

[33] J. Vardakas, I. Moscholios, M. Logothetis, and V. Stylianakis, "An analytical approach for dynamic wavelength allocation in WDM-TDMA PONs servicing ON-OFF traffic", *IEEE/OSA J. Optical Commun. Netw.*, vol. 3, no. 4, pp. 347–358, 2011.

[34] Y. Deng and P. Prucnal, "Performance analysis of heterogeneous optical CDMA networks with bursty traffic and variable power control", *IEEE/OSA J. Optical Commun. Netw*., vol. 3, no. 6, pp. 487–492, 2011.

[35] N. Jara and A. Beghelli, "Blocking probability evaluation of end-to-end dynamic WDM networks", *Photonic Netw. Commun*., vol. 24, no. 1, pp. 29–38, 2012.

[36] J. Vardakas, I. Moscholios, M. Logothetis, and V. Stylianakis, "Blocking performance of multi-rate OCDMA PONs with QoS guarantee", *Int. J. Adv. Telecommun*., vol. 5, no. 3–4, pp. 120–130, 2012.

[37] J. Vardakas, I. Moscholios, M. Logothetis, and V. Stylianakis, "Performance analysis of OCDMA PONs supporting multi-rate bursty traffic", *IEEE Trans. Commun*., vol. 61, no. 8, pp. 3374–3384, 2013.

[38] K. Ross, *Multiservice Loss Models for Broadband Telecommunication Networks*. Springer, 1995.

[39] J. Ni, D. Tsang, S. Tatikonda, and B. Bensaou, "Threshold and reservation based call admission control policies for multiservice resource-sharing systems", in *Proc. IEEE 24th Annual Joint. Conf. Comp. Commun. Soc. INFOCOM 2005*, Miami, FL, USA, 2005.

[40] J. Ni, D. Tsang, S. Tatikonda, and B. Bensaou, "Optimal and structured call admission control policies for resource-sharing systems", *IEEE Trans. Commun*., vol. 55, no. 1, pp. 158–170, 2007.

[41] D. Tsang and K. Ross, "Algorithms to determine exact blocking probabilities for multirate tree networks", *IEEE Trans. Commun*., vol. 38, no. 8, pp. 1266–1271, 1990.

[42] F. Cruz-Perez, J. Vazquez-Avila, and L. Ortigoza-Guerrero, "Recurrent formulas for the multiple fractional channel reservation strategy in multi-service mobile cellular networks", *IEEE Commun. Lett.*, vol. 8, no. 10, pp. 629–631, 2004.

[43] I. Moscholios, M. Logothetis, and M. Koukias, "A state-dependent multi-rate loss model of finite sources with QoS guarantee for wireless networks", *Mediterranean J. Comp. Netw*., vol. 2, no. 1, pp. 10–20, 2006.

[44] M. Głąbowski, A. Kaliszan, and M. Stasiak, "Asymmetric convolution algorithm for blocking probability calculation in full-availability group with bandwidth reservation", *IET Circ. Devices & Syst*., vol. 2, no. 1, pp. 87–94, 2008.

[45] S. Miyata and K. Yamaoka, "Flow-admission control based on equality of heterogeneous traffic (two-type flow model)", *IEICE Trans. Commun*., vol. E93-B, no. 12, pp. 3564–3576, 2010.

[46] V. Vassilakis, I. Moscholios, and M. Logothetis, "The extended connection-dependent threshold model for call-level performance analysis of multi-rate loss systems under the bandwidth reservation policy", *Int. J. Commun. Syst*., vol. 25, no. 7, pp. 849–873, 2012.

[47] I. Moscholios, M. Katsiva, G. Kallos, V. Vassilakis, and M. Logothetis, "Equalization of congestion probabilities in a W-CDMA cell supporting calls of finite sources with interference cancellation", in *Proc. IEEE/IET Int. Symp. Commun. Syst., Netw. & Digit. Sig.l Process. CSNDSP 2014*, Manchester, United Kingdom, 2014.

[48] M. Stasiak, M. Głąbowski, A. Wisniewski, and P. Zwierzykowski, *Modeling and Dimensioning of Mobile Networks*. Chichester: Wiley, 2011.

[49] I. Moscholios, M. Logothetis, J. Vardakas, and A. Boucouvalas, "Performance metrics of a multirate resource sharing teletraffic model with finite sources under the threshold and bandwidth reservation policies", *IET Networks*, vol. 4, no. 3, pp. 195–208, 2015.

[50] Simscript III [Online]. Available: http://www.simscript.com (accessed: Oct. 2015).

Ioannis D. Moscholios received the Dipl.-Eng. degree in Electrical and Computer Engineering from the University of Patras, Patras, Greece, in 1999, the M.Sc. degree in Spacecraft Technology and Satellite Communications from the University College London, UK, in 2000 and the Ph.D. degree in Electrical and Computer Engineering from the University of Patras, in 2005. From 2005 to 2009 he was a Research Associate at the Wire Communications Laboratory, Dept. of Electrical and Computer Engineering, University of Patras. From 2009 to 2013 he was a Lecturer in the Dept. of Telecommunications Science and Technology, University of Peloponnese, Tripolis, Greece. Currently, he is an Assistant Professor in the Dept. of Informatics and Telecommunications, University of Peloponnese, Tripolis, Greece. His research interests include simulation and performance analysis of communication networks. He has published over 100 papers in international journals/conferences.

E-mail: idm@uop.gr
Department of Informatics and Telecommunications
University of Peloponnese
221 00 Tripolis, Greece

Michael D. Logothetis received his Dipl.-Eng. degree and Ph.D. in Electrical Engineering, both from the University of Patras, Patras, Greece, in 1981 and 1990, respectively. From 1982 to 1990, he was a Teaching and Research Assistant at the Laboratory of Wire Communications, University of Patras, and participated in many national and EU research programmes, dealing with telecommunication networks, as well as with office automation. From 1991 to 1992 he was Research Associate in NTT's Telecommunication Networks Laboratories, Tokyo, Japan. Afterwards, he was a Lecturer in the Dept. of Electrical and Computer Engineering of the University of Patras, and recently he has been elected Professor in the same Department. His research interests include teletraffic theory and engineering, traffic/network control, simulation and performance optimization of telecommunications networks. He has published over 180 conference/journal papers and has over 530 third-part citations.

E-mail: mlogo@upatras.gr
Wire Communications Laboratory
Department of Electrical and Computer Engineering
University of Patras
265 04 Patras, Greece

Anthony C. Boucouvalas received the B.Sc. degree in Electrical and Electronic Engineering from Newcastle upon Tyne University, U.K., in 1978, the M.Sc. and D.I.C. degrees in Communications Engineering from Imperial College, University of London, U.K., in 1979, and the Ph.D. degree in fiber optics from Imperial College, in 1982. Subsequently, he joined the GEC Hirst Research Center working on fiber-optic components, measurements, and sensors until 1987, when he joined Hewlett Packard Laboratories (HP) as a Project Manager. At HP, he worked in the areas of optical communication systems, optical networks, and instrumentation, until 1994, when he joined Bournemouth University, Bournemouth, U.K. In 1996, he became a Professor in Multimedia Communications, and in 1999 he became Director of the Microelectronics and Multimedia research Center. Currently, he is a Professor in the Dept. of Informatics and Telecommunications, University of Peloponnese, Greece. His current research interests span the fields of wireless communications, optical fiber communications and components and multimedia communications where he has published over 300 papers.

E-mail: acb@uop.gr
Department of Informatics and
Telecommunications
University of Peloponnese
221 00 Tripolis, Greece

Vassilios G. Vassilakis received his Ph.D. degree in Electrical and Computer Engineering from the University of Patras, Greece in 2011. From 2011 to 2013 he was with the Network Convergence Laboratory (NCL), University of Essex, and conducting research on information centric networking and network security. From 2013 to 2015 he was with the Institute for Communication Systems (ICS), University of Surrey, and conducting research on wireless networks. In 2015 he joined the Computer Laboratory, University of Cambridge. He has been involved in EU, UK, and industry funded R&D projects related to the design of future mobile networks and new Internet architectures. His main research interests are in the areas of next-generation wireless and mobile networks, future Internet technologies, software-defined networks, and network security. He has published over 60 journal/conference papers.

E-mail: vv274@cl.cam.ac.uk
Computer Laboratory
University of Cambridge
CB3 0FD, Cambridge, United Kingdom

Paper

Estimation of Network Disordering Effects by In-depth Analysis of the Resequencing Buffer Contents in Steady-state

Alexander Pechinkin and Rostislav Razumchik

Institute of Informatics Problems, Federal Research Center "Computer Science and Control", Russian Academy of Sciences, Moscow, Russia

Abstract—The paper is devoted to the analytic analysis of resequencing issue, which is common in packet networks, using queueing-theoretic approach. The authors propose the mathematical model, which describes the simplest setting of packet resequencing, but which allows one to make the first step in the in-depth-analysis of the queues dynamics in the resequencing buffer. Specifically consideration is given to N-server queueing system ($N > 3$) with single infinite capacity buffer and resequencing, which may serve as a model of packet reordering in packet networks. Customers arrive at the system according to Poisson flow, occupy one place in the buffer and receive service from one of the servers, which is exponentially distributed with the same parameter. The order of customers upon arrival has to be preserved upon departure. Customers, which violated the order are kept in resequencing buffer which also has infinite capacity. It is shown that the resequencing buffer can be considered as consisting of n, $1 \leq n \leq N-1$, interconnected queues, depending on the number of busy servers, with i-th queue containing customers, which have to wait for i service completions before they can leave the system. Recursive algorithm for computation of the joint stationary distribution of the number of customers in the buffer and servers, and each queue in resequencing buffer are being obtained. Numerical examples, which show the dynamics of the characteristics of the queues in resequencing buffer are given.

Keywords—infinite capacity, joint distribution, queueing system, resequencing.

1. Introduction

It is well-known that performance of multi-node simultaneous processing systems can suffer from the resequencing issue, i.e. when the order of arriving customers (packets, jobs, items, etc.) is violated due to disordering, which may be introduced by service process or other external/internal factors. As a consequence of disordering, some customers have to wait for other customers before they are allowed to leave the system. So far various analytical methods and models have been proposed to study the impacts of resequencing. Survey on the resequencing problem that covers the early period up to 1997 and review of queueing theoretic methods and early models for the modeling and analysis of parallel and distributed systems, including network systems, with resequencing can be found in [1] and [2]. Queueing-theoretic approach to the resequencing problem implies that the system under consideration is represented as interconnected queueing systems/networks, where the disordering of customers takes place. The system is followed with resequencing buffer, where the order of customers is recovered. When the system under consideration is the packet network, then the disordering may take place in the core network and the resequencing buffer is, for example, the de-jitter buffer in the end node. In [3] there was proposed to group existing papers on resequencing into two categories: papers that characterize the disordering process using single queueing system with several servers sharing a single queue (see e.g. [4]) and papers where disordering is modeled by a queueing system with several parallel servers and queues, and each server has its own dedicated queue (see e.g. [5]). Paper [3] contains the survey of papers belonging to these two categories.

In this paper, authors consider the system belonging to the first one. Up to now various problems setting have been considered and solved including calculation of the distribution of number of packets in resequencing buffer and in system under different assumptions about arrival and service process, calculation of the distribution of the resequencing delay, and optimal allocation of customers (see e.g. [1], [2], [5]–[15]). The resequencing effects can be estimated by calculation one or several parameters of the resequencing buffer (say, mean buffer size). Clearly the less mean buffer size is observed, the less packet resequencing is required in the system.

Here authors propose to dig deeper in the resequencing issue by giving a more thorough analysis of the resequencing buffer. It is probably the simplest problem setting but it gives a general view of the approach and method of the analysis. It is important to notice that the proposed method heavily relies on the fact that the servers are homogeneous and its extension to the heterogeneous case is a question of further research.

Specifically the network is modeled, where disordering takes place, as a $M|M|N|\infty$ queue ($N > 3$). Here each server may represent the link (or group of links) in the network. Transmission times (service times) are exponentially distributed with the same parameter. The elimination of the

disordering effect (i.e. recovery of the packets' sequence) takes place in the resequencing buffer. The sketch of the system can be seen in Fig. 1. Packets arrive according to Poisson flow and are stored in the infinite capacity buffer before entering the network, where from they are chosen for transmission according to First Come First Served (FCFS) or Last Come First Served (LCFS) or Random discipline. Customers, which violated the arrival order are kept in the resequencing buffer (RB) of infinite capacity before each of them can leave the system. As it was noticed in [16], in such $M|M|N|\infty$ resequencing queue with $N > 2$ servers, the resequencing buffer can be thought of either as a single queue, where all customers which violated arrival order reside together (Fig. 1a) or as a collection of several separate interconnected queues (Fig. 1b). In the latter case i-th queue contains those customers, which have to wait for i service completions before they can leave the system. Notice that the number of service completions needed by a customer in the RB to leave the system cannot be greater than $N-1$.

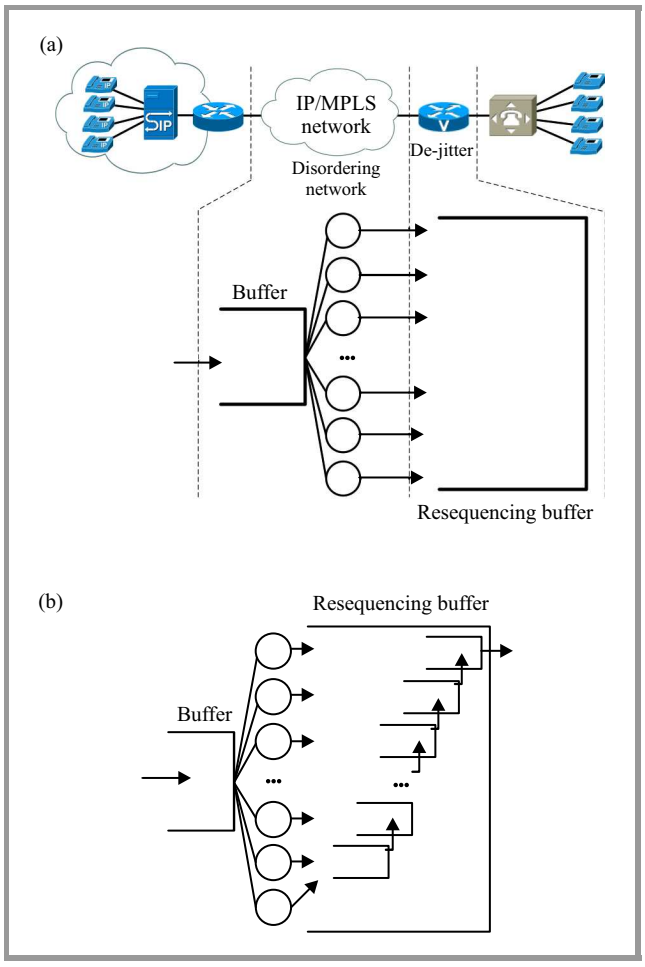


Fig. 1. (a) example of the resequencing issue in the VoIP scenario, (b) sketch of the multiserver resequencing queue with separate interconnected queues in the resequencing buffer.

The proposed point of view of the in-depth-dynamics of the RB can be probably best described by an example. Consider the network modeled by $M|M|4|\infty$ queueing system (where disordering takes place) and a resequencing queue at the exit from the network (see Fig. 1a). Without loss of generality authors suppose that packets (customers) upon entering the network (system) obtain a sequential number. The sequence starts from 1 and coincides with the row of natural numbers. Let us assume that at some time instant network occupancy is as depicted in Fig. 2a. Each square represents one packet and number in the square is its sequential number.

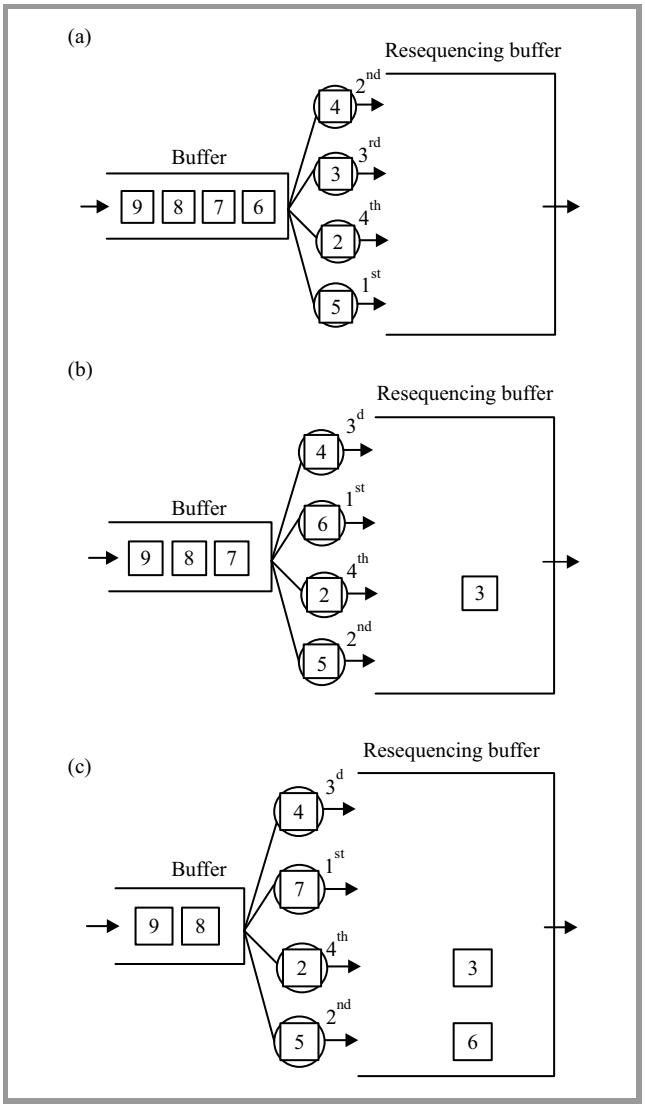


Fig. 2. Example of how resequencing system's content may evolve in one step.

After each service completion let one label customers in servers according to the order in which they occupied servers. Let us refer to the customer, which was the last to enter server as the 1st level customer. Customer which entered server just before the 1st level customer is referred to as the 2nd level level customer. The 3rd level customer is the one which entered server just before the 2nd level customer. Finally the 4th level customer was the first (among

other three in service) to enter server. In Fig. 2a one can see the corresponding labeling.

Assume the next service to happen is the completion of service of the 3rd level customer. It will not leave the system but occupy one place in the resequencing buffer, and customer from the buffer with the sequential number 6 will occupy free server (Fig. 2b). At this time instant authors have to re-label customers in servers because the order in which they occupied servers had changed. Now customer with the sequential number 6 becomes the 1st level customer. New labeling can be seen in Fig. 2b. If the next service completion is the service completion of the 1st level customer, then it joins the resequencing buffer and customer from the buffer with sequential number 7 occupies free server. From Fig. 2c it can be seen, that though two customers reside together in the resequencing buffer and can constitute a single queue, time until each of them leaves the system is different. Indeed customer with the sequential number 3 has to wait only for one customer (one service completion) before it can leave the system and customer with the sequential number 6 has to wait for three customers before it may depart from the system. By this attribute – number of service completions, which customer residing in resequencing buffer has to wait for before it can leave the system – by which the single queue in resequencing buffer can be partitioned into several separate interconnected queues (see Fig. 1b).

One may continue the example further and arrive, for example, to the network occupancy as depicted in Fig. 3. In figure one can see how packets in RB are distributed among different queues. Partitioning of the RB into several queues gives a more detailed view of its dynamics and leads to number of interesting questions:

- what is the joint stationary distribution of all queues in the system?
- are there any dependencies between queues' sizes?
- what happens with queues in the RB if N grows without bound?
- what influence does service rate (distribution) has on queues' sizes in the RB, etc?

In this paper the authors focus on the first two questions.

In system with $N \geq 2$ servers, if all of them are busy, then resequencing buffer can be partitioned into $N-1$ queues (see Fig. 3 as example for $N = 4$). If the number of busy servers is less than N, then the number of queues in the resequencing buffer is equal to the number of busy servers. The analysis of the joint stationary distribution of number of customers even in simple cases with Poisson flow and homogeneous exponential servers turns out to be a challenging task. In [16] for $M/M/3/\infty$ queue followed with infinite resequencing buffer one obtains expressions for joint stationary distribution of number of customers in buffer and servers, and number of customers in each of two queues in resequencing buffer both in explicit form and in terms

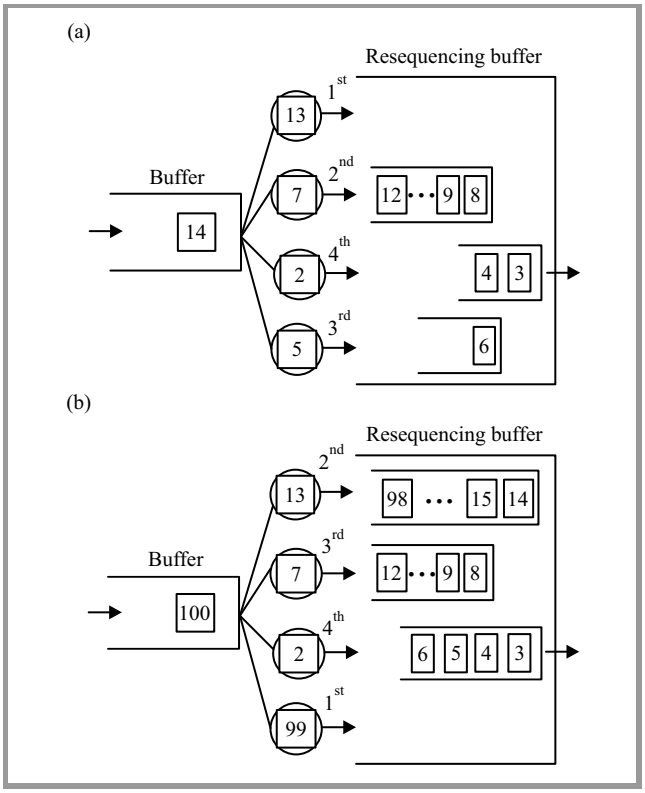


Fig. 3. Examples of resequencing system's contents at two different time instants.

of generating functions. In [17] for $M/M/N/\infty$ queue followed with infinite resequencing buffer there was obtained algorithm for recursive computation joint stationary distribution of number of customers in buffer and servers, and sum of number of customers in two, three, ..., and $N-1$ queues in resequencing buffer.

In this paper by modeling the disordering of packets by $M/M/N/\infty$ queue followed with the RB of infinite capacity we propose the methodology for computation of joint stationary distribution of number of customers in buffer and servers, and number of customers in each queue in the RB. Here it is shown that in the general case $N > 3$ the joint stationary distribution can be computed recursively. The special case of this methodology has already been used in [16]. The authors note that the joint distribution for the general case can be also obtained algorithmically in terms of the generating functions (as it is shown in [18]), but that results are, as usual, hardly applicable for the computation of the joint distribution itself.

The next Section 2 is devoted to the description of the system and the necessary notation. In Section 3 it is shown how one can obtain the system of equilibrium equations for joint stationary distribution of number of customers in buffer and servers, and number of customers in each queue in resequencing buffer. The description of the solution algorithm comes after. Several numerical examples are given in Section 4. In the conclusion, one provides a short discussion of obtained results and outlines possible directions of further research.

2. System Description and Notation

Consider a queueing system with $3 < N < \infty$ servers, infinite capacity buffer, incoming Poisson flow of customers of intensity λ, exponential service time distribution in each server with parameter μ and RB of infinite capacity. Customer upon entering the system obtains a sequential number and joins the buffer. Without the loss of generality authors suppose that the sequence starts from 1 and coincides with the row of natural numbers, i.e. customer upon entering the empty system receives number 1, the next one – number 2 and so on and so forth. Customers leave the system strictly in the order of their arrival. Thus, after customer's arrival it enters server (if there are any idle) or remains in the buffer for some time and then receives service from one of the servers. If at the moment of its service completion there are no customers in the system or all other customers present at that moment in the buffer and in all other servers have greater sequential numbers it leaves the system. Otherwise, it occupies a place in the RB. Each customer from the RB leaves it if and only if its sequential number is less than sequential numbers of all other customers present in the system. It may be noticed that the customers may leave the RB in groups. For example, in Fig. 3a if customer with sequential number 2 is the next to finish service then it leaves the system at one together with customer number 3 and 4.

In order to correctly define the partitioning of the RB into several queues the following approach is used. Assume there are n, $n = \overline{1,N}$, busy servers in the system. Each time any server becomes free or busy the customers in servers are labeled according to the order in which they occupied servers. Let us refer to the customer which was the last to enter server as the 1st level customer. Customer, which entered server right before the 1st level customer, is referred to as the 2nd level level customer. The 3rd level customer is the one which entered server before the 2nd level customer. Proceeding in similar manner customer, which was the first (among n) to enter server, is referred to as the n^{th} level the customer. Customers which reside in the RB form $(n-1)$ separate queues in the following way. Customers which entered the RB between the 1st level and the 2nd level customer form queue #1. Customers which entered the RB between the 2nd level and the 3rd level customer form queue #2 and so on. Customers which entered the RB between the $(n-1)$ level and the n^{th} level customer form queue $\#(n-1)$. Example of such partitioning of the RB into separate queues in case when $N = 4$ is given in Fig. 3.

Let us denote by $\xi(t)$ – the number of customers in buffer and servers at instant t, and by $\eta_i(t)$ — the number of customers in i-th queue in resequencing buffer at instant t. Then the Markov process $\zeta(t)$, describing the stochastic behavior of the system, is

$$\zeta(t) = \{(\xi(t), \eta_1(t), \eta_2(t), \ldots, \eta_{N-1}(t)), \quad t \geq 0\}.$$

In case $\xi(t) = 0$, all components of the process $\zeta(t)$ except for the first one are omitted; in case $\xi(t) = n$, $n = \overline{1,N-2}$, last $N-1-n$ components are omitted. The state space of the process $\zeta(t)$ has the form

$$\mathscr{X} = \{0\} \cup \{(1,i_1),\ i_1 \geq 0\} \cup \{(2,i_1,i_2),\ i_1,i_2 \geq 0\} \cup \ldots \\ \cup \{(n,i_1,i_2,\ldots,i_{N-1}),\ n \geq N-1, i_1,i_2,\ldots,i_{N-1} \geq 0\}.$$

Let us denote by p_n, $n \geq 0$, the stationary probabilities of the fact, that there are n customer in buffer and servers (customers in the RB are not taken into account), i.e.

$$p_n = \lim_{t\to\infty} \mathbf{P}\{\xi(t) = n\}.$$

One can notice that p_n, $n \geq 0$, are determined by the same equations as in the simple $M/M/N/\infty$ queue (see e.g. [19]):

$$p_0 = \left(\sum_{i=0}^{N-1} \frac{\rho^i}{i!} + \frac{\rho^N}{(N-1)!\,(N-\rho)} \right)^{-1},\ \rho = \lambda/\mu, \quad (1)$$

$$p_i = \frac{\rho^i}{i!} p_0, \quad i = \overline{1,N}, \quad (2)$$

$$p_i = \frac{\rho^i}{N!N^{i-N}} p_0 = \tilde{\rho}^{i-N} p_N,\ \tilde{\rho} = \rho/N, \quad i \geq N+1. \quad (3)$$

It can be observed that for the stationary probabilities of the considered system with resequencing to exist it is necessary and sufficient that the condition (necessary and sufficient) for the existence of probabilities p_n is fulfilled, i.e. $\rho/N < 1$ must hold.

Let us denote by $p_{n;i_1,\ldots,i_m}$, $m = \overline{1,N-1}$, $i_1,\ldots,i_m \geq 0$, the stationary probability of the fact that there are $n \geq N$ customers in buffer and servers, and in the RB there are i_1 customers in queue #1, i_2 customers in queue #2, …, i_m customers in queue #m, that is

$$p_{n;i_1,\ldots,i_m} = \lim_{t\to\infty} \mathbf{P}\{\xi(t) = n, \eta_1(t) = i_1, \ldots, \eta_m(t) = i_m\}, \\ m = \overline{1,N-1},\ n \geq N,\ i_1,\ldots,i_m \geq 0.$$

If the number of busy servers is $n < N$, then we denote by $p_{n;i_1,\ldots,i_m}$, $m = \overline{1,n}$, $i_1,\ldots,i_m \geq 0$, the stationary probability of same fact, that is

$$p_{n;i_1,\ldots,i_m} = \lim_{t\to\infty} \mathbf{P}\{\xi(t) = n, \eta_1(t) = i_1, \ldots, \eta_m(t) = i_m\}, \\ n = \overline{1,N-1},\ m = \overline{1,n},\ i_1,\ldots,i_m \geq 0.$$

The only difference between cases $n \geq N$ and $n < N$ is that in the former case number of queues in RB may vary from 1 to $N-1$ and in the latter case it may vary only from 1 to n. From the definition of the joint probabilities it follows that the stationary distribution p_n, $n \geq 1$, can be calculated from $p_{n;i_1,\ldots,i_m}$ by summation

$$p_n = \mathbf{P}\left\{ \zeta(t) \in \bigcup_{i_1,\ldots,i_n \geq 0}^{\infty} (n,i_1,i_2,\ldots,i_n) \right\} \\ = \sum_{i_1,\ldots,i_n=0}^{\infty} p_{n;i_1,\ldots,i_n},\ n = \overline{1,N-2},$$

$$p_n = \mathbf{P}\left\{ \zeta(t) \in \bigcup_{i_1,\ldots,i_{N-1} \geq 0}^{\infty} (n,i_1,i_2,\ldots,i_{N-1}) \right\} \\ = \sum_{i_1,\ldots,i_{N-1}=0}^{\infty} p_{n;i_1,\ldots,i_{N-1}},\ n \geq N-1.$$

3. System of Equilibrium Equations

In order to obtain the balance equations let us consider step-by-step different partitions of the state space and use rate-in-rate-out principle (local balance). Notice that if one sums up, say the probability $p_{N;i_1,\ldots,i_{N-1}}$, over all possible values of $i_2,\ldots,i_{N-1}$, then one obtains probability of the state set

$$\bigcup_{i_2,\ldots,i_{N-1}\geq 0}^{\infty}(N,i_1,i_2,\ldots,i_{N-1})\,,$$

i.e. probability of the fact that there are N customers in buffer and servers, and queue #1 contains i_1 customers (irrespectively of the number of customer in the queues #2, #3 ... #$(N-1)$ in the RB). For the probabilities of such state sets it is possible to analyse one-step transitions and write out the balance equations, that eventually lead to the determination of the whole joint distribution.

Denote by $p_{n;i_1,\ldots,i_m}$, $n\geq 2$, $m=\overline{1,\min(n-1,N-2)}$, $i_1,\ldots,i_m\geq 0$, the probability of the fact that there are n customers in the queue and servers, and in the RB there are i_1 customers in queue #1, i_2 customers in queue #2, ..., i_m customers in queue #m, that is

$$p_{n;i_1,\ldots,i_m}=\sum_{i_{m+1},\ldots,i_n=0}^{\infty}p_{n;i_1,\ldots,i_m,i_{m+1},\ldots,i_n}\,, \quad (4)$$
$$n=\overline{2,N-2},\ m=\overline{1,n-1},\ i_1,\ldots,i_m\geq 0,$$

$$p_{n;i_1,\ldots,i_m}=\sum_{i_{m+1},\ldots,i_{N-1}=0}^{\infty}p_{n;i_1,\ldots,i_m,i_{m+1},\ldots,i_{N-1}}\,, \quad (5)$$
$$n\geq N-1,\ m=\overline{1,N-2},\ i_1,\ldots,i_m\geq 0\,.$$

Notice that Eqs. (4) and (5) define the probabilities not of a single state of the system but of the set of states. For example, probability $p_{N-2;i_1}$ defined by (4) is the probability of the fact that there are $N-2$ busy servers, the buffer is empty, and there are $i_1\geq 0$ customers in queue #1 in RB.

Balance equations for $p_{n;i_1,\ldots,i_m}$ will be written out in the following way. Firstly, one establishes equations for $p_{n;i_1}$, $n\geq N$, $i_1\geq 0$ and then for $p_{n;i_1}$, $n=\overline{N-1,1}$, $i_1\geq 0$. Secondly, one finds equations for $p_{n;i_1,i_2}$, $n\geq N$, $i_1,i_2\geq 0$ and then for $p_{n;i_1,i_2}$, $n=\overline{N-1,2}$, $i_1,i_2\geq 0$. After that one proceeds to $p_{n;i_1,i_2,i_3}$, $n\geq N$, $i_1,i_2,i_3\geq 0$ and $p_{n;i_1,i_2,i_3}$, $n=\overline{N-1,3}$, $i_1,i_2,i_3\geq 0$. This procedure continues until one arrives to $p_{n;i_1,\ldots,i_m}$, $n\geq N$, $i_1,\ldots,i_m\geq 0$ and $p_{N-1;i_1,\ldots,i_m}$, $i_1,\ldots,i_m\geq 0$.

For probabilities $p_{n;i_1}$, $n\geq N$, $i_1\geq 0$, the following equations hold

$$p_{n;0}(\lambda+N\mu)=p_{n-1;0}\lambda+p_{n+1}(N-1)\mu,\quad n\geq N, \quad (6)$$

$$p_{n;i_1}(\lambda+N\mu)=p_{n-1;i_1}\lambda+p_{n+1;i_1-1}\mu,\ n\geq N,\ i_1\geq 1\,. \quad (7)$$

Equation (6) is derived as follows. Assume that the system is in one of the states when there are $n\geq N$ customers in the buffer and servers and queue #1 in the RB is empty. The considered state set is $\bigcup_{i_2,\ldots,i_{N-1}\geq 0}^{\infty}(n,0,i_2,\ldots,i_{N-1})$ and the probability of this state set is $p_{n;0}$ according to Eq. (5). The system can leave this state set if the service completion or arrival occurs, i.e. the rate-out flow is $p_{n;i_1}(\lambda+N\mu)$. The system can enter this state set if:

- there were $n+1$ customers in the buffer and servers (which happens with probability p_n) and service completion of any of the N customers except for the 1^{st} *level* customer occurred, which happens with rate $(N\mu)\frac{(N-1)}{N}=(N-1)\mu$;
- there were $n-1$ customers in the buffer and servers and queue #1 in the RB was empty, which happens with the probability $p_{n+1;0}$ according to Eq. (5), and an arrival occurred.

Thus the rate-in flow is $p_{n-1;0}\lambda+p_{n+1}(N-1)\mu$. By equating rate-out and rate-in flows one obtains Eq. (6).

In order to explain Eq. (7) assume that the system is in one of the states when there are $n\geq N$ customers in the buffer and servers and there are $i_1\geq 1$ customers in queue #1 in the RB. The considered state set is $\bigcup_{i_2,\ldots,i_{N-1}\geq 0}^{\infty}(n,i_1,i_2,\ldots,i_{N-1})$ and the probability of this state set is $p_{n;i_1}$ according to Eq. (5). The rate-out flow from this state set equals $p_{n;i_1}(\lambda+N\mu)$. The system can enter this state set with an arrival if there were $n-1$ customers in the buffer and servers and i_1 customers in queue #1 in the RB, which happens with the probability $p_{n-1;i_1}$ according to Eq. (5). The system can also enter this state set with a service completion from state set when there were $n+1$ customers in the queue and servers, and queue #1 in the RB contained i_1-1 customers, which happens with the probability $p_{n+1;i_1-1}$ according to Eq. (5) and service completion of the 1st level customer occurred (which happens with rate $(N\mu)\frac{1}{N}=\mu$). By equating rate-out and rate-in flows one obtains Eq. (7).

Probabilities $p_{N-1;i_1}$, $i_1\geq 0$, are governed by the following equations

$$p_{N-1;0}\big[\lambda+(N-1)\mu\big]=p_{N-2}\lambda+p_N(N-1)\mu\,, \quad (8)$$

$$p_{N-1;i_1}\big[\lambda+(N-1)\mu\big]=p_{N;i_1-1}\mu,\quad i_1\geq 1\,. \quad (9)$$

Probabilities $p_{n;i_1}$, $n=\overline{1,N-2}$, $i_1\geq 0$, are given by

$$p_{n;0}(\lambda+n\mu)=p_{n-1}\lambda+p_{n+1;0}n\mu,\quad n=\overline{1,N-2}, \quad (10)$$

$$p_{n;i_1}(\lambda+n\mu)=p_{n+1;i_1}n\mu\quad+\sum_{j=0}^{i_1-1}p_{n+1;i_1-j-1,j}\mu\,,$$
$$n=\overline{1,N-2},\quad i_1\geq 1\,. \quad (11)$$

For probabilities $p_{n;i_1,\ldots,i_m}$, $m=\overline{2,N-1}$, $n\geq m$, $i_1,\ldots,i_{N-1}\geq 0$, one can write out the system of balance equations in the general form. It holds

$$p_{n;0,i_2,\ldots,i_m}(\lambda+N\mu)=p_{n-1;0,i_2,\ldots,i_m}\lambda+$$
$$+p_{n+1;i_2,\ldots,i_m}(N-m)\mu+\sum_{j=0}^{i_2-1}p_{n+1;j,i_2-j-1,i_3,\ldots,i_m}\mu+\ldots$$
$$+\sum_{j=0}^{i_m-1}p_{n+1;i_2,\ldots,i_{m-1},j,i_m-j-1}\mu,\ \ n\geq N,\ i_2,\ldots,i_m\geq 0\,, \quad (12)$$

$$p_{n;i_1,\ldots,i_m}(\lambda+N\mu)=p_{n-1;i_1,\ldots,i_m}\lambda+p_{n+1;i_1-1,i_2,\ldots,i_m}\mu,$$
$$n\geq N,\quad i_1\geq 1,\quad i_2,\ldots,i_m\geq 0, \quad (13)$$

$$p_{N-1;0,i_2,\ldots,i_m}\big[\lambda+(N-1)\mu\big]=p_{N-2;i_2,\ldots,i_m}\lambda+$$
$$+p_{N;i_2,\ldots,i_m}(N-m)\mu+\sum_{j=0}^{i_2-1}p_{N;j,i_2-j-1,i_3,\ldots,i_m}\mu+\ldots$$
$$+\sum_{j=0}^{i_m-1}p_{N;i_2,\ldots,i_{m-1},j,i_m-j-1}\mu,\quad i_2,\ldots,i_m\geq 0, \quad (14)$$

$$p_{N-1;i_1,\ldots,i_m}\big[\lambda+(N-1)\mu\big]=p_{N;i_1-1,i_2,\ldots,i_m}\mu,$$
$$i_1\geq 1,\quad i_2,\ldots,i_m\geq 0, \quad (15)$$

$$p_{n;0,i_2,\ldots,i_m}(\lambda+n\mu)=p_{n+1;0,i_2,\ldots,i_m}(n-m+1)\mu+$$
$$+p_{n-1;i_2,\ldots,i_m}\lambda+\sum_{j=0}^{i_2-1}p_{n+1;0,j,i_2-j-1,i_3,\ldots,i_m}\mu+\ldots+$$
$$+\sum_{j=0}^{i_m-1}p_{n+1;0,i_2,\ldots,i_{m-1},j,i_m-j-1}\mu,$$
$$m\neq N-1,\quad n=\overline{m,N-2},\quad i_2,\ldots,i_m\geq 0, \quad (16)$$

$$p_{n;i_1,\ldots,i_m}(\lambda+n\mu)=p_{n+1;i_1,\ldots,i_m}(n-m+1)\mu+$$
$$+\sum_{j=0}^{i_1-1}p_{n+1;j,i_1-j-1,i_2,\ldots,i_m}\mu+\ldots$$
$$+\sum_{j=0}^{i_m-1}p_{n+1;i_1,\ldots,i_{m-1},j,i_m-j-1,}\mu,$$
$$m\neq N-1, n=\overline{m,N-2},\quad i_1\geq 1,\quad i_2,\ldots,i_m\geq 0. \quad (17)$$

In Eqs. (12)–(17) for the sake of brevity agreement is used that $\sum_{i=0}^{-1}a_i=0$. The system of Eqs. (12)–(17) is derived using the same argumentation, which is used above for Eqs. (6)–(7).

For the fixed value of N system, Eqs. (6)–(17) can be solved recursively. Computation of $p_{n;i_1,\ldots,i_m}$ consists of $N-1$ steps. The first step consists of the following sequential computations. Firstly one computes probabilities p_n, $n\geq 0$ using Eqs. (1)–(3). Then one finds probability $p_{N-1;0}$ from Eq. (8), probabilities $p_{n;0}$, $n=\overline{N-2,1}$, from Eq. (10) and then probabilities $p_{n;0}$, $n\geq N$, from Eq. (6). Secondly one computes probability $p_{N-1;0,0}$ from Eq. (14), probabilities $p_{n;0,0}$, $n\geq N$, from Eq. (12), and probabilities $p_{n;0,0}$, $n=\overline{N-2,2}$ from Eq. (16). Thirdly for each $i\geq 1$ using Eqs. (9) and (7) one finds probabilities $p_{n;i}$, $n\geq N-1$.

The second step starts with computation of probability $p_{N-2-k;1-k}$, $k=\overline{0,\min(0,N-1)}$ from Eq. (11). Then starting from $i_1=0$ one computes probabilities $p_{N-1;i_1,i_2}$, $i_1+i_2=1$, from Eqs. (14) and (15). Finally, starting from $i_1=1$, one finds probabilities $p_{n;i_1,i_2}$, $n\geq N$, $i_1+i_2=1$, from Eq. (13).

The third step starts with computation of probabilities $p_{N-2-k;2-k}$, $k=\overline{0,\min(1,N-2)}$, from Eq. (11). Then starting from $i_1=0$, using Eqs. (14) and (15) one finds probabilities $p_{N-1;i_1,i_2}$, $i_1+i_2=2$. After that starting from $i_1=2$, one computes probabilities $p_{n;i_1,i_2}$, $n\geq N$, $i_1+i_2=2$, from Eq. (13). Finally using Eqs. (16) and (17) one obtains probabilities $p_{N-2;i_1,i_2}$, $i_1+i_2=1$, starting from $i_1=0$ and then from Eqs. (12) and (13), firstly, one computes probabilities $p_{N-1;i_1,i_2,i_3}$, $n\geq N-1$, $i_1+i_2+i_3=1$, starting from $i_3=1$ and, secondly, one computes probabilities $p_{n;i_1,i_2,i_3}$, $n\geq N$, $i_1+i_2+i_3=1$, starting from $i_3=1$.

The fourth step starts with computation of probabilities $p_{N-2-k;3-k}$, $k=\overline{0,\min(2,N-3)}$, from Eq. (11), which is followed by computation of probabilities $p_{N-1;i_1,i_2}$, $i_1+i_2=3$, starting from $i_1=0$, etc.

The algorithm for the computation of the whole joint stationary distribution, wherefrom the general pattern can be seen, is given below in pseudo code.

Algorithm 1: Computation of the joint stationary distribution

```
for c ≥ 0 do
    Compute p_{N-2-k;c+1-k}, k = 0,min(c,N-c-1)
     using Eq. (11).
    Compute p_{N-1;i_1,i_2}, i_1+i_2 = c+1, starting from
    i_2 = c+1 using Eqs. (14) and (15).
    Compute p_{n;i_1,i_2}, n ≥ N, i_1+i_2 = c+1, from Eq. (13).
    if c = 1 then
        Compute p_{N-2;i_1,i_2}, i_1+i_2 = c, starting from
        i_2 = c using Eqs. (16) and (17).
        Compute p_{N-1;i_1,i_2,i_3}, i_1+i_2+i_3 = c, starting
        from i_3 = c using Eqs. (12) and (13).
        Compute p_{n;i_1,i_2,i_3}, n ≥ N, i_1+i_2+i_3 = c, starting
        from i_3 = c using Eqs. (12) and (13).
    end if
    if c = 2 then
        Compute p_{N-3;i_1,i_2}, i_1+i_2 = c-1, using Eq. (16)
        and (17).
        Compute p_{N-2;i_1,i_2,i_3}, i_1+i_2+i_3 = c-1, starting
        from i_3 = c-1 using Eq. (16) and (17).
        Compute p_{N-1;i_1,i_2,i_3,i_4}, i_1+i_2+i_3+i_4 = c-1,
        starting from i_4 = c-1 using Eq. (12) and (13).
        Compute p_{n;i_1,i_2,i_3,i_4}, n ≥ N, i_1+i_2+i_3+i_4 =
        c-1, starting from i_4 = c-1 using Eqs. (12)
        and 13).
    end if
    if c = 3 then
        ...
    end if
    ...
end for
```

4. Numerical Examples

Extensive numerical experiments were carried out with recursive algorithm described in the previous section, which involved computation of the joint stationary distribution of number of customers in buffer and servers, and number

of customers in queues in the RB, as well as several important performance characteristics. The complexity of the algorithm grows very fast as number of servers increases the computation of the whole joint stationary distribution becomes very slow.

Below several numerical results are given, which show different aspects of the in-depth-behavior of the queues in the RB.

It is assumed that number of servers is $N = 4$ and the service rate is $\mu = 1$. The mean and variance of the number of customers in the RB and correlation coefficient of the number of customers in the buffer and each queue in the RB, as functions of the system's load ρ/N, are depicted in Figs. 4 and 5.

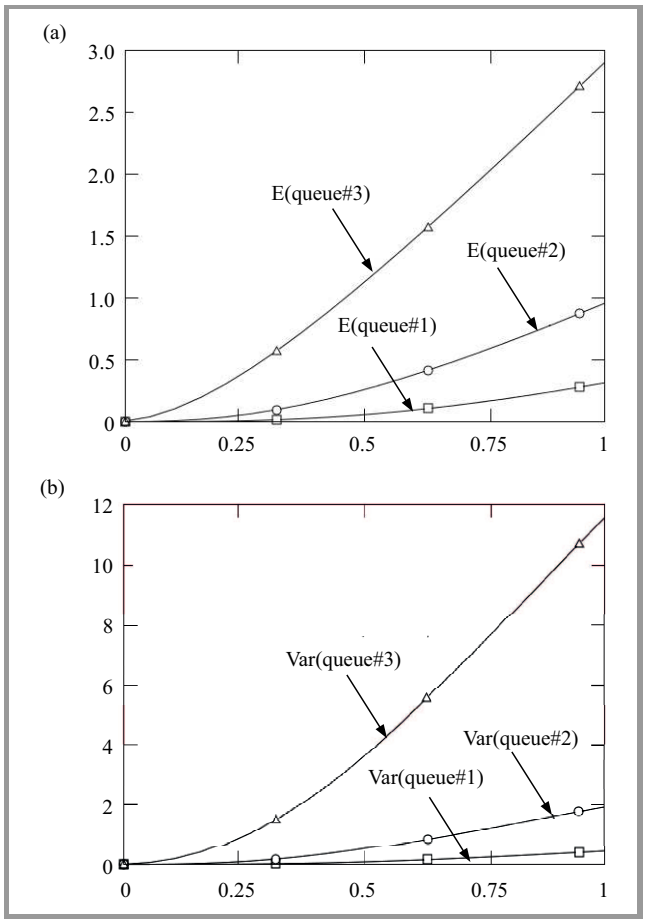


Fig. 4. Dependency of: (a) mean number of customers in each queue in the RB, (b) variance of the number of customers in each queue in the RB, on the system's load ρ/N.

From Fig. 5 it follows that number of customers in queues are weakly correlated and become uncorrelated as the value of load approaches critical value of 1. Conducted experiments show that the same result holds when one considers more general model with MAP arrivals and PH service times (for $N = 2$). From Fig. 4 it can be also observed that the mean lengths of queues in the RB are finite which follows from Little's law. In fact all the moments of the lengths of the queues in the RB are finite. The mean queue

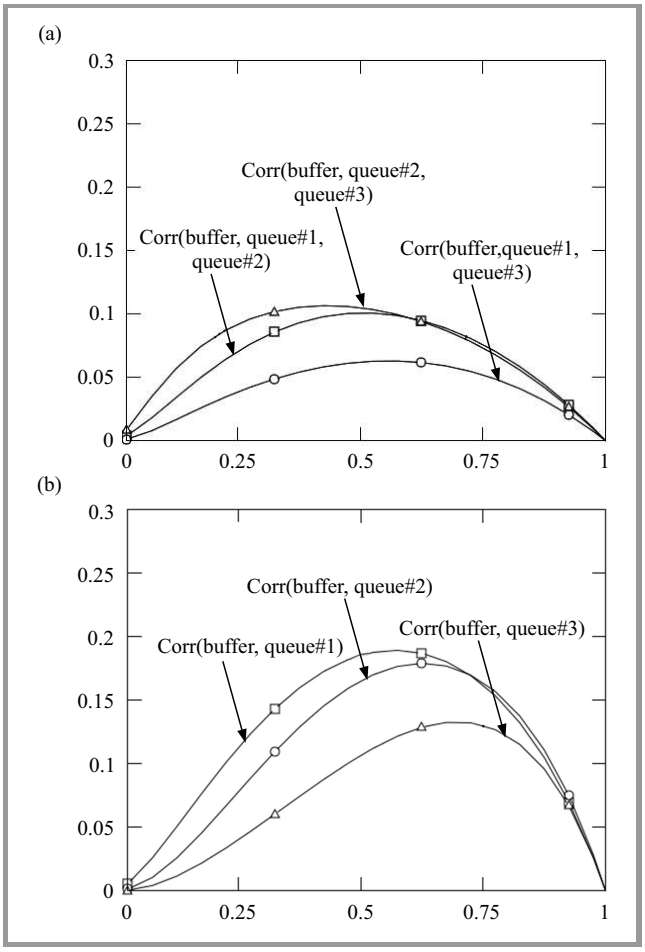


Fig. 5. Dependency of correlation coefficient: (a) number of customers in queues in the RB (pairwise), (b) number of customers in the queue and *each* queue in the RB (pairwise), on the system's load ρ/N.

sizes in the RB are related to each other by inequalities E(queue #3) $>$ E(queue #2) $>$ E(queue #1). The same holds for the variances and, in general, for any $N \geq 3$ such inequalities hold. Intuitively this can be explained by the fact that queue #1 exists in the RB only when the number of busy servers is at least $N-1$, whereas queue $\#(N-1)$ already appears when two servers become busy. The mean queue size in the RB if one sees it as a single queue is the sum of mean queue sizes of queue #1, queue #2, ... and queue $\#(N-1)$. This suggests that the moments of queue $\#(N-1)$ size, say mean, may serve as another performance characteristic of the system with resequencing because eventually its dynamics shows how much disordering is incurred by the network.

5. Conclusion

In this paper the authors have considered probably the simplest model for the resequencing issue using queueing-theoretic approach, which allowed one to look "deeper" into the dynamics of the RB. It turns out that the joint stationary distribution of all queues can be computed recursively and,

as expected, queues in the RB are not equivalent, although surprisingly weakly correlated. The mechanism according to which the queues in the RB are built allows one to use such characteristic as queue-size moments of the queue $\#(N-1)$ in the RB as another performance indicator of the whole system with resequencing. There are many possible ramifications of the system, which may make it more suitable for practical needs. Probably the Poisson arrival (and exponential service) assumption should not be the first ones to be relaxed, because, for example, in $MAP|PH|2|\infty$ queue followed with resequencing buffer joint stationary distribution can be also found in recursive way and the weak correlation of queue-sizes is preserved. The introduction of heterogeneity and rule for choosing idle servers (say, i-th server with probability p_i, or i-th server with probability $p_{i,j}$ if j servers are busy) is the more promising direction of research.

Acknowledgements

The research is supported by the Russian Foundation for Basic Research (project 15-07-03007 and 13-07-00223).

References

[1] O. Boxma, G. Koole, and Z. Liu, "Queueing-theoretic solution methods for models of parallel and distributed systems", in *Performance Evaluation of Parallel and Distributed Systems: Solution Methods. Proceedings of the Third QMIPS Workshop CWI*. Tract 105 & 106. Amsterdam, Netherlands: Centrum voor Wiskunde en Informatica, 1994, pp. 1–24.

[2] B. Dimitrov, D. Green Jr., V. Rykov, and P. Stanchev, "On performance evaluation and optimization problems in queues with resequencing", in *Advances in Stochastic Modelling*, J. Artalejo and A. Krishnamoorty, Eds. New Jersey: Notable Publications Inc., 2002, pp. 55–72.

[3] K. Leung and V. O. K. Li, "A resequencing model for high-speed packet-switching networks", *J. Comp. Commun*., vol. 33, no. 4, pp. 443–453, 2010.

[4] S. Agrawal and R. Ramaswamy, "Analysis of the resequencing delay for M/M/m systems", in *Proc. ACM Sigmetrics Conf. Measur. Model. Comp. Syst*, Banff, Alberta, Canada, 1987, vol. 15, pp. 27–35.

[5] Y. Xia and D. N. C. Tse, "On the large deviations of resequencing queue size: 2-M/M/1 Case", *IEEE Trans. Inform. Theory*, vol. 54, no. 9, pp. 4107–4118, 2008.

[6] F. Baccelli, E. Gelenbe, and B. Plateau, "An end to end approach to the sequencing problem", Rapports de Recherche, no. RR-0097, INRIA, Nov. 1981 [Online]. Available: https://hal.inria.fr/inria-00076464

[7] S. Chowdhury, "Distribution of the total delay of packets in virtual circuits", in *Proc. 10th Ann. Joint Conf. IEEE Comp. Commun. Soc. INFOCOM'91*, Bal Harbour, FL, USA, 1991, vol. 2. pp. 911–918.

[8] S. Chakravarthy, S. Chukova, and B. Dimitrov, "Analysis of MAP/M/2/K queueing model with infinite resequencing buffer", *J. Perform. Eval.*, vol. 31. no. 3–4. pp. 211–228, 1998.

[9] C. De Nicola, A. Pechinkin, and R. Razumchik, "Stationary characteristics of homogenous Geo/Geo/2 queue with resequencing in discrete time", in *Proc. 27th Eur. Conf. Model. Simul. ECMS 2013*, Alesund, Norway, 2013, pp. 594–600.

[10] R. Gogate and S. Panwar, "Assigning customers to two parallel servers with resequencing", *IEEE Commun. Lett.*, vol. 3, no. 4, pp. 119–121, 1999.

[11] T. Huisman and R. J. Boucherie, "The sojourn time distribution in an infinite server resequencing queue with dependent interarrival and service times", *J. Appl. Probab*., vol. 39, no. 3, pp. 590–603, 2002.

[12] M. Jain and G. C. Sharma, "Nopassing multiserver queue with additional heterogeneous servers and inter-dependent rates", in joint conference *5th Canadian Conf. in Applied Statist. STATISTICS 2011* and *20th Conf. Forum for Interdiscip. Mathem. "Interdisciplinary Mathematical & Statistical Techniques" IMST 2011*, Montreal, Quebec, Canada, 2011.

[13] M. Lelarge, "Packet reordering in networks with heavy-tailed delays", *Mathem. Methods Oper. Res*., vol. 67, no. 2, pp. 341–371, 2008.

[14] I. Caraccio, A. V. Pechinkin, and R. V. Razumchik, "Stationary characteristics of MAP—PH—2 resequencing queue", in *Proc. 1st Eur. Conf. on Queueing Theory ECQT 2014*, Ghent, Belgium, 2014, pp. 42–46.

[15] T. Takine, J. Ren, and T. Hasegawa, "Analysis of the resequencing buffer in a homogeneous M/M/2 Queue", *Perform. Eval*., vol. 19, no. 4, pp. 353–366, 1994.

[16] A. V. Pechinkin, I. Caraccio, and R. V. Razumchik, "Joint stationary distribution of queues in homogenous M/M/3 queue with resequencing", in *Proc. 28th Eur. Conf. Model. Simul. ECMS 2014*, Brescia, Italy, 2014, pp. 558–564.

[17] I. Caraccio, A. V. Pechinkin, and R. V. Razumchik, "On joint stationary distribution in exponential multiserver reordering queue", in *Proc. 12th Int. Conf. Numerical Anal. Appl. Mathem. ICNAAM 2014*, Rhodes, Greece, 2014.

[18] A. V. Pechinkin and R. V. Razumchik, "Joint stationary distribution of m queues in the n-server queueing system with reordering", *Inform. and its Appl*., vol. 9, no. 3, pp. 26–32, 2015 (in Russian).

[19] J. Riordan, *Stochastic Service Systems*. New York: Wiley, 1962.

Alexander Pechinkin (1946–2014) held the Ph.D. of Sciences in Physics and Mathematics and has principal scientist at the Institute of Informatics Problems of the Russian Academy of Sciences. He held a Professor position at the Peoples' Friendship University of Russia. He was the author of more than 200 papers in the field of theoretical and applied probability theory.

Rostislav Razumchik received his Ph.D. in Physics and Mathematics in 2011. At present he is a senior scientist at Institute of Informatics Problems of FRC CSC RAS and also holds associate professor position at the Peoples' Friendship University of Russia. His current research activities focus on queueing theory and scheduling.

E-mail: rrazumchik@ipiran.ru
Institute of Informatics Problems
Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences,
Vavilova, 44-2
119333 Moscow, Russia

Paper

Multicast Connections in Wireless Sensor Networks with Topology Control

Maciej Piechowiak[1], Krzysztof Stachowiak[2], and Tomasz Bartczak[2]

[1] *Institute of Mechanics and Applied Computer Science, Kazimierz Wielki University, Bydgoszcz, Poland*
[2] *Faculty of Electronics and Telecommunications, Poznan University of Technology, Poznan, Poland*

Abstract—The article explores the quality of multicast trees constructed by heuristic routing algorithms in wireless sensor networks where topology control protocols operate. Network topology planning and performance analysis are crucial challenges for wire and wireless network designers. They are also involved in the research on routing algorithms, and protocols for these networks. In addition, it is worth to emphasize that the generation of realistic network topologies makes it possible to construct and study routing algorithms, protocols and traffic characteristics for WSN networks.

Keywords—multicasting, routing, wireless sensor networks.

1. Introduction

Wireless sensor networks (WSN) are communication networks composed of the several autonomous devices that use sensors to monitor physical or environmental conditions, such as temperature, vibration, pressure, stress, etc. WSN network nodes are equipped with sensors, microprocessors and transmitting and receiving devices with short-range transmit power that exchange values of measured parameters. The nodes create a global knowledge base of the examined parameters in monitored area. The user has an access to the database through one or more nodes constituting the network gateways.

Most of the problems associated with the implementation of services operating in the wireless sensor networks coincides with the challenges of all the ad hoc network. In the case of WSN networks, the energy consumption reduction by nodes becomes a priority. Devices that are members of WSN are up to miniaturized, resulting in relatively low battery capacity. Requirements for these networks relate to long lifetime. In most applications, charging or replacing batteries in such devices is impossible. The efficient use of energy resources available to sensor network nodes is one of the fundamental tasks for network designers [1]. Reduction of the energy consumed by radio communication is an important issue. Topology control mechanisms allow to maintain the lowest energy requirements of nodes and the maximum network throughput.

Due to a dynamic nature of ad hoc networks, traditional network routing protocols are not viable. Thus, nodes act both as the end system (transmitting and receiving data) and the router (allowing traffic to pass through), which results in multihop routing. Networks are *in motion*, i.e. nodes are mobile and may go out of range of other nodes in the network [2]. Nodes in these networks generate traffic to be forwarded to some other nodes (unicast) or a group of nodes (multicast) [3], [4]. Routing is then a challenging task due to the specific characteristics that distinguish wireless sensor networks from other wireless networks (i.e. mobile ad hoc networks or cellular networks).

The communication model for multicast connections provides an opportunity to reduce traffic by transmitting single packets through routers from the sender to the locations where hosts interested in receiving the data are located. Such a communication model requires special routing algorithms to be applied. These algorithms construct distribution trees (also known as multicast trees) so that packet transmission in the network can be executed.

Constrained Minimal Steiner Tree Problem (CMSTP) [5], [6] involves connecting a single source with multiple destinations in such way that one of the multiple metrics of the structure is minimal, under the restriction that the others do not violate required constraints. Therefore, when comparing different algorithms, one has to examine the costs of the multicast tree found in a given graph for given input parameters. The evaluation of the result is a non-trivial task. The metric which is to be minimized, should obviously be the lowest, but the constrained metrics may be of greater or lesser importance depending on assumed goals. The CMSTP problem can be considered both in wired and wireless networks (ad hoc, mesh, WSN, etc.).

The analysis of routing algorithms for multicast connections involves a concomitant definition of the way the network in which the algorithms are to be implemented will be represented. The problem of the appropriate representation of the network and its influence upon the efficiency and effectiveness of the algorithms under scrutiny is analyzed in [1], [7]. Reference [8] proves that in networks in which nodes are arranged and connected randomly, the effectiveness of multicast algorithms is at least twofold lower than that in hierarchical networks that reflect the properties of the internet network.

The article focuses on the quality of trees constructed by multicast routing algorithms in WSN networks that use topology control mechanisms. It starts with an overview of the available algorithms and evaluation techniques in

Section 2. Section 3 defines topology control mechanisms and basic parameters describing network topology while Section 4 contains simulation study and research methodology. In Section 5, the results of the simulation of the implemented topology control protocols along with their interpretation are described. Finally, Section 6 concludes the article.

2. Algorithms Description

2.1. Aggr_MLARAC Algorithm

The Aggregated Multi-dimensional LAgrangian Relaxation based Aggregated Cost (MLARAC) [9] is a variant of the multi-criterial unicast algorithm adopted for a multicast problem by performing an aggregation of the unicast results (paths from the source node to each of the destination nodes) into a multicast tree (a tree that spans all of the multicast group members). The MLARAC algorithm is on the other hand a multidimensional generalization of the LARAC algorithm [10].

The LARAC algorithm is a technique that utilizes Lagrangian relaxation in path optimization problem with a single constraint. The foundation of the Lagrangian relaxation is the maximization of the Lagrangian dual function. The merit of solving the Lagrangian relaxation problem is finding a maximum to a concave, piecewise linear function, which in the two criterion optimization boils down to a set of the segments of linear functions. The technique used in the LARAC algorithm boils down to finding consecutive approximations of the maximum by finding intersections of the pairs of the linear functions, which are guaranteed to intersect in the maximum neighborhood. The difficulty of finding the maximum is that the function is also piecewise linear, and thus the extreme cannot be found in the analytical way.

In the LARAC algorithm two distant segments of the function are found and based on the intersections of the lines to which they belong an approximation of the optimum is found. Based on the approximation, another segment, closer to the optimum is determined and used to find another intersection. This procedure is repeated, and after each step, a better approximation is obtained. The algorithm is guaranteed to find the optimum after finite number of steps.

The MLARAC algorithm is a generalization of the problem to multiple dimensions. Increasing the number of the optimization criteria increases the number of the dimensions of the Lagrangian dual function. In the MLARAC algorithm the intersection of lines has been replaced with the intersection of the hyperplanes. Also two problems that appear in the multidimensional space have been heuristically solved: the definition of the initial hyper-segments to intersect, and handling of the determined approximation. In the first case the one dimensional optimization is easier, because there are two sides of the hill of which the peak is to be found. There exists a robust way of selecting segments from the two sides of the hill. In the multidimensional case there is no straightforward equivalent method to determine the initial conditions. When the intersection of the hyperplanes is found presenting the new approximation of the result, there exists a condition that defines precisely, how it should be used in the consecutive intersections, but the exact equivalent for the multiple dimensions have not been found.

The aggregation of the results in the Aggregated MLARAC is performed by performing a union operation of the paths obtained from multiple MLARAC passes, from the source node to each of the destination nodes, which produces a subgraph containing all the multicast participants. Such structure is then pruned using the Prim algorithm [11]. A similar technique has been used earlier in [12].

2.2. HMCMC Algorithm

The Heuristic Multi-Constrained MultiCast (HMCMC) algorithm [13] is a relatively simple heuristic that has combines two main ideas. One is to handle the multiple criteria by aggregating them utilizing a nonlinear function:

$$m_{aggr}(t) = \max\left\{\frac{m_1(t)}{c_1}, \frac{m_2(t)}{c_2}, \ldots\right\}. \tag{1}$$

The second concept behind the HMCMC algorithm is performing the Dijkstra's algorithm multiple times [8] with the application of the metric aggregation. It defines the multicast participants as the source and the destination nodes separately. The Dijkstra's algorithm is performed from the source first, and if the shortest paths to all destinations that are obtained this way fulfill the constraints defined in the problem they are accepted as the result. Otherwise the Dijkstra's algorithm is performed from all the destinations towards which the constraints have not been met.

When relaxing the graph from the destination node towards the source node, the information from the initial algorithm pass is used to heuristically improve the quality of the selected path. Such an approach is computationaly cheap as the number of times that the Dijkstra's algorithm needs to be performed is the same as the number of the multicast participants. The experiments have shown that it also provides a feasible result in many cases.

2.3. RDP Algorithm

The RDP algorithm [14], named after the concept of the RenDezvouz Point, is an algorithm based on a simulation semantics applied a modified version of the Dijkstra's algorithm. The first of the two variations from the original algorithm is the multi-source approach. It is based on a slight change that the relaxation is initialized in multiple sources rather than one. As the result the labeling of the costs of reaching particular nodes is performed from different sources. The costs of reaching the nodes are stored separately so they don't override each other. This way if the relaxation is performed for the entire graph, the cost labels for each of the graph's nodes will store the information about reaching the given node from each of the initial nodes. If the initial nodes are the same as the multicast participants, then these cost labels may play role of

a weighted routing tables for each of the graph nodes. It is worth noting that in order to deal with multiple metric the same metric aggregation is utilized as in the HMCMC algorithm.
The second variation consists in the renaming of the original Dijkstra's algorithm's operations. It is performed in such a way that instead of describing the graph relaxation a simulation of the signal propagation in the graph is described. Introducing the notion of time into the consideration presents us with a means to define simultaneously of the node analysis operations.
Combining these two variations creates a context in which it is possible to treat the relaxations performed from the different sources as concurrently performed signal propagation processes. Therefore, it is possible to state that at a certain point of the simulation time the signals propagating from all of the sources have reached a given node. In such conditions the given node is said to be equally or similarly close (in the topological metric) to all of the source nodes. The thesis behind the RDP algorithm is that such nodes (further referred to as the *rendez vouz points* or the RDPs) may be considered as the middle points for the multicast trees with a considerable probability.
In [15] two variants of the above technique have been presented and analyzed with the regard to quality of the obtained results. The quality is defined as the costs of the obtained multicast trees. The research has shown that there was no significant difference between the variants therefore the more performant algorithm should be used as the representative implementation of the general RDP technique.

3. Topology Control in Wireless Sensor Networks

Topology control is the art of controlling decision-making mechanisms of network nodes, taking into account their transmission range, that aims at a generation of networks with specific properties. Unlike the wired networks with fixed network topologies each node in wireless sensor network is capable of changing network topology by adjusting its transmission range and choosing the neighboring nodes through which data will be directed. Thus the main goal of topology control mechanism implemented in wireless sensor networks is to keep the connectivity between nodes (and therefore routing) while maintaining the lowest energy requirements of nodes and the maximum throughput of the network.
Topology control mechanisms are used to ensure that certain parameters in the whole network are secure. Decisions in nodes are made locally to achieve a global goal. Both centralized and distributed techniques of topology control can be classified as topology control mechanisms.

3.1. Network Model

The wireless sensor network can be represented by unit disc graph and consist of set of nodes distributed in a two-dimensional plane. Each sensor is equipped in omni-directional antenna thus the transmission between nodes is possible only when they are in each other's transmission ranges (they can communicate directly) or two far away nodes can communicate through multi-hop wireless links using intermediate nodes. Such a graph is represented by an undirected, connected graph $G = (V, E)$, where V is a set of nodes and E is a set of links. The existence of the link $e = (u, v)$ between node u and v entails the existence of the link $e' = (v, u)$ for any $u, v \in V$ (corresponding to two-way links in communications networks). In the most common power-attenuation model, the power needed to support a link $e = (u, v)$ is $p(e) = ||u, v||^{\beta}$, where $||u, v||$ is the Euclidean distance between u and v, and β is a real constant between 2 and 5 dependent on the wireless transmission environment (path loss model) [1].

3.2. Protocols of Distributed Topology Control

A practical approach to topology control requires a creation of distributed protocols that operate locally, without the knowledge of the global state of the network, and generate topologies close to the optimal. Topology graphs should provide desirable properties of a network using symmetric edges and should be consistent (if these properties are satisfied in the graph of the maximum power that contains the edges resulting from the maximum transmit power of the nodes) [16]. It is desirable then to build a graph of the least degrees of nodes, which reduces the probability of interference in the network. It is also desirable to create optimal topology based on inaccurate information. Providing accurate information on the nodes is often too expensive, because it requires GPS receiver in each node of the network.
Topology control protocols based on the knowledge of the position of the nodes (called *location-based topology control*) are based on the assumption of available information to the nodes with a very precise location of the neighboring nodes. The easiest way to satisfy this condition is to equip the nodes with GPS receivers, which are expensive, but provide reliable and accurate information. An alternative solution is to use techniques that make an approximation of the position based on messages received from its neighbors possible. A few nodes equipped with a GPS receiver communicating with neighboring nodes may enable them to calculate position. This solution is less expensive to implement, but is associated with the generation of additional traffic on the network [17].
Local Minimum Spanning Tree (LMST) protocol calculates the local approximation of the minimum spanning tree [18]. It is performed in three, or optionally four, stages.
The first stage is the exchange of information. All nodes send messages to their visible neighbors containing their identities and locations (visible neighbor nodes that are within range when transmitting at the maximum power).
In the second stage of topology creation, each node performs locally Prim's algorithm [11] taking their Euclidean length of edge as cost – the minimum spanning tree $T_u = (VN_u, E_u)$ contains all visible neighbors of node u (VN_u) in the max-power graph $G_{\varepsilon} = (N, V_{\varepsilon})$. Then, each node defines a set of neighbors.

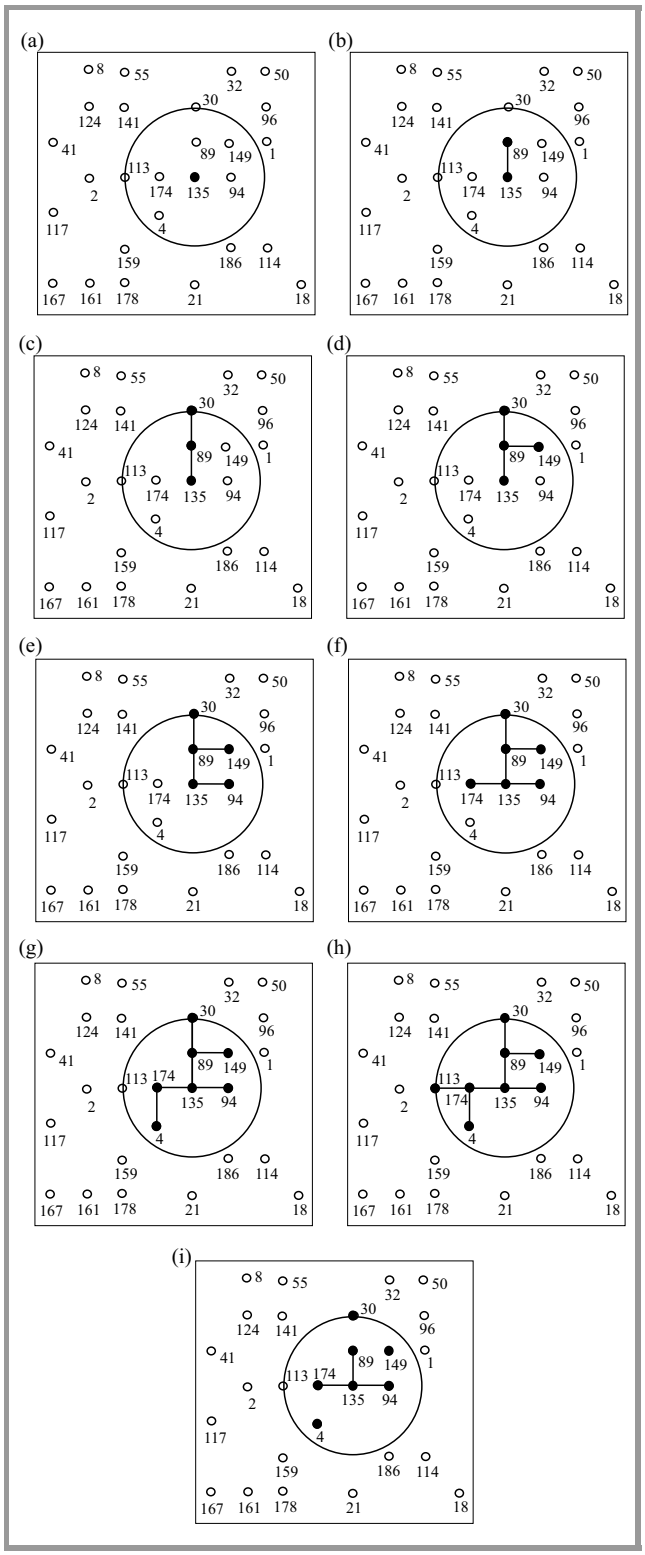


Fig. 1. The steps for generating network topology with an application of the LMST model for exemplary node deployments.

The node v is treated as a neighbor of node u ($u \to v$) if a node v is within range of node u and is available in one step in a minimum spanning tree computed in this node $T_u = (VN_u, E_u)$:

$$u \longrightarrow v \Longleftrightarrow (u,v) \in E_u. \tag{2}$$

A set of neighbors of node u is defined as:

$$N(u) = \{v \in VN_u | u \longrightarrow v\}. \tag{3}$$

Network topology defined in the LMST protocol is represented by a directed graph $G_{LMST} = (N, E_{LMST})$, where directed edge $(u,v) \in E_{LMST}$ exists only if $u \longrightarrow v$ (Fig. 1). In the last (required) step of the protocol, power levels of signals required for the communication with neighboring nodes are calculated. This can be obtained by measuring the power of incoming messages sent to the nodes in the first stage of protocol with the maximum power received from the visible neighbors.

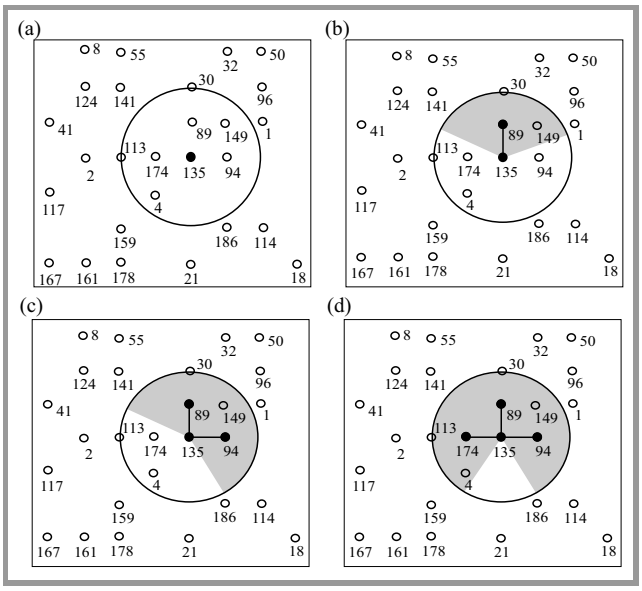


Fig. 2. The steps for generating network topology with an application of the DistRNG model for exemplary node placements.

The fourth (optional) step creates a topology with symmetric links. This is achieved either by replacing the asymmetric edges of symmetric ones or by removing asymmetric edges.

Distributed Relative Neighborhood Graphs (DistRNG) protocol [7] constructs a RNG graph built on a set of nodes N that has an edge between a pair of nodes $u, v \in N$ if and only if there is a node $w \in N$ such that:

$$\max\{\delta(u,w), \delta(v,w)\} \leq \delta(u,v). \tag{4}$$

The DistRNG protocol uses the concept of *coverage area*. If node v is a neighbor of node u, the coverage area of node v: $Cov_u(v)$ is defined as the clipping plane with the center at node u and width $a\hat{u}b$, where a and b are the points of intersection of the circles with the radius $\delta(u,v)$ and midpoints in the nodes of u and v. The total coverage area of node u is the sum of the areas of all of its neighbors (Fig. 2).

4. Simulation Study

To support the study of routing algorithms, the topology generator for ad hoc networks has been proposed. The

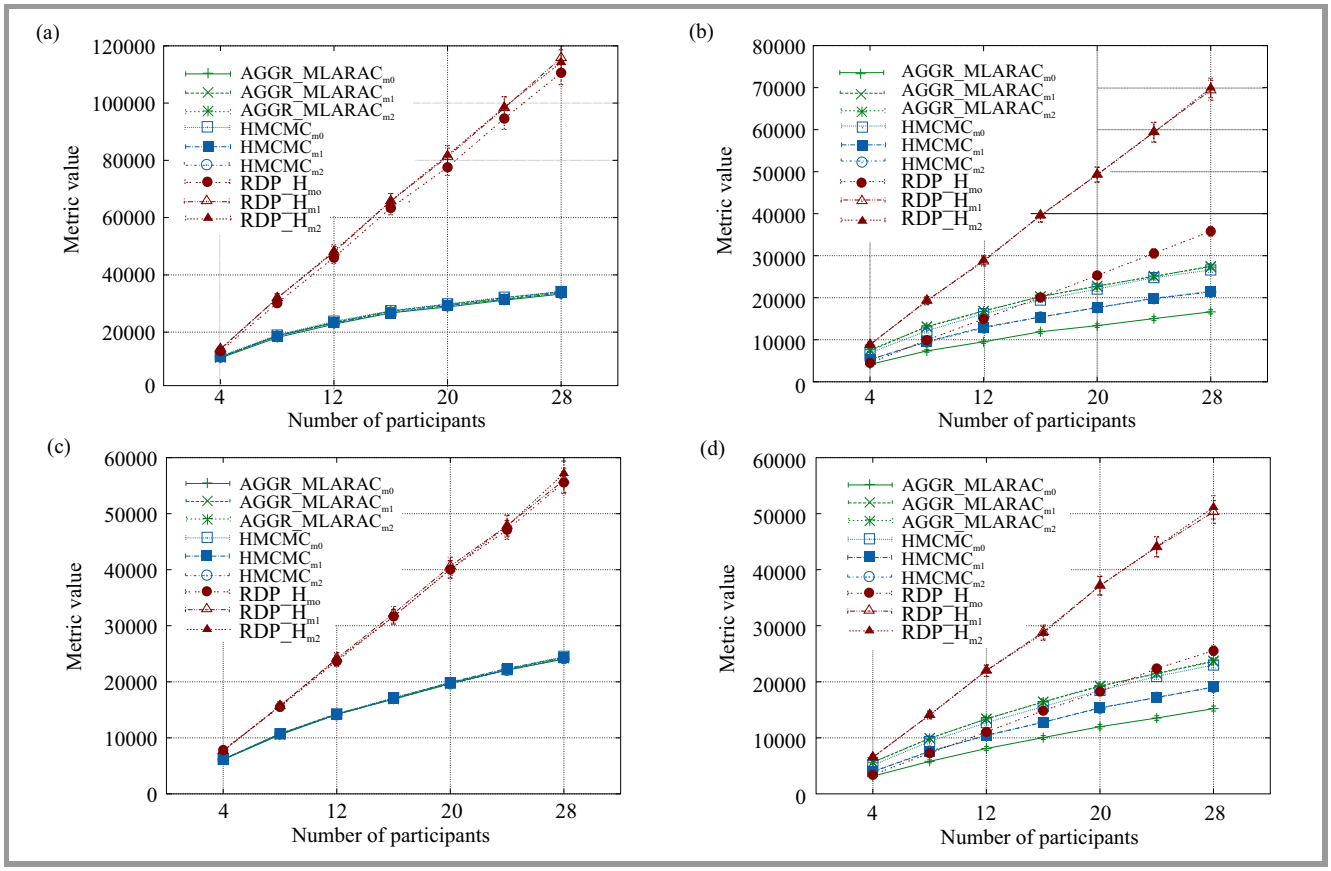


Fig. 3. Average cost of constrained multicast trees obtained in networks with 200 nodes generated according to: (a) LMST protocol, (b) DistRNG protocol, (c) Waxman model with $k = 100$, and (d) Waxman model with $k = 200$.

generator was created based on the structure and the methods that support the process of topology generation of the BRITE application [19]. Its flexibility and functionality to generate the topology of wired networks was preserved. Its capabilities were additionally extended by creating new classes supporting the process of generation of ad hoc network topologies [20].

The BRITE generator was equipped with tools needed to generate the topologies according to the two basic topology control protocols described in Section 3. Protocols based on the knowledge of the position and direction were selected. These protocols are widely used in existing ad hoc networks and their usefulness in the simulation of theoretical network models is beyond dispute. Implementation of distributed protocols is associated with a relatively high computational complexity and, consequently, with significant power requirements from the processor and memory demands from the generator. Each node in the network has limited knowledge about the entire network topology. For this reason, a creation of optimal topology is generally not possible in realistic scenarios. Hence, reflecting this problem in generative models is desirable.

During application development, additional classes extending the functionality of the generator were created. The purpose of these structures was to represent ad hoc network basis in a format determined by the BRITE application. In this way, the application was extended by additional tools that mainly supported the visualization of network topologies and the presentation of data obtained in the simulation.

A comparative analysis of the most important parameters of the topology generated by the implemented method were conducted. The topologies generated by models based on the DistRNG and LMST protocols and situated in the square plane with a side length of $Size = 1000$ were compared. Nodes in all models assumed the value of the maximum transmission range of $RangeMax = 250$.

Distributed topology control protocols do not guarantee the consistency of the generated graph. Calculations of topologies diameters were performed only for nodes forming coherent graphs.

The aim of research study is to analyze the cost of the trees as a function of the number of multicast group members. The simulation process uses 1000 topologies that model ad hoc networks with LMST and DistRNG topology control mechanisms. With a constant value of the number of nodes ($n = 100$) and the maximum transmission range ($RangeMax = 250$), the LMST protocol generates network topologies with the average number of edges $k = 100$, while DistRNG – about 200.

The simulation process also uses network topologies represented by random graphs generated by the application of the Waxman method. In order to guarantee the consistency

of the graph and create short edges between nodes, boundary values of the Waxman method parameters have been set up ($\alpha = 0.15$, $\beta = 0.05$). The aim of the authors was to investigate whether the results of multicast algorithms in ad-hoc networks are comparable with results obtained in random graphs with such short edges such as ad hoc networks. Therefore, they used network topologies generated by Waxman node with an average node degree of $D_{av} = 2$ ($k = 100$) and $D_{av} = 4$ ($k = 200$).

5. Experimental Results

The comparison of the multicriterial algorithms is a hard task not only because of the complexity of the algorithms themselves, but also because of the multitude of detail involved in the performance of the simulation, let alone its initiation. Thus, in [21] an innovative method of multicast algorithms evaluation based on a fuzzy system was introduced. It shows usefulness of imprecise analysis in routing algorithms comparison.

In a simulation study authors compared the cost of the multicast trees obtained in different network topologies for routing algorithms without constraint (m_0), with one constraint (m_1) and two constraints (m_2).

Simulations were performed for the sets of graphs of 200 nodes generated with LMST and DistRNG protocol, and compared with Waxman model in two scenarios: with $k = 100$ edges and $k = 200$ edges. In order to achieve the high statistical quality of the results 1000 graphs were generated for each of the topology model. Three metrics (constraints) were randomly generated from the range $\langle 1, 1000 \rangle$ for each edge in the graph. Each of the generated topologies was tested for connecting $4, 8, \ldots, 28$ multicast nodes. The technique presented in [22] was used to pick the constraints for the MCMST problem.

The results presented in Fig. 3 show a comparison of Aggr_MLARAC, HMCMC and RDP_H algorithms in relation to a number of multicast nodes m in the networks obtained with the above-mentioned methods. The results show that the average cost of multicast trees increases with the increase of the number of multicast nodes in the network, with a defined maximum delay value along the path in the tree ($\Delta = 1000$). The influence of different network topologies is observable. The costs of obtained trees are smallest in ad hoc networks with LMST protocol for each examined algorithms. Aggr_MLARAC and HMCMC multicast algorithms have the best performance in LMST ad hoc networks.

Analysis of the results presented in Fig. 3 indicate strong similarities in the results obtained with the algorithms generated network topologies using a LMST protocol and Waxman model ($k = 100$), as well as the protocol DistRNG and Waxman model ($k = 200$). In the second case, the costs of obtained trees are comparable and smallest for each examined algorithms. Aggr_MLARAC and HMCMC multicast algorithms have the best performance in DistRNG ad-hoc networks and networks generated with an application of Waxmax model ($k = 200$). This leads to the conclusion that in simulations studies on ad hoc networks it is possible to use fast methods that generate random graphs.

6. Conclusion

Multicriterial constrained multicast routing problems presents a non-trivial level of complexity. An additional criterion of comparing algorithms is the network topology and topology control mechanisms. Following this concept, a need for a broad analysis techniques spectrum arises.

It has been shown that exploring not only the space of the algorithms, but also the space of their comparison is worth an increased amount of effort as the conclusions may render different algorithms useful in different situations. It is also observable that for certain parameters complex network topologies obtained by the topology control protocols can be modeled by random methods. In addition, the stability of the algorithms against changes in different conditions can be shown with the use of the innovative and non-standard analysis.

The authors are still developing optimization methods for multicast connections. A new method based on innovative model of imprecise calculations called *Ordered Fuzzy Numbers* [23], [24] seems to be an interesting idea in future works.

References

[1] P. Santi, *Topology Control in Wireless Ad Hoc and Sensor Networks*. Chichester, UK: Wiley, 2005.

[2] P. Santi, *Mobility Models for Next Generation Wireles Networks: Ad Hoc, Vehicular, and Mesh Networks*. Chichester, UK: Wiley, 2012.

[3] "Wireless sensors and integrated wireless sensor networks", Frost & Sullivan Technical Insights, 2004.

[4] M. Głąbowski, B. Musznicki, P. Nowak, and P. Zwierzykowski, "An algorithm for finding shortest path tree using ant colony optimization metaheuristic", in *Image Processing and Communications Challenges 5*, R. S. Choraś, Ed. Advances in Intelligent Systems and Computing, vol. 233, pp. 317–326, 2014.

[5] S. Chen and K. Nahrstedt, "An overview of quality of service routing for next-generation high-speed networks: problems and solutions", *IEEE Network*, vol. 12, pp. 64–79, 1998.

[6] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.

[7] S. Borbash and E. Jennings, "Distributed topology control algorithm for multihop wireless networks", in *Proc. 2002 World Congr. Comput. Intell. WCCI 2002*, Honolulu, Hawaii, USA, 2002, pp. 355–360.

[8] E. Dijkstra, "A note on two problems in connexion with graphs", *Numerische Mathematik*, vol. 1, pp. 269–271, 1959.

[9] K. Stachowiak, J. Weissenberg, and P. Zwierzykowski, "Lagrangian relaxation in the multicriterial routing", in *Proc. IEEE AFRICON 2011*, Livingstone, Zambia, 2011, pp. 1–6.

[10] A. Jüttner, B. Szviatovszki, I. Mecs, and Z. Rajko, "Lagrange relaxation based method for the QoS routing problem", in *Proc. 20th Ann. Joint Conf. IEEE Comp. Commun. Soc. INFOCOM 2001*, Anchorage, Alaska USA, 2001.

[11] R. Prim, "Shortest connection networks and some generalizations", *Bell Systems Tech. J.*, vol. 36, pp. 1389–1401, 1957.

[12] M. Piechowiak and P. Zwierzykowski, "A new delay-constrained multicast routing algorithm for packet networks", in *Proc. IEEE AFRICON 2009*, Nairobi, Kenya, 2009, pp. 1–5.

[13] F. Gang, "A multi-constrained multicast QoS routing algorithm", *Comp. Commun.*, vol. 29, no. 10, pp. 1811–1822, 2006.

[14] K. Stachowiak and P. Zwierzykowski, "Rendezvous point based approach to the multi-constrained multicast routing problem", *AEU – Int. J. Electron. Commun.*, vol. 68, no. 6, pp. 561–564, 2014.

[15] K. Stachowiak and P. Zwierzykowski, "Innovative method of the evaluation of multicriterial multicast routing algorithms", *J. Telecommun. Inform. Technol.*, no. 1, pp. 49–55, 2013.

[16] R. Rajaraman, "Topology control and routing in ad hoc networks: a survey", *ACM SIGACT News*, vol. 33, no. 2, pp. 60–73, 2002.

[17] P. Santi, "Topology control in wireless ad hoc and sensor networks", *ACM Comput. Surveys*, vol. 37, no. 2, pp. 164–194, 2005.

[18] N. Li, J. Hou, and L. Sha, "Design and analysis of an MST-based topology control algorithm", in *Proc. 22th Ann. Joint Conf. IEEE Comp. Commun. Soc. INFOCOM 2002*, San Francisco, CA, USA, 2003, pp. 1702–1712.

[19] A. Medina, A. Lakhina, I. Matta, and J. Byers, "BRITE: An approach to universal topology generation", in *Proc. 9th Int. Worksh. Model. Anal. Simul. Comp. Telecommun. Syst. MASCOTS 2001*, Cincinnati, OH, USA, 2001, pp. 346–356.

[20] A. Zamożniewicz, "Methods for generating topologies of ad hoc networks", *Master thesis*, Poznan University of Technology, 2009 (in Polish).

[21] P. Prokopowicz, M. Piechowiak, and P. Kotlarz, "The linguistic modeling of fuzzy system as multicriteria evaluator for the multicast routing algorithms", in *Artificial Intelligence and Soft Computing*, L. Rutkowski *et al.*, Eds. Proc. of ICAISC 2014, Zakopane, Poland, Part II. *LNAI*, vol. 8468, pp. 665–675. Springer, 2014.

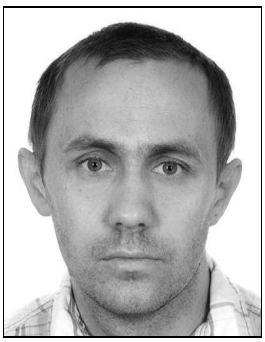
[22] F. Gang, "The revisit of QoS routing based on non-linear Lagrange relaxation", *Int. J. Commun. Syst.*, vol. 20, no. 1, pp. 9–22, 2007.

[23] P. Prokopowicz, "Flexible and simple methods of calculations on fuzzy numbers with the ordered fuzzy numbers model", *Artificial Intelligence and Soft Computing*, L. Rutkowski *et al.*, Eds. Proc. of ICAISC 2013, Zakopane, Poland, Part I. *LNAI*, vol. 7894, pp. 365–375. Springer, 2013.

[24] J. M. Czerniak, W. Dobrosielski, Ł. Apiecionek, and D. Ewald, "Representation of a trend in OFN during fuzzy observance of the water level from the crisis control center", in Proc. 2015 Federated Conf. Comp. Sci. & Inform. Syst. FedCSIS 2015, Łódź, Poland, 2015, *Annals of Computer Science and Information Systems*, vol. 5, pp. 443–447 (doi: 10.15439/2015F217).

Maciej Piechowiak received his M.Sc. degree from the University of Technology and Life Sciences, Bydgoszcz, Poland in 2002 and his Ph.D. degree from the Poznan University of Technology, Poznan, Poland in 2010. He is currently an assistant professor in the Department of Mechanics and Applied Computer Science at the Kazimierz Wielki University, Bydgoszcz, Poland. Dr. Piechowiak is an author and co-author of dozens articles published in journals and conference proceedings (several conference awards). He has served as a Guest Editor and Editorial Board of two international journals and TPC member of several international conferences. His main research fields are: routing algorithms and protocols, optimization techniques in networks and modeling of network topologies. He is a member of Institute of Electronics, Information and Communication Engineers (IEICE) and Polish Information Processing Society.
E-mail: mpiech@ukw.edu.pl
Institute of Mechanics and Applied Computer Science
Kazimierz Wielki University
Kopernika st 1
85-172 Bydgoszcz, Poland

Krzysztof Stachowiak received his M.Sc. degree in Telecommunications from Poznan University of Technology, Poland in 2009. Since 2009 he has been pursuing a Ph.D. degree at the Poznan University of Technology in the faculty of Electronics and Telecommunications. He is involved in the research regarding multicast routing in the packet switching networks which mainly consists in the analysis of the existing QoS routing algorithms as well as inventing new proposals. The main subject of his research is multicriterial optimization which has resulted in a series of papers on the subject of the linear and non-linear Lagrangian relaxation. He has taken part in the organization of two international scientific conferences, and has been a participant of several others. Besides the scientific research he is also in charge of the professional training center, coordinating and conducting courses on the usage of the Linux kernel based operating systems.
E-mail: krzysiek.stachowiak@gmail.com
Chair of Communication and Computer Networks
Faculty of Electronics and Telecommunications
Poznan University of Technology
Pl. Marii Sklodowskiej-Curie 5
60-965 Poznan, Poland



Tomasz Bartczak received his M.Sc. degree in Telecommunications from Poznan University of Technology, Poland in 2003. During the last 2 years, he has been working for Dolby Systems. Since 2005, he has been cooperating with Chair of Communications and Computer networks at the Poznan University of Technology. He is co-author of over 20 papers mostly related to multicast optimization algorithms and protocols.
E-mail: tbartcz@gmail.com
Chair of Communication and Computer Networks
Faculty of Electronics and Telecommunications
Poznan University of Technology
Pl. Marii Sklodowskiej-Curie 5
60-965 Poznan, Poland

Paper

LDAOR – Location and Direction Aware Opportunistic Routing in Vehicular Ad hoc Networks

Marziyeh Barootkar[1], Akbar Ghaffarpour Rahbar*[1], and Masoud Sabaei[2]

[1] *Computer Networks Research Lab, Electrical Engineering Technologies Research Center, Sahand University of Technology Tabriz, Iran*
[2] *Computer Engineering and Information Technology Department, Amirkabir University of Technology Tehran, Iran*
* *Corresponding Author*

Abstract—Routing in Vehicular Ad hoc Networks (VANETs) has found significant attention because of its unique features such as lack of energy constraints and high-speed vehicles applications. Besides, since these networks are highly dynamic, design process of routing algorithms suitable for an urban environment is extremely challenging. Appropriate algorithms could be opportunistic routing (OR) where traffic transmission is performed using the store-carry-forward mechanism. An efficient OR mechanism, called Location and Direction Aware Opportunistic Routing (LDAOR), is proposed in this paper. It is based on the best neighbor node selection by using vehicles positions, vehicles directions, and prioritization of messages from buffers, based on contact histories and positions of neighbor nodes to destination. In LDAOR, when multiple nodes make contact with a carrier node, the closest neighbor node to destination is selected as the best forwarder. However, when only one node makes contact with the carrier node, the message is delivered to it if it moves toward the destination. Using the ONE simulator, the obtained performance evaluation results show that the LDAOR operates better than conventional OR algorithms. The LDAOR not only increases delivery rate, but also reduces network overhead, traffic loss, and number of aborted messages.

Keywords—carry-and-forward mechanism, contact history knowledge, direction and location aware routing, opportunistic routing, vehicular ad hoc networks.

1. Introduction

The routing algorithms designed for Mobile Ad Hoc Networks could not be appropriate for Vehicular Ad Hoc Networks (VANETs) since they do not consider inherent features of VANET such as high mobility of vehicles (that leads to frequent topology changes and unstable links), and short-time connections among vehicles [1], [2]. Therefore, new routing algorithms must be designed in such a way that no packet is lost when connections are disconnected. To solve this problem, opportunistic routing (OR) algorithms have been proposed for VANETs [3], where packets are buffered in nodes when a disconnection occurs between two nodes and there is no continuous path available between them, thus increasing packets delay [4]. The store-carry and then forward mechanism is used in these algorithms when a connection is established, thus increasing delivery ratio and considerably reducing the data loss rate [5]. This mechanism keeps messages in node buffers during disconnection time, and takes the advantage of node mobility feature to find appropriate nodes within different partitions in order to route messages toward their destinations.

New routing algorithms must benefit from history of node contacts and status of nodes in a network topology for routing decisions in order to achieve efficient decisions. Due to the lack of stable links in opportunistic networks, the memory overhead is high. In addition, since permanent links do not exist in these networks, the bandwidth of links becomes an important resource that must be fully utilized when a contact is made. Therefore, identifying potential intermediate carrier nodes based on the network knowledge is essential for messages. By efficiently utilizing the bandwidth for a given message, the opportunity to transmit other messages in the network can increase. These issues motivate us to design a new routing algorithm that resolves them.

The authors objective is to propose the Location and Direction Aware Opportunistic Routing (LDAOR) algorithm that considers position of vehicles to avoid flooding messages to all contacts and to limit replication rate. The aim is on reducing the overhead in addition to improving delivery ratio, delay, drop ratio, and the number of aborted messages. In addition, the angle between motion vector of adjacent nodes and the distance vector from neighbor node to destination node is considered to avoid sending a message to the vehicles moving in the opposite direction of the message destination.

The authors contribution is the proposal of LDAOR as a location-based and history-based knowledge OR technique that chooses the best forwarder node to carry a message to its destination based on the position and direction of the forwarder node with respect to the destination. In addition, it picks messages for transmission based on their assigned priorities to quickly forward messages to their destinations before overflowing of limited buffers used in nodes.

2. Related Works

The OR algorithms have several prominent functions as:

1. multi-copy,
2. single copy,
3. location-based,

4. history-based,
5. special node-based [6], [7].

Each one of the OR algorithms can take the advantages of one or several functions for making their routing decisions.

The first two functions are often used in flooding-based OR algorithms. Under the first function [8], [9], the copy of a message is given to all intermediate nodes connected to a carrier node. This causes a message to move towards its destination through many directions, thus increasing delivery rate. However, a message may be given to the vehicles not moving towards the destination node of the message making buffers full in nodes, losing a group of messages, and increasing network overhead. In addition, this function can seriously reduce the network efficiency under low network resources [7]. On the other hand, using the single copy function, the number of copies of a message is limited. Either a carrier node or the first node that communicates with carrier nodes attempts to directly deliver the copy of a message to its desired destination [10], [11]. Some techniques dynamically determine the number of required copies of a message according to network conditions [12]–[15]. The location-based algorithms take the advantage of physical location of vehicles for routing during establishing a connection [16]–[20]. These methods decide regardless of the status of nodes in the past, but their advantages in using updated information in nodes are consistent with network conditions. Based on the fourth function, the history of contacts is used for making routing decisions [9], [21], in which routing decisions are assessed according to general network conditions within the entire network and during all the time. However, right decisions may not be made due to old information. Finally, in OR based on special nodes [22]–[24], certain nodes are used to deliver messages to destination nodes.

Many algorithms have been introduced to reduce the flooding effects, e.g., [12]–[15], [25], [26], by forwarding a message to high-quality nodes that have better chance for delivering the message to its destination. The quality of a node can be defined by various metrics such as the frequency that a node meets other nodes, the frequency that a node meets the destination, the last contact time of a node with other nodes, and the last contact time of a node with the destination. For example, MaxProp [9] is a flooding-based protocol [12], [27] since a carrier node sends messages to all contacts without distinguishing between them. Besides, MaxProp is based on history [7] since it takes the advantage of history of contact nodes for prioritization of messages for transmission and for removing from node buffers. It prioritizes each message based on two criteria: hop-count, i.e., number of nodes a message has traveled since its generation, and delivery probability to destination. When sorted based on hop-count, the messages with the hop-count smaller than a given threshold have high priority for transmission. On the other hand, those messages with hop-counts exceeding the threshold are sorted based on delivery probability to their destinations. Then, the messages with small chance of delivery to their destinations have the highest priority to be deleted from the buffer when the buffer is becoming full.

Some OR protocols are location-based that can provide better performance results than [28]–[30]. Instead of links statuses, routing decisions are based on the positions of source node, destination node and adjacent nodes at contact time. Since location-based protocols do not require the overhead of saving the information of previous nodes, they can achieve the desired goal with minimal overhead. For example, in Packet-Oriented Routing (POR) [18], messages belonging to far destinations have higher priority for transmission compared to the messages belonging to close destinations. There is no limit on buffering messages in nodes under POR. The POR decides only based on the distance of an adjacent node to the destination node and does not consider the history of contacts at all. Using POR, a carrier node only selects the best forwarder node among adjacent nodes for all its messages. After prioritizing them, POR sends the messages in sequence to the new forwarder node.

The Prophet protocol [21] benefits from the history of contacts of nodes to destination node besides the multi-copy function. By this history, the delivery probability of a message to its destination through adjacent nodes can be computed by a carrier node. Then, the message will be given to those nodes that have visited the destination node more than the carrier node itself.

In this paper, the best forwarder nodes are selected using their statuses at the time of contact to prevent from flooding. In addition, based on the history of contacts and position of nodes in the network, priorities are assigned to messages for sending and removing them from buffers. This can improve performance parameters in the whole network.

3. The LDAOR Protocol

In the following, the LDAOR protocol shall be described after stating network model.

3.1. Network Model, Data Structures, and Performance Parameters

The following shows presented network model and its relevant assumptions:

- The network focuses on vehicle to vehicle communications in an urban area with many junctions in which roads are two-way. For example, Helsinki city includes these features.
- The proposed VANET includes different vehicles such as privately-owned vehicles (POV), buses and taxis with special mobility patterns. Among the vehicles, several cars are randomly determined as destination of messages and other are selected as source vehicles. Destination nodes are considered stationary

(fixed). This assumption is suitable for applications such as delivering a message to a base point as access point.

- Since the speed of nodes cannot exceed a limit in a city scenario, each node can find the location of a destination node using a suitable location server, e.g. the method presented in [32]. Location servers can provide lookup and publish algorithms to exchange information about geographic positions of nodes in a network. In order to select the best forwarder nodes, a carrier node needs to calculate the direction of any candidate node with respect to the destination node and its distance to the destination node.
- Each vehicle is equipped with Geographical Position Systems (GPS) to obtain its current position.
- Since there is no resource without limitation, each node has a limited buffer.
- Transmission rates in all vehicles are all the same.
- There is no faulty node or link in the network.
- There are two reasons for having low collision in opportunistic networks. First, the number of neighbor nodes that can correctly receive and send messages is low because of instability links. Second, since the positions of nodes are different, the nodes do not receive a request message from a carrier node at the same time.

The following list shows the data structures required for the LDAOR protocol:

- Node A uses five fields to keep its status as:
  - **Node ID** – identification code of node A;
  - **Delivery probability list** – the list of delivery probabilities relevant to delivering of messages from node A to each one of other nodes. For example, if the value of probability in carrier node B to node A is 0.25, node B has made contact to node A with probability 0.25 so far. There is only one delivery probability in node A to each one of other nodes in the network. When a contact is made between nodes A and B, delivery probabilities are updated in both nodes based on the method presented in MaxProp. In the same way, a carrier node may evaluate the probability of nodes to the destination node for finding a node on a path with a high contact probability. To calculate delivery probability, whenever node X makes contact to node Y, the value of probability in all nodes is increased by one, and then the probabilities of all nodes are re-normalized based on the rule provided in Section 3 of [9] so that sum of probabilities in all nodes becomes 1. Thus, delivery probability list values in a node indicate that this node moves in either a crowed path or a sparse path;
  - **List of MsgIDs already sent** – this list keeps the ID of transmitted messages to other nodes. For example, consider node A has sent messages M_1 and M_2 to node B; and M_3, M_4, and M_5 to node C. Then, node A keeps {B, (M_1, M_2)} and {C, (M_3, M_4, M_5)} in this field. In the next contact, node A can easily find out that node B has previously received message M_2, and it will avoid sending message M_2 to node B for the second time. This list is scanned at some time intervals and the old IDs are removed from the list;
  - **Current Location**(x, y) – at any time, current location of node A is obtained from the GPS system and saved in these fields;
  - **Previous location**(x, y) – previous location of node A;
  - **Average transmitted bytes per transfer opportunity** – when node A makes a contact with node B, and then sends its messages to node B within the contact period, total number of transmitted bytes in node A is updated by the size of transmitted messages.
- Each message M includes the following fields:
  - **MsgID** – identification code of message M given by the source node. This ID is a combination of node ID and a sequence number generated by the node;
  - **Source** – the ID of the source node that has generated message M;
  - **Destination** – the ID of destination node of message M;
  - **Hop list** – when message M passes through different intermediate nodes, the IDs of the intermediate nodes are recorded in this list;
  - **TTL** – time to live for message M. When TTL of the message expires, the message should be deleted.
  - **Hop-count** – This field is set to zero when message M is generated. When this message arrives at an intermediate node, the hop-count field is incremented by 1;
  - **Message text** – the text of message M.

In a Hello message, there is a field called AckedMsgs. To avoid propagating a message already reached its destination, the destination node adds the ID of the delivered message to the AckedMsgs field of a Hello message and broadcasts it

in the network. Then, each node receiving this Hello message removes the acknowledged messages from its buffer. The following shows the performance parameters defined based on the ONE simulator [35]:

- **Aborted messages** – the number of aborted transmissions between nodes divided by total number of generated messages. A message is aborted when a receiver node cannot receive the message from a transmitting node because of small contact duration.
- **Loss in buffers** – number of dropped messages (including replicated messages) in buffers due to buffer overflow. This loss occurs when a receiving node does not have enough room in its buffer.
- **Delivery ratio** – message delivery probability defined by

$$Delivery\ ratio = \frac{Number\ of\ delivered\ messages}{Number\ of\ generated\ messages}.$$

- **Overhead ratio** – assessment of bandwidth efficiency defined by

$$Overhead = \frac{Number\ of\ relayed\ messages\ including\ replicated\ message}{Number\ of\ delivered\ messages} - \\ - \frac{Number\ of\ delivered\ messages}{Number\ of\ delivered\ messages}.$$

  Since in all OR algorithms request messages and reply messages should be communicated between intermediate nodes, their overhead has not been considered in the overhead ratio. According to the above formula, only those data messages that cannot be delivered to their destinations are accounted for overhead ratio.
- End-to-end delay – average delay from generation time of a message until successfully delivering to its destination.

3.2. The LDAOR Protocol

The pseudo code of LDAOR is shown in Algorithm 1 and Algorithm 2. It includes two phases: selecting the best neighbor nodes in order to store-carry and then forward a message toward its destination (see Subsection 3.2.1), and prioritization of messages according to MaxProp and then sorting the messages based on the minimum distance between neighbor nodes and destination nodes (see Subsection 3.2.2). To reduce overhead, LDAOR limits the rate of message replication with respect to physical location and direction of neighbor nodes with the destination node. The LDAOR utilizes both node history information and node information at the time of contact. The general parameters used in LDAOR are depicted in Table 1.

Algorithm 1: LDAOR, phase 1 – neighbor selection on contact event

Step 1: Exchange the status of connection to each other
Step 2: Delete the acknowledged messages from the buffer
Step 3: Direct delivery: if the neighbor node is the destination of any message in the buffer, then deliver it first
Step 4: $CM = \{\}$
for each message M_k in carrier node c_i **do**
  n_k = number of neighbors that have not received M_k
  if $(n_k = 0)$ **then**
    M_k must still remain in buffer
  if $(n_k = 1)$ **then**
    g_n = best forwarder node based on the angle-based method
  else if $(n_k > 1)$ **then**
    g_n = best forwarder node based on the distance-based method
  end if
    if $(n_k \neq 0)$ **then** add (g_n, M_k) to set CM
end for
return CM

Algorithm 2: LDAOR, phase 2 – determine priorities of messages for transmission

Step 5: Determine threshold H // similar to MaxProp
Step 6: Split CM into two sections – sorting messages with hop-count lower than threshold H and messages with hop-count greater than threshold H
Step 7: **for** $k = 2$ to size (CM) **do** // take all messages in CM
  Take messages M_k and M_{k-1} from set CM
  if $(h_k < H$ **and** $h_{k-1} \geq H)$ **then**
  send M_k
  else if $(h_{k-1}$ **and** $h_k \geq H)$ **then**
  1. if $(h_k < H$ **and** $h_{k-1} < H)$ **then**
    if $(DI_k < DI_{k-1})$ **then** Send M_k
    else Send M_{k-1}
    end if
  2. if $(h_k \geq H$ and $h_{k-1} \geq H)$ **then**
    if $(dp_k > dp_{k-1})$ **then** Send M_k
    else if $(dp_k < dp_{k-1})$ **then** Send M_{k-1}
      else // the same dp for forwarders of M_k and M_{k-1}
        if $(DI_k < DI_{k-1})$ then Send M_k
        else Send M_{k-1}
      end if
  end if
end for

3.2.1. Neighbor Node Selection (Phase 1)

Whenever a carrier node wants to find a forwarder node, it should select the best neighbor node as a forwarder node in phase 1.

Step 1 of phase 1. The operation of the first phase of the algorithm is as follows. Since LDAOR decides for each message individually, it is necessary to obtain required information from its adjacent nodes in order to select the

Table 1
Notations used in LDAOR

Notation	Description
$angle$	Angle between two location coordinates for a neighbor node
b	Buffer size
c_i	Carrier node i
c_n	Candidate neighbor node
CM	Array of selected forwarder nodes with messages for transmission
DI_k	Distance array between neighbor nodes k and destination node
di_c	Distance value between candidate node c and destination node
d_k	The destination of the k-th message
dp	Delivery probability for a neighbor node
g_n	Good neighbor
H	Threshold on hop-count of a message
h_k	Hop-count of the k-th message
(lc_x, lc_y)	Location coordinate for neighbor node at current time
(lp_x, lp_y)	Location coordinate for candidate neighbor node at previous time
M_k	The k-th message in a buffer
$\overrightarrow{ND}$	Distance vector from neighbor node to destination M_k
n_k	Number of neighbors that have not received M_k
p	Portion of buffer
speed	Speed of neighbor node c_n
S_T	Average transmitted bytes
$\vec{V}_N$	Neighbor node velocity vector
(v_x, v_y)	Velocity coordinate of a neighbor node
θ	A direction angle of a neighbor node

best forwarder node among them when a contact happens. This information includes: the acknowledged messages in order to avoid decision making once again and resending them again, and the location and direction of neighbor nodes in order to check their statuses with respect to the destination of messages, where LDAOR obtains this information using location server. These operations are carried out based on the information available in nodes (see Subsection 3.1).

Step 2 of phase 1. After receiving the acknowledge message for a transmitted message, a carrier node removes the message from its buffer. Notice that relative mobility between two vehicles may be high. Then, there will be a delay in receiving acknowledge of messages. In this case, the carrier node removes the acknowledged messages from its buffer using the following mechanism. If the carrier node has not received the acknowledge for a transmitted message, it assumes that the message has not been delivered to the relevant destination node yet. Therefore, it tries to find a forwarder node for carrying the message by sending a request message (in order to resubmit the message) to neighbor nodes. It is likely that some of the neighbor nodes have already received the acknowledge of the message. Hence, they avoid receiving the duplicate message, and notify the carrier node (with a reply message) that the message has already been delivered to its destination. In the worst case, the message may be delivered to a node that has not received its acknowledge. Nevertheless, it is likely that its neighbor nodes have already received the acknowledge.

Step 3 of phase 1. If there is a neighbor node which is the destination of a message in the buffer, LDAOR directly delivers the message to the neighbor node. By this way, the number of messages from the buffer will reduce.

Step 4 of phase 1. If there is no destination node for a message among the neighbor nodes, LDAOR enters the decision making step for selecting the best neighbor nodes for carrying the remaining messages inside the buffer of the carrier node (see Fig. 1).

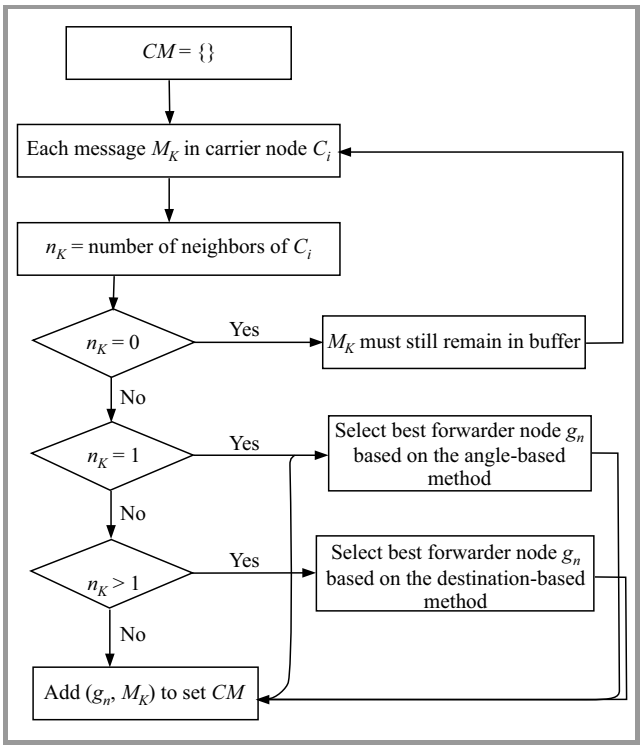


Fig. 1. Step 4 in phase 1.

The neighbor node selection mechanism in LDAOR operates as follows, in which only one forwarder node can be selected for each message. For a given message, when there is only one contact that has not previously received the message, the carrier node uses the angle-based method. On the other hand, when the carrier node has several contacts that neither of them has previously received the message, the distance-based method is utilized for forwarder node selec-

tion. Note that $n_k = 1$ in a region indicates that the number of candidate nodes is low and there is the probability of partitioning. This is because in this region, only one node has been candidate to receive the message. Thus, sending the message to this node needs more precision. Hence, if the message is sent to a node that moves away from the destination node (in the worst case, it moves in the opposite direction to the destination node), the chance of sending the message to the best other node is low, and therefore, this transmission may be vain in the network. However, $n_k > 1$ indicates that congestion of nodes is relatively high in the region, and the closer node to the destination node for sending the message is better. In this case, even if this node moves away from the destination node, there are some chances for delivering the message to the other best node. At the end, set *CM* is provided, where each item in this set includes two entries as the selected neighbor node and the relevant message.

To illustrate the importance of neighbor node selection, consider the scenario displayed in Fig. 2a. Suppose carrier node A has messages M_1, M_2 and M_3 in its buffer, respectively, with destinations D_1, D_2 and D_3. Node A has made contact with three nodes B, C and F. Node A has received the status of the nodes in response to Hello messages for carrying message M_1. Then, based on the status of nodes, node A is noticed that node B has previously received message M_1. Thus, either node C or node F should be selected to carry message M_1. Using the distance-based method, node F with minimum distance to destination D_1 is chosen as the forwarder node for message M_1. Based on the statuses of adjacent nodes, node A is also noticed that only node C has not previously received message M_2. Hence, node A uses an angle-based method to check whether node C moves toward destination D_2 or not. After calculating the angle between motion vector of neighbor node C and distance vector of neighbor node C to D_2, node A realizes that node C moves completely in opposite direction to destination D_2. Thus, M_2 has no chance to be delivered to D_2 by node C. Therefore, node A must still keep M_2 in its buffer until the setup of an appropriate contact. This avoids useless saving of a message in buffer of node C. Similarly, the carrier node chooses the best forwarder node for other messages in its buffer. After preparing the set of ready contacts for receiving messages (as depicted in Fig. 2b), the phase 2 is started. The carrier node sends the messages based on their priorities to the appropriate nodes selected in the first phase (see Subsection 3.2.2).

When a carrier node wants to select the best forwarder node, it uses the location service for receiving the location of the destination node. The carrier node decides to select either the distance based method or the angle-based method according to the number of candidate nodes. In the distance based method, the carrier node obtains the position of the destination node by location service in order to compute the distance of the candidate node to the destination node. In the angle-based method, the carrier node obtains the position of the destination node by location service in order to compute the angle between the distance vector to the destination node and the motion vector of the candidate node.

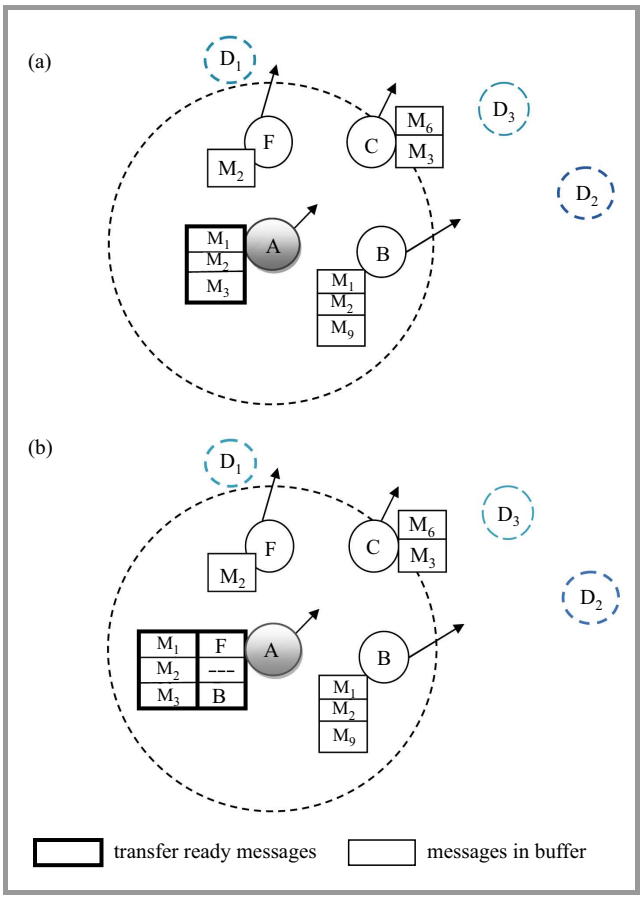


Fig. 2. Neighbor selection scenarios: (a) neighbor node selection scenario in LDAOR and (b) contacts with messages ready to transmit.

In angle-based method an exactly prediction of direction of vehicles is more complex in city scenarios due to high number of branches. Therefore, LDAOR determines the direction of a neighbor node by approximating the angle between the velocity vector of the neighbor node and distance vector from the neighbor node to the destination node (see Algorithm 3). If there exists any calculation error for the angle, it surely depends on the received information from the nodes that may depend on the error originated from GPS. Therefore, if there is a GPS with low error, LDAOR can calculate the angle according to the Eq. (1) very well.

In this case, if the estimated angle is acute angle (i.e. below 90°), then the neighbor vehicle is likely moving toward the destination node. Actually, in LDAOR, vehicle direction with respect to the destination node is important, not its direction in each junction.

Let vehicle V be the only candidate node for receiving the message. Then, the carrier node uses the angle-based method for approximating the direction of vehicle V with respect to the message destination. In this case, the mes-

Algorithm 3: Angle-based method to evaluate a neighbor node c_n for a given message M

1: $angle = \tan^{-1}\left(\frac{lc_y - lp_x}{lc_x - lp_x}\right)$
2: $v_x = speed \times \cos(angle)$
3: $v_y = speed \times \sin(angle)$
4: $|\vec{V}_N| = \sqrt{v_x^2 + v_y^2}$
5: Compute based on destination of message M
6: $\theta = \cos^{-1}\left(\frac{\overrightarrow{ND} \times \vec{V}_N}{|\overrightarrow{ND}| \times |\vec{V}_N|}\right)$
7: **if** $(\theta < 90^\circ)$ **then**
8: c_n is chosen as forwarder node for the message
9: **else**
10: The message should still be kept in buffer
11. **end if**

sage is sent to a neighbor node V only if it moves toward the message destination. Determining whether node V is moving toward the message destination or not can be obtained by calculating θ:

$$\theta = \cos^{-1}\left(\frac{\overrightarrow{ND} \times \vec{V}_N}{|\overrightarrow{ND}| \times |\vec{V}_N|}\right), \qquad (1)$$

where $\overrightarrow{ND}$ is the distance vector from node V to the destination node of the message, and $\vec{V}_N$ is the velocity vector of node V. If $\theta < 90^\circ$, node V is likely moving toward the message destination, i.e., the message can be delivered to it. Otherwise, when $\theta \geq 90^\circ$, node V moves away from the destination node, and therefore, the message should be held in the carrier node buffer until finding a better forwarder node. This mechanism avoids delivering messages to the nodes that do not move toward the destination. Therefore, the network traffic decreases and buffers are not filled with those messages that do not have any chance to be delivered to their destinations.

Algorithm 4: Neighbor selection by distance-based method for message M_k with destination d_k

1: D = infinity
2: **for** each neighbor node c_n **do**
3: $di_c = \text{distance}(c_n, d_k)$
4: **if** $(di_c < D)$ then
5: $D = di_c$
6: $g = c_n$
7: **end if**
8: **end for**
9: **return** g

In distance-based method the multiple nodes are candidate for receiving the message. In this case, the carrier node uses the distance-based method (see Algorithm 4) for selecting the best forwarder node. Hence, each candidate neighbor node notifies its physical location inside a reply message to the carrier node. Then, the carrier node computes the distance of each candidate vehicle from its current position to the destination node and selects the node with the smallest distance to the destination node as the forwarder node. Then, the carrier node delivers the message to the selected forwarder node.

In short, among multiple candidate neighbor nodes, a node is selected with the minimum distance to the destination node of a given message. This method can relatively remove redundancy created by the flooding methods. In addition, delivering the message to the nodes that are closer to the message destination can reduce delay.

3.2.2. Management of Buffer in a Carrier Node (Phase 2)

Since it is assumed that nodes have limited buffers, their overflows may happen and some important messages may be lost. Hence, a mechanism should be provided for buffer management so that the messages that are more likely to reach their destinations are processed faster. On the other hand, those messages with minimum chance of delivery to their destinations should be removed from buffers when overflowing, i.e. the messages with minimum delivery probabilities with hop counts exceeding a threshold. These deleted messages are counted as lost messages. Therefore, messages should be prioritized in buffer of each node. By this, more space can be provided for future coming messages. In LDAOR, transmission opportunity is the same for all messages at the beginning. However, when message M_1 has been transmitted for a number of times (i.e. a message with high hop-count), message M_1 should have less priority in re-transmission compared with newly arrived message M_2. This is because message M_1 has already used network bandwidth and has not been successfully delivered yet. Hence, it is fair to service message M_2. Still message M_1 has transmission opportunity in future. The LDAOR tries to provide fairness for message transmission opportunity. When two messages have the same hop count, decision is made based on the status of the transmitter node (as stated in the following). Therefore, messages will not encounter bandwidth starvation under LDAOR.

Step 5 of phase 2. Consider a carrier node has made contact with a given node. Under LDAOR, the carrier node can send its traffic to the given node as long as the contact is setup. Whenever the contact is shut down, the carrier node computes the volume of transmitted traffic in that contact in bytes. Note that, within the contact period, the carrier node may transmit a number of messages or only a part of a message. Then, the carrier node computes average transmitted traffic S_T among contacts as following. Consider a sliding window of 10 last contacts set up by the carrier node. Let S_i be the volume of whole traffic transmitted in bytes within the i-th contact in the sliding window, computed at the end of the contact i. Let r (where $r \leq 10$) denote the number of contacts made within the sliding window. Then, S_T is computed by

$$S_T = \frac{1}{r}\sum_{i=1}^{r} S_i, \quad \text{if } S_i \neq 0\,.$$

Prioritization of messages in LDAOR is performed by the following rules. For prioritization of messages, we need to

define parameter H as a threshold on hop-count of messages in a given carrier node, which is the same for all messages inside the carrier node buffer. Note that each node has its own H at any time. This threshold is computed based on the messages available in the carrier node buffer. Similar to MaxProp, the LDAOR uses average transmitted bytes S_T and buffer size b to adjust threshold H. For this purpose, the carrier node calculates parameter p using Eq. (2) (in bytes) [9]:

$$p = \begin{cases} S_T & S_T < \dfrac{b}{2} \\ \min(S_T, b - S_T) & \dfrac{b}{2} \leq S_T < b \\ 0 & b < S_T \end{cases} . \qquad (2)$$

A carrier node computes parameter p in Eq. (2) under two situations since average number of transmitted bytes S_T may have changed:

- When the carrier node wants to send a message, it must calculate p according to average transmitted bytes S_T. Then, it sorts the messages in its buffer based on their hop counts. Finally, the node calculates threshold value H;
- When the carrier node wants to receive a message while its buffer is full, it needs to re-calculate p. Then, considering the threshold value, it removes a low-priority message from its full buffer in order to receive the new message.

Then, z items in set *CM* (obtained in phase 1) are sorted (in ascending order) based on the hop-count of the messages. Starting from the beginning of the sorted list *CM*, denote the size of z messages by $N_1, N_2, \ldots, N_z$. Now consider the j-th message satisfies the condition $\sum_{i=1}^{j} N_i > p$, where $\sum_{i=1}^{j-1} N_i \leq p$. Then, threshold H is set to the hop-count of the j-th message plus 1.

After computing threshold H, the messages in *CM* are split into two sections (similar to MaxProp): messages with hop-count $< H$ and messages with hop-count $\geq H$. The LDAOR mechanism determines the priorities for transmission of messages and deleting messages from the carrier node buffer as displayed in Fig. 3. Notice at the left part of Fig. 3, messages are first sorted based on hop-count, and then based on distance of neighbor nodes to destinations. In other words, if hop-counts of few messages are the same, they are sorted based on the distance of their neighbor nodes to their destination nodes. On the other hand, the right part in Fig. 3 is sorted based on delivery probability. When delivery probability is the same for a few messages, they are sorted based on distance of neighbor nodes to destination nodes, as displayed in Fig. 3.

As stated in Subsection 3.1, hop-count of a message shows the number of nodes the message has visited. By sorting messages based on their hop-counts in a carrier node, authors give high priority of transmission to those messages that have been generated newly and have visited small

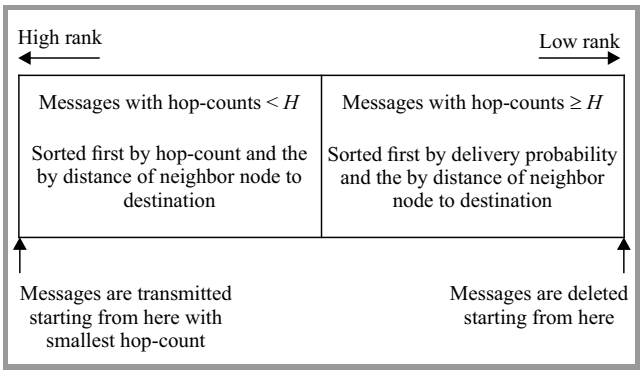


Fig. 3. The LDAOR priority mechanism for splitting messages in *CM*.

number of nodes. Notice a message with high hop-count shows that the network has attempted for a number of times to deliver the message to its destination by passing through a high number of nodes. However, the message has found less chance of delivery and less importance.

Delivery probability for a forwarder node shows the number of frequencies that the forwarder node has made contacts with other nodes. When delivery probability is high for a forwarder node, it shows that the delivering of messages through that forwarder node is high. When a contact is made between two nodes, delivery probabilities are updated in both nodes, based on the method presented in [9].

Step 7 of phase 2. In this step, di_c shows the distance between the selected forwarder node and the destination node of message M_k, the notation h_k denotes the hop-count of message M_k, and dp_k refers to the delivery probability of the forwarder node assigned to message M_k. In each loop, Step 7 evaluates two consecutive messages M_k and M_{k-1} in *CM*. If hop-counts of both messages are smaller than threshold H, their priorities are determined based on the distance-based method, i.e. high priority for transmission is given to the message that its relevant forwarder node has smaller distance to its destination. The advantage of this mechanism is that when a forwarder node has the smallest distance to the destination, its relevant message is transmitted at first. Then, the next messages are transmitted in sequence according to distances of their associated forwarders to their destinations. This mechanism causes a message to be located closer to its destination before its corresponding forwarder node goes in a situation that is out of the communication range of the destination. For example, consider there are three messages M_1, M_2, and M_3 in *CM*, where forwarder nodes c_{n1}, c_{n2}, and c_{n3} have been assigned to carry them, respectively. Consider hop-counts of messages M_2 and M_3 are smaller than H, and hop-count of message M_1 is larger than H. Assume c_{n3} to be closer to destination of message M_3 than c_{n2} to destination of message M_2. Therefore, M_3 is transmitted sooner than M_2. After sending M_2 and M_3, the carrier node sends M_1 due to its hop-count greater than H. This mechanism can increase delivery rate and can reduce delay in delivering messages to their destinations.

Table 2
Example for message transmission

Hop count	8	2	12	18	5	3	8	10	11	16	9	3
msgID	2	9	11	8	4	13	5	7	1	19	16	23
Message size	32099	55999	16558	432233	486776	52641	375477	189610	300743	336101	86929	254886

Table 3
Sorted messages

Hop count	2	3	3	5	8	8	9	10	11	12	16	18
msgID	9	13	23	4	2	5	16	7	1	11	19	8
Message size	55999	52641	254886	486776	32099	375477	86929	189610	300743	16558	**336101**	432233

On the other hand, for those messages with hop-counts exceeding threshold H, the LDAOR assesses the delivery probabilities of forwarder nodes to destination nodes. Then, the message with the highest delivery probability is sent to its forwarder node at first. Then, other messages with smaller delivery probabilities are sent in sequence. If delivery probability of forwarder nodes of both messages is the same, high priority for transmission is given to the message that its relevant forwarder node has smaller distance to its destination than the other forwarder node.

Let us consider as an example that there are 12 messages available in the buffer of a carrier node (see Table 2), and the carrier node has decide to send a message with the highest priority. Let average transmitted bytes be $S_T = 3,047,877$, and buffer size $b = 5,000,000$ bytes. Since $b > S_T > \frac{b}{2}$, we have $p = \min(S_T, b - S_T) = 1,952,123$ bytes. Then, the messages are sorted based on their hop-counts (see Table 3). Since the summation of messages sizes until the message with msgID=19 is greater than p, the hop-count of this message is chosen as our threshold with $H = 16 + 1 = 17$. Now all the messages in the buffer are split into two groups as depicted in Table 3, messages with hop-count smaller than H at the left side of buffer, and the messages with hop-count greater or equal than H at the right side of buffer.

3.2.3. Complexity Analysis of LDAOR and other OR Algorithms

The amount of transmitted information between nodes in LDAOR is almost similar to MaxProp because it uses the same mechanism for determining threshold as MaxProp, but LDAOR considers location of nodes as well. Prophet considers only the history of nodes contacts with the destination node for determining both the best forwarder node and priority of sending messages. The POR evaluates only the location of each node with respect to the destination node for determining both the best forwarder node and priority of messages.

Based on the distance-based and angle-based methods, a carrier node at first evaluates only the value of current location(x, y) and previous location(x, y). If a candidate node is chosen as the best forwarder node, the carrier node uses average transmitted bytes per transfer opportunity and delivery probability list for prioritizing messages. If a carrier node selects a candidate node, it enters the message prioritizing step and evaluates these two parameters. These values are based on the history of contacts and history of sending messages by this carrier node. It does not depend on a special time, and therefore, it cannot be outdated. The List of MsgIDs already sent is only kept in a node and it is not transmitted between nodes.

The computational complexity of Epidemic is O(1) [8] due to the lack of using any knowledge from the network. There is no forwarder node selection in MaxProp and it only prioritizes messages for transmission or removing from buffer. Hence, the complexity of MaxProp is $O(m \times \log_2(m))$, where m is the number of messages available in a node buffer. The MaxProp performs sorting twice for determining priority of messages. First for all nodes based on hop-count, and then based on delivery probability for the right part of a buffer.

Due to the process of selecting neighbor nodes in LDAOR, its complexity is $O(d)$ at phase 1, where d is the number of neighbor nodes for a message. LDAOR also adjusts prioritization for sending and removing a message with complexity $O(m \times \log_2(m))$ in phase 2. Notice LDAOR performs sorting twice; first, for all nodes based on hop-count, and then based on delivery probability for the right part of a buffer. Since for each message, only one neighbor node is selected, the complexity of LDAOR becomes $O(d \times m \times \log_2(m))$. Number of adjacent nodes d is small and this means that the nodes that can send or receive messages correctly is small because of instability of links [33], [34]. This is the most important distinction in opportunistic networks compared to other networks. In other words, the number of contacts is low, i.e. the number of neighbor nodes, and not the number of nodes in total [33], [34]. This is why there is a store-carry and forward mechanism in opportunistic networks.

Similar to LDAOR, the POR checks each one of its adjacent nodes for selecting the best forwarder node and also determines priorities for messages. Thus, the complexity of POR is the same as LDAOR. The Prophet counts the number of times adjacent nodes have visited a given destination node. Then, those nodes that have met the destination node of a given message more than the carrier node

Table 4
Complexity of opportunistic routing algorithms

LDAOR	MaxProp	POR	Epidemic	Prophet
$O(m \times log_2(m) \times d)$	$O(m \times log_2(m))$	$O(m \times log_2(m) \times d)$	O(1)	$O(m \times log_2(m) \times d)$

itself are chosen as forwarder nodes. Next, the messages are sorted in a descending order based on the number of visits whose forwarders have met their destinations. Finally, the messages are transmitted from the sorted list. For example, consider there are three messages M_1, M_2, M_3 in the buffer of a carrier node, and the carrier node has found that neighbor nodes c_{n1}, c_{n2}, and c_{n3} have met the destinations of M_1, M_2, and M_3, respectively, three, two, and five times more than the carrier node itself. Then, the carrier node first sends M_3, followed by M_1 and M_2. Hence, the complexity of Prophet is also $O(d \times m \times log_2(m))$. The complexities of the algorithms are shown in Table 4.

Notice that updating transmitted bytes and computing the threshold value is performed by simple arithmetic operations such as comparison and subtraction with complexity O(1). In addition, the complexity of sorting messages in a buffer is $O(m \times \log_2(m))$, where m is number of messages in the buffer. For example, average contact ratio per hour (obtained from 30 times simulation replications) is 396.1364 in LDAOR and 395.7455 in MaxProp. Even in the worst case, if all contacts send messages or delete messages, it is not so big complexity for today's advanced processors.

Although the complexity of LDAOR seems to be higher than the complexity of MaxProp and Epidemic, the number of adjacent nodes is usually small in opportunistic networks and $O(d \times m \times \log_2(m))$ can be approximately considered as $O(m \times \log_2(m))$, especially at low density traffic.

4. Performance Evaluation

The performance of LDAOR is compared with Prophet [21], POR [18], Epidemic, and MaxProp [9] in urban environments. The conducted performance evaluation is based on the network model stated in Subsection 3.1. Simulations are performed using the Opportunistic Network Environment (ONE) simulator [35], an open source Java-based simulator designed for evaluation of opportunistic networks and DTN routing algorithms.

Fig. 4. The Helsinki city scenario in [33].

To approach a real environment, three different types of vehicles (private vehicle, bus, and taxi) with specific mobility patterns are considered in the Helsinki map (see Fig. 4). This map matches to considered urban features stated in Subsection 3.1. Since Helsinki is one of the cities that have good features including more branches (junctions), it is widely used as benchmarking city in many articles, e.g. [31], [36]–[39]. Private vehicles move based on the Map-Based model developed in the ONE simulator. In this model, each private vehicle randomly chooses a path based on the city map to reach its destinations. Buses follow predefined routes based on the Bus Movement model so that when a bus reaches the end of its path, it moves back to the beginning of the path. Similar to buses, taxis move on predefined routes. Unlike a bus, a taxi can choose the shortest path between the source and destination. Recall several vehicles are randomly selected as destination of messages and other vehicles are selected as source vehicles. Note that destination nodes can be considered as

Table 5
Parameter settings in ONE simulator

Network simulator	ONE
Simulation area	4500×3400 m
Simulation duration	12 h
Message size	Uniform (800 B, 500 KB)
Buffer size	5 MB
Data rate	2 Mb/s
Message TTL	300 min
Transmit range	200 m
Average speed	Uniform (10, 50) km/h
Number of nodes (taxi, bus, POV)	Dynamic (see Table 5)
Mobility model	Taxi: MapRoute Movement
	Bus: Bus Movement
	POV: MapBased Movement

Table 6
Number of different vehicles

Number of nodes	20	40	60	80	100
Bus	1	2	3	4	5
Taxi	2	4	6	8	10
POV	17	24	51	68	85

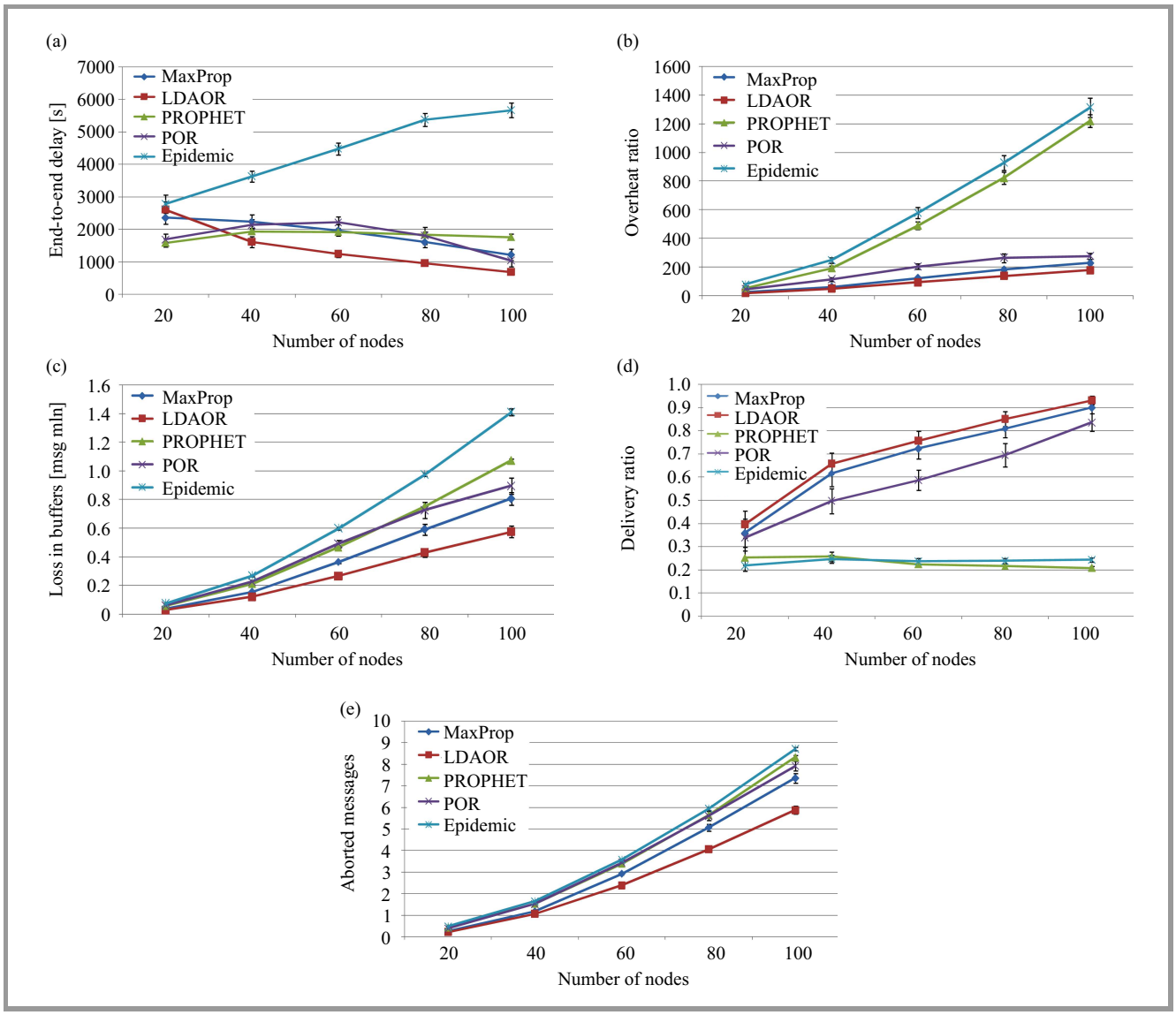


Fig. 5. Traffic load of 1 packet/ Uniform(5, 15) sec under different densities: (a) end-to-end delay, (b) overhead ratio, (c) loss in buffers, (d) delivery ratio, (e) aborted messages.

intermediate nodes for sending or receiving a message to other destination nodes, and therefore, the number of intermediate nodes is not less in the network.

Simulation parameters are shown in Table 5 and the numbers of different nodes are depicted in Table 6. In each diagram, simulation results are plotted with 95% confidence interval, where for each point 30 simulation replications have been done. Performance diagrams shown in the following are all measured within simulation period of 12 hours.

Note that the original POR paper has not considered any limitation for buffer of nodes. But, for simulating the algorithms under the same situation in presented simulations, authors consider buffer limitation for POR.

In the following evaluations, traffic load is expressed as the

$$\frac{\text{Arrival of } x \text{ packets}}{\text{Uniform } (y,\ z)}.$$

For example, in traffic load of one packet per Uniform (5, 15), inter-arrivals are computed from Uniform (5, 15) and in each inter-arrival one packet arrives at a node. Similarly in traffic load of 5 packets per Uniform (1, 2), inter-arrivals follow the distribution of Uniform (1, 2) sec and in each inter-arrival five packets arrive as a burst in a node.

The diagrams in Fig. 4 show the performance results under traffic load of one packet/Uniform (5, 15) sec. As one can see in Fig. 5a, when the density of the network is very small, LDAOR has more end-to-end delay than Prophet, POR, and MaxProp. This is because the number of nodes that meet the criteria defined by LDAOR is small. However, by increasing the density of vehicles in the network, end-to-end delay of LDAOR declines so that LDAOR achieves the best end-to-end delay when number of nodes becomes greater than 40. Figure 4b shows that not only LDAOR has the smallest overhead, but also its overhead is relatively

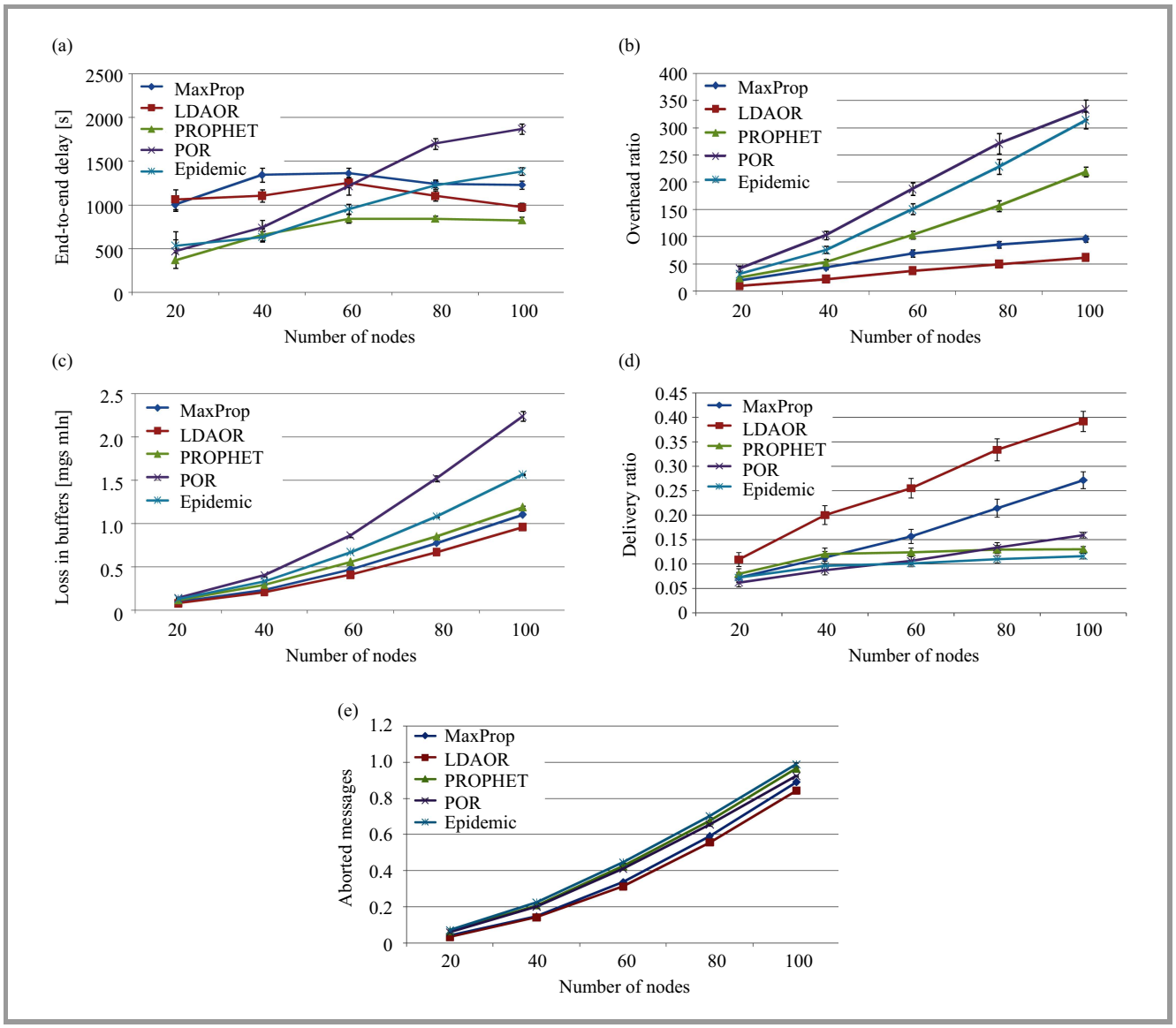


Fig. 6. Traffic load of one packet/Uniform(1, 2) sec under different densities: (a) end-to-end delay, (b) overhead ratio, (c) loss in buffers, (d) delivery ratio, (e) aborted messages.

flat. However, when number of nodes is very small, there is almost no difference among the overhead of LDAOR, POR, and MaxProp algorithms. This improvement on overhead is due to the criteria defined by LDAOR to avoid flooding.

Results in Fig. 5c illustrate the traffic loss from buffers when they overflow. The loss in LDAOR is less than other routing algorithms under different network densities. This is because of the fact that LDAOR mostly sends a message to those nodes that have much chance of delivering the message to its destination. Hence, the buffers are not completely filled in vain. Therefore, there will be enough space in buffers for saving those messages that have chance of delivery to their destinations.

Notice that number of relayed messages (and as a result the number of aborted messages) is more than the number of generated messages due to message replication. In addition, the number of dropped messages (loss in buffers) includes replicated messages. Hence, the values displayed in Fig. 5a,c,e are high.

At the first sight it seems that flooding-based OR algorithms should provide the highest delivery rate. However, as Fig. 5d depicts, LDAOR not only has increased the delivery rate but also has avoided flooding of messages. The delivery rate is rising when increasing network density. The main reasons for increasing the delivery rate under LDAOR compared to other algorithms are:

- a message in a buffer is only sent to those neighbor nodes located in better positions with respect to the message destination. Therefore, by preventing from additional transmissions and receiving of messages, bandwidth can be efficiently utilized and the opportunity for transmitting messages increases;
- priority for transmission of messages is provided based on the physical locations of new forwarder

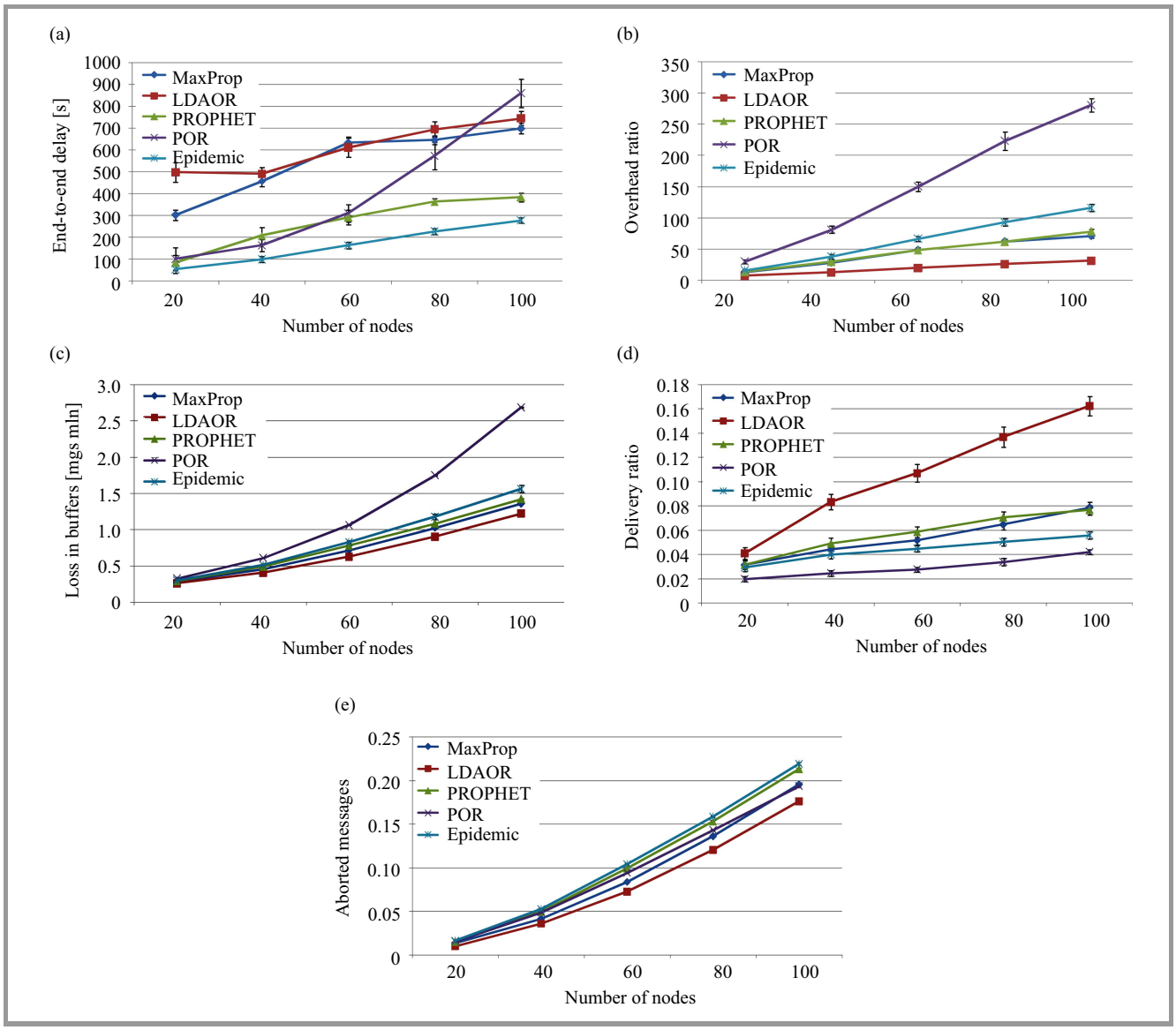


Fig. 7. Traffic load of 5 packets/Uniform(1, 2) sec under different densities: (a) end-to-end delay, (b) overhead ratio, (c) loss in buffers, (d) delivery ratio, (e) aborted messages.

nodes in addition to the history of contacts. By this, neighbor nodes (with high delivery probability) closer to the destination of a message can be selected as forwarder nodes of that message.

In the following, performance evaluation is performed at relatively higher traffic loads compared with Fig. 5. The diagrams in Fig. 6 and Fig. 7 show performance results under traffic load of one packet/Uniform(1, 2) sec and 5 packets/Uniform(1, 2) sec. By increasing traffic load, significant differences can be achieved compared to Fig. 5. For all routing algorithms, Fig. 6a and Fig. 7a depict significant reduction in end-to-end delay compared to Fig. 5a. Notice that by increasing traffic load, opportunity for transmission of all messages saved in a buffer decreases, thus reducing message delivery rate. Since message delivery ratio decreases, only those messages that are easy to be delivered quickly arrive at their destinations, thus reducing delay. Recall end-to-end delay is only averaged over successfully delivered messages. As it can be observed, end-to-end delay under LDAOR is more than some protocols at high traffic loads because determining a suitable node for each message leads to more waiting time in buffers.

Results in Fig. 6b and Fig. 7b show that POR has more overhead than other protocols. This is because POR only chooses one forwarder node for all messages in a carrier node buffer while their destinations could be different. Although many messages are sent under POR, there may be no chance for successfully delivering some of them by the selected forwarder node. The LDAOR has the lowest overhead since the messages are only delivered to appropriate nodes, thus avoiding additional transmissions of messages. Hence, buffer of nodes are less occupied and traffic loss in buffers reduces in LDAOR as shown in both Fig. 6c and Fig. 7c. As shown in Fig. 6c and Fig. 7c, by increasing traf-

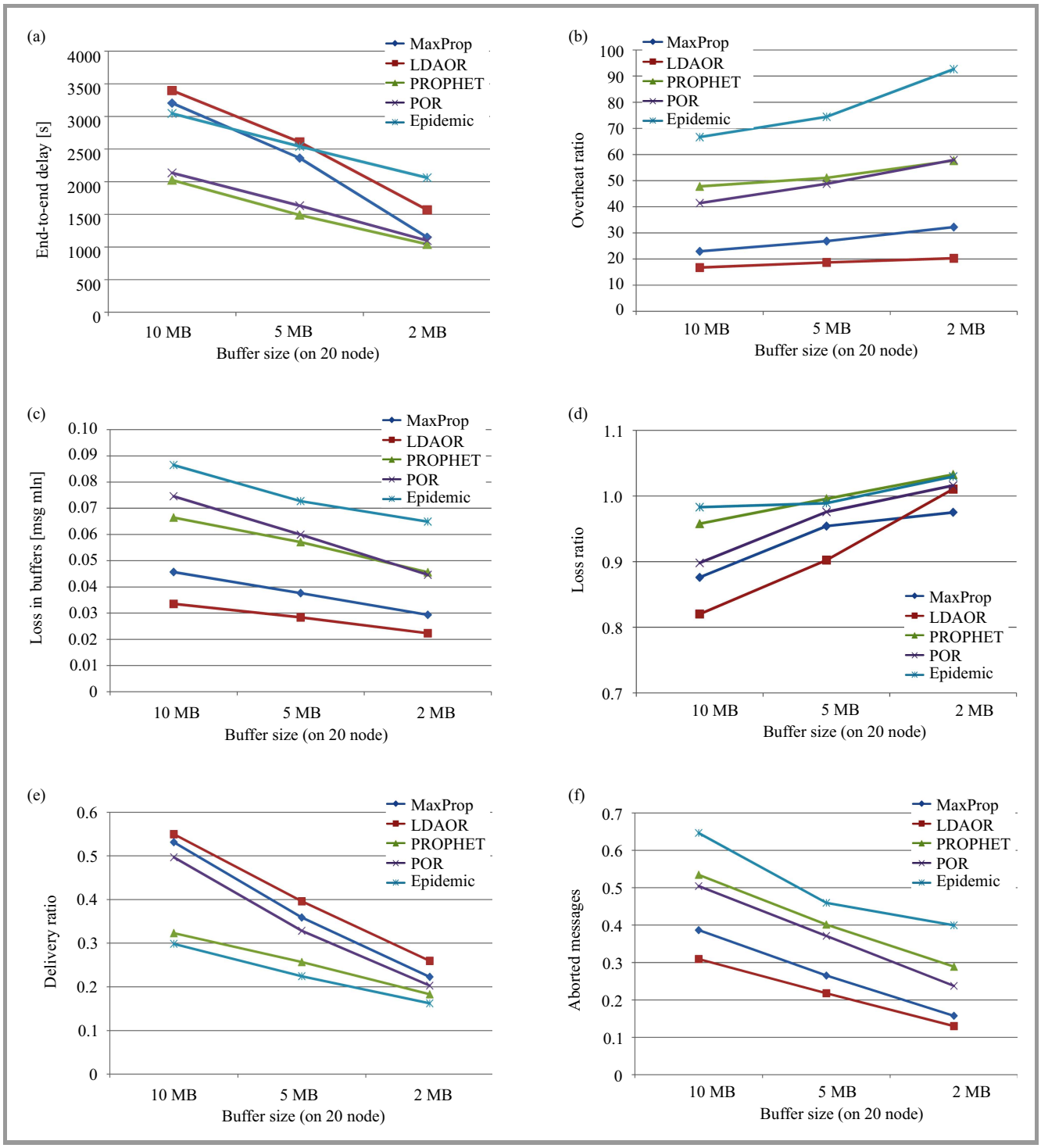


Fig. 8. Traffic load of 1 packets/Uniform(5, 15) sec under different buffer sizes: (a) end-to-end delay, (b) overhead ratio, (c) loss in buffers, (d) loss ratio, (e) delivery ratio, (f) aborted messages.

fic load, loss in buffers increases compared to Fig. 5c. This is because those messages that cannot be delivered should still remain in buffers of nodes. On the other hand, new traffic is always generated. Hence, buffers will overflow, thus resulting in dropping more messages.

As aforementioned, message delivery rate reduces by increasing traffic load (see Fig. 6d and Fig. 7d). As a result, in order to fully utilize the maximum bandwidth, selecting the best forwarder node for messages and their prioritization finds importance. As it can be observed, the LDAOR has the highest message delivery rate compared to other protocols, as shown in Fig. 6d and Fig. 7d.

By limited number of transmitted messages compared to more generated messages at high traffic load, the number of aborted messages goes down (compare Fig. 5e with Fig. 6e and Fig. 7e). The LDAOR still experiences the

least number of aborted messages compared to other routing protocols even at high traffic load, as shown in Fig. 6e and Fig. 7e.

Figure 8 shows performance of network under different buffer sizes in a VANET with 20 nodes. As it can be observed, reducing the buffer size reduces delivery ratio due to high limitation on buffer size. Notice when a buffer is full, messages should be removed from the buffer. As a result, a transmitted message may be removed from a buffer before being delivered to its destination. This issue increases overhead and decreases delivery ratio. When the buffer size reduces, the number of relayed messages reduces as well since small number of messages can be saved in buffers. Note the ratio of the number of dropped messages over the number of relayed message increases, thus increasing loss ratio (as shown in Fig. 8d). Since end-to-end delay is computed based on successfully delivered messages to their destinations, end-to-end delay also decreases because of decreasing the delivery ratio.

5. Conclusion

The LDAOR method has been proposed for opportunistic VANET in order to improve the performance of routing. The idea behind this approach is to consider physical location and direction of vehicles for choosing the best forwarder node among multiple neighbor nodes. It has been shown that LDAOR can provide better performances compared to other conventional routing protocols even when resources such as buffers are limited and traffic density is high. The LDAOR reduces traffic loss, aborted messages, and overhead ratio. On the other hand, it increases the probability of successful message delivery. The LDAOR provides smaller end-to-end delay at low traffic loads as well. Although the delivery ratio and overhead in LDAOR is not significantly different from MaxProp, but the differences between LDAOR and MaxProp in terms of end-to-end delay, loss in buffers and aborted messages are considerable. The complexity of LDAOR depends on the number of neighbor nodes in each contact. However, the number of neighbor nodes in opportunistic networks is not high in practice.

Acknowledgements

We would like to thank Mr. Ari Keränen for his assistance in using the ONE simulator.

References

[1] K. C. Lee and M. Gerla, "Opportunistic vehicular routing", in *Proc. Eur. Wirel. Conf. EW 2010*, Lucca, Italy, 2010, pp. 873–880.

[2] A. Casteigts, A. Nayak, and I. Stojmenovic, "Communication protocols for vehicular ad hoc networks", *Wirel. Commun. Mob. Comput.*, vol. 11, no. 5, pp. 567–582, 2011.

[3] R. C. Shah, S. Wietholter, J. Rabaey, and A. Wolisz, "When does opportunistic routing make sense?", in *Proc. 3rd IEEE Int. Conf. Pervasive Comput. Commun. Worksh. PerCom 2005*, Kauai Island, HI, USA, 2005, pp. 350–356.

[4] Z. Zhang and Q. Zhang, "Delay/disruption tolerant mobile ad hoc networks: latest developments", *Wirel. Commun. Mob. Comput.*, vol. 7, no. 10, pp. 1219–1232, 2007.

[5] W. Chen, R. K. Guha, T. J. Kwon, J. Lee, and Y.-Y. Hsu, "A survey and challenges in routing and data dissemination in vehicular ad hoc networks", *Wirel. Commun. Mob. Comput.*, vol. 11, no. 7, pp. 787–795, 2011.

[6] X. Zhao, "An adaptive approach for optimized opportunistic routing over delay tolerant mobile ad hoc networks", Ph.D. thesis, Computer Science Department, Rhodes University, Grahamstown, South Africa, 2007.

[7] R. J. D'Souza and J. Jose, "Routing approaches in delay tolerant networks: A survey", *Int. J. Comp. Appl.*, vol. 1, no. 17, pp. 8–14, 2010.

[8] A. Vahdat and D. Becker, "Epidemic routing for partially connected ad hoc networks", Tech. Rep. CS-200006, Duke University, 2000.

[9] J. Burgess, B. Gallagher, D. Jensen, and B. N. Levine, "MaxProp: routing for vehicle-based disruption tolerant networks", in *Proc. 25th IEEE Int. Conf. Comp. Commun. INFOCOM 2006*, Barcelona, Spain, 2006, pp. 1–11.

[10] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Single-copy routing in intermittently connected mobile networks", *IEEE/ACM Trans. on Netw.*, vol. 16, no. 1, pp. 63–76, 2008.

[11] T. Spyropoulos, K. Psounis, and C. Raghavendra, "Spray and wait: an efficient routing scheme for intermittently connected mobile networks", in *Proc. ACM SIGCOMM Worksh. Delay-tolerant Netw. WDTN'05*, Philadelphia, PA, USA, 2005, pp. 252–259.

[12] S. C. Nelson, M. Bakht, R. Kravets, and A. F. Harris, "Encounter based routing in DTNs", in *Proc. 28th Conf. Comp. Commun. IEEE INFOCOM 2009*, Rio de Janeiro, Brazil, 2009, pp. 846– 854.

[13] H. Kang and D. Kim, "Vector routing for delay tolerant networks", in *Proc. 68th IEEE Veh. Technol. Conf. VTC 2008-Fall*, Calgary, Canada, 2008, pp. 122–135.

[14] E. Bulut, Z. Wang, and B. K. Szymański, "Cost-effective multi period spraying for routing in delay-tolerant networks", *IEEE/ACM Trans. on Netw.*, vol. 18, no. 5, pp. 1530–1543, 2010.

[15] K. A. Harras and K. C. Almeroth, "Controlled flooding in disconnected sparse mobile networks", *Wirel. Commun. Mob. Comput.*, vol. 9, no. 1, pp. 21–33, 2009.

[16] L. M. Kiah, L. K. Qabajeh, and M. M. Qabajeh, Unicast position-based routing protocols for ad-hoc networks", *Acta Polytechnica Hungarica*, vol. 7, no. 5, pp. 19–46, 2010.

[17] H. Takagi and L. Kleinrock, "Optimal transmission ranges for randomly distributed packet radio terminals", *IEEE Trans. on Commun.*, vol. 32, no. 3, pp. 246–257, 1984.

[18] X. Li, M. Li, W Shu, and M. Y. Wu, "Packet-oriented routing in delay-tolerant vehicular sensor networks", *J. Inform. Sci. Engin. (JISE)*, vol. 25, no. 6, pp. 1803–1817, 2009.

[19] P.-C. Cheng, K. C. Lee, M. Gerla, and J. Härri, "GeoDTN+Nav: Geographic DTN routing with navigator prediction for urban vehicular environments", *Mob. Netw. Appl.*, vol. 15, no. 1, pp. 61–82, 2010.

[20] I. Jang, W. Choi, and H. Lim, "An opportunistic forwarding protocol with relay acknowledgment for vehicular ad hoc networks", *Wirel. Commun. Mob. Comput.*, vol. 11, no. 7, pp. 939–953, 2011.

[21] A. Lindgren, A. Doria, and O. Schelen, "Probabilistic routing in intermittently connected networks", *ACM SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 7, no. 3, pp. 19–20, 2003.

[22] D. Yu and Y.-B. Ko, "FFRDV: Fastest-ferry routing in DTN-enabled vehicular ad hoc networks", in *Proc. 11th Int. Conf. Adv. Commun. Technol. ICACT 2009*, Phoenix Park, Korea, 2009, vol. 2, pp. 1410–1414.

[23] Y. Ding, C. Wang, and L. Xiao, "A static-node assisted adaptive routing protocol in vehicular networks", in *Proc. 4th ACM Int. Worksh. Veh. Ad Hoc Netw. VANET'07*, Montreal, Quebec, Canada, 2007, pp. 59–68.

[24] I. Leontiadis and C. Mascolo, "GeOpps: Geographical opportunistic routing for vehicular networks", in *Proc. IEEE Int. Symp. World of Wirel., Mob. Multimed. Netw. WoWMoM 2007*, Helsinki, Finland, 2007.

[25] C. Liu and J. Wu, "An optimal probabilistically forwarding protocol in delay tolerant networks", in *Proc. 10th ACM Int. Symp. Mob. Ad Hoc Netw. Comput. ACM MobiHoc 2009*, New Orleans, Louisiana, USA, 2009, pp. 105–114.

[26] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Spray and focus: effcient mobility-assisted routing for heterogeneous and correlated mobility", in *Proc. 5th IEEE Int. Conf. Pervasive Comput. Commun. Worksh. PerCom 2007*, White Plains, New York, USA, 2007.

[27] Routing in delay-tolerant networking [Online]. Available: http://en.wikipedia.org/wiki/Routing_in_delay-tolerant_networking

[28] H. Fubler, M. Mauve, H. Hartenstein, D. Vollmer, and M. Kasemann, "Location-based routing for vehicular ad-hoc networks", *ACM SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 7, no. 1, pp. 47–49, 2003.

[29] M. Jerbi, S.-M. Senouci, T. Rasheed, and Y. M. Ghamri-Doudane, "Towards efficient geographic routing in urban vehicular networks", *IEEE Trans. Veh. Technol.*, vol. 58, no. 9, pp. 5048–5059, 2009.

[30] F. Li and Y. Wang, "Routing in vehicular ad hoc networks: A survey", *IEEE Veh. Technol. Mag.*, vol. 2, no. 2, pp. 12–22, 2007.

[31] J. M. Soares, M. Franceschinis, R. M. Rocha, W. Zhang, and M. A. Spirito, "Opportunistic data collection in sparse wireless sensor networks", *EURASIP J. Wirel. Commun. Netw.*, vol. 2011, pp. 6:1–6:20, 2011.

[32] G. Owen and M. Adda, "SOLS: Self organizing distributed location server for wireless ad hoc networks", *Int. J. Comp. Netw. & Commun. (IJCNC)*, vol. 1, no. 1, pp. 17–30, 2009.

[33] S. Farrell *et al.*, "Report on an arctic summer DTN 2010 trial", Tech. rep., Trinity College, Dublin, Ireland, 2011 (draft 2011-05-18 Work-in-progress) [Online]. Available: http://dtn.dsg.cs.tcd.ie/n4c-summer10/summer10.pdf

[34] J. A. B. Link, C. Wollgarten, S. Schupp, and K. Wehrle, "Perfect difference sets for neighbor discovery energy efficient and fair", in *Proc. 3rd ACM Extreme Conf. on Commun. ExtremeCom 2011*, Manaus, Brazil, 2011.

[35] Opportunistic Network Environment (ONE) homepage, ver. 1.4.1 [Online]. Available: http://www.netlab.tkk.?/%7Ejo/dtn/ (accessed June 2011).

[36] V. Soares, J. Rodrigues, P. Salvador, and A. Nogueira, "Improvement of messages delivery time on vehicular delay-tolerant networks", in *Proc. Int. Worksh. Next Gener. Wirel. Mob. Netw. NGWMN 2009*, Vienna, Austria, 2009.

[37] A. Keränen, T. Kärkkäinen, and J. Ott, "Simulating mobility and DTNs with the ONE", *J. Commun.*, vol. 5, no. 2, pp. 92–105, 2010.

[38] V. Soares, F. Farahmand, and J. R. Rodrigue, "Improving vehicular delay-tolerant network performance with relay nodes", in *Proc. 5th Euro-Ngi Conf. Next Gener. Internet Netw. NGI 2009*, Aveiro, Portugal, 2009.

[39] F. Alnajjar and T. Saadawi, "Performance analysis of routing protocols in delay/disruption tolerant mobile ad hoc networks", in *Proc. 10th WSEAS Int. Conf. on Elec., Hardw., Wirel. Opt. Commun. EHAC'11, and 10th WSEAS Int. Conf. on Sig. Process., Robot. and Autom. ISPRA'11, and 3rd WSEAS Int. Conf. Nanotechnol. NANOTECHNOLOGY'11*, Cambridge, UK, 2011, pp. 407–417.

Marziyeh Barootkar received her B.S. degree in Software Engineering in 2008 from Zanjan University, Zanjan, Iran, and M.Sc. degree in Information Technology from Sahand University of Technology in 2011. Her research interests include wireless, MANET and VANET networks and performance evaluation.

E-mail: m_barootkar@sut.ac.ir
Computer Networks Research Lab
Electrical Engineering Technologies Research Center Sahand University of Technology
Tabriz, Iran

Akbar Ghaffarpour Rahbar received the B.Sc. and M.Sc. degrees in computer hardware and computer architecture from the Iran University of Science and Technology, Tehran, Iran, in 1992 and 1995, respectively, and the Ph.D. degree in computer science from the University of Ottawa, Ottawa, Canada, in 2006. He is currently a Professor with the Electrical Engineering Department, Sahand University of Technology, Sahand New Town, Tabriz, Iran. He is the director of the Computer Networks Research Laboratory, Sahand University. Dr. Rahbar is a senior member of the IEEE. He is currently on the Editorial Board of the Wiley Transactions on Emerging Telecommunications Technologies Journal and the Journal of Convergence Information Technology. He is editor-in-chief of Journal of Nonlinear Systems in Electrical Engineering. His current research interests include optical networks, optical packet switching, scheduling, PON, IPTV, network modeling, analysis and performance evaluation, the results of which can be found in over 110 technical papers.
E-mail: ghaffarpour@sut.ac.ir
Computer Networks Research Lab
Electrical Engineering Technologies Research Center
Sahand University of Technology
Tabriz, Iran

Masoud Sabaei received his B.Sc. degree from Isfahan University of Technology, Isfahan, Iran, and his M.Sc. and Ph.D. form Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran, all in the field of Computer Engineering in 1992, 1995 and 2000, respectively. Dr. Sabaei has been professor of Computer Engineering Department, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran since 2002. His research interests are software defined networking, Internet of Things, wireless networks and telecommunication network management.
E-mail: sabaei@aut.ac.ir
Computer Engineering and Information Technology Department
Amirkabir University of Technology
Tehran, Iran

Paper

A Novel Technique of Optimization for the COCOMO II Model Parameters using Teaching-Learning-Based Optimization Algorithm

Thanh Tung Khuat and My Hanh Le

The University of Danang, University of Science and Technology, Danang, Vietnam

Abstract—Software cost estimation is a critical activity in the development life cycle for controlling risks and planning project schedules. Accurate estimation of the cost before the start-up of a project is essential for both the developers and the customers. Therefore, many models were proposed to address this issue, in which COCOMO II has been being widely employed in actual software projects. Good estimation models, such as COCOMO II, can avoid insufficient resources being allocated to a project. However, parameters for estimation formula in this model have not been optimized yet, and so the estimated results are not close to the actual results. In this paper, a novel technique to optimize the coefficients for COCOMO II model by using teaching-learning-based optimization (TLBO) algorithm is proposed. The performance of the model after optimizing parameters was tested on NASA software project dataset. The obtained results indicated that the improvement of parameters provided a better estimation capabilities compared to the original COCOMO II model.

Keywords— COCOMO II, cost estimation, NASA software, optimization, teaching-learning-based optimization algorithm.

1. Introduction

Effort and cost estimation process in any software engineering project is an extremely important component. The success or failure of projects depends greatly on the accuracy of effort and schedule estimations. Errors in the cost estimation process can result in the serious issues [1]. Underestimating the costs may result in management approving proposed systems that then exceed their budgets, with underdeveloped functions and poor quality, and failure to complete on time. Overestimating may result in too many resources committed to the project, or, during contract bidding, result in not winning the contract, which can lead to the loss of jobs. Therefore, it is desired to find out the method to estimate the effort for software projects accurately. The introduction of the COCOMO II model has contributed significantly to the enhancement of accuracy in the software cost estimation process and currently this is one of the most commonly used models. COCOMO II has three sub-models including the Application Composition, the early design and the post-architecture (PA) models. The application composition model is used to estimate effort and schedule on projects that use integrated computer aided software engineering tools for rapid application development. The early design and the PA models are employed in estimating effort and schedule on application generator, system integration, or infrastructure developments [2]. In this work, we take into account the PA model, which is a detailed model being used once the project is ready to develop and sustain a fielded system.

Although COCOMO II is an efficient software cost estimation model, the accuracy of the model's output still relies on several constant values in the parametric-based estimation equations. These constants have not been optimized yet, and thus the accuracy of estimations on projects is not high in comparison with the actual effort and time. In this work, the constant values of COCOMO II model are optimized by using teaching-learning-based optimization (TLBO) algorithm. The proposed approach increases the efficiency of COCOMO II model when experimenting on "NASA 93" projects [3]. The test results showed that COCOMO II with optimized parameters had better performance in the software project cost estimation compared to the original COCOMO II and there was also smaller magnitude of relative error (MRE).

The remainder of paper is organized as follows. Section 2 introduces the COCOMO II model. Section 3 represents the teaching-learning-based optimization algorithm and its application into software cost estimation issues. The experiments are shown in Section 4 and finally in Section 5, the conclusion and future works are presented.

2. COCOMO II Model

COnstructive COst MOdel II (COCOMO II) [4], which was developed in 1995, is a model that allows one to estimate the cost, effort, and schedule when planning a new software development activity. It takes qualitative inputs and produces quantitative results. In COCOMO II, the effort is represented as person-months (PMs). A person-month is the amount of time one person spends working on the software development project for one month [5]. The

Table 1
Cost drivers for COCOMO-II PA model

Driver	Symbol	Very low	Low	Nominal	High	Very high	Extra high
RELY	EM_1	0.82	0.92	1.00	1.10	1.26	–
DATA	EM_2	–	0.90	1.00	1.14	1.28	–
CPLX	EM_3	0.73	0.87	1.00	1.17	1.34	1.74
RUSE	EM_4	–	0.95	1.00	1.07	1.15	1.24
DOCU	EM_5	0.81	0.91	1.00	1.11	1.23	–
TIME	EM_6	–	–	1.00	1.11	1.29	1.63
STOR	EM_7	–	–	1.00	1.05	1.17	1.46
PVOL	EM_8	–	0.87	1.00	1.15	1.30	–
ACAP	EM_9	1.42	1.19	1.00	0.85	0.71	–
PCAP	EM_{10}	1.34	1.15	1.00	0.88	0.76	–
PCON	EM_{11}	1.29	1.12	1.00	0.90	0.81	–
APEX	EM_{12}	1.22	1.10	1.00	0.88	0.81	–
PLEX	EM_{13}	1.19	1.09	1.00	0.91	0.85	–
LTEX	EM_{14}	1.20	1.09	1.00	0.91	0.84	–
TOOL	EM_{15}	1.17	1.09	1.00	0.90	0.78	–
SITE	EM_{16}	1.22	1.09	1.00	0.93	0.86	0.80
SCED	EM_{17}	1.43	1.14	1.00	1.00	1.00	–

Table 2
Scale factor values for COCOMO II model

Scale factors	Symbol	Very low	Low	Nominal	High	Very high	Extra high
PREC	SF_1	6.20	4.96	3.72	2.48	1.24	0.00
FLEX	SF_2	5.07	4.05	3.04	2.03	1.01	0.00
RESL	SF_3	7.07	5.65	4.24	2.83	1.41	0.00
TEAM	SF_4	5.48	4.38	3.29	2.19	1.10	0.00
PMAT	SF_5	7.80	6.24	4.68	3.12	1.56	0.00

COCOMO II model predicts the software development effort by using the formula shown in Eq. 1.

$$PM = A \cdot Size^E \cdot \prod_{i=1}^{17} EM_i, \tag{1}$$

where A is a multiplicative constant having the value of 2.94, *Size*, which is the estimated size of software development, is the most important factor in calculating the effort of the software project and it is measured in kilo line of code (KLOC). EM_i is one of a set of effort multipliers shown in Table 1. This is the seventeen PA effort multipliers (EM) are used in the COCOMO II model to adjust the nominal effort. These multipliers are values of rating level of every multiplicative cost driver used to capture features of the software development affecting the effort to complete the project [5].

The exponent E in Eq. 1 is an aggregation of five scale factors (SF) that account for the relative economies or diseconomies of scale encountered for software projects of different sizes [4] and is computed as the following formula:

$$E = B + 0.01 \cdot \sum_{j=1}^{5} SF_j, \tag{2}$$

where B is a constant having the value of 0.91. Each scale factor has a range of rating levels, from very low to extra high. Each rating level has a weight which is presented in Table 2.

In addition to the effort, the software companies are also more interested in calculating the development time (TDEV) for projects [6]. It is derived from the effort according to the following equations:

$$TDEV = C \cdot PM^F, \tag{3}$$

$$F = D + 0.2 \cdot 0.01 \cdot \sum_{i=1}^{5} SF_j. \tag{4}$$

The values of C and D for the COCOMO II schedule equation are obtained by calibration to the actual schedule values for the 161 project currently in the COCOMO II database and results are C = 3.67 and D = 0.28.

Mean of MRE (MMRE) and prediction level (PRED) are usually used as an accurate reference value in the study of the software effort estimation. COCOMO's performance is often gauged in terms of PRED(30) [7]. PRED(30) is computed from the relative error (RE), which is the relative size of the difference between the actual and estimated values:

$$RE_i = \frac{estimate_i - actual_i}{actual_i}. \tag{5}$$

After that, the MMRE is the percentage of the absolute values of the relative errors, averaged over the T projects in the test dataset.

$$MRE_i = |RE_i|\,, \tag{6}$$

$$MMRE = \frac{100}{T} \cdot \sum_{i=1}^{T} MRE_i\,. \tag{7}$$

PRED(N) reports the average percentage of estimates that were within $N\%$ of the actual values:

$$PRED\,(N) = \frac{100}{T} \cdot \sum_{i=1}^{T} \begin{cases} 1, & \text{if } MRE_i \leq \frac{N}{100} \\ 0, & \text{otherwise} \end{cases}\,. \tag{8}$$

3. Teaching-Learning-Based Optimization Algorithm

In the COCOMO II model, the values of A, B, C, and D are constant and they are not tuned following the actual effort and time of new software projects. Therefore, the accuracy of estimated activities for projects is not exact. In this paper, the authors propose a novel approach to optimize these parameters of COCOMO II by using the historical software projects and TLBO algorithm.

3.1. Fitness Function for the Software Cost Estimation Problem

In the effort and time estimation issue for software projects, if the estimated cost roughly matches the actual end cost then the project is completed successfully. This means that the lower of the value of MMRE, the higher accuracy of the estimated cost is. Therefore, this paper uses the value of MMRE on training datasets of historical projects to assess the quality of cost estimations. The fitness function is the sum of time MMRE and effort MMRE as follows:

$$f = MMRE(Time) + MMRE(Effort)\,. \tag{9}$$

3.2. Teaching-Learning-Based Optimization Algorithm

Teaching-learning-based optimization algorithm which proposed by Rao *et al.* [8] is one of the most recently developed meta-heuristics. This algorithm is the population-based algorithm inspired by learning process in a classroom. For the TLBO, the population is considered as a group of learners or a class of learners. The search process contains two phases: teacher phase and learner phase.

3.2.1. Teacher Phase

In the teacher phase, learners get knowledge from a teacher. In the entire population, the best solution is considered as the teacher ($\vec{X}_{teacher}$). In this phase, the teacher tries to improve the results of other individuals ($\vec{X}_i$) by increasing the average result of the classroom ($\vec{X}_{mean}$) towards his/her level [8]. The solution is updated according to the difference between the existing and the new mean given by:

$$\vec{X}_{new} = \vec{X}_i + r_i \cdot (\vec{X}_{teacher} - T_f \cdot \vec{X}_{mean})\,, \tag{10}$$

where T_f is a teaching factor that decides the value of mean to be changed, and r_i is a random number in the range of $0 \ldots 1$. The value of T_f can be either 1 or 2, which is again a heuristic step. Moreover, $\vec{X}_{new}$ and $\vec{X}_i$ are the new and existing solutions of the i-th learner, respectively.

3.2.2. Learner Phase

In the learner phase, learners try to increase their knowledge by interacting with others. A learner interacts randomly with other learners with the help of group discussions, presentations, formal communications, etc. [8]. A learner learns something new if another learner has more knowledge than him or her. The modification of the learner is represented as follows:

$$\vec{X}_{new} = \vec{X}_i + r_i \cdot (\vec{X}_j - \vec{X}_k) \qquad \text{if} \quad f(\vec{X}_j) < f(\vec{X}_k)\,, \tag{11}$$

$$\vec{X}_{new} = \vec{X}_i + r_i \cdot (\vec{X}_k - \vec{X}_j) \qquad \text{if} \quad f(\vec{X}_k) < f(\vec{X}_j)\,, \tag{12}$$

Algorithm 1: The TLBO pseudo code

Input:
- d is the number of variables of problems
- n is the number of students
- G is the maximal number of generations

Output: The best individual in the population: $\vec{x}_{best} = \{x^1_{best}, x^2_{best}, x^D_{best}\}$.

Generate n initial students of the classroom randomly.
Calculate fitness function $f(\vec{X}_i)$ for whole students of the classroom.
$id = 0$
while $id < n$ && all $f(\vec{X}_i) \neq 0$ **do**
  Calculate the mean of each variable $\vec{X}_{mean}$
  Identify the best solution (teacher)
  for $i = 1$ **to** n **do**
    Find teaching factor $T_f = round[1 + rand(0,1)\{2-1\}]$
    Modify solution based on teacher:
      $\vec{X}_{new,i} = \vec{X}_i + rand(0,1) \cdot (\vec{X}_{teacher} - T_f \cdot \vec{X}_{mean})$
    Calculate fitness function for new student $f(\vec{X}_{new,i})$
    if ($\vec{X}_{new,i}$ is better than $\vec{X}_i$) **then**
      $\vec{X}_i = \vec{X}_{new,i}$
    end if
    Randomly select two learners $\vec{X}_j$ and $\vec{X}_k$ ($j \neq k$)
    if ($\vec{X}_j$ is better than $\vec{X}_k$) **then**
      $\vec{X}_{new,i} = \vec{X}_i + rand(0,1) \cdot (\vec{X}_j - \vec{X}_k)$
    else
      $\vec{X}_{new,i} = \vec{X}_i + rand(0,1) \cdot (\vec{X}_k - \vec{X}_j)$
    end if
    if ($\vec{X}_{new,i}$ is better than $\vec{X}_i$) **then**
      $\vec{X}_i = \vec{X}_{new,i}$
    end if
  end for
  $id++$
end while

where $\vec{X}_k$ and $\vec{X}_j$ ($j \neq k$) are two students chosen randomly in the population, and f is the fitness function.

If the new solution $\vec{X}_{new}$ is better, it is accepted in the population. The algorithm will continue until the termination condition is met. The Algorithm 1 shows the pseudo code of TLBO algorithm step by step.

4. Experimentation

The main objective of the experiment carried out is to reduce the uncertainty of current COCOMO II post architecture coefficients (A, B, C and D) and to get the best software effort estimation results being equivalent to the actual effort by using the TLBO algorithm. Experiments have been conducted on "NASA 93" dataset [3], in which 65 projects were used as training data to optimize the parameters for COCOMO II model and the other 28 projects were used for testing the performance of this model after optimizing coefficients. In this experiment, the configuration parameters for the TLBO are that the number of students is 200 and the number of generations is 2000.

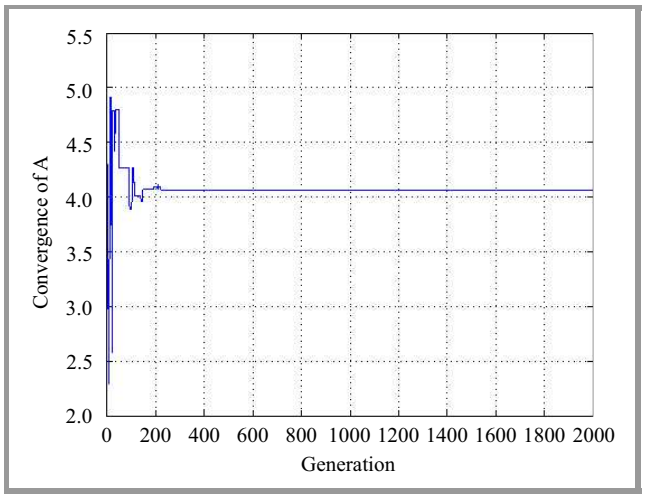


Fig. 1. Convergence of the model parameter A.

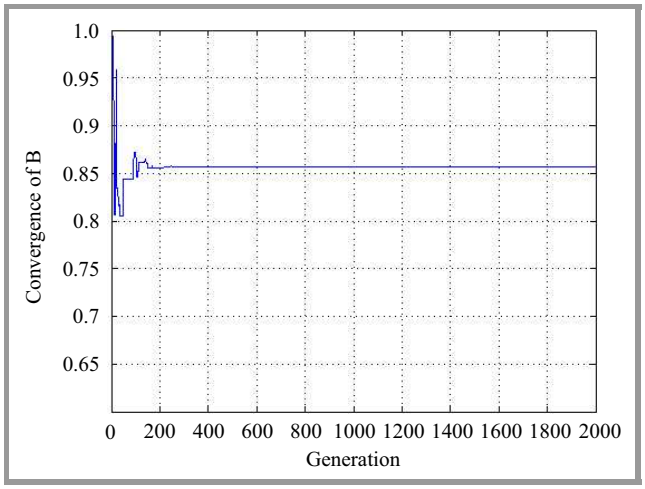


Fig. 2. Convergence of the model parameter B.

The optimized COCOMO II PA coefficients by using the TLBO are A = 4.064, B = 0.857, C = 2.938 and D = 0.357.

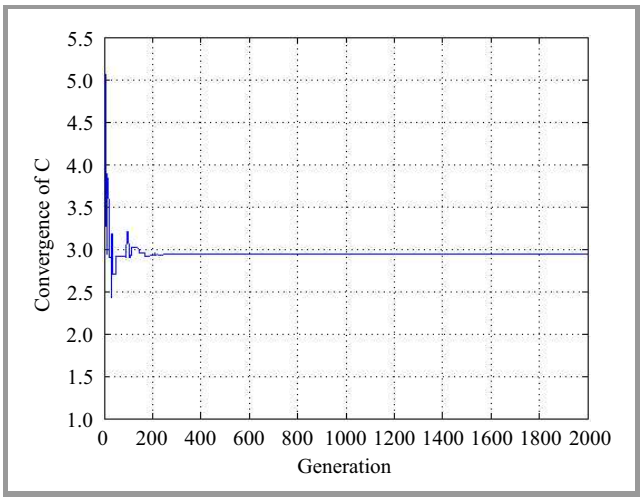


Fig. 3. Convergence of the model parameter C.

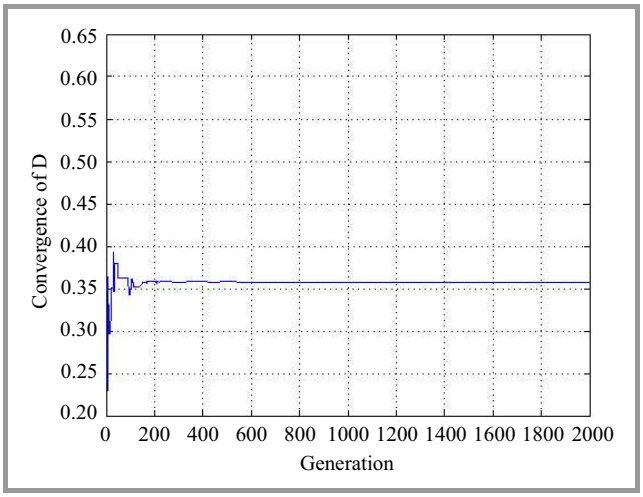


Fig. 4. Convergence of the model parameter D.

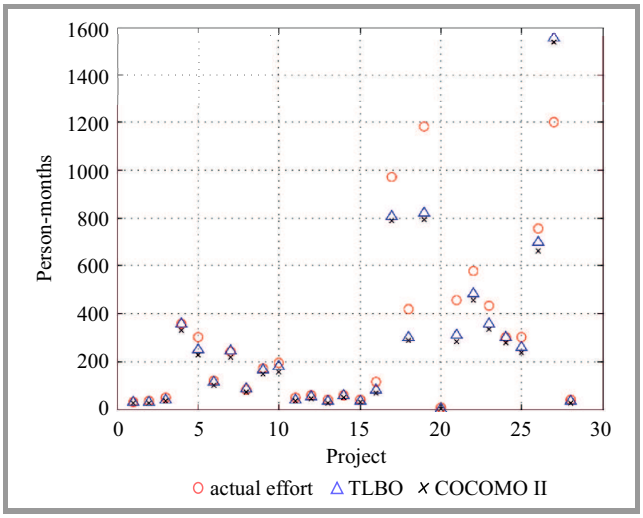


Fig. 5. Actual effort and estimated effort using TLBO and COCOMO II.

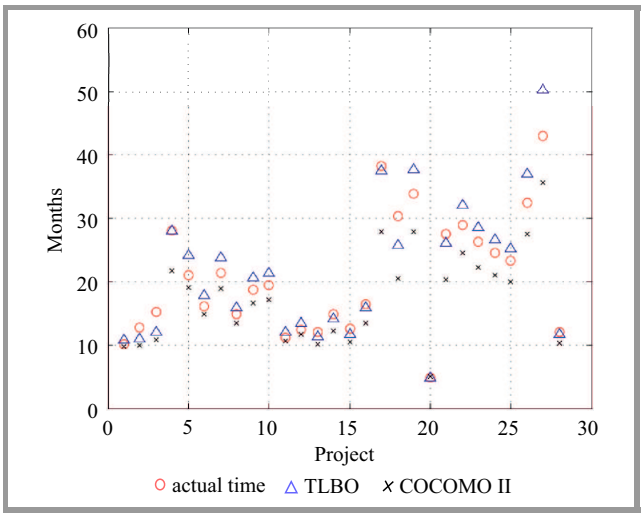


Fig. 6. Actual time and estimated time using TLBO and COCOMO II.

The convergence of the model parameters after each generation is described in Figs. 1–4.

Table 3
MRE values for estimations using TBLO and COCOMO II

Project ID	MRE of effort		MRE of time	
	TLBO	COCOMO II	TLBO	COCOMO II
3	0.0085	0.2008	0.0722	0.0367
13	0.1477	0.2989	0.1475	0.2294
15	0.1744	0.3000	0.2029	0.2870
16	0.0009	0.0774	0.0009	0.2244
22	0.1593	0.2403	0.1449	0.0940
23	0.0481	0.1760	0.1120	0.0768
28	0.0302	0.0845	0.1133	0.1219
29	0.0397	0.1279	0.0757	0.0953
31	0.0158	0.1436	0.1000	0.1118
32	0.0533	0.1725	0.1000	0.1163
34	0.1712	0.3212	0.0805	0.0462
35	0.0778	0.2327	0.0893	0.0551
36	0.2082	0.3675	0.0622	0.1650
37	0.0005	0.1716	0.0468	0.1798
39	0.1163	0.2862	0.0610	0.1682
40	0.2831	0.3993	0.0315	0.1835
44	0.1688	0.1887	0.0172	0.2723
47	0.2810	0.3131	0.1502	0.3256
56	0.3031	0.3279	0.1152	0.1783
58	0.5435	0.6716	0.0006	0.0187
61	0.3202	0.3841	0.0528	0.2619
69	0.1571	0.2065	0.1130	0.1525
70	0.1758	0.2221	0.0853	0.1529
72	0.0003	0.0778	0.0906	0.1435
73	0.1333	0.2066	0.0855	0.1402
76	0.0748	0.1236	0.1392	0.1549
77	0.2956	0.2789	0.1714	0.1724
93	0.0373	0.2618	0.0273	0.1510
MMRE	**14.38%**	**24.51%**	**8.89%**	**15.41%**

The graph in Fig. 5 illustrates the results of the effort estimation using the parameters optimized by TLBO and the original coefficients of COCOMO II compared with the actual effort. Figure 6 is the graph of values of estimated time by employing the parameters optimized by the TLBO and the original coefficients of COCOMO II in comparison with the actual time.

Based on these results, it can be seen that the COCOMO II with optimized parameters by the TLBO gived the higher estimated results compared to the original one because the estimated effort and time of the improved COCOMO II were more close to actual effort and time than the original model.

Table 3 shows the comparison of MRE between the improved COCOMO II model with optimized parameters by the TLBO and original model in terms of effort and time for 28 projects from NASA software project datasets. The obtained results indicated that the improved model have had lower MRE error compared to the original COCOMO model. As also can be seen that the model with optimized parameters has reduced MMRE error value for both the effort and time and it can be said that these are helpful methods for the software cost estimation process.

Another criterion to assess the effectiveness of the improved model is the value of PRED. From Table 3, the values of PRED(30) by using Eq. (8) for models as presented in Table 4 can be computed.

Table 4
The values of PRED(30) using TBLO and COCOMO II

	Time	Effort
TLBO	100%	89.29%
COCOMO II	96.43%	75%

Actually, the proposed method has considerably enhanced the accuracy of the software cost estimation in terms of effort and time.

5. Conclusion and Future Work

Accurate software cost estimation is a critical activity in the project planning. The authors found that the use of TLBO Algorithm to optimize the parameters of the COCOMO II model has resulted in the predicted effort and time of this model closing to the real effort. Thus, the proposed algorithm has effectively addressed the complicated optimization problem and achieved more accurate results by optimizing the coefficients of the COCOMO II model. The obtained results will contribute to the development of software projects within time and budgets.

However, there still exists some drawbacks in presented study. Experiments are only carried out on NASA projects which are characterized by lines of code, a number of scale factors and effort multipliers. The obtained results indicate that the improved model is more accurate on NASA

projects than traditional COCOMO II. Authors firmly believe that the proposed model is also more efficient than the conventional COCOMO II model for non-NASA projects influenced by factors as mentioned above. Due to the difficult in the project dataset collection, this has not yet been proven by experiments. Therefore, authors intend to apply the improved model for experimental studies on non-NASA projects in the future.

COCOMO II expands the capabilities of the original model and can estimate applications using modern development methods [9]. In the report of Jones [10], he pointed out that COCOMO II was one of the most widely used estimation tools in 2013. In [11], Menzies *et al.* analyzed the experiments and compared COCOMO II to other software effort estimation models to find the answer for the question "*Are the old parametric calibrations relevant to more recent projects*?". Authors concluded that COCOMO II calibration is relevant to more recent projects. These figures indicate that the improved COCOMO II model still counts in estimating the effort for contemporary software projects. Therefore, the proposed model in this paper might be utilized for predicting the effort of the current software projects. Authors plan to carry out experiments to verify the effectiveness of the improved COCOMO II on the modern projects. This is an important area that requires further research.

In the future work, authors also intend to apply the TLBO Algorithm for Agile Software Effort Estimation. The various nature-inspired algorithms will be employed to optimize the parameters of the COCOMO II model as well.

References

[1] C. Jones, "Why flawed software projects are not cancelled in time", *Cutter IT J.*, vol. 10, no. 12, pp. 12–17, 2003.

[2] B. Clark, S. Devnani-Chulani, and B. Boehm, "Calibrating the COCOMO II post-architecture model", in *Proc. 20th Int. Conf. Softw. Engin. ICSE*, Kyoto, Japan, 1998, pp. 477–480.

[3] T. Menzies, The Tera-PROMISE Repository for COCOMO 93, 2015 [Online]. Available: http://openscience.us/repo/effort/cocomo/nasa93.html

[4] B. Boehm, B. Clark, E. Horowitz, C. Westland, R. Madachy, and R. Selby, "Cost Models for Future Software Life Cycle Processes: COCOMO 2.0", *Annals of Softw. Engin.*, vol. 1, no. 1, pp. 57–94, 1995.

[5] C. Abts, B. Clark, S. Devnani-Chulani, E. Horowitz, R. Madachy, D. Reifer, R. Selby, and B. Steece, "COCOMO II Model Definition Manual", Tech. Rep., Center for Software Engineering, University of Southern California, Los Angeles, CA, USA, 1998.

[6] J. Kaur and R. Sindhu, "Parameter estimation of COCOMO II using tabu search", *Int. J. Comp. Sci. Inform. Technol.*, vol. 5, no. 3, pp. 4463–4465, 2014.

[7] Z. Chen, T. Menzies, D. Port, and B. Boehm, "Feature Subset Selection Can Improve Software Cost Estimation Accuracy", in *Proc. Int. Worksh. Predic. Models in Softw. Engin. PROMISE 2005*, St. Louis, MO, USA, 2005, pp. 1–6.

[8] R. V. Rao, V. J. Savsani, and D. P. Vakharia, "Teaching-Learning-Based optimization: A novel method for constrained mechanical design optimization problems", *Computer-Aided Design*, vol. 43, pp. 303–315, 2011.

[9] B. Boehm, C. Abts, and S. Chulani, "Software development cost estimation approaches – A survey", *Annals of Softw. Engin.*, vol. 10, no. 1–4, pp. 177-205, 2000.

[10] C. Jones, "A short history of software estimation tools", Tech. Rep., VP and CTO, Namcook Analytics LLC, Narragansett, RI, USA, 2013.

[11] T. Menzies, B. Boehm, Y. Yang, J. Hihn, and N. Lekkalapudi, "Just how good is COCOMO and parametric estimation", in *Proc. 29th Int. Forum on COCOMO and Syst. Softw. Cost Model.*, Los Angeles, CA, USA, 2014 [Online]. Available: http://csse.usc.edu/new/events/cocomo-2014/program

Thanh Tung Khuat completed the B.Sc. degree in Software Engineering from University of Science and Technology, Danang, Vietnam, in 2014. Currently, he is participating in the research team at DATIC Laboratory, University of Science and Technology, Danang. His research interests focus on software engineering, software testing, evolutionary computation, intelligent optimization techniques and applications in software engineering.

Email: thanhtung09t2@gmail.com
The University of Danang
University of Science and Technology
54 Nguyen Luong Bang, Lien Chieu
Danang, Vietnam

My Hanh Le is currently a lecturer of the Information Technology Faculty, University of Science and Technology, Danang, Vietnam. She gained M.Sc. degree in 2004 and completed the Ph.D. program in Computer Science at the University of Danang in 2015. Her research interests are about software testing and more generally application of heuristic techniques to problems in software engineering.

Email: ltmhanh@dut.udn.vn
The University of Danang
University of Science and Technology
54 Nguyen Luong Bang, Lien Chieu
Danang, Vietnam

Paper

100 Gb/s Data Link Layer – from a Simulation to FPGA Implementation

Łukasz Łopaciński[1], Marcin Brzozowski[2], Rolf Kraemer[2], Steffen Buechner[1], and Jörg Nolte[1]

[1] *Brandenburg University of Technology Cottbus-Senftenberg, Cottbus, Germany*
[2] *Innovations for High Performance Microelectronics GmbH, Frankfurt (Oder), Germany*

Abstract—In this paper, a simulation and hardware implementation of a data link layer for 100 Gb/s terahertz wireless communications is presented. In this solution the overhead of protocols and coding should be reduced to a minimum. This is especially important for high-speed networks, where a small degradation of efficiency will lower the user data throughput by several gigabytes per second. The following aspects are explained: an acknowledge frame compression, the optimal frame segmentation and aggregation, Reed-Solomon forward error correction, an algorithm to control the transmitted data redundancy (link adaptation), and FPGA implementation of a demonstrator. The most important conclusion is that changing the segment size influences the uncoded transmissions mostly, and the FPGA memory footprint can be significantly reduced when the hybrid automatic repeat request type II is replaced by the type I with a link adaptation. Additionally, an algorithm for controlling the Reed-Solomon redundancy is presented. Hardware implementation is demonstrated, and the device achieves net data rate of 97 Gb/s.

Keywords—ARQ, FEC, frame aggregation, HARQ, link adaptation, Reed-Solomon FEC, segmentation.

1. Introduction

Within the last two years, a few new approaches for 100 Gb/s wireless communication have been proposed. Research on physical transceivers and baseband processing changed the state of the art in the targeted area. Components required to modulate the 100 Gb/s wireless signal in the terahertz band are close to release in engineering samples. In [1] a 100 Gb/s baseband signal has been sent over a 237.5 GHz link. Similar results are shown in [2]. More teraherz (THz) communication activity on the physical layer is documented in [3]–[6]. In this paper, a data link layer for a wireless 100 Gb/s system is proposed. The designed solution is 14 times faster than the state-of-the-art 802.11ac (5 GHz) and 802.11ad (60 GHz) WLANs shown in [7]. Even if the achievement in 100 Gb/s wireless communication is impressive, the PHY circuit, baseband processing, and data link layer have not been integrated yet. To the authors best knowledge, the fully functional data link layer dedicated for 100 Gb/s wireless THz application has not been shown yet.

2. Related Work

Many research efforts have been addressed to highly efficient wireless protocols. A data link layer goodput analysis is a very popular topic, especially for WLAN. Presented methodology for frame segmentation is very similar to efforts presented by T. Li *et al*. in [8], where segmentation is deeply investigated. Li proves that a frame fragmentation may increase protocol efficiency. There are many authors, who publish papers similar work, for example [9]–[11]. They consider possible improvements for the WLANs, mostly by using fragmentation and aggregation. The main difference is that in this paper, authors are strongly focused on ad hoc connections for short distances with the highest possible efficiency (over 95%), and 100 Gb/s data rate.

Another deeply investigated topic is an automatic repeat request (ARQ). Similar work can be found in [12], [13], but this work focused on the ARQ concatenated with forward error correcting codes (FEC) [14]. Such technique is called hybrid-ARQ (HARQ) [15].

There are only a few wireless transceivers working at high-speed data rates. For example, paper [16] introduces a system for wireless communication working at the 60 GHz band. However, the supported data rate of 4 Gb/s is still much lower than authors' goal: 100 Gb/s wireless communication. The core task of this paper is to test adaptation algorithms for forward error correction. This allows controlling the redundant data in view of the channel quality.

3. Work Details

In this section, authors explain how the results are generated. Next, the employed simulation environment and the emulated wireless channel are explained. After that, all implemented techniques used in the research are described. At the end, the FPGA prototype is presented.

3.1. Simulation Model

The Matlab simulations of the planned system were performed, before the real demonstrator was implemented. The simulations are using the same algorithms to the

solutions implemented in the hardware. The field programmable gate arrays (FPGAs) are used for the final demonstrator.

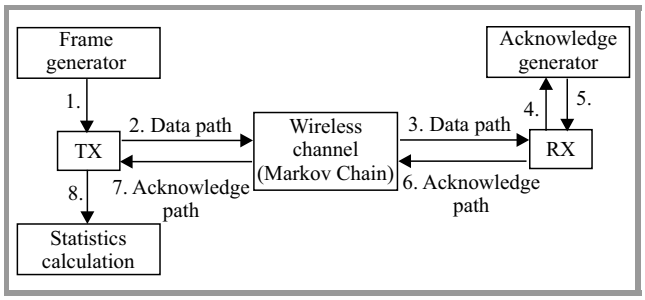


Fig. 1. A Matlab model is used to generate transmission statistics. The receiver uses an acknowledge generator to build the ACK-frame. The transmitter uses the frame for retransmissions and statistic calculations.

Figure 1 explains the simulation model. Two devices are communicating by an emulated wireless channel. They are exchanging data frames (the data path) and confirmation messages (the acknowledge path). Every successfully received data frame is confirmed by the receiver device (RX). That makes the data exchange process reliable, because the transmitter (TX) can repeat all lost frames. This process is called an automatic repeat request (ARQ). The core function of the ARQ process is generation of the acknowledge frame (ACK) and sending it to the transmitter device. Additionally, the TX device can calculate communication statistics. Such a mechanism allows estimating the efficiency of the implemented algorithm.

3.2. Wireless Channel Emulation

In this subsection, the implementation of the wireless channel used in the simulation (according to Fig. 1) is introduced. Such a two state Markov chain for errors emulation are used, because this solution requires two transition statistics, which defines the channel. The probabilities of the transition define a bit error rate and error length in bits. It does not use any physical aspects of the wireless transmission. For testing the data link layer it is acceptable, because only the characteristic and distribution of the errors are necessary. The cause is unimportant, until the parameters describe the channel moreover correctly. A detailed description of the Markov chain can be found in [11].

3.3. Frame Segmentation and Aggregation

A frame size and a bit error rate (BER) have a significant impact on the wireless communication efficiency. When the payload is longer in the frame, then less overhead is generated by the headers and checksums. Transmission is more efficient. Unfortunately, long frames are more vulnerable to transmission errors. This is explained in Fig. 2. If the frames become longer, then the probability that some bits in the frame will be corrupted is higher. The frame

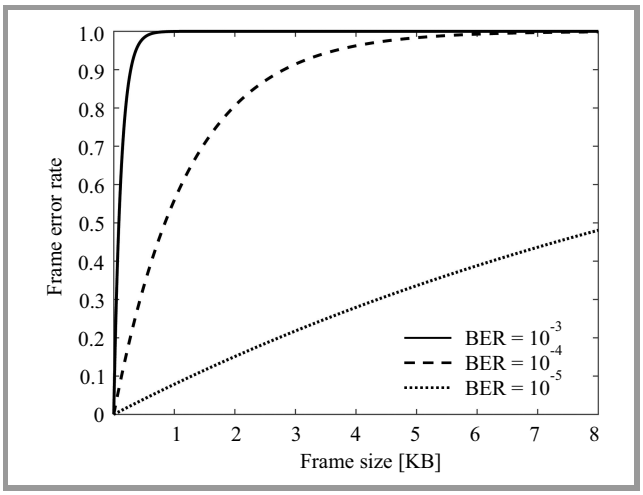


Fig. 2. A frame error rate in view of the frame size. If frame is longer, then higher is the probability that an error will occur during the transmission and the frame will be lost. Due to this aspect, shorter frames are preferred in a noisy wireless environment.

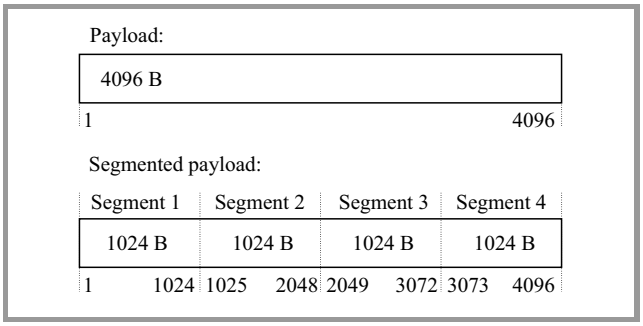


Fig. 3. An example of segmentation for a frame payload. The example payload is chopped to four segments of equal length. The shorter segments are more efficient during a transmission in a noisy channel.

can be split to independent segments, to improve the robustness and the communication efficiency. The splitting process is explained in Fig. 3. In the example, a single 4 KB

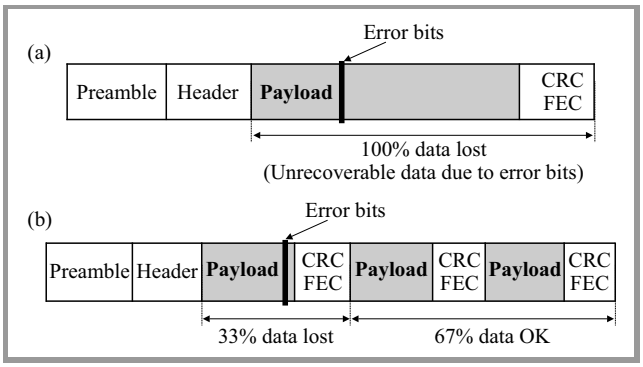


Fig. 4. An explanation how the segmented frame can improve the total efficiency: (a) classical frame, (b) frame with subframes. In a case of bit errors in a classical frame, the whole payload has to be retransmitted. If the segmented frame is used, then only invalid data part is rejected and there is no need to retransmit the whole frame, but just only the defected segment.

frame is split to four 1 KB segments. Now, the individual segments are acting like subframes (frame fragmentation). Every segment is using an individual header and checksum, but the preamble is shared (frame aggregation). It means that the errors in one segment do not influence the payload in the other segments. That improves the communication efficiency (Fig. 4). In case of a bit error, only the defected part must be repeated but not the complete frame. In this case, the default frame size is 64 KB, and is segmented to 64 fragments. In a single ARQ session 64 frames are transported (4 MB). The FPGA implementation allows changing the frame settings in the fly, and only the on-chip memory buffers are limiting the flexibility of the frame format.

3.4. Automatic Repeat Request Process

As was already mentioned, the TX and RX devices are working in a closed feedback loop. This loop is called ARQ. Every frame sent by the TX device is locally copied to the TX ARQ buffer (Fig. 5). If the RX will not acknowledge all of the sent frames, then the TX reads the lost frame from the buffer and makes retransmission. The retransmission process repeats until the positive ACK for the frame is received. If the ACK frame is lost, then the transmitter sends an ACK-request frame after a predefined timeout. In this case, this procedure have been adopted to proposed implementation. Instead of acknowledging of full frames, every single segment (subframe) is acknowledged. It means that the ARQ process works on frame fragments but not on full-frames. For designed FPGA prototype, an additional future is used. The implementation uses a zero-copy approach. The transmitted data is not copied to a dedicated buffer, but a pointer to a memory segment is requested from a higher layer in a case of retransmission. This saves energy and reduces memory footprint for the FPGA.

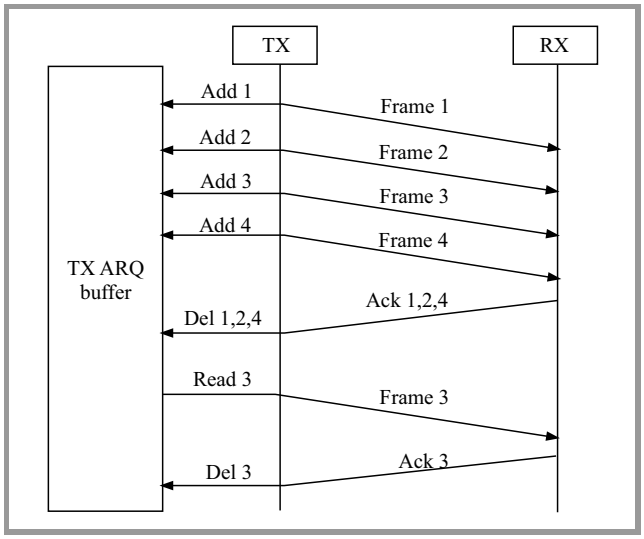


Fig. 5. An automatic repeat request process (ARQ). All transmitted frames are copied to the temporary TX ARQ buffer. If any frame will be lost during the transmission, then the transmitter reads the lost data from the buffer and starts the retransmission.

3.5. Forward Error Correction

The FEC algorithms are reducing the number of retransmitted frames in the ARQ process. That significantly improves the transmission efficiency. The transmitter is sending the data with some redundant bytes. In this work, the Reed-Solomon (RS) codes are used, because of relatively high throughput. Due to many complicated aspects, the detailed introduction to FEC is omitted in this work, and in-depth details can be found in [18], [19]. The authors will just explain how the RS is building the blocks. It is important to understand results of our paper. In the simulation three RS flavors are used: RS(255, 249), RS(255, 239), and RS(255, 223). The numbers are defining the RS block size (255 bytes in this case) and the payload size (249, 239 or 223 bytes). It means that the redundant information is 6, 16, or 32 bytes long. This is explained in Fig. 6. The redundant bytes are used for error corrections. If more redundant data is produced, then more error symbols can be corrected. The RS(255, 249) can correct up to 3 bytes in the block, RS(255, 239) 8 bytes, and the RS(255, 223) 16 bytes [18]. The aim is to find a trade-off between the redundancy and the payload, so the transmission process is efficient. The VHDL implemented FEC engine for the FPGA is more flexible, and more RS flavors is available. The implemented FPGA FEC engine is supporting any coding in a range of 2–18 redundancy bytes per a single RS block. It means that the following coding schemes are supported: (255, 237), (255, 239), (255, 241), (255, 243), (255, 245), (255, 247), (255, 249), (255, 251), and (255, 253). Coding can be adjusted on the fly, and this feature is used by the proposed adaptation algorithm to choose the optimal coding for the current wireless channel condition. In presented case, the higher coding granularity improves the overall performance.

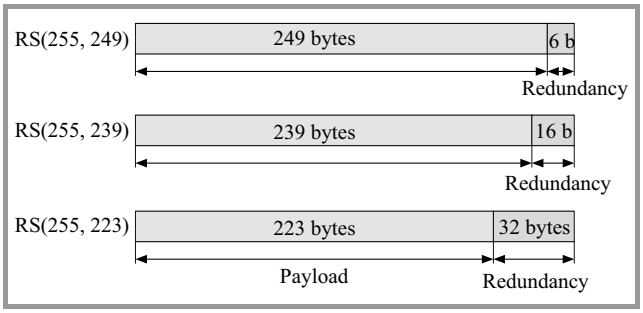


Fig. 6. The Reed-Solomon (RS) blocks. The algorithm is building the blocks of size of 255 bytes in presented case. The redundancy is adjustable. If more redundancy bytes are used, then less payload is carried by the segments. More redundancy bytes allow correcting more errors after the transmission.

The RS calculation is the most calculation demanding operation performed in the FPGA logic. The encoders and decoders occupy 55% of the FPGA logic resources. To support the targeted 100 Gb/s stream, eighty encoders and eighty decoders are in use.

3.6. Hybrid ARQ

Any combination of the ARQ and FEC is called Hybrid-ARQ (HARQ). Two mainly investigated in the paper HARQ methods are HARQ type I and II. The HARQ-I adds error detection code and FEC to every packet at every condition. The HARQ-II sends the FEC data during the retransmission only. In such a case, the error correction data is not overloading the link during the regular transmission (Figs. 7 and 8). This can introduce some improvements in efficiency. We answer in the next paragraph, which strategy is better for our protocol. A detailed description of the HARQ-I and II can be found in [18] and [20].

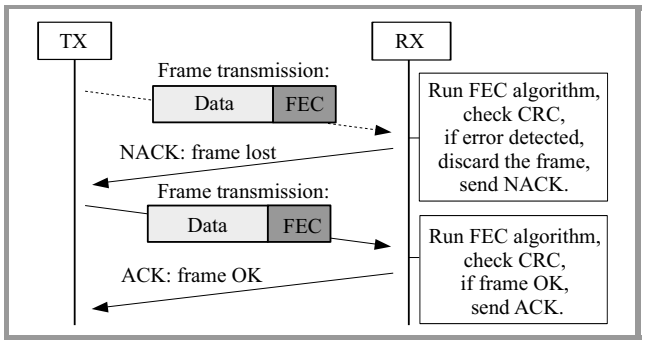


Fig. 7. The HARQ-I scheme. The transmitter always sends the frame with a forward error correction data. The retransmitted frame is a mirror copy of the original frame.

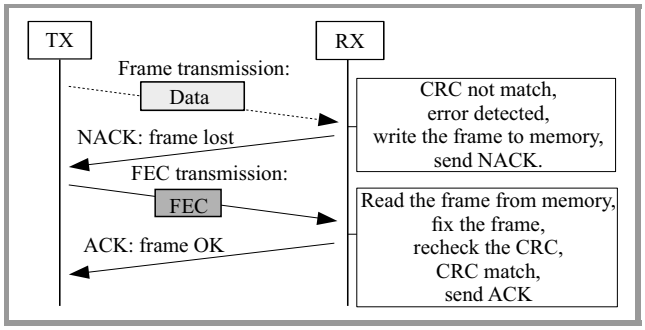


Fig. 8. The HARQ-II scheme. The transmitter usually sends the frame without forward error correction data. The standard frame is not extended by the FEC field. In a case when the frame is lost, then the transmitter sends the FEC only. The frame data is not retransmitted. The HARQ-II reduces the retransmission overhead in compare to the HARQ-I.

3.7. FPGA Demonstrator

The hardware demonstrator consists of two hardware boards (Fig. 9). The Tilera server is a dedicated 72 cores processor employed for frames segmentation and fast memory access. The FPGA is a calculation coprocessor supporting CRC, FEC calculations, and frames aggregation. The main state machine responsible for data link layer is run on the Tilera server. The FPGAs and sever are connected with 10 Gb/s Ethernet optical fiber. For now, the architecture supports up to 80 Gb/s with two FPGA boards (interfaces

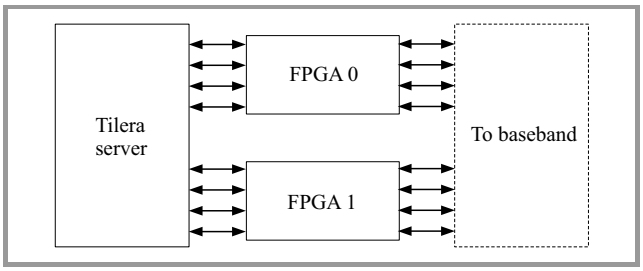


Fig. 9. The demonstrator overview.

constraints). Generally, the Virtex 7 FPGA can process up to 100 Gb/s in a back-to-back connection (Fig. 10). The baseband processor is not finished yet. Thus, authors can test the processor only in a loopback mode. A single, logical FPGA processing pipeline (lane) is shown in Fig. 11.

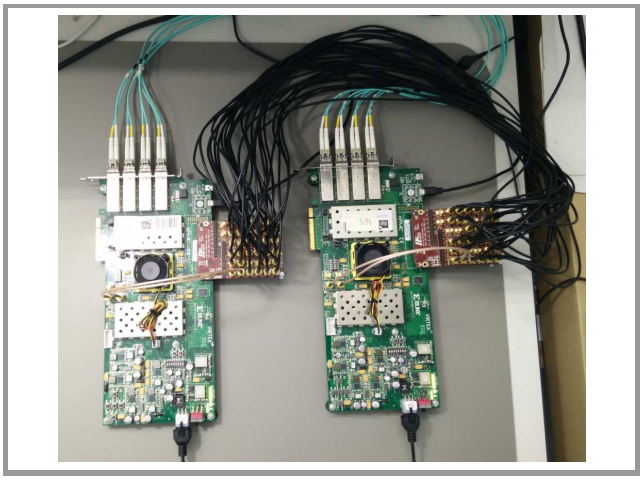


Fig. 10. The FPGA demonstrator.

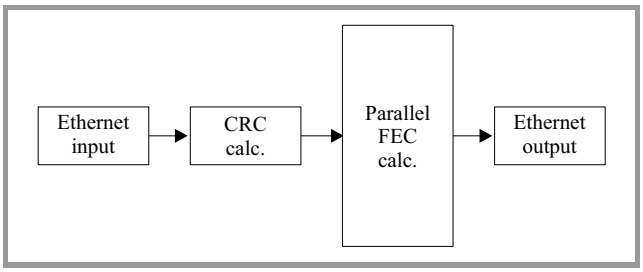


Fig. 11. A single processing lane (logical pipeline).

3.8. Parallel FPGA Processing

There is no possibility to process the 100 Gb/s stream in a single processing pipeline (lane) [14]. Even if one of the fastest FPGA developments kit is used, the stream processing have to be divided and calculated in parallel. For that purpose, a parallel calculation array is implemented. The array calculates 640 bits @ 156.25 MHz. Internally the 640-bits-word is organized in ten sub-words processed by ten calculation lanes (Fig. 12). Every lane runs at 10 Gb/s, and is connected to two 10 Gb/s Ethernet ports (data input and data output). Such a processor uses 294115 lookup-tables and 239019 flip-flops. It is

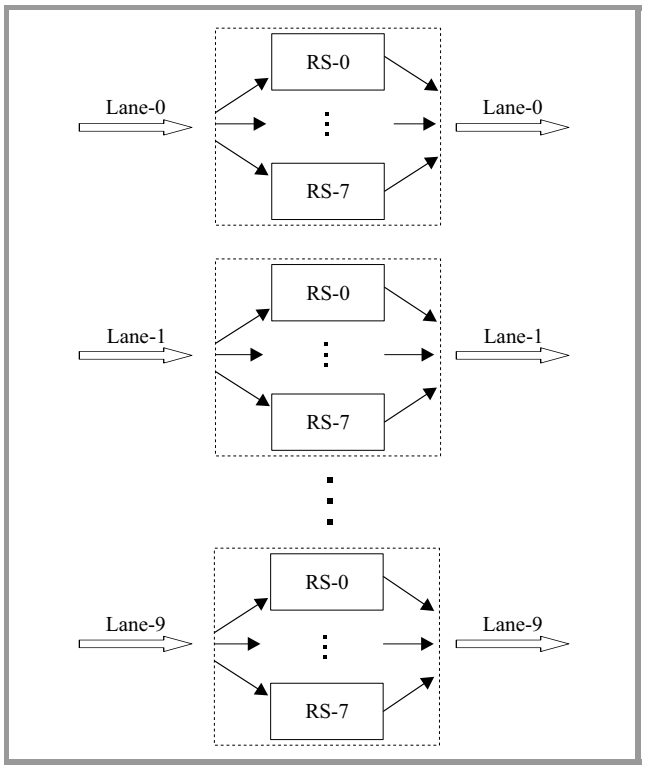


Fig. 12. The parallel FEC calculation array implemented in the FPGA logic.

respectively 65% and 27% of the total resources available in the Virtex 7-690T FPGA. The slices occupation is equal to 80%.

4. Results

4.1. Transmission Limiting Factors

The authors have performed transmission experiments and recorded the most important parameters (the overall efficiency, the percentage of successfully received segments, the percentage of successfully received frame headers, the total number of acknowledge frames, the number of timeouts, and the total number of physical layer turnarounds). That allows to investigate, which factors reduce throughput in test system. Additionally, the retransmission segment size can be adjusted in a range of 32 to 65536 bytes. A following assumption can be done after analysis of the results. The ACK-frame has to be as short as it is possible and always encoded with robust coding. Practically it means that the ACK-frame should be encoded with a code rate lower that the code rate of the data segments (a lower code rate means improved error correction). This reduces the total number of lost ACK-frames, timeouts, and PHY turnarounds. After that, only the loss of the data segments limits the throughput. Intensive FEC coding and segmentation for the data segments makes no sense without improved reliability of the ACK-frame. Figures 13 and 14 demonstrate the used methodology for uncoded and encoded transmissions. In both cases the throughput is limited by lose of the data segments but not by the ACK-frames. The total number of timeouts and the PHY turnarounds are relatively low during the simulation.

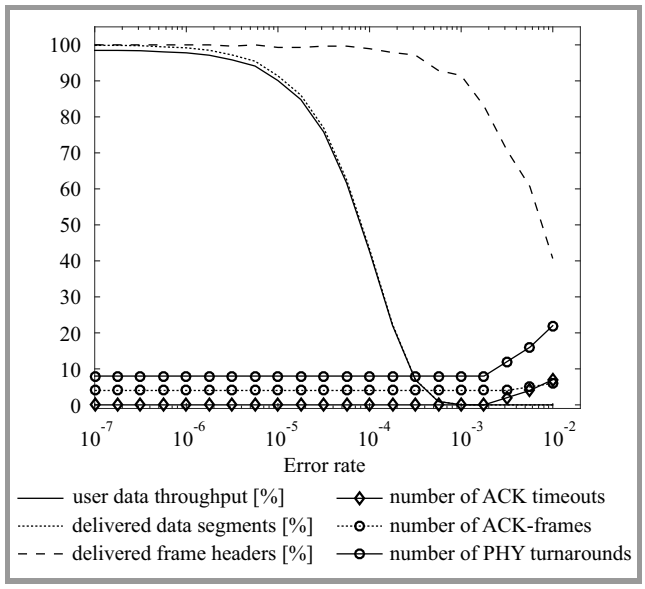


Fig. 13. Limiting factors of the transmission. The data segments are uncoded. The frame headers are delivered with a relatively low error rate. The goodput is limited by lose of the data segments.

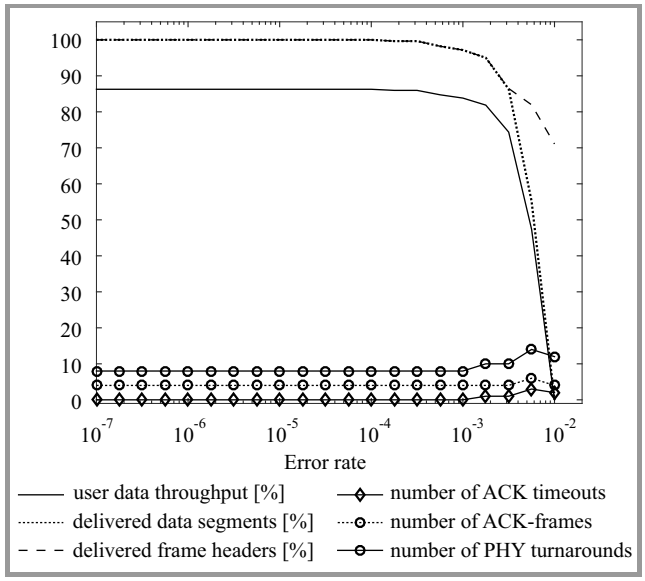


Fig. 14. Limiting factors of the transmission. The data segments are coded with RS(255, 223). The error rate of the data segment is strongly reduced as compared to uncoded transmission simulation.

The ACK-frame length is depended from the total number of successfully received segments in a single ARQ session (positive acknowledgment). If the data frame segmentation is increased, then many small parts have to be sent and acknowledged. That increases the ACK-frame size. Unfortunately, too long ACK-frames cannot be delivered errorless and are limiting the throughput. Instead of the efficiency improvement, degradation is observed. The ideal solution

is to keep the ACK-frame size smaller than the size of the data segment. An ACK-frame compression is needed to achieve that in our case. The three solutions were considered: a bit map coding, and two versions of a sequence number range coding. A single *uint16* value and a bit map are sent in the bit map scheme. The uint16 value defines the first acknowledged segment number, and the bit map defines all next values. The bit position defines an offset and the bit value defines if the segment is acknowledged or not.

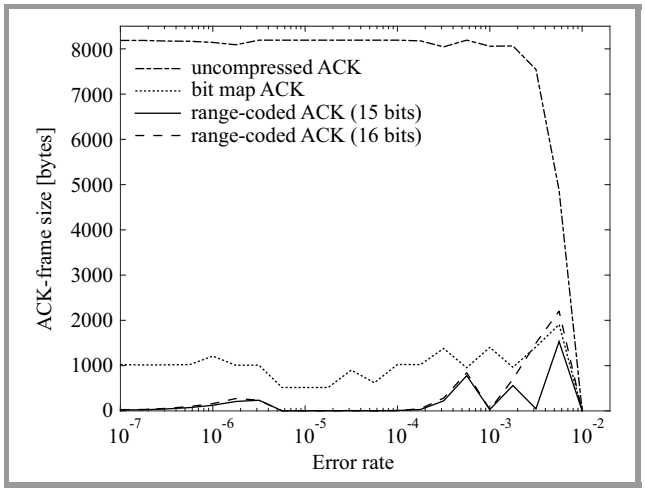


Fig. 15. The maximal ACK-frame sizes during the simulation. Three types of the ACK-frame compression methods are presented. The compressed ACK-frame is significantly shorter and is much more robust during the transmission.

The second and third methods send only a range of addresses of the acknowledged segments. In some cases that may lead to an extended frame size. All three methods were investigated, and the results are shown in Fig. 15.

4.2. Optimal Segment Size

If the problem of the disadvantageous ACK-frame size is reduced, then additional improvements for the data segments can be done. First of all, the influence of the segment size is considered. By reducing the segment size, the efficiency can be improved on "bad" channels. From the other side, more segments have to be sent to transmit the same data. Every segment is equipped with an individual header and checksum. This induces overhead. Additionally, enabling the FEC introduces some additional issues. This happens because block codes are used (in this case the RS block size is equal to 255 bytes). This introduces additional indirect-segmenting. The errors in each RS block are corrected individually, and each RS block acts like an independent sub-segment. In Fig. 16 the data segment size is investigated. It can be observed that the optimal segment size for error rates below 10^{-6} is in the range of 2 to 4 KB (16 to 32 segments for a 64 KB frame size). When the error rate increases, then the segment size should be reduced. Five hundred and more segments are required for links with an

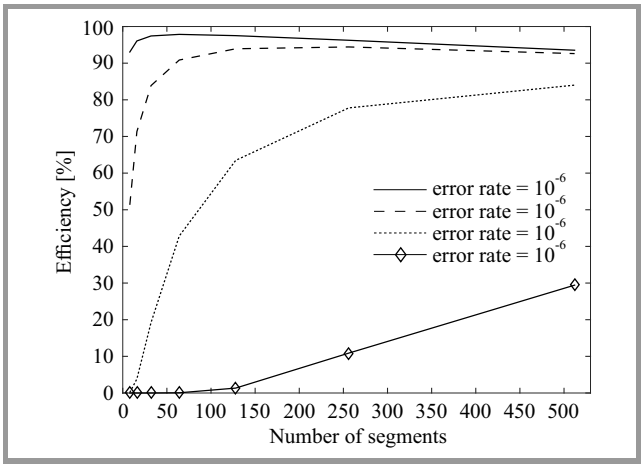


Fig. 16. The data link layer efficiency vs. the data segment size vs. an error rate. The data segments are uncoded. If the error rate increases, then smaller segments are preferred.

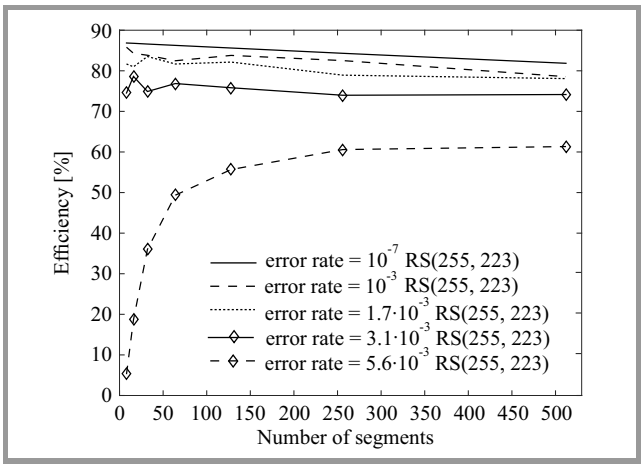


Fig. 17. The data link layer efficiency vs. the data segment size vs.an error rate. The data segments are coded with the RS(255, 223). The RS encoded frames are less sensitive to the segment size that the uncoded frames.

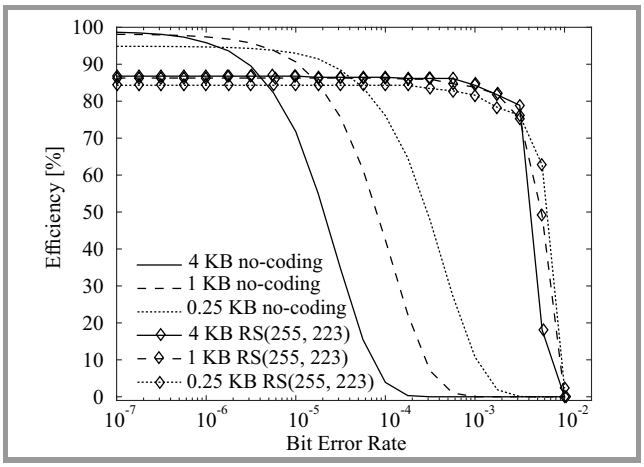


Fig. 18. The data link layer efficiency vs. the data segment size vs. an error rate. The uncoded and RS coded transmissions are plotted in one figure. The RS encoded frames are less sensitive to the segment size than the uncoded frames.

error rate higher than 10^{-5}. It means that the transmission without coding is very sensitive to the segment size. Dynamic change of this parameter can introduce some significant improvements to the efficiency. Slightly different situation can be observed, when RS coding is used. This situation is shown in Fig. 17. The transmission with RS coding is less sensitive to the segment size. That means that, advantages of the variable segment size can be reduced after enabling the coding. In presented FPGA demonstrator, the implementation of this feature in the first iteration is omitted. Authors presume that the block-FEC can be a good substitute of the variable segment size. To get better feeling of this observation, more simulations were performed (Fig. 18). The improvement of the variable segment size for the RS-coded transmissions is marginal.

4.3. Dynamic FEC Redundancy

In this subsection, a dynamic algorithm to find a trade-off between the FEC coding and the demanded error correction performance is proposed. The algorithm analyses the number of successfully delivered data segments and the number of corrected errors in the RS blocks. If the efficiency is degraded by loses of the data segments, then the algorithm increases the FEC coding. This solution is uncomplicated, but it is important to define a threshold, when the FEC mode should be changed. In this paper, the thresholds are set to $249/255 \approx 97.6\%$, $239/255 \approx 93.7\%$, and $223/255 \approx 87.5\%$. If the data delivery efficiency is below the given values, then the corresponding RS code is used. It tries to find a compromise between the RS overhead and the rate of the lost segments. The thresholds correspond to the code rates and define upper bounding of the goodput. In this solution, an error statistics of all decoded RS blocks are calculated, and all corrupted segments are categorized to some groups. Every error category can be corrected by a different RS code. If the statistic is known, then the best RS code can be chosen for all future transmissions. Results of the algorithm are shown in Fig. 19. It may happen that the channel changes so rapidly that this solution will work too slowly. To minimize this factor, HARQ-II can be applied. After that, any mistake of the adaption algorithm can be corrected by the FEC data sent in the next ARQ session. An additional simulation was performed to test how the algorithm performs in rapidly changing environment (Fig. 20). Results of this simulation are presented in Table 1. Nine algorithms were tested. The best performance is achieved by using adaptive redundancy with the HARQ-I scheme. This algorithm is relatively easy to implement in the FPGA hardware. The HARQ-II scheme is giving similar results, but the complexity of the HARQ-II is higher. The HARQ-I algorithm achieves also quite high efficiency. It is possible to improve the switching logic in the future. For example, a proportional-integral-derivative (PID) or a fuzzy logic controller can be employed. These controllers can rely not only on the instantaneous value, but can track more parameters on a longer period.

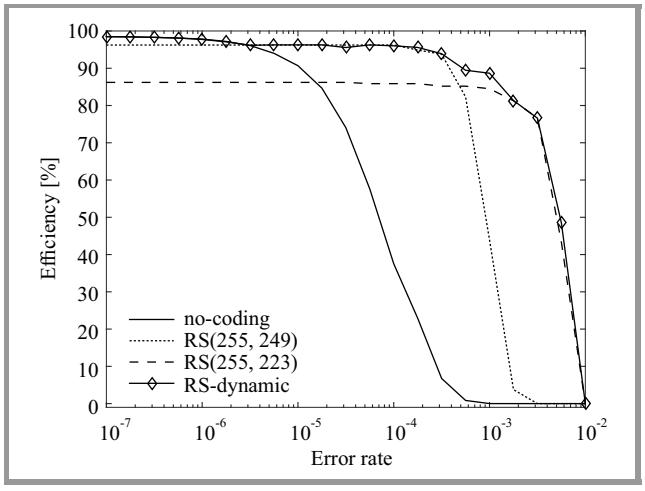


Fig. 19. The dynamic FEC algorithm results. The probability P(C) is equal to 0.5. The adaptive algorithm chooses the optimal coding and maximizes the goodput.

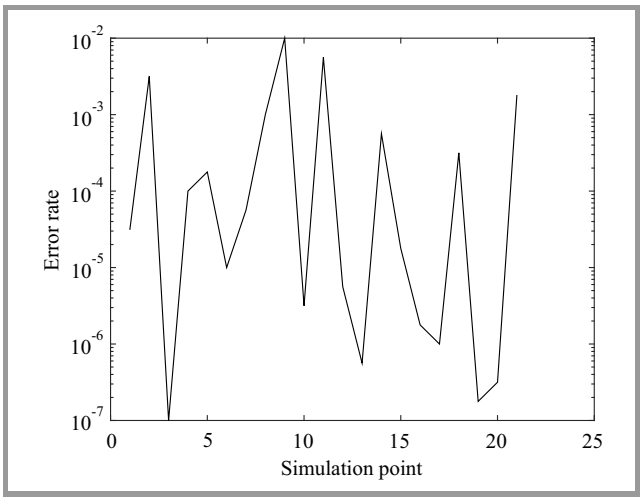


Fig. 20. An example characteristic used for evaluation of adaptive algorithms. The error rate is a permutation in a range of $[10^{-7}; 10^{-2}]$.

Table 1
Different algorithms vs. the rapidly changing channel (Fig. 20)

Algorithm	Average efficiency [%]	Peak efficiency [%]
No coding (ARQ)	54.72	98.47
RS(255, 249) (HARQ I)	74.69	96.23
RS(255, 239) (HARQ I)	79.84	92.43
RS(255, 223) (HARQ I)	79.19	86.20
Adaptive RS(HARQ I)	79.66	98.33
HARQ II with RS(255, 223)	74.82	98.33
Adaptive RS with HARQ II	79.57	98.13
Adaptive RS (modified)	83.05	96.28
Adaptive RS with HARQ II (modified)	82.46	96.10

4.4. Performance of the FPGA Implementation

The back-to-back connected FPGAs and the implemented wireless channel emulator are used to test the algorithms in a real hardware. Presented FPGA-implementation accepts a BER up to $2 \cdot 10^{-3}$. Above this value, the RS engine cannot fix errors in the stream, and the performance rapidly drops. In some cases, the wireless channel may produce BER higher than $2 \cdot 10^{-3}$. The hardware-implemented data link layer cannot operate in such conditions, and the device will lose the link. To improve the error correction results, an extended version of FEC engine is proposed.

In the simulation, the authors presume that the engine must support at least the RS(255, 223) with code rate $R \approx 0.875$. The used RS VHDL-implementation cannot support a lower coding than the RS(255, 237) with code rate $R \approx 0.929$. Thus, the achieved FPGA results are worse than simulated. There is a possibility to use shortened RS codes to decrease the code rate of the produced stream. That is the easiest approach to deal with the problem. The second solution is to redesign the implemented RS entity, that it can natively support the RS(255, 223) [21]. The both approaches are compared in terms of consumed logic area (Fig. 21) and error correction performance (Figs. 22 and 23).

The implementation of the RS(127, 109, 8-bit symbol) is realized by shortening the RS(255, 237) by removing 128 symbols from the message part of the codeword. Practically, it is achieved by using two hardware entities of the default code, and by multiplexing/switching the data input and output interfaces (Fig. 24). Every coder calculates half of the data block, and the rest of the symbols are filled with zeroes.

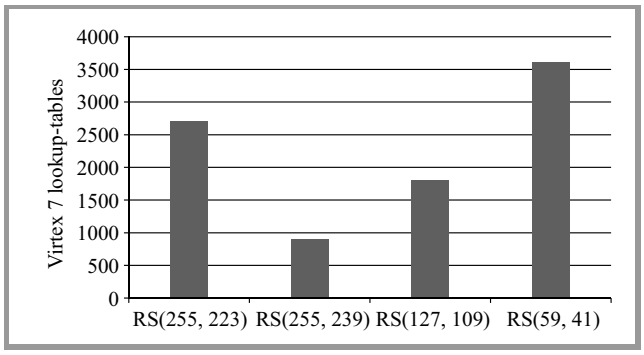


Fig. 21. Consumed logic area by the proposed solutions.

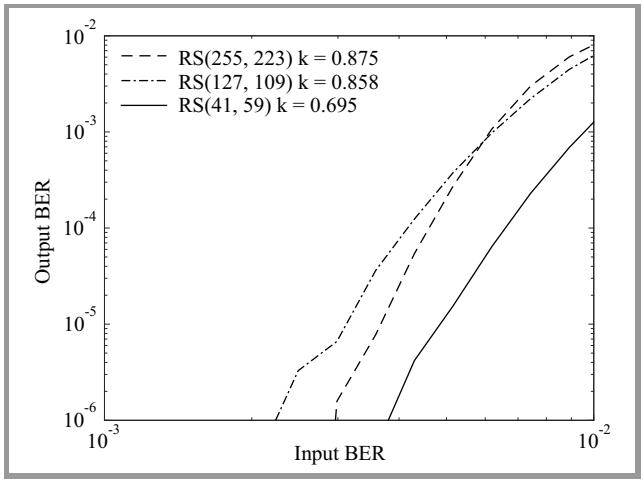


Fig. 22. Error correction performance of the proposed coding schemes.

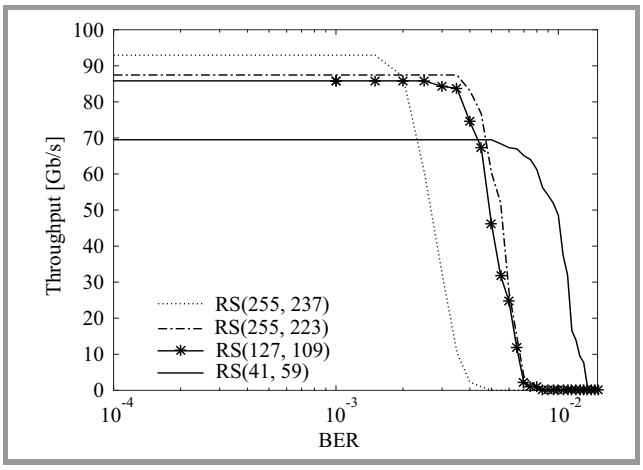


Fig. 23. Simulated throughput of the FPGA demonstrator with the improved FEC engine.

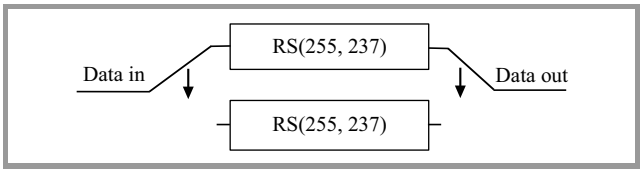


Fig. 24. Implemented code shortening.

The RS(59, 41, 8-bit symbol) is a shortened version of the RS(255, 237) by removing 196 data symbols. Practically those are four calculation entities switched four times during a single codeword coding. These uncomplicated operations improve the error correction performance without any significant system complications. Especially, the resources used for implementation of the RS(255, 223) can be reduced by around 33%, when the coding is replaced with the proposed RS(127, 109). This simplification causes a small loss of the error correction performance. The redundancy symbols are spread more uniformly over the frame and the redundancy cannot be used as flexible as during processing of the full-length codewords. The error correction process is focused on shorter blocks, and some of the redundant information is not used efficiently.

5. Future Work

We experiment with interleaving and multiplexing matrixes to increase error correction performance of our implementation (Fig. 25). The assumption is to improve decoding performance of the RS(255, 239) and to achieve error correction performance similar to the RS(255, 223). The proposed decoder must be mathematically analysed and

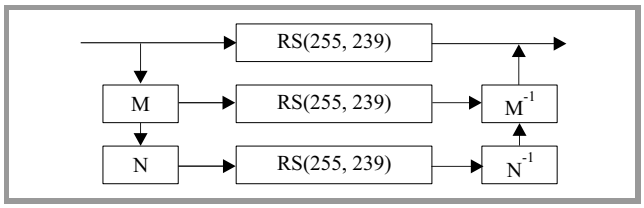


Fig. 25. Experimental RS decoder.

it have to be proven, that the proposed structure requires less calculation operations than the RS(255, 223). Up to now, achieved results are disappointing. The solution is inefficient against single, uniformly distributed bit errors (e.g. AWGN channel). In such a case, authors cannot tune the structure to get any optimistic results. Usually, more power than a single RS(255, 239) decoder is consumed, and the increase of the error correction performance is marginal.

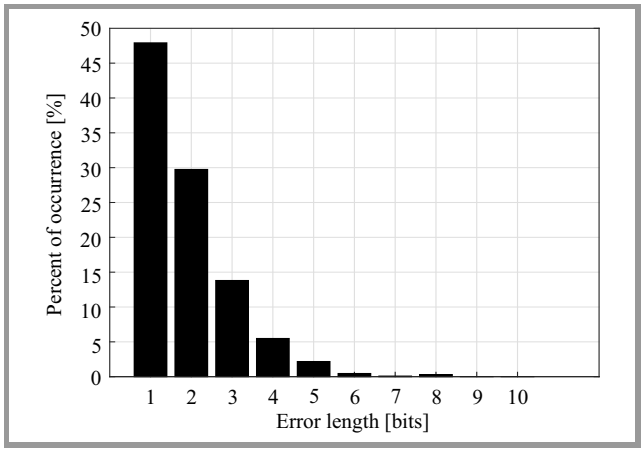


Fig. 26. Example error characteristic.

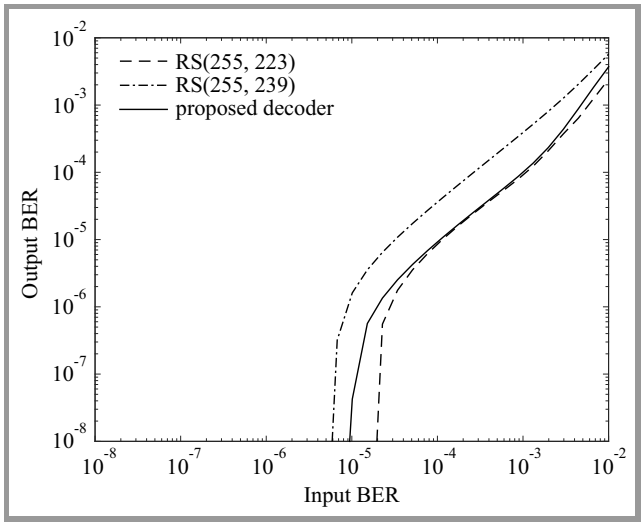


Fig. 27. Error correction results of the proposed decoder.

Better results are achieved if the structure is run against burst errors. In Fig. 26, an example error characteristic is presented. The proposed characteristic to test presented structure (Fig. 27) is used. The solution achieves very good BER performance, but this is not the most important statistic. Number of bit errors in individual blocks is significantly reduced, but the total number of fully recovered blocks is lower than after the RS(255, 223) decoding (Fig. 28).

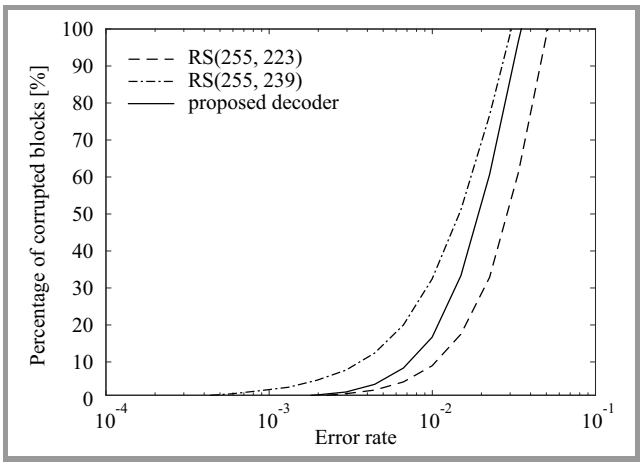


Fig. 28. Block correction results of the proposed decoder.

An additional disadvantage is that the M and N matrixes (Fig. 25) are dependent from the error characteristic, and are not universal for all burst error lengths. If the length of the typical error produced by a channel is changing, then also the matrixes have to be adopted.
The proposed scheme can be run iteratively. Authors do not consider an iterative mode of operation due to energy consumption and latency. For now, the presented solution cannot be considered as a substitute of the typical RS decoder.

6. Conclusion

In the paper, three major aspects of the 100 Gb/s data link layer are explained. Firstly, the limiting factors of the implementation are analyzed. The data link layer robustness is improved after introducing the ACK-frame compression

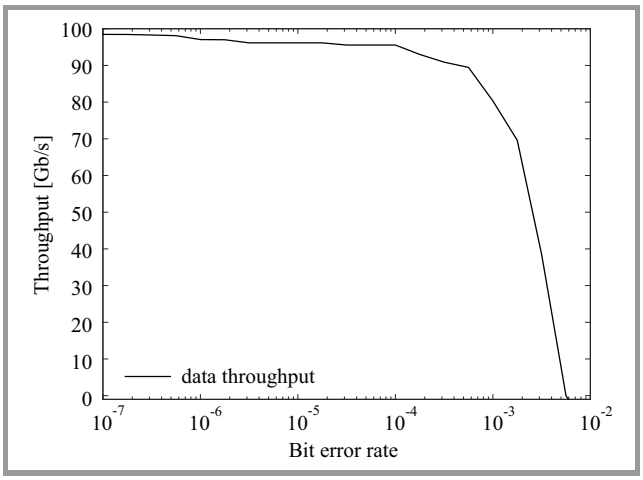


Fig. 29. Performance of the FPGA implementation in view of a bit error rate.

and coding. This reduces the total number of timeouts in the system simulation. After that, the data segmentation can be investigated. The most important observation is that the segmentation has more influence on the uncoded transmissions, than for the transmissions coded with the RS block codes. Because of this reason, authors skip the implementation of a variable segment size in the first iteration. Instead of it, authors focus on the FEC algorithms and a solution to manage the FEC overhead against the transmission requirements. The goal is to use as little overhead as possible and maximize the efficiency.

Link adaptation used with the HARQ-I simplifies the FPGA design, and it is a good substitute of the more complicated HARQ-II method. That allows removing buffers from the design, due to the fact that broken frames do not have to be buffered.

All presented results are validated on the Xilinx VC709 Virtex 7 FPGA platform. The implementation supports a net data rate of 97 Gb/s on the real FPGA-hardware (Fig. 29).

Acknowledgements

This paper is related to the End2End100 project and cooperates with other proposed projects of the DFG Special Priority Program 1655 (SPP1655) on "Wireless 100 Gb/s and beyond", e.g. the Real100G.COM and Real100G.RF. This group of projects will investigate a complete wireless 100 Gb/s system at ultra-high frequencies (240 GHz).

References

[1] S. Koenig *et al.*, "Wireless sub-THz communication system with high data rate", *Nature Photonics*, vol. 7, no. 12, pp. 977–981, 2013.

[2] F. Boes, T. Messinger, J. Antes, D. Meier, A. Tessmann, and I. Kallfass, "Ultra-broadband MMIC-based wireless link at 240 GHz enabled by 64 GS/s DAC", in *Proc. 39th Int. Conf. Infrared, Millim., & Terahertz Waves IRMMW-THz 2014*, Tucson, AZ, USA, 2014.

[3] H. Wang, W. Yuan, B. Zhang, H. Li, Z. Zhang, X. Yang, and W. Shi, "The design, test, and application of the front end in 0.3 THz wireless communication systems", in *Proc. Selec. Proc. Photoelec. Technol. Committee Conf. SPIE held June-July 2015*, vol. 9795, 2015 (doi: 10.1117/12.2214175).

[4] T. Nagatsuma, K. Kato, and J. Hesler, "Enabling technologies for real-time 50-Gbit/s wireless transmission at 300 GHz", in *Proc. ACM Int. Conf. Nanoscale Comput. & Commun. ACM NanoCom 2015*, Boston, MA, USA, 2015

[5] I. T. Monroy, "Photonic techniques for sub-Terahertz wireless data transmission", in *Proc. Photonic Networks and Devices (Networks) 2015*, Boston, MA, 2015 (doi:10.1364/NETWORKS.2015.NeT1D.1).

[6] K. KrishneGowda, T. Messinger, A. C. Wolf, R. Kraemer, I. Kallfass, and J. C. Scheytt, "Towards 100 Gbps wireless communication in THz Band with PSSS modulation: A promising hardware in the loop experiment", in *Proc. IEEE Int. Conf. Ubiquit. Wirel. Broadb. ICUWB 2015*, Montreal, Canada, 2015.

[7] 802.11ad-2012 – 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 3: Enhancements for Very High Throughput in the 60 GHz Band, IEEE Standard Association, 12.2012 [Online]. Available: http://www.standards.ieee.org

[8] T. Li, Q. Ni, D. Malone, D. Leith, Y. Xiao, and T. Turletti, "Aggregation with fragment retransmission for very high-speed WLANs", *IEEE/ACM Trans. Networ. (TON)*, vol. 17, no. 2, pp. 591–604, 2009.

[9] D. Qiao, S. Choi, and K. G. Shin, "Goodput analysis and link adaptation for IEEE 802.11 a wireless LANs", *IEEE Trans. Mob. Comput.*, vol. 1, no. 4, pp. 278–292, 2002.

[10] D. Skordoulis, Q. Ni, H.-H. Chen, A. P. Stephens, C. Liu, and A. Jamalipour, "IEEE 802.11n MAC frame aggregation mechanisms for next-generation high-throughput WLANs", *IEEE Wirel. Commun.*, vol. 15, no. 1, pp. 40–47, 2008.

[11] E. H. Ong, J. Kneckt, O. Alanen, Z. Chang, T. Huovinen, and T. Nihtil, "IEEE 802.11ac: Enhancements for very high throughput WLANs", in *Proc. IEEE 22nd Int. Symp. Personal Indoor & Mob. Radio Commun. PIMRC 2011*, Toronto, Canada, 2011.

[12] S. Choi and K. Shin, "A class of adaptive hybrid ARQ schemes for wireless links", *IEEE Trans. Veh. Technol.*, vol. 50, no. 3, pp. 777–790, 2001.

[13] L. Badia, N. Baldo, M. Levorato, and M. Zorzi, "A Markov framework for error control techniques based on selective retransmission in video transmission over wireless channels", *IEEE J. Selec. Areas Commun.*, vol. 28, no. 3, pp. 488–500, 2010.

[14] M. A. Ingale, "Error correcting codes in optical communication systems", Master Thesis, School of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden, 2003.

[15] S. Falahati and A. Svensson, "Hybrid type-II ARQ schemes with adaptive modulation systems for wireless channels", in *IEEE VTS 50th Veh. Technol. Conf. VTC 1999-Fall*, Amsterdam, The Netherlands, 1999.

[16] M. Ehrig and M. Petri, "60 GHz broadband MAC system design for cable replacement in machine vision applications", *AEU-Int. J. Elec. Commun.*, vol. 67, no. 12, pp. 1118–1128, 2013.

[17] E. Esteves, P. J. Black, and M. I. Gurelli, "Link adaptation techniques for high-speed packet data in third generation cellular systems", in *Proc. Eur. Wirel. Conf.*, Florence, Italy, 2002.

[18] S. Lin and D. Costello, Error Control Coding: Fundamentals and Applications, New Jersey: Prentice-Hall, 1983.

[19] Ł. Łopaciński, M. Brzozowski, R. Kraemer, and J. Nolte, "100 Gbps wireless – challenges to the data link layer", in *IEICE Inform. & Commun. Technol. Forum IEICE ICTF 2014*, Poznań, Poland, 2014.

[20] H. Chen, R. G. Maunder, and L. Hanzo, "A survey and tutorial on low-complexity turbo coding techniques and a holistic hybrid ARQ design example", *IEEE Commun. Surv. & Tutor.*, vol. 15, no. 4, pp. 1546–1566, 2013 (doi: 10.1109/SURV.2013.013013.00079).

[21] M. Marinkovic, M. Krstic, E. Grass, and M. Piz, "Performance and complexity analysis of channel coding schemes for multi-Gbps wireless communications", in *Proc. IEEE 23rd Int. Symp. Personal Indoor and Mob. Radio Commun. PIMRC 2012*, Sydney, Australia, 2012.

Łukasz Łopaciński received his M.Sc. degree in Computer Science from West Pomeranian University of Technology, Szczecin, Poland, in 2009. Since 2007, he worked in industrial companies in field of embedded systems and wireless communication. Currently he is working in BTU Cottbus (Germany).

E-mail: lukasz.lopacinski@b-tu.de
Brandenburg University of Technology
Cottbus-Senftenberg
Platz der Deutschen Einheit 1
03046 Cottbus, Germany

Marcin Brzozowski received his M.Sc. and Ph.D. degrees in Computer Science from BTU Cottbus, Germany, in 2006 and 2012, respectively. Currently he is working with networking and embedded systems in IHP Germany. His research interests include computer networks and operating systems.

E-mail: brzozowski@ihp-microelectronics.com
IHP Microelectronics GmbH
Im Technologiepark 25
15236 Frankfurt (Oder), Germany

Rolf Kraemer received his M.Sc. and Ph.D. from RWTH Aachen in electrical engineering and computer-science in 1979 and 1985. He joined the Philips research laboratories in 1985 where he worked in different positions and responsibilities. In 1998 he became professor at the technical university of Cottbus with the joined appointment of the department head of wireless systems at the IHP in Frankfurt (Oder). In the IHP he leads a research department with approximately 70 researchers in topics of high speed wireless communication, context aware middleware, sensor networks as well as embedded processors for encryption, and protocol acceleration.

E-mail: kraemer@ihp-microelectronics.com
IHP Microelectronics GmbH
Im Technologiepark 25
15236 Frankfurt (Oder), Germany

Steffen Buechner received his M.Sc. in Computer Science from the BTU Cottbus in 2011. After that, he participated at several research projects at the Distributed Systems/Operating Systems Group of the BTU Cottbus in the fields wireless sensor networks and embedded systems. Currently he is working on a parallel event stream processing concept for utilizing the processing power of embedded many cores for ultra-high data rate wireless communication protocol handling. His research interests include networking, embedded distributed systems, and operating systems.

E-mail: Steffen.Buechner@b-tu.de
Brandenburg University of Technology
Cottbus-Senftenberg
Platz der Deutschen Einheit 1
03046 Cottbus, Germany

Jörg Nolte is professor for distributed systems and operating systems at the Brandenburg University of Technology in Cottbus (Germany). He received his M.Sc. (1988) and Ph.D. (1994) in Computer Science from the Technical University of Berlin. He was a principal member and finally the vice-head of the PEACE group at GMD FIRST (Berlin) that developed the operating system for Germany's first massively parallel supercomputer. In the 90s he was a post-doc fellow of the Real World Computing Partnership (RWCP) in Tsukuba Science City, Japan. His major research interests are operating systems, middleware and programming languages for parallel, distributed and embedded systems.

E-mail: Joerg.Nolte@b-tu.de
Brandenburg University of Technology
Cottbus-Senftenberg
Platz der Deutschen Einheit 1
03046 Cottbus, Germany

Paper

DS-UWB and TH-UWB Energy Consumption Comparison

Adil Elabboubi[1,3], Fouzia Elbahhar[1,3], Marc Heddebaut[1,3], and Yassin Elhillali[2,3]

[1] *IFSTTAR, Villeneuve d'Ascq, France*
[2] *IEMN-DOAE, Valenciennes, France*
[3] *Université Lille Nord de France, France*

Abstract—The energy consumption of the wireless communication systems is starting to be unaffordable. One way to improve the power consumption is the optimization of the communication techniques used by the communication networks and devices. In order to develop an energy efficient UWB multi-user communication system, the choice of modulation and multi access technique is important. This paper compares two Ultra-wideband multi-user techniques, i.e. the DS-UWB and the TH-UWB in the case of the Nakagami-m fading channel. For the DS-UWB technique, the orthogonal (T-OVSF, ZCD) and non-orthogonal (Kasami) codes are used. For TH-UWB, authors consider different modulations (PPM, PSM, PAM). This comparison allows choosing the best solution in terms of energy consumption, data rate and communication range. Two different studies are realized to find the most efficient technique to use. In the first study, the same number of users for the different type of codes (data rate values) is chosen and the total energy consumption for several distances and path-loss coefficient is computed. In the second one, the multiusers effects (same data rate) for various values of distances and path-loss are evaluated.

Keywords—energy consumption, multi-user techniques, Nakagami-m fading channel, TH time hopping, UWB.

1. Introduction

Ultra-wideband (UWB) [1] is a radio technique that offers the ability to create efficient energy and low complexity communication systems, at limited cost. This energy efficiency is called green communication [2]. Researchers have already explored various subjects related to energy efficiency from system design to network protocols. In [3], the authors present an energy-efficient low complexity pulse generator for an UWB system in CMOS technology. In [4], a low power UWB transmitter with the digital pulse generator and binary phase modulator also built in CMOS chip is proposed. Authors in [5] describe a new IR-UWB receiver capable of achieving remarkable energy saving. They are operating in the lower, 3.1 to 4.5 GHz, UWB frequency band allocated in some regions of the world. Reference [6] introduces and analyses an efficient energy adaptive transmission protocol called ATP-UWSN for UWB Wireless Sensor Networks. This protocol adapts the error-control code rate and the spreading code length to match the Channel State Information (CSI), therefore reaching optimum communication parameters. Paper [7] considers an energy-aware and link-adaptive strategy for UWB WSNs to introduce different routing metrics. In [8], the authors take advantage of the positioning capabilities of UWB to propose an energy efficient routing algorithm. This algorithm is developed to search for energy efficient routes with respect to the Quality of Service (QoS) of the system. In [9] and [10] two well-known UWB modulations are compared, M-ary Pulse Position Modulation (MPPM) and M-ary Pulse Amplitude Modulation (MPAM), in the AWGN channel and the Nakagami-m fading channel respectively. The authors showed in both that the efficient technique has to be determined with respect to the transmission scenario and that generally MPPM is less energy consuming than MPAM for high constellations and long communication ranges. These results were obtained for a communication link with only one active user.

As a supplementary contribution to [9], [10], in this paper, authors study the energy efficiency of multi-user techniques for UWB systems. For this comparison, two largely used UWB multi-users techniques are selected, i.e. the Direct Sequence UWB (DS-UWB) and the Time Hopping UWB (TH-UWB). So, the purpose of this work is to compare two multi-user methods to identify the less energy consuming technique according to some operational constraints. Moreover, in this work, an energetic model is developed for two UWB multi-user techniques in Nakagami-m fading radio channels respectively. The aim is to find the most efficient UWB-multi-users technique for different transmission scenarios as a function of the communication range and of the associated path-loss attenuation. Nakagami-m distribution is used in UWB to model the multi-path components (MPCs) in the IEEE 802.15.4a [11]. To authors' knowledge there is no energetic model based on this distribution to estimate the energy consumption of UWB multiuser systems.

The paper is organized as follows. In Section 2 a brief definition of direct spreading DS-UWB and time hoping TH-UWB is presented. The system model and the channel description are shown in Section 3. The different multiple user techniques are presented in Section 4. Section 5 compares the results. Section 6 concludes the paper and gives some perspectives.

2. UWB Multi-User Techniques Description

To support multi-users transmission in UWB, the two well-known techniques are used, which are the DS-UWB and

the TH-UWB. In the DS-UWB, authors often use BPSK to modulate the transmitted signal. However, for the TH-UWB, several techniques such as PPM, PAM and PSM or BPSK-PSM (combination of two modulations) are used. In the following, the two multi-access techniques with these respective modulations are described.

2.1. DS-UWB

An UWB DS-UWB signal can be written as [10]

$$S_{DS}^{(k)}(t) = \sqrt{\frac{E_b}{N_c}} \sum_{j=-\infty}^{+\infty} \sum_{n=0}^{N_c-1} d_j^{(k)} c_n^{(k)} p\left(t - jT_f - nT_c\right). \quad (1)$$

In Eq. (1), $S_{DS}^{(k)}(t)$ represents the signal of the k-th user, $p(t)$ is the UWB pulse, which is normalized to satisfy $\int_{-\infty}^{+\infty} p^2(t)dt = 1$. The other parameters in the signal definition can be described as:

- $\sqrt{\frac{E_b}{N_c}}$ is a normalization factor to make all considered systems having the same per bit energy noted E_b;
- N_c is the number of chips used to transmit one bit of information;
- $c_n^{(k)}$ is the unique spreading sequence allocated to each k-th user; its values are set from $-1 \ldots 1$;
- T_f is the frame duration and is the chip duration satisfying $T_f = N_c T_c$;
- T_b is the bit duration; in this paper, $T_b = T_f$;
- $d_j^{(k)}$ is the data bit information; its values are from $-1 \ldots 1$ (the BPSK for modulation is used).

In this work, the different orthogonal codes are used: the ternary Orthogonal Variable Spreading Factor (T-OVSF) [11], Zero Correlation Duration (ZCD) [12] and non-orthogonal codes Kasami Type 1 [13] as spreading sequence techniques.

2.2. TH-UWB

TH-Pulse Position Modulation (PPM), an UWB TH-PPM signal can be written as [10]:

$$S_{TH\text{-}PPM}^{(k)}(t) = \sqrt{\frac{E_b}{N_s}} \sum_{j=-\infty}^{+\infty} p\left(t - jT_f - c_j^{(k)} T_c - d_{j/N_s}^{(k)} \delta\right). \quad (2)$$

In Eq. (2) $S_{TH\text{-}PPM}^{(k)}(t)$ represents the signal of the k-th user, $p(t)$ is the UWB pulse, which is normalized to satisfy $\int_{-\infty}^{+\infty} p^2(t)dt = 1$. The other parameters in the signal definition are:

- N_s is the number of pulses used to transmit an information bit; it is also called the repetition code;
- $C_j^{(k)}$ represents the time hopping (TH) code; it is a pseudorandom variable bounded by $0 \leq c_j^{(k)} \leq N_h$, where N_h is the hop count;
- T_c is the hop width satisfying $T_f = N_h T_c$;
- $T_b = N_h T_f$ in TH-UWB systems;
- δ is the PPM modulation parameter and $d_{j/N_s}^{(k)}$ is the data bit information that takes values from range $0 \ldots 1$.

In TH-BPSK-Hybrid BPSK and Pulse Shape Modulation (PSM) an UWB TH-BPSK-PSM signal can be written as:

$$S_{TH\text{-}BPSK\text{-}PPM}^{(k)}(t) = \sqrt{\frac{E_b}{N_s}} \sum_{j=-\infty}^{+\infty} d_{j/N_s}^{(k)} p_{MHP}\left(t - jT_f - c_j^{(k)} T_c\right),$$

where:

- $p_{MHP}(t)$ is one of the possible Modified Hermite Pulses (MHP) defined in [16]. To generate MHP waveforms, the following recurrence relation is used:

$$h_n(t) = (-1)^n e^{\frac{t^2}{4}} \frac{d^n}{dt^n} e^{\frac{-t^2}{2}}.$$

Therefore, MHPs for $n = 0, \ldots, 3$ are:

$$h_0(t) = e^{\frac{-t^2}{4}},$$
$$h_1(t) = te^{\frac{-t^2}{4}},$$
$$h_2(t) = (t^2 - 1)e^{\frac{-t^2}{4}},$$
$$h_3(t) = (t^3 - t)e^{\frac{-t^2}{4}}.$$

- $d_{j/N_s}^{(k)}$ is the data bit information and takes values from $-1 \ldots 1$.

In systems using TH-Pulse Amplitude Modulation (PAM) an UWB TH-PAM signal can be written as:

$$S_{TH\text{-}PAM}^{(k)}(t) = \sqrt{\frac{E_b}{N_s}} \sum_{j=-\infty}^{+\infty} p\left(t - jT_f - c_j^{(k)} T_c\right).$$

3. Model and Channel Description

3.1. The Model

In order to study the energy consumption of a UWB system, the architecture of the adopted transmitter and receiver shown in Fig. 1 was used. The P_{pg}, P_{pa}, P_{filt}, P_{ADC}, P_{LNA}, P_{mix}, P_{int} and P_{filr} are the power levels of pulse generator, power amplifier, transmitter filter, digital to analogue converter, analogue to digital converter, low noise amplifier, mixer, integrator, and receiver filter respectively.

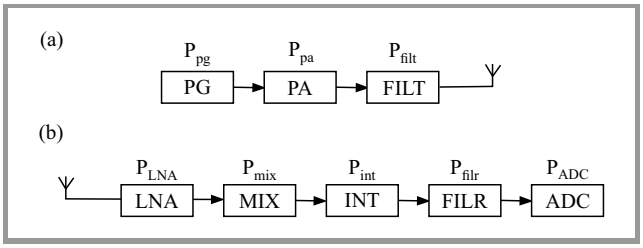


Fig. 1. Block diagram of used: (a) transmitter, (b) receiver.

Sending a sequence of N bits requires T time. Usually, the transceiver is assumed to work according to three different operating modes:

- active mode – in this mode the information is transmitted. The time spent by the transceiver in this active mode is noted T_{ac}. All the components are consuming power;
- sleep mode – when there is no information to convey, the system enters a sleep mode. The time spent by the transceiver in this sleep mode is noted T_{sl}. In this case, a restricted number of components are active;
- transient mode – this corresponds to the transition mode between the sleep and the active modes. The time spent by the transceiver in this transient mode is noted T_{tr}.

Thus, T can be written as:

$$T = T_{ac} + T_{sl} + T_{tr}\,. \tag{3}$$

In paper [17], the authors did not take into account the time to shift from active to sleep mode because it is usually very fast compared to the shift from sleep to active mode. However, they still consider the amount of time spent by the system to shift from sleep to active mode due to the use of a frequency synthesizer, which is energy consuming when the transition takes place. For the UWB Impulse Radio technique, the synthesizer is not used. So, this transient time is not considered. Therefore, the energy needed to transmit N bits is given by Eq. (4):

$$E = P_{ac}T_{ac} + P_{sl}T_{sl}\,, \tag{4}$$

where P_{ac} and P_{sl} represent the power consumption during the active mode and the sleep mode respectively. The power consumption of the active mode includes the transmission power P_t and the electronic circuitry power consumption P_c. P_c combines the receiver power consumption noted P_{cr}, and the transmitter power consumption noted P_{ct}. In the transmission part, the power consumption of Power Amplifier (PA) is linked to the overall transmission power by: $P_{pa} = \alpha P_t$, where $\alpha = \frac{\xi}{\eta - 1}$, ξ is the average of peak to ratio and η is the drain efficiency of the PA. These variables depend on the class of the amplifier and the selected modulation scheme. As compared to the power consumption in the active mode, the power consumption in the sleep mode is very low. Therefore, in this study, $P_{sl} = 0$ i assumed. However, it could be considered for specific applications necessitating long periods in sleep mode. Finally, the amount of energy needed to transmit one bit of information is given in Eq. (5):

$$E_T = \frac{\lfloor (1+\alpha)P_t + (P_c - P_{pa}) \rfloor T_{ac}}{N}\,. \tag{5}$$

3.2. Channel Description

For the sake of simplification, a number of parameters of the IEEE 802.15.4a standard [10] is used and a Nakagami-m distribution for the radio propagation channel amplitudes is adopted. This was integrated in the standard with the UWB path-loss model as described below. Usually the IEEE 802.15.4a uses a clustering behavior for the MPCs. In each cluster there are several rays, the arrivals of the clusters and the rays are modeled as Poisson process. In this model, a deterministic model is used where the MPCs are spaced by the same duration of time τ, and the channel can be written as

$$h(t) = \sum_{l=1}^{L} \alpha_l \delta(t - l\tau)\,,$$

where α_l is the path amplitude. It was shown in [18] that the two models provide almost the same results. This channel model proves to be sufficient for presented analysis. Moreover, the channel gain for a z-th path-loss channel using a distance d between the transmitter and the receiver is

$$G_d = \frac{P_t}{P_s} = G_1 M_l d^z\,.$$

In this expression:

- P_s and P_t are the received power (energy per symbol) and the transmitted power respectively,
- M_l is the gain margin,
- $G_1 = \frac{(4\pi)^2}{g_t g_r \lambda^2}$ is the channel gain for $d = 1$ m which can be obtained from the transmitter and the receiver antenna gains g_t and g_r and from the free space wavelength λ.

Thus, the instantaneous SNR is $\gamma_l = \dfrac{\alpha_l P_t}{G_d N_0 B}$, and the average SNR is $\overline{\gamma_l} = \dfrac{\Omega_l P_t}{G_d N_0 B}$ with $\Omega_l = E\left\lfloor |\alpha_l^2| \right\rfloor$.
In this case, the power density function of the instantaneous SNR is:

$$f_\gamma(\gamma_l) = \frac{m^m \gamma_l^{m-1}}{\overline{\gamma_l}^m \Gamma(m)} e^{\frac{-m\gamma_l}{\overline{\gamma_l}}}\,.$$

The path-loss in UWB is modeled as a normal distribution $z(\mu_z, \sigma_z)$ with $z = \mu_z + n\sigma_z$ and $n = -0.75 \ldots 0.75$ [19]. The path-loss mean and variance values change if the transmitter and the receiver antennas are within Line of Sight (LOS) or within Non Line of Sight (NLOS) conditions. The corresponding statistics are presented in Table 1.

Table 1
Path-loss statictics

	LOS		NLOS	
	μ_z	σ_z	μ_z	σ_z
z	1.7	0.3	3.5	0.98

4. Energy Efficiency Analysis

4.1. DS-UWB with Orthogonal Codes

The chip duration is $T_c = aT_p$ where T_p is the pulse duration and bit duration $T_b = T_f = N_c T_c$. The duration of the active mode $T_{ac} = NT_b$, where N is the length of the data sequence.

For the sake of simplicity, it is assumed that presented system is perfectly synchronized. Therefore, the codes are orthogonal and, as a consequence, their cross-correlations are nil.
For such a system, the conditioned BER on the instantaneous SNR is [20]:

$$P_e(\gamma_l) = Q\left(\sqrt{2\sum_{l=1}^{L}\gamma_l}\right),$$

where $SNR = \gamma$. Thus, the average BER for the DS-UWB system using orthogonal codes can be upper bounded by [20]:

$$P_e = \int\int\int_0^{+\infty} P_e(\gamma_l) f_\gamma(\gamma_l) d\gamma_1 \ldots d\gamma_L \leq \frac{1}{2}\prod_{l=1}^{L} I(\overline{\gamma_l}),$$

where $I(\overline{\gamma_l}) = \left(1+\frac{\overline{\gamma_l}}{m}\right)^{-m}$.
The Maximum Ratio Combining (MRC) technique [21] is used for detection, which is mandatory in a rake receiver. The fading amplitudes are independent and not necessarily identical. In the special case where the L channels are identically distributed with the same average SNR, $\overline{\gamma}$, the BER is simply written as:

$$P_e \approx \frac{1}{2}\left(1+\frac{\overline{\gamma}}{m}\right)^{-mL}.$$

So, the average SNR can be expressed as $\overline{\gamma}=m\left((2P_e)^{\frac{-1}{mL}}-1\right)$ with $\overline{\gamma} = \dfrac{\Omega P_t}{G_d N_0 B}$.
Hence, $P_t T_{ac} = am\left((2P_e)^{\frac{-1}{mL}} - 1\right)\dfrac{N_0 G_d N N_c}{\Omega}$.
Then, the total energy consumption of the DS-UWB system using orthogonal codes is:

$$E_{DS\text{-}OC} = (1+\alpha)am\left((2P_e)^{\frac{-1}{mL}}-1\right)\frac{N_0 G_d N_c}{\Omega} + \frac{(P_c - P_{pa})T_{ac}}{N}. \quad (6)$$

For transmission, the same transmitter is used as in the case of an AWGN channel but, for reception, a partial rake receiver using the MRC technique and the following characteristics $P_c = P_{ct} + P_{cr}$ was chosen:

- transmission power: $P_{ct} = P_{pg} + P_{pa} + P_{filt}$,
- reception power: $P_{cr} = P_{LNA} + L(P_{int} + P_{mix}) + P_{filr} + P_{ADC}$,

where L is the number of the rake receiver branches.

4.2. DS-UWB with Non-orthogonal Codes

The difference between non-orthogonal and orthogonal codes relies on the fact that the cross-correlation values are not nil. Thus, the BER cannot be identical to the mono-user case. For a predefined BER $P_e = 10^{-3}$, the interferences caused by the other users can be approximated by a Gaussian distribution. In the literature, the Gaussian approximation does not hold for all SNR values [12], [22], [23] but, in [12] the approximation is almost Gaussian for $P_e = 10^{-3}$. Hence, the signal to interference plus noise ratio (SINR) can be written as follows [11]:

$$SINR = \frac{2\sum_{l=1}^{L}\gamma_l}{1+\dfrac{N_u-1}{N_c}\sigma_{DS}\sum_{l=1}^{L}\overline{\gamma_l}},$$

where $\sigma_{DS} = \dfrac{1}{T_c}\displaystyle\int\int_{-\infty}^{+\infty}\left[w_{rec}(t)w_{rec}(t-s)\right]^2 ds$,
$w_{rec} = \left[1-4\pi\left(\dfrac{t}{T_p}\right)^2\right]e^{-2\pi\left(\frac{t}{T_p}\right)^2}$, N_u is number of users, and N_c is number of pulses.
The conditional BER on the instantaneous SNR can be written as (a special case where the L channels are identically distributed with the same average): $P_e(\gamma_l) = Q\sqrt{2SINR}$.
Thus, the average BER for the DS-UWB system using non-orthogonal codes can be upper bounded by:

$$P_e = \int\int\ldots\int_0^{+\infty} P_e(\gamma_l) f_\gamma(\gamma_l) d\gamma \ldots d\gamma_L \leq \frac{1}{2}\prod_{l=1}^{L} I(\overline{\gamma_l}),$$

where $I(\overline{\gamma_l}) = \left(1+\dfrac{\frac{1}{m}}{\frac{1}{\overline{\gamma_l}}+\frac{N_u-1}{N_c}L\sigma_{DS}}\right)^{-m}$.
So,

$$P_e \approx \frac{1}{2}\left(1+\frac{\frac{1}{m}}{\frac{1}{\overline{\gamma}}+\frac{N_u-1}{N_c}L\sigma_{DS}}\right)^{-mL}$$

and

$$\overline{\gamma} = \frac{1}{\dfrac{1}{m\left(2P_e^{\frac{-1}{mL}}-1\right)} - \frac{N_u-1}{N_c}L\sigma_{DS}}.$$

Hence,

$$P_t T_{ac} = \frac{\frac{N_0 G_d N_c N}{\Omega}}{\dfrac{1}{m\left(2P_e^{\frac{-1}{mL}}-1\right)} - \frac{N_u-1}{N_c}L\sigma_{DS}}.$$

The total energy consumption of a DS-UWB system using non-orthogonal codes in Nakagami-m channel with path-loss can be written as:

$$E_{DS\text{-}NOC} = (1+\alpha)\frac{\frac{aN_0 G_d N_c}{\Omega}}{\dfrac{1}{m\left(2P_e^{\frac{-1}{mL}}-1\right)} - \frac{N_u-1}{N_c}L\sigma_{DS}} + \frac{(P_c-P_{pa})T_{ac}}{N}.$$

The consumed powers of the transmission and the reception circuitries are identical to the system using the orthogonal codes.

4.3. TH-BPSK-PSM

In order to evaluate the performance of MHPs, in the same conditions as other multiple user techniques, two different waveforms at least must be used. Therefore, the first MHP

is assigned to the first group of 8 users and the second MHP is assigned to the second group of 8 users. TH-BPSK is also the technique used for assuring simultaneous transmission. As the MHP are orthogonal pulses, the interference will mainly be caused by the users belonging to the same group. As in the previous cases, the interference can be approximated as a Gaussian variable for $P_e = 10^{-3}$. The SINR can be written as follows:

$$SINR = \frac{2\sum_{l=1}^{L}\gamma_l}{1+\frac{\frac{N_h}{2}-1}{N_s}\sigma_{BP}\sum_{l=1}^{L}\overline{\gamma_l}},$$

where $\sigma_{BP} = \frac{1}{T_f}\int\int_{-\infty}^{+\infty}\left[P_{MHP}(t)P_{MHP}(t-s)\right]^2 ds.$

The conditional BER on the instantaneous SNR can be written as (special case where the L channels are identically distributed with the same average): $P_e(\gamma_l) = Q\sqrt{2SINR}$.

The BER can be upper bounded of this system by writing:

$$P_e = \int\int\ldots\int_0^{+\infty} P_e(\gamma_l) f_\gamma(\gamma_l) d\gamma_1 \ldots d\gamma_L \leq \frac{1}{2}\prod_{l=1}^{L} I(\overline{\gamma_l}),$$

$$P_e \approx \frac{1}{2}\left(1+\frac{\frac{1}{m}}{\frac{1}{\overline{\gamma}}+\frac{\frac{N_h}{2}-1}{N_s}L\sigma_{BP}}\right)^{-mL},$$

$$\overline{\gamma} = \frac{1}{\frac{1}{m\left(2P_e^{\frac{-1}{mL}}-1\right)}-\frac{\frac{N_h}{2}-1}{N_s}L\sigma_{BP}}.$$

Hence,

$$P_t T_{ac} = \frac{\frac{N_0 G_d N_s N_h N}{\Omega}}{\frac{1}{m\left(2P_e^{\frac{-1}{mL}}-1\right)}-\frac{\frac{N_h}{2}-1}{N_s}L\sigma_{BP}}.$$

The total energy consumption by a system using MHP in a Nakagami-m channel is:

$$E_{MHP} = (1+\alpha)\frac{\frac{aN_0 G_d N_s N_h}{\Omega}}{\frac{1}{m\left(2P_e^{\frac{-1}{mL}}-1\right)}-\frac{\frac{N_h}{2}-1}{N_s}L\sigma_{BP}} + \frac{(P_c-P_{pa})T_{ac}}{N}.$$

4.4. TH-PAM

For a BER $P_e = 10^{-3}$, the multi-user interference is approximated by a Gaussian distribution. In this case the SINR is:

$$SINR = \frac{2\sum_{l=1}^{L}\gamma_l}{1+\frac{N_h-1}{N_s}\sigma_{PAM}\sum_{l=1}^{L}\overline{\gamma_l}},$$

where $\sigma_{PAM} = \frac{1}{T_f}\int\int_{-\infty}^{+\infty}\left[w_{rec}(t)w_{rec}(t-s)\right]^2 ds.$

The conditional BER on the instantaneous SNR as (a special case where the L channels are identically distributed with the same average) is: $P_e(\gamma_l) = Q\sqrt{2SINR}$.
Then,

$$P_e \approx \frac{1}{2}\left(1+\frac{\frac{1}{m}}{\frac{1}{\overline{\gamma}}+\frac{N_h-1}{N_s}L\sigma_{PAM}}\right)^{-mL},$$

$$\overline{\gamma} = \frac{1}{\frac{1}{m\left(2P_e^{\frac{-1}{mL}}-1\right)}-\frac{N_h-1}{N_s}L\sigma_{PAM}},$$

$$P_t T_{ac} = \frac{\frac{N_0 G_d N_s N_h N}{\Omega}}{\frac{1}{m\left(2P_e^{\frac{-1}{mL}}-1\right)}-\frac{N_h-1}{N_s}L\sigma_{PAM}}.$$

The total energy of a TH-PAM system in a Nakagami-m channel with a path-loss is:

$$E_{TH\text{-}PAM} = (1+\alpha)\frac{\frac{aN_0 G_d N_s N_h}{\Omega}}{\frac{1}{m\left(2P_e^{\frac{-1}{mL}}-1\right)}-\frac{\frac{N_h}{2}-1}{N_s}L\sigma_{PAM}} + \frac{(P_c-P_{pa})T_{ac}}{N}.$$

4.5. TH-PPM

In this case the same scheme as the TH-PAM case is used, but for modulation $P_e(\gamma_l) = Q\sqrt{SINR}$. The total energy consumption of a TH-PPM system in a Nakagami-m channel with a path-loss is:

$$E_{TH\text{-}PPM} = (1+\alpha)\frac{\frac{aN_0 G_d N_s N_h}{\Omega}}{\frac{1}{2m\left(2P_e^{\frac{-1}{mL}}-1\right)}-\frac{\frac{N_h}{2}-1}{N_s}L\sigma_{PPM}} + \frac{(P_c-P_{pa})T_{ac}}{N}.$$

where

$$\sigma_{PPM} = \frac{1}{T_f}\int\int_{\infty}^{+\infty}\left[w_{rec}(t)\left(w_{rec}(t-s)-w_{rec}(t-s-\delta)\right)\right]^2 ds.$$

5. Simulation Results

In this section, the simulation results of the total energy needed to transmit one bit using the different access techniques. To compare the energy efficiency of the previously analyzed multi-user techniques, the different sets of parameters are explored. It is assumed that the communication bandwidth has a minimum value $B = 500$ MHz with center at $f_c = 3.46$ GHz situated in the low part of the 3.1–10.6 GHz UWB band. The all the system parameters chosen as being representative of realistic equipment using

the current CMOS technology: $f_c - 3.46$ GHz, $\alpha = 0.78$, $N_0 = -170$ dBm/Hz, $B = 500$ MHz, $N = 10^6$, $P_{pg} = 25.2$ mW, $P_{LNA} = 7.68$ mW, $P_{mix} = 15$ mW, $P_{pg} = 2.5$ mW, $P_{ADC} = 7.6$ mW, $P_{filt} = P_{filr} = 2.5$ mW, $L = 4$, $P_e = 10^{-3}$, $M_l = 40$ dB, $G_1 = 33.2$ dB, $\Omega = 1$ ([3] and [4]).
All the experiments are performed in an NLOS propagation channel with $m = 0.7$, which is the depth of fading valid for all the following test configurations. The authors successively consider two different simulation scenarios. The first one maintains the same number of users, the second one imposes the same rate.

5.1. Using with Same Number of Users

In Fig. 2, the requested transmission energy DS-UWB with orthogonal (T-OVSF, ZCD) and non-orthogonal (Kasami I) codes and TH-BPSK-PSM described above versus the distance $d = 5 \ldots 50$ m, for $P_e = 10^{-3}$ and $z = 3.5$ is depicted. The purpose of this initial test is to check the compliance of the selected UWB system power transmission regarding the FCC authorized standard.

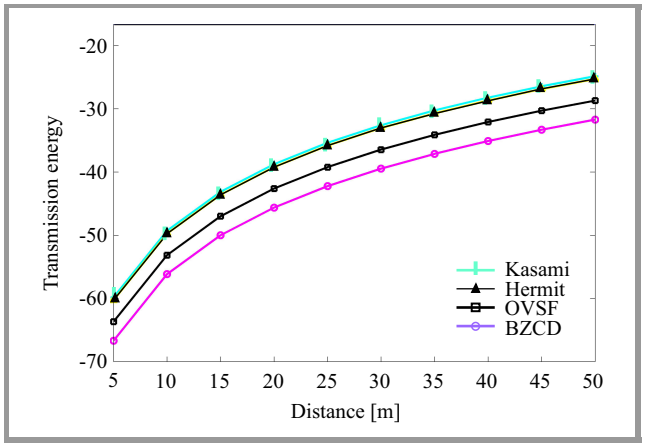


Fig. 2. Transmission energy for different multi-users techniques.

All the techniques studied are compliant with the UWB mask for the whole set of communication ranges. In addition the Kasami technique requires more transmission energy as compared to other techniques. Due to the need for a high number of pulses to handle 16 users, Kasami consumes more energy than that needed for the orthogonal codes and the MHP.
In Fig. 3, the total energy consumption for different multi-user access technique versus distance is presented where in Nakagami-m fading channel and same condition as in Fig. 2.
From the beginning to the end of the, BZCD is the most efficient access technique. It consumes less energy to handle 16 users because it has the advantage of using the least number of pulses compared to the other techniques. While, BZCD codes need 64 pulses to handle 16 users OVSF codes need 128 and the Kasami codes 255. In simulation 8 pulses for the MHP were used, but since the frame duration of MHP to manage 16 users is higher than the frame duration

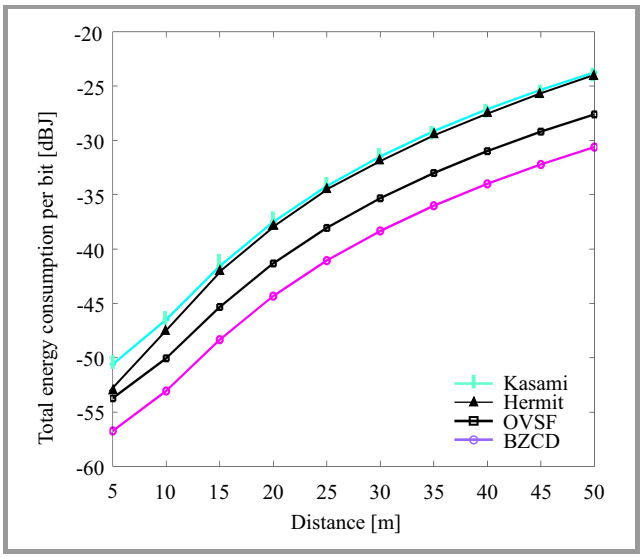


Fig. 3. Total energy for different multi-users techniques using in Nakagami-m fading channel.

of BZCD, MHP technique consumes more than the BZCD one, even with less pulses. The Kasami codes are not very efficient on the whole set of ranges due to the huge number of pulses needed for supporting 16 users. The number of pulses plays a major role in the overall energy consumption since the circuitry required energy and the transmission energy becomes higher and higher with the increase of the number of pulses to be generated.
In Fig. 4, the total energy consumption against the path-loss z is shown using parameters $d = 30$ m, $P_e = 10^{-3}$ and $z = 2.8 \ldots 4.2$ in the Nagakami-m fading channel.

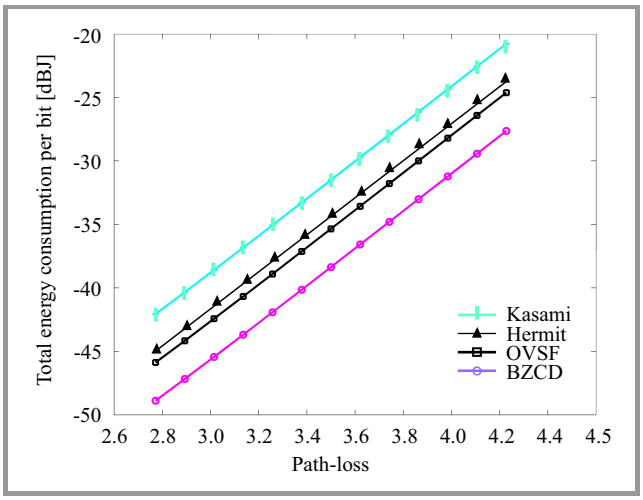


Fig. 4. Total energy for different multi-users techniques vs. path loss z.

Using these settings, the BZCD is the most efficient access technique and that Kasami is the less efficient one. For this particular considered distance and path-loss range, using fewer pulses to transmit the data bits is helpful to achieve an energy efficient system.

5.2. Using the Same Rate

In these experiments, the TH-PPM is compared with other techniques. The latter needs to double the signal processing time, therefore it is suitable to add the power consumption of a microcontroller. It is assumed that 20% of the energy is consumed by the reception circuitries.
In Fig. 5, the transmission energy for four of the multi-user techniques described above versus the distance are depicted, in Nakagami-m fading channel.

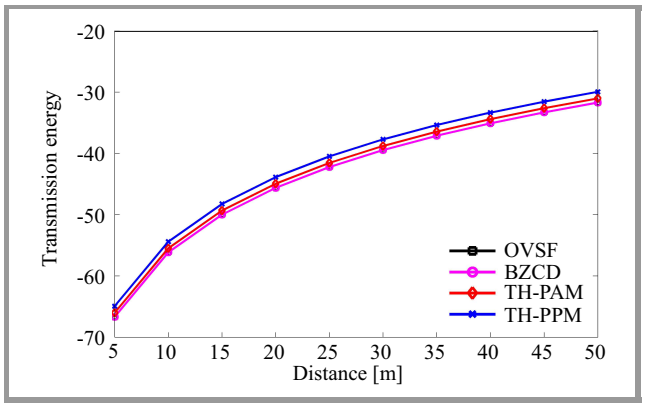


Fig. 5. Transmission energy for different multi-users techniques in Nakagami-m fading channel.

The orthogonal codes request the lowest transmission energy. The highest transmission energy is required by the TH-PPM technique.
In Fig. 6, the total energy consumption for different multi-user access techniques versus the distance is shown using $d = 5 \ldots 50$ m, $P_e - 10^{-3}$ and $z = 3$ in the Nakagami-m channel. To maintain the same rate, $N_c = N_h N_s$ is assumed.

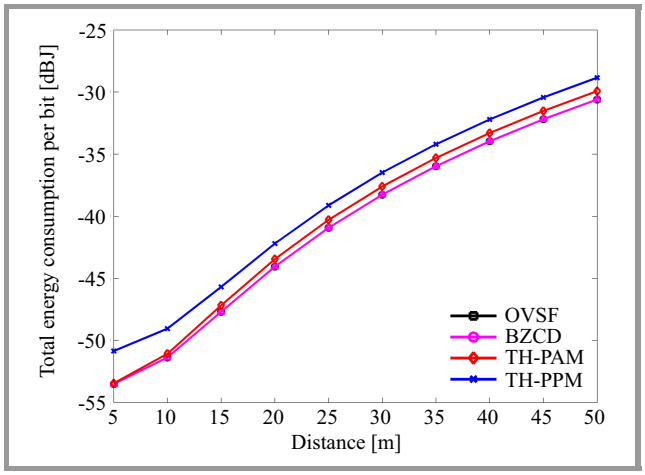


Fig. 6. Total energy for different multi-users techniques vs. distance in Nakagami-m channel.

The simulation results show that T-OVSF and BZCD are the most efficient multi-access techniques from the beginning of the selected communication range up to its end. Despite the use of fewer pulses, TH-PAM and TH-PPM are less efficient than both orthogonal codes. However, TH-PAM is better than TH-PPM energetically speaking. This is due to the robustness of TH-PAM against the multi-user interference. Even if T-OVSF and BZCD consume the same amount of energy for the same number of pulses used, BZCD can handle up to 16 users unlike T-OVSF codes, which are limited to 8 (in this setting). This makes BZCD a better choice. Some tradeoff between the number of users and the energy consumed may be performed if the application demands it.
Figure 7 shows the total energy consumption for different multi-user technique against the path-loss z ($d = 30$ m, $P_e = 10^{-3}$, and $z = 2.8 \ldots 4.2$) in the Nakagami-m channel.

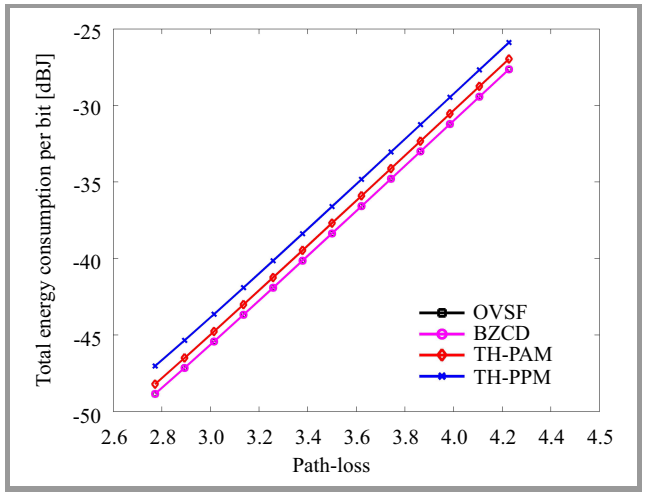


Fig. 7. Total energy for different multi-users techniques vs. path-loss in Nakagami-m channel.

In this considered range, T-OVSF and BZCD are the most efficient techniques over the whole range of path-loss. For the same duration frame, the orthogonal codes are better choice than the time hopping one. This energy efficiency is linked to the influence of the multi-user scenario. When the interference variance increases the total energy consumed increases.

6. Conclusion

In order to develop an energy efficient UWB multi-user, high bit rate, communication system, this paper compared DS-UWB using orthogonal or non-orthogonal codes and different TH-UWB techniques. They were evaluated in Nagakami-m fading propagation channels. In the first part, the different theoretical expressions of the consumed energy were established. Then, in the second part, simulations were run to effectively compare the consumed energy. In the first step, a fixed number of users were considered. We found that the BZCD are more efficient for the range of distances considered and for the whole interval of path-loss coefficient. This first experiment showed that the efficient system is the system using the least number of pulses to transmit the data bit. In the second set of experiments, the transmission rate was fixed. The T-OVSF and BZCD are the less energy consuming for the whole range of dis-

tances. Besides, when the distance has a middle value and the path-loss varies, T-OVSF and BZCD remain less energy consuming than the other techniques considered. The second experiment showed that the least efficient technique is the technique having the highest interference variance. To conclude, this energetic model can be further used as a green communication tool to determine the best multi-user techniques regarding energy consumption in a given transmission conditions.

Acknowledgment

The authors would like to thank the Railenium Technological Research Institute for supporting this work.

References

[1] F. Chuan and L. Anqing, "Key techniques in green communication", in *Proc. Int. Con. Consumer Electron., Commun. & Networks CECNet 2011*, XianNing, China, 2011, pp. 1360–1363.

[2] M. Z. Win and R. A. Scholtz, "Impulse radio: how it works", *IEEE Commun. Lett.*, vol. 2, no. 2, pp. 36–38.

[3] A. T. Phan, J. Lee, V. Krizhanovskii, Q. Le, S.-K. Han, and S.-G. Lee, "Energy-efficient low-complexity CMOS pulse generator for multiband UWB impulse radio", *IEEE Trans. Circuits & Syst. I: Reg. Papers*, vol. 55, no. 11, pp. 3552–3563, 2008.

[4] T. Yuan, Y. Zheng, K. S. Yeo, C. C. Boon, and A. V. Do, "A CMOS energy efficient UWB transmitter module", in *Proc. IEEE SoC Design Conf. ISOCC 2009*, Busan, South Korea, 2009, pp. 25–28.

[5] J. Hu, Y. Zhu, S. Wang, and H. Wu, "An energy-efficient IR-UWB receiver based on distributed pulse correlator", *IEEE Trans. Microw. Theory Techn.*, vol. 61, no. 6, pp. 2447–2459, 2013.

[6] N. Riaz and M. Ghavami, "An energy-efficient adaptive transmission protocol for ultra-wideband wireless sensor networks", *IEEE Trans. on Veh. Technol.*, vol. 58, no. 7, pp. 3647–3660, 2009.

[7] X. Jinghao, B. Peric, and B. Vojcic, "Energy-aware and link-adaptive routing metrics for ultra-wideband sensor networks", in *IEEE Worksh. Ultra-Wideband for Sensor Networks*, Rome, Italy, 2005.

[8] X. An and K. Kwak, "An energy-efficient routing scheme for UWB sensor networks", in *Proc. Asia-Pacific Conf. Commun. APCC 2006*, Busan, South Korea, 2006, pp. 1–5.

[9] A. Elabboubi, F. Elbahhar, M. Heddebaut, and Y. Elhillali, "An energy efficiency modulation comparative study for a railway beacon", in *Proc. Int. Symp. signal Image Video & Commun. ISIVC 2014*, Marrakech, Morocco, 2014.

[10] A. Elabboubi, F. Elbahhar, M. Heddebaut, and Y. Elhillali, "Comparison of UWB modulations over Nakagami-m Fading Channels with Path-loss for an energy efficient railway balise application", in *Proc. IEEE Int. Conf. Ultra-Wideband ICUWB 2014*, Paris, France, 2014, pp. 427–432, 2014.

[11] A. F. Molisch *et al.*, "A comprehensive model for ultrawideband propagation channels of UWB system proposals standard for these applications", in *Proc. IEEE Global Telecommun. Conf. GlobeCom 2005*, St. Louis, MO, USA, 2005.

[12] B. Hu and N. C. Beaulieu, "Accurate performance evaluation of time-hopping and direct-sequence UWB systems in multi-user interference", *IEEE Trans. Commun.*, vol. 53, no. 6, pp. 1053–1062, 2005.

[13] D. Wu, P. Spasojevi, and I. Seskar, "Ternary zero-correlation zone sequences for multiple code UWB", in Proc. *38th Ann. Conf. Inform. Sci. & Syst. CISS 2004*, Princeton, NJ, USA, 2004, pp. 939–943.

[14] J. Cha, N. Hur, K. Moon, and C. H. Lee, "ZCD-UWB system using enhanced ZCD codes", in *Proc. Int. Worksh. on Ultra Wideband Syst. Joint with Conf. on Ultrawideband Syst. & Technol. Joint UWBST & IWUWBS 2004*, Kyoto, Japan, 2004, pp. 371–375.

[15] Y. R. Tsai and X. S. Li, "Kasami code-shift-keying modulation for ultra-wideband communication systems", in *Proc. IEEE Int. Conf. Ultra-Wideband*, Waltham, MA, USA, 2006, pp. 37–42.

[16] C. J. Mitchell, G. T. F. de Abreu, and R. Kohno, "Combined pulse shape and pulse position modulation for high data rate transmissions in ultra-wideband communications", *Int. J. Wirel. Inform. Netw.*, vol. 10, no. 4, pp. 167–178, 2003.

[17] S. Cui, A. J. Goldsmith, and A. Bahai, "Energy-constrained modulation optimization", *IEEE Trans. Wirel. Commun.*, vol. 4, no. 5, pp. 2349–2360, 2005.

[18] L. J. Greenstein, S. S. Ghassemzadeh, S.-C. Hong, and V. Tarokh, "Comparison study of UWB indoor channel models", *IEEE Trans. Wirel. Commun.*, vol. 6, no. 1, pp. 128–135, 2007.

[19] S. S. Ghassemzadeh and V. Tarokh, "The ultra-wideband indoor path loss model", Tech. Rep., IEEE P802.15 Working Group for Wireless Personal Area Networks (WPANs), June 2002.

[20] J. G. Proakis, *Digital Communications*. New York: McGraw-Hill, 2001, ch. 5, p. 257.

[21] M. K. Simon and M. S. Alouini, "A unified approach to the performance analysis of digital communication over generalized fading channels", *Proc. IEEE*, vol. 86, no. 5, pp. 1860–1877, 1998.

[22] F. Kharrat-Kammoun, C. Le Martret, and P. Ciblat, "Performance analysis of IR-UWB in multi-user environment", *IEEE Trans. Wirel. Commun.*, vol. 8, no. 11, pp. 5552–5563, 2009.

[23] N. C. Beaulieu and D. J. Young, , "Designing time-hopping ultra-wide bandwidth receivers for multiuser interference environments", *Proc. IEEE*, vol. 97, no. 2, pp. 255–284, 2009.

Adil Elabboubi received his engineering degree in Communication Systems from Telecom Lille 1 (France) in 2011. He is currently a Ph.D. student at the University of Valenciennes (France). He is doing his research works at the French National Institute for Transportation and Safety Research (IFSTTAR). His current fields of research are UWB systems, energy aware communication systems and green communication.
E-mail: adil.elabboubi@ ifsttar.fr
French National Institute for Transportation
and Safety Research (IFSTTAR)
20 rue Élisée Reclus BP 70317
F-59666, Villeneuve d'Ascq, France

Fouzia Elbahhar received the M.Sc. and Ph.D. degrees from the University of Valenciennes (France) in 2000 and 2004, respectively. She is actually employed as researcher at IFSTTAR/LEOST, Villeneuve d'Ascq, France. She participates at many national and European projects dedicated to transport applications. She serves also

as a reviewer for several journals and international conference. She is involved in signal processing especially ultra-wide band technology. Her major research interests are land transportation like communication vehicle-to-vehicle and vehicle-to-infrastructure, and localization.
E-mail: fouzia.boukour@ifsttar.fr
French National Institute for Transportation
and Safety Research (IFSTTAR)
20 rue Élisée Reclus BP 70317
F-59666, Villeneuve d'Ascq, France

Marc Heddebaut received the M.Sc. and Ph.D. degrees in Electronics from the University of Lille France in 1980 and 1983, respectively. He joined the French National Institute for Transportation and Safety Research (INRETS now IFSTTAR) in 1983 and became a senior researcher in 1988. Since 1979, he has been working in the field of land mobile communication and electromagnetic compatibility. His primary interests are telecommunication systems dedicated to land transport, mobile localization and command control of automated vehicles.
E-mail: marc.heddebaut@ifsttar.fr
French National Institute for Transportation
and Safety Research (IFSTTAR)
20 rue Élisée Reclus BP 70317
F-59666, Villeneuve d'Ascq, France

Yassin El Hillali received the M.Sc. and Ph.D. degrees in Electronic and Communication Systems from the University of Valenciennes (France) in 2002 and 2005, respectively. He is actually an associate professor at the University of Valenciennes. His areas of interest are embedded systems, digital signal processing, wireless sensor networks, radar systems and ultra-wide band systems.
E-mail: yassin.elhillali@univ-valenciennes.fr
IEMN-DOAE
Valenciennes, 59300 France

Paper

Monetary Fair Battery-based Load Hiding Scheme for Multiple Households in Automatic Meter Reading System

Ryota Negishi, Shuichiro Haruta, Chihiro Inamura, Kentaroh Toyoda, and Iwao Sasase

Dept. of Information and Computer Science, Keio University, Hiyoshi, Kohoku, Yokohama, Kanagawa, Japan

Abstract—Automatic Meter Reading (AMR) system is expected to be used for real time load monitoring to optimize power generation and energy efficiency. Recently, it has been a serious problem that user's lifestyle may be revealed by a tool to estimate consumer's lifestyle from a real-time load profile. In order to solve this issue, Battery-based Load Hiding (BLH) algorithms are proposed to obfuscate an actual load profile by charging and discharging. Although such BLH algorithms have already been studied, it is important to consider multiple households case where one battery is shared among them due to its high cost. In this paper, a monetary fair BLH algorithm for multiple households is proposed. In presented scheme, the core unit calculates the difference between the charged amount and discharged one for each household. If the difference is bigger than the predefined threshold (monetary unfair occurs), the most disadvantageous and advantageous households are given priority to discharge and charge the battery and other households should charge to achieve monetary fairness. The efficiency of the scheme is demonstrated through the computer simulation with a real dataset.

Keywords—Automatic Meter Reading, Battery Load Hiding, Privacy for Smart Grid.

1. Introduction

In recent years, smart meters have gained much popularly with growing support from the electric power company and governments. However, smart meters pose substantial threat to the privacy of individuals [1]. Smart meters use measurement circuits that can record the load profile by a second or minute order. Recently, it has been a serious problem that user's lifestyle may be revealed by a tool, which is called Non-Intrusive Load Monitoring (NILM), to estimate consumer's lifestyle from a real-time load profile [2]–[4]. The most of NILM techniques are to detect edges in a load profile [5]–[7]. Batra *et al*. publish an open source toolkit of NILM named NILMTK [8]. However, NILM gives rise to serious user privacy concerns. Multiple studies have shown that smart meters are vulnerable to an attack that could leak fine grained usage data to third parties, e.g. an electric power industry [9]. In order to preserve individual's privacy, a Battery-based Load Hiding (BLH) technique is proposed to avoid the information leakage by NILM [10]–[14]. The basic concept of BLH is to hide actual load by wisely charging/discharging a battery. For example, in Best Effort (BE), the core unit, which is a battery controller for BLH, charges/discharges a battery to flatten the metered load [10]. Another novel work is Non-Intrusive Load Leveling (NILL) algorithm [11]. In NILL algorithm, the core unit aims to flatten the metered load and controls the residual energy of the battery in order to continue BLH [11]. However, these schemes disclose true demand when the battery is almost empty or full. In order to solve this problem, Stepping Framework (SF) is proposed to step a metered load instead of flattening it by considering the current energy consumption level of the appliances [12].

Although many BLH algorithms have been studied in the literature, most of them do not consider the multiple households case. Privacy leakage problem is related with all regions where a real-time load measuring system is offered. According to [15], countries all over the world, e.g., US, Canada, United Kingdom, France, Spain, China and Japan, have taken the decision to roll out smart metering system. Irrespective of country, one may feel that it is expensive because a battery of 1 kWh might cost at least $1,200 [16]. Therefore it has a great importance to realize a BLH where a battery is shared among multiple households. A realistic case of the shared battery is an apartment, condominium or a set of houses [17]. In this case, inhabitants who want to avoid the privacy leakage by smart meter may cover the expenses of the development and maintenance of such a battery system. Vilardebo *et al*. propose a BLH scheme for multiple households, however, they do not consider monetary fairness [13]. That is, an unfair situation may occur when households pay a money to charge a battery by BLH but they do not use the same amount of the charged energy from it. Therefore, it is necessary to propose a monetary fair BLH scheme for multiple households.

In this paper, a monetary fair BLH scheme for multiple households by using only one battery is proposed. Authors first present a monetary fair BLH scheme for two households. In the scheme, the core unit chooses one of the following three modes based on monetary loss and residual energy on the battery: the stabilization mode, fairness mode, and normal mode. In the stabilization mode, the core unit controls the amount of residual energy and avoids the situation where BLH cannot be executed. In the fairness mode, the core unit lets an overcharged household discharge, while it lets the other charge in order to solve monetary unfairness. Finally, in the normal mode, the core unit

calculates each household's metered load at time t against every possible case and chooses the case where the residual energy approaches almost the half of battery capacity.

Authors further extend proposed algorithm to deal with more than two households is applied. If original algorithm for multiple households, the core unit would have to calculate all patterns in the normal mode and it would require heavy computation on the core unit – the order is $O(2^N)$ where N denotes the number of households. Therefore, authors propose an extended algorithm to deal with multiple households by approximating the algorithm in the normal mode. More specifically, the core unit first decides the number of charging (or discharging) households so that residual energy approaches to the target energy level (more specifically 55% of the maximum capacity). If the residual energy is less than that value, more households charge battery. Then which household charges/discharges is assigned. The efficiency of proposed scheme is shown by the computer simulation. The evaluation metrics are mutual information, which is a major indicator of how much information is leaked by BLH, and monetary loss. Authors also clarify how many households can be covered with proposed algorithm. A real electric loads dataset called Wiki-Energy is used [18] to obtain reasonable outcome.

The remainder of this paper is organized as follows. Related work regarding BLH and its shortcomings is summarized in Section 2. The proposed scheme with discussion is described in Section 3. Simulation results are shown in Section 4. The paper is concluded in Section 5.

2. Related Work

2.1. Summary of Battery Load Hiding Schemes

To protect a privacy for smart meter users, many researchers have proposed BLH algorithms considering various constraints on the battery such as capacity to minimize the amount of information leakage [10]–[14]. In BLH algorithm, the operation system controls the battery based on the demand load and previous time energy consumption observed by a smart meter (the metered load) in order to control the currently metered load.

Current BLH algorithms basically aim to flatten the metered load by wisely charging/discharging a battery. The main difference among these algorithms is how to react when the residual energy of a battery is in almost empty or full. In the BE [10], when the energy level reaches the minimum or maximum, the core unit determines whether it should be charged or discharged at the maximum rate. In the NILL [11], the core unit chooses a charging/discharging rate with respect to the energy consumption of appliances. Yang *et al*. analyze the above two algorithms and show that these two algorithms disclose the true energy consumption when the battery is too low or too high. To solve this problem, they propose SF-LS2 [12]. In SF-LS2 , instead of trying to maintain a constant load, the core unit monitors the current energy consumption level of the appliances and chooses a target load value from a set of predefined values. Yang *et al*. verify the tradeoff between the privacy and the electricity bill and propose an online algorithm that can optimally control the battery to protect the smart meter data privacy and cut down the electricity bill [14]. Vilardebo *et al*. propose a BLH scheme that operates over multiple users by defining privacy-power function [13].

2.2. Shortcomings in Conventional BLH Schemes

Although there are many BLH algorithms, most of algorithms do not consider using one battery for multiple households. One may feel that it is expensive since a battery of a 1 kWh might cost at least $1,200 [16]. Therefore it has a great importance to realize a BLH, where a battery is shared among multiple households. Vilardebo *et al*. propose such a BLH scheme with a single battery, however, they do not consider monetary fairness (cost/profit balance between users) [13]. Without considering it for the multiple households case, one might gain or lose money by executing BLH. Here, monetary fairness denotes that the charged amount for BLH must be same as the discharged amount for each household. However, it is difficult to achieve the monetary fairness because of two constraints on the battery. The first constraint is that the battery has a limit on charge and discharge rate. The core unit needs to choose, which user and how much energy should be charged or discharged. The second one is that BLH is limited by the battery capacity. When the system deals with multiple households with one battery, it is challenging to appropriately execute BLH for each one.

3. Proposed Scheme

The paper proposed a monetary fair BLH algorithms for multiple households. Firstly, a BLH scheme for two households with a battery and then extend it to deal with more than two by approximating the computationally heaviest part in the algorithm is shown. Figure 1 shows the system

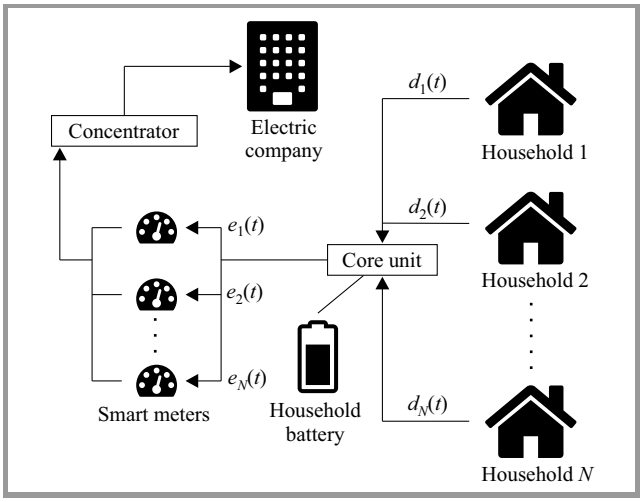


Fig. 1. The system model of BLH scheme.

Table 1
Notations used in presented scheme

Parameters	Definition
i	Index of a household
β	Quantization width
β'	$\frac{\beta}{N}$
$E_{rest}(t)$	Ratio of residual energy to the battery capacity at time t [%]
$l_i(t)$	Monetary loss caused by charging and discharging within household i
l_{th}	Threshold of $l_i(t)$
$d_i(t)$	Demand load in household i
$s_i(t)$	Charging signal. If $s_i(t)=1$, the core unit quantizes household i's load by charging. Otherwise, the core unit quantizes household i's load by discharging.
$e_i(t)$	Metered load (the load after BLH) in household i at time t
$C_{\max}$	Battery capacity
E_{fine}	Fine level of the battery. $E_{fine}=0.55C_{\max}=\frac{0.9+0.2}{2}C_{\max}$
p	The household which most charged during the period from 0 to $t-1$.
q	The household which most discharged during the period from 0 to $t-1$.
K_1	Difference between $0.9C_{\max}$ and $E_{rest}(t)$
K_2	Difference between $E_{rest}(t)$ and $0.2C_{\max}$
N_C	Number of charging households other than q at time t
N_D	Number of discharging households other than q at time t

model of BLH scheme. $d_i(t)$ denotes the total electric load demanded by the appliances in a household i at time t. In contrast, $e_i(t)$ is the summation of $d_i(t)$ and load charged/discharged by BLH at time t (see Table 1). In order to realize BLH for multiple households, each household's $e_i(t)$ must be calculated based on $d_i(t)$. After deciding $e_i(t)$, the core unit controls the battery in order to output $e_i(t)$ to each smart meter. After that the core unit sends each smart meter to $e_i(t)$. When each smart meter receives $e_i(t)$, each smart meter sends $e_i(t)$ to the concentrator and the concentrator sends $e_i(t)$ to the electric company.

The threshold l_{th} is defined that determines the upper bound of instantaneous monetary unfairness. When the difference between the charged amount and discharged one exceeds the predefined threshold, the core unit lets the overcharged household discharge, or vice versa. This scheme consists of three modes: stabilization, fairness, and normal mode. The control unit changes its mode based on the residual energy and the amount of loss caused by BLH. When the residual energy is almost empty or full, the core unit transits to the stabilization mode, which is based on the state-of-the-art BLH scheme SF-LS2 [12] to avoid the situation where BLH cannot be executed. If a household charges too much, the core unit transits to the fairness mode to solve monetary unfairness. Otherwise, the core unit executes the normal mode so that the residual energy approaches almost half of its capacity. After deciding its mode, the core unit decides each household's metered load $e_i(t)$ with a quantization band β, where β is a quantization bandwidth for household i's demand load $d_i(t)$. β indicates how coarsely hides a demand load and it is given by taking into account to the battery capacity and charging/discharging rate, where charging/discharging rate denotes how much energy the battery can charge/discharge within a time unit. Finally, the core unit charges or discharges by the calculated amount.

Moreover, authors extend the algorithm to deal with more than two households. In the extended version, the core unit first decides the number of households that executes BLH by charging (and discharging), which is denoted as N_C (and N_D), based on the current residual energy $E_{rest}(t)$. After deciding the number of charging and discharging households, the core unit then assigns each household to charging or discharging group.

3.1. Three Modes of BLH Algorithm

Deciding mode. Algorithm 1 shows an algorithm for the core unit to select its operating mode. First, if the residual energy is almost empty – $E_{rest}(t-1) \leq 20\%$ or full $E_{rest}(t-1) \geq 90\%$, the stabilization mode is chosen to avoid the situation where the residual energy gets empty or full. If either of households overcharges, i.e., the charged amount is beyond the pre-defined threshold l_{th}, the control unit transits to the fairness mode to achieve monetary fairness. Otherwise, the control unit chooses the normal mode so that the residual energy approaches almost half of the battery capacity $C_{\max}$, where $C_{\max}$ is maximum battery capacity.

Algorithm 1: Deciding mode

```
Input E_rest(t-1)
if E_rest(t-1) is almost empty ∪ E_rest(t-1) is almost full then
    mode ← Stabilization
else if |l_i(t)| ≥ l_th for i = 1 and/or 2 then
    mode ← Fairness
else
    mode ← Normal
end if
Return mode
```

Stabilization mode. In the stabilization mode (Algorithm 2), the core unit lets each household charge ($s_1(t) \leftarrow 1$, $s_2(t) \leftarrow 1$) when the residual energy is almost empty (under 20%). On the other hand, the core unit lets each house-

Algorithm 2: Stabilization mode

Input $E_{rest}(t-1)$
for $i \in 1:2$ **do**
  if $E_{rest}(t-1) \leq 20\%$ **then**
    $s_1(t) \leftarrow 1$
    $s_2(t) \leftarrow 1$
  else if $E_{rest}(t-1) \geq 90\%$ **then**
    $s_1(t) \leftarrow 0$
    $s_2(t) \leftarrow 0$
  end if
  $\beta' \leftarrow \frac{\beta}{2}$
  if $s_i(t) = 1$ **then**
    $e_i(t) \leftarrow \left\lceil \frac{d_i(t)}{\beta'} \right\rceil \beta'$
  else if $d_i(t) \bmod \beta \neq 0$ **then**
    $e_i(t) \leftarrow \left\lfloor \frac{d_i(t)}{\beta'} \right\rfloor \beta'$
  else
    $e_i(t) \leftarrow \left(\frac{d_i(t)}{\beta'} - 1 \right) \beta'$
  end if
end for
Return $e_1(t)$ and $e_2(t)$

hold discharge – $s_1(t) \leftarrow 0$, $s_2(t) \leftarrow 0$, when the residual energy is almost full (over 90%). Here, $s_i(t)$ denotes whether household i hides its load by charging or discharging at time t. That is, $s_i(t) = 1$ indicates that the core unit lets household i charge, while $s_i(t) = 0$ indicates that the core unit lets household i discharge. Then, the core unit calculates a target quantized load $e_i(t)$ for each household according to $s_i(t)$. Here, β' is set as $\frac{\beta}{2}$ so that each household equally charges/discharges.

Fairness mode. In the fairness mode (Algorithm 3), the core unit lets an overcharged household i.e. $l_i(t-1) \geq l_{th}$ discharge and lets the other charge to solve monetary unfairness, where $l_i(t)$ denotes the difference between charged and discharged amount of energy for a household i at time t. Then, the core unit calculates a target quantized load $e_i(t)$ for each household according to $s_i(t)$.

Algorithm 3: Fairness mode

Input $l_1(t-1)$ and $l_2(t-1)$
if $l_1(t-1) \leq l_2(t-1)$ **then**
  $s_1(t) \leftarrow 1$
  $s_2(t) \leftarrow 0$
else
  $s_1(t) \leftarrow 0$
  $s_2(t) \leftarrow 1$
end if
for $i \in 1:2$ **do**
  if $s_i(t) = 1$ **then**
    $e_i(t) \leftarrow \left\lceil \frac{d_i(t)}{\beta} \right\rceil \beta$
  else if $d_i(t) \bmod \beta \neq 0$ **then**
    $e_i(t) \leftarrow \left\lfloor \frac{d_i(t)}{\beta} \right\rfloor \beta$
  else
    $e_i(t) \leftarrow \left(\frac{d_i(t)}{\beta} - 1 \right) \beta$
  end if
end for
Return $e_1(t)$ and $e_2(t)$

Normal mode. Algorithm 4 shows the algorithm of the normal mode. The fine level E_{fine} of the battery is defined and set

$$E_{fine} = 0.55 C_{\max} = \frac{0.9 + 0.2}{2} C_{\max}.$$

In the normal mode, the core unit calculates each household's metered load at time t for every possible case, i.e. $\{s_1(t),\ s_2(t)\}$ in $\{\{0,0\},\{0,1\},\{1,0\},\{1,1\}\}$. Then, the core unit chooses the case where the residual energy most approaches E_{fine}.

Algorithm 4: Normal mode

for $\{s_1(t),\ s_2(t)\} \in \{\{0,0\},\{0,1\},\{1,0\},\{1,1\}\}$ **do**
  for $i \in 1:2$ **do**
    if $s_i(t) = 1$ **then**
      $e_{i,s_i(t)}(t) \leftarrow \left\lceil \frac{d_i(t)}{\beta} \right\rceil \beta$
    else if $d_i(t) \bmod \beta \neq 0$ **then**
      $e_{i,s_i(t)}(t) \leftarrow \left\lfloor \frac{d_i(t)}{\beta} \right\rfloor \beta$
    else
      $e_{i,s_i(t)}(t) \leftarrow \left(\frac{d_i(t)}{\beta} - 1 \right) \beta$
    end if
  end for
  if the combination of $e_{1,s_1(t)}(t)$ and $e_{2,s_2(t)}(t)$ more approaches $E_{rest}(t) = 55\%$ **then**
    $e_1(t) \leftarrow e_{1,s_1(t)}(t)$
    $e_2(t) \leftarrow e_{1,s_2(t)}(t)$
  end if
end for
Return $e_1(t)$ and $e_2(t)$

3.2. Extended Algorithm for Multiple Households

In the next step the algorithm was extended for more than two households. Although the modes in the extended algorithm are almost same with the algorithm for two households, each mode needs to be slightly modified to deal

Algorithm 5: Deciding mode in multiple households case

Input $E_{rest}(t-1)$
if $E_{rest}(t-1)$ is almost empty $\cup$ $E_{rest}(t-1)$ is almost full **then**
  $mode \leftarrow Extended\ Stabilization$
else if i exists that satisfies $|l_i(t)| \geq l_{th}$ **then**
  $mode \leftarrow Extended\ Fairness$
else
  $mode \leftarrow Extended\ Normal$
end if
Return *mode*

with more households due to the following two reasons. The first one is to require large computational complexity. The second one is the possibility that BLH cannot be executed when the battery is fully charged or empty gets higher in presence of multiple households. Therefore, some parts of operated modes are modified to take into account these difficulties.

Deciding mode. Algorithm 5 shows the algorithm to decide the operating mode. First, the core unit checks the residual energy with the same way of the deciding mode in two households. Then, if there exist households whose loss or profit is more than l_{th}, the core unit chooses the extended fairness mode to solve monetary unfairness. Otherwise, the core unit transits to the extended normal mode.

Extended normal mode. In Algorithm 6, the normal mode for two households decides each $s_i(t)$ for every possible case and thus the computation complexity is $O(2^N)$, where N denotes the total number of households. It is necessary to decrease the computation complexity when the system deals with more than two households. Therefore, authors take an approximate measure to decide each $s_i(t)$ and $e_i(t)$ in the extended normal mode. First N_C was set, which is the number of $s_i(t) = 1$, i.e. the number of households that execute BLH by charging, by taking into account the residual energy $E_{rest}(t)$. For the ease of discussion, first it is assumed each household consumes the same amount of energy. Intuitively, more households must charge when the residual energy $E_{rest}(t)$ is below the target E_{fine}. More specifically, N_C was corrected by the ratio of $0.9C_{\max} - E_{rest}(t)$ to $E_{rest}(t) - 0.2C_{\max}$. Next, N_D was set, which is the number of $s_i(t) = 0$, as $N_D = N - N_C$. Therefore the number of charging households N_C and that of discharging households N_D are calculated by:

$$N_C = round\left(\frac{K_1}{K_1+K_2}\right)N, \tag{1}$$

$$N_D = N - N_C. \tag{2}$$

Figure 2 shows an example of calculating N_C and N_D. In Figure 2, we can obtain $K_1 = 0.9C_{\max} - 0.4C_{\max} = 0.5C_{\max}$ and $K_2 = 0.4C_{\max} - 0.2,\ C_{\max} = 0.2,\ C_{\max}$, thus $\frac{K_1}{K_1+K_2} = \frac{5}{5+2}$. Hence, when there are 7 households, the number of charging households N_C is 5 and that of discharging households N_D is 2.

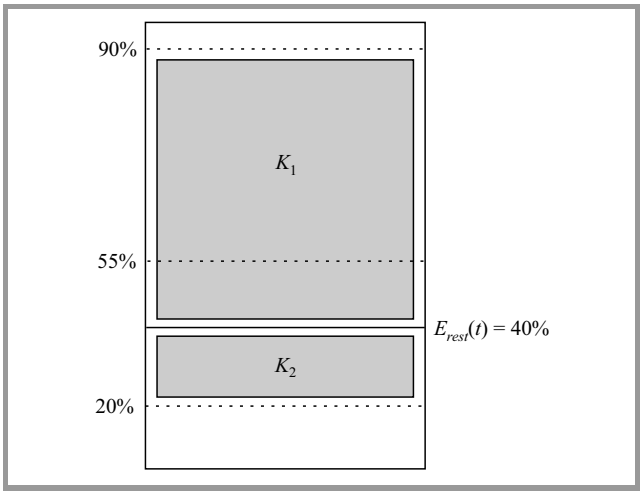


Fig. 2. Example of calculating K_1 and K_2.

Algorithm 6: Extended normal mode

$K_1 \leftarrow 0.9C_{\max} - E_{rest}(t)$
$K_2 \leftarrow E_{rest}(t) - 0.2C_{\max}$
$N_C \leftarrow round\left(\frac{K_1}{K_1+K_2}\right)N$
$N_D \leftarrow N - N_C$
for $i \in 1:N$ **do**
  $\text{Diff}_i \leftarrow \left\lceil \frac{d_i(t)}{\beta} \right\rceil \beta - d_i(t)$
end for
if $E_{rest}(t) \leq E_{fine}$ **then**
  indices $\leftarrow$ the indices $\{i\}$ of top N_C households that have largest Diff_i.
else
  indices $\leftarrow$ the indices $\{i\}$ of top N_C households that have smallest Diff_i.
end if
for $i \in 1:N$ **do**
  if $i \in$ indices **then**
    $e_i(t) \leftarrow \left\lceil \frac{d_i(t)}{\beta} \right\rceil \beta$
  else if $d_i(t) \bmod \beta \neq 0$ **then**
    $e_i(t) \leftarrow \left\lfloor \frac{d_i(t)}{\beta} \right\rfloor \beta$
  else
    $e_i(t) \leftarrow \left(\frac{d_i(t)}{\beta} - 1\right)\beta$
  end if
end for
Return $e_i(t)$

After determining the number of charging/discharging households, the core unit selects, which households should charge/discharge. This is because when the residual energy is less than $E_{rest}(t)$, the more energy should be charged in order to keep the normal mode. So, the households are selected, which will charge more energy to the battery by taking difference between the quantized demand value $e_i(t)$ and the demand load in household i. In order to calculate the amount of charged energy, the core unit checks the quantized demand value $e_i(t)$ when assuming all $s_i(t) = 1$. The core unit can expect each amount of charged energy by:

$$\text{Diff}_i = \left\lceil \frac{d_i(t)}{\beta} \right\rceil \beta - d_i(t), \tag{3}$$

where Diff_i is the amount of charged energy by a household i. When the residual battery is less than E_{fine}, N_C charging households which have larger Diff_i are chosen, because more energy should be charged to keep residual energy around E_{fine}. When the residual energy is more than E_{fine}, and vice versa. Algorithm 6 shows the algorithm of the extended normal mode.

Algorithm 7: Extended fairness mode

Input $l_p(t-1)$ and $l_q(t-1)$
$s_p(t) \leftarrow 0$
$s_q(t) \leftarrow 1$
$K_1 \leftarrow 0.9C_{\max} - E_{rest}(t)$
$K_2 \leftarrow E_{rest}(t) - 0.2C_{\max}$
$N_C \leftarrow round\left(\frac{K_1}{K_1+K_2}\right)(N-2)$
$N_D \leftarrow (N-2) - N_C$
for $i \in 1 : N-2$ **do**
  $\text{Diff}_i \leftarrow \left\lceil \frac{d_i(t)}{\beta} \right\rceil \beta - d_i(t)$
end for
if $E_{rest}(t) \leq E_{fine}$ **then**
  indices $\leftarrow$ the indices $\{i\}$ of top N_C households that have the largest Diff_i.
else
  indices $\leftarrow$ the indices $\{i\}$ of top N_C households that have the smallest Diff_i.
end if
for $i \in 1 : N-2$ **do**
  if $i \in indices$ **then**
    $e_i(t) \leftarrow \left\lceil \frac{d_i(t)}{\beta} \right\rceil \beta$
  else if $d_i(t) \bmod \beta \neq 0$ **then**
    $e_i(t) \leftarrow \left\lfloor \frac{d_i(t)}{\beta} \right\rfloor \beta$
  else
    $e_i(t) \leftarrow \left(\frac{d_i(t)}{\beta} - 1\right)\beta$
  end if
end for
Return $e_i(t)$

Extended fairness mode. Algorithm 7 shows the algorithm of the extended fairness mode. p and q denote the households that most charged and discharged during the period from 0 to $t-1$, respectively. Therefore, $l_p(t-1)$ and $l_q(t-1)$ denote the difference between charged and discharged amount of most charged household p and least charged household q during the period from 0 to $t-1$, respectively. In the extended fairness mode, the core unit first allots $s_i(t) \leftarrow 0$ to the most overcharged household, and $s_i(t) \leftarrow 1$ to the least charged household. The core unit then decides other $N-2$ households' $s_i(t)$ and $e_i(t)$ with the same way of the extended normal mode. When there are some households with loss more than l_{th}, the core unit chooses only two households, which have largest and smallest loss to reduce the complexity. The core unit checks each household's loss by interval measurements so that each loss is converged in l_{th}.

Extended stabilization mode. In the extended stabilization mode, the algorithm is almost same to the stabilization mode in two households except for changing β' to $\frac{\beta}{N}$.

3.3. Discussion

Other energy sources for BLH – although presented scheme assumes that only a battery is used for BLH, other sources such as a solar panel can also be used together with a battery. In this case, energy produced by other sources should also be taken into account for BLH. Authors do not consider the use of other energy sources in this research because the charged amount depends on the nature, which is typically difficult to model or estimate.

Initial cost to introduce BLH – a 1 kWh Li-ion battery costs at least $1,200 [16]. By using presented scheme and sharing one battery with more than two households, the installation cost for each household can be divided.

Limitation of our scheme – the monetary fairness between two households can be reduced by the fairness mode. However, proposed scheme cannot exactly get rid of monetary unfairness between multiple households even if the core unit sets l_{th} to 0. This is because the scheme solves the monetary unfairness after observing the previous outcome of BLH.

Privacy Concern – third parties cannot estimate both household's demand loads because they cannot obtain the residual energy on real time. However, when the system deals with two households, one household may estimate the other household's demand load in real time if each household knows its own demand load, metered load, and the residual energy on real time. Household 1 can calculate the household 2's load demand $d_2(t)$ as follows:

$$d_2(t) = e_2(t) + e_1(t) - d_1(t) - (E_{rest}(t) - E_{rest}(t-1)). \quad (4)$$

To satisfy the privacy of households using proposed scheme, both households must have cooperative relationships. This issue is not important when the number of households is more than two, because a household which tries to estimate other households' demand loads needs more demand load information from several households and this could be infeasible.

4. Simulation Results

4.1. Simulation Model

During simulation a mutual information and the monetary loss are evaluated. Based on the definition in [19], the "mutual information" of household i when $t = T$ is defined as the following equation with the set of output $E = \{e_i(t)\}$ and raw measurements $D = \{d_i(t)\}$:

$$I_i(E;D) = \sum_{e \in E} \sum_{d \in D} p(e,d) \log\left(\frac{p(e,d)}{p(e)p(d)}\right), \quad (5)$$

where $p(e,d)$ denotes a joint distribution of e and $d_i(t)$ and $p(e_i(t))$ and $p(d_i(t))$ are marginal distributions of $e_i(t)$ and $d_i(t)$, respectively. Intuitively, $I_i(E;D)$ represents how much information is shared between $e_i(t)$ and $d_i(t)$ for $1 \leq t \leq T$. Therefore, if good BLH is realized, the two

variables E and D are not correlated and thus $I_i(E;D)$ will take small value. On the contrary, if the BLH is not good, the two variables E and D share similar values and thus $I_i(E;D)$ will take large value. Mutual information between two variables $e_i(t)$ and $d_i(t)$ indicates how $e_i(t)$ and $d_i(t)$ are related. If $e_i(t)$ and $d_i(t)$ are totally independent, $e_i(t)$ does not give any information about $d_i(t)$, so their mutual information is zero [12]. The monetary loss indicates the absolute value of household's loss or gain at the end of simulation.

In used simulator, electric demands $d_i(t)$ of N households are extracted from the datasets and are input into the function of core unit that to output $e_i(t)$. After all $d_i(t)$ are processed, mutual information is calculated for each household with a package "infotheo" [20]. If not stated otherwise, the simulation parameters specified in Table 2 are used and a one-minute resolution dataset named Wiki-Energy [18] for evaluation. This dataset includes totally 722 houses' data collected in the USA from 2012 to 2014: 631 in Texas, 49 in Colorado, and 42 in California [21]. The detail of 722 households is as follows: 501 single-family homes, 183 apartments, 35 town homes and 3 mobile homes. Randomly sampled 100 households' electricity data measured for one month in April 2014 was used. For the evaluation of two households case, every combination of two households from randomly sampled 100 households in the dataset were used. By referring to [12], assume the maximum battery capacity $C_{\max}$ is 1.0 kWh and its charging and discharging rate β is 1.0 kW, which means that the battery can be fully exhausted or charged in an hour.

Table 2
Parameters used in simulation

Parameters	Definition
Dataset	Wiki-Energy [18]
Interval between measurements	1 minute
Simulation duration	30 days
Maximum battery capacity $C_{\max}$	1.0 kWh
Quantization width β	1.0 kW
Electric rate	16.341 cent/kWh
Threshold l_{th}	1, 5, 10, 25, and ∞ cent
Number of households N	2, 4, 8, 16, 32, and 64

The same flat electric rate 16.341 [cent/kWh] for all households was considered. This electric rate is cited from the one actually used in Pacific Gas and Electric Company [22]. In the two households case, the scheme was compared with SF-LS2 with the same battery capacity $C_{\max} = 1$ kWh and for l_{th} as $l_{th} = 1,\ 5,\ 10,\ 25,$ and ∞ [cent]. Furthermore, both mutual information and monetary loss for extended algorithm were also evaluated by varying the number of households N as $N = 2,\ 4,\ 8,\ 16,\ 32,$ and 64.

4.2. Comparison of Mutual Information

Mutual information for two households. Table 3 shows mutual information against both SF-LS2 and proposed scheme in two households case. There is no significant difference between SF-LS2 and the scheme irrespective of the chosen threshold l_{th}. However, there is the difference between the best case and the worst case in SF-LS2 and this scheme. This comes from the difference in total demand for one month. That is, the total demand load is 175 kWh in the best case, whereas 2097 kWh in the worst case. This follows the intuition that more information leaks when a household uses more appliances. Here, "information leaks" means that real demand load is leaked to the electric company. From this result, one can see that the larger power a household consumes, the more difficult to realize BLH due to the limitation of battery capacity.

Table 3
Mutual information of SF-LS2 and our scheme

Scheme		Mutual information		
		Average	Best	Worst
SF-LS2		0.0135	0.0018	0.0317
Authors' scheme	$l_{th} = 1$	0.0134	0.0014	0.0368
	$l_{th} = 5$	0.0128	0.0008	0.0325
	$l_{th} = 10$	0.0127	0.0008	0.0329
	$l_{th} = 25$	0.0127	0.0007	0.0330
	$l_{th} = \infty$	0.0132	0.0007	0.0409

Mutual information for multiple households. Figure 3 shows mutual information versus N when $l_{th} = 10$. The confidence intervals represent the standard deviation of all measurements. Since every combination was simulatedby choosing 2 out of 100 households, the number of measurements was $4{,}950 = \binom{100}{2}$ and the standard deviation was calculated from them. In Figure 3, "without BLH" indicates mutual information calculated without using BLH. One can see that as the number of households N increases, mutual information linearly increases. This is because as

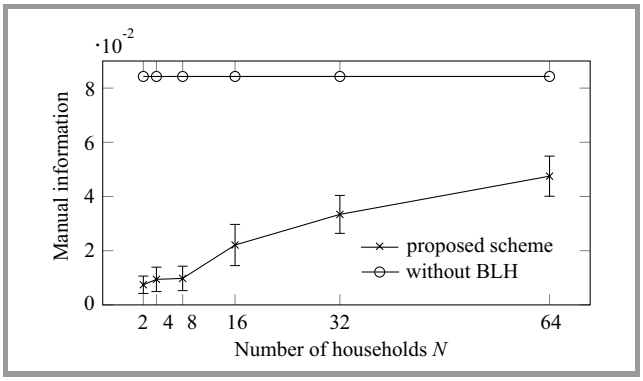


Fig. 3. Mutual information versus N.

N increases, the quantization width β' gets narrow, i.e. $\beta' = \frac{\beta}{N}$. However, the scheme still decreases mutual information by 44% even when $N = 64$. Therefore, our scheme is still effective against $N = 64$ with the battery capacity $C_{\max} = 1$ kWh.

4.3. Comparison of Monetary Loss

Monetary loss for two households. Table 4 shows the monetary loss caused by the scheme against l_{th}. In Table 4, average, best, and worst indicate the averaged, minimum, and maximum of the instantaneous loss for each l_{th} through the simulation, respectively. The average values of the monetary loss are calculated from every pair of household, i.e. 4,950 combinations, after BLH has been done. One can see that if we set $l_{th} = \infty$, which indicates the case where no

Table 4
Instantaneous loss versus l_{th}

l_{th}	Monetary loss [cent]		
	Average	Best	Worst
1	3.41	1.21	3.54
5	5.26	5.19	7.08
10	10.3	10.2	10.3
25	25.3	25.2	25.3
∞	$2.44 \cdot 10^3$	65.3	$6.78 \cdot 10^3$

Table 5
Details of processed modes when $l_{th} = 1$.

Pattern	Stabilization [%]	Fairness [%]	Normal [%]
Best	0	34.8	65.2
Worst	20.1	66.3	13.7

monetary fairness is considered, the average loss is nearly \$24.46. This situation cannot be tolerant in the real case. On the other hand, when l_{th} is set as a certain value, the loss can be controlled almost within l_{th}. However, when $l_{th} = 1$, the loss is 1.22 in the best case but 3.41 on average. This indicates that even if witht $l_{th} = 1$, the core unit cannot reduce the loss by nearly 1 in most cases. Table 5 shows the details of operated modes for the best case and the worst case. When the ratio of the stabilization mode is low or that of the normal mode is high, the loss results in a small value. On the other hand, when the ratio of the stabilization mode is high or that of the normal mode is low, the loss becomes large. This is caused by the similarity of demand loads between household 1 and 2. Figure 4 shows the time series of the loss for two households in the best and worst cases when $l_{th} = 1$, respectively. Figure 4a shows that their loss values are almost symmetry. On the other hand, from Fig. 4b, their losses are asymmetric in the worst case. From this result, when the system deals with two households, the combination of buddy households gives the difference of monetary loss.

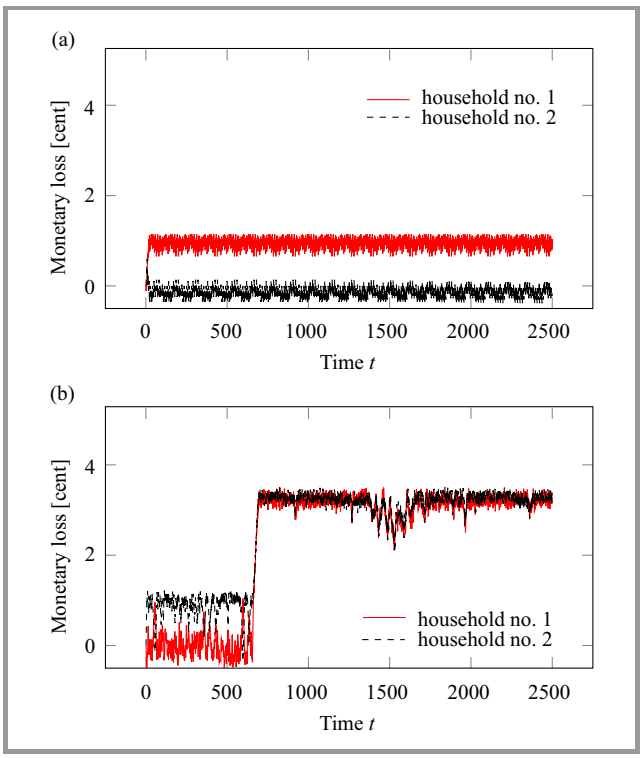


Fig. 4. Instantaneous loss versus time t ($l_{th} = 1$) for: (a) best case and (b) worst case.
(See color pictures online at www.nit.eu/publications/journal-jtit)

Table 6
Maximum instantaneous loss versus N when $l_{th} = 10$

N	Maximum loss [cent]		
	Average	Best	Worst
2	6.63	3.66	8.55
4	5.49	3.84	7.56
8	5.39	4.13	6.99
16	$1.21 \cdot 10^3$	33.2	$1.98 \cdot 10^3$
32	$2.02 \cdot 10^3$	$1.17 \cdot 10^3$	$2.42 \cdot 10^3$
64	$1.41 \cdot 10^3$	$1.08 \cdot 10^3$	$1.71 \cdot 10^3$

Monetary loss for multiple households. Table 6 shows the maximum loss caused by the scheme versus N when $l_{th} = 10$. From Table 6, one can see that the scheme maintains the monetary loss within the threshold $l_{th} = 10$ when the system deals with less than or equal to 8 households. However, when the number of households is more than 8, monetary loss suddenly exceeds $l_{th} = 10$. In order to clarify this reason, authors investigate operated modes for BLH. Table 7 shows the details of operated modes versus the number of households N when $l_{th} = 10$. As the number of households N is more than or equal to 16, the core unit

Table 7
Details of processed modes when l_{th} = 10

N	Stabilization [%]	Fairness [%]	Normal [%]
2	0	10.1	88.7
4	0.88	18.7	80.3
8	5.6	35.2	59.1
16	42.7	46.5	10.7
32	62.5	36.4	0.99
64	78.0	20.0	0.57

more frequently chooses the stabilization mode. From this result, when using a 1 kWh battery for more than 8 households, the core unit does not have much room to consider monetary fairness.

5. Conclusion

The paper presents a monetary fair BLH scheme for multiple households with one battery. Authors show BLH algorithm for more than two households. The proposed BLH scheme consists of three modes: the stabilization, fairness, and normal mode and the core unit changes its mode based on the residual energy and the amount of loss caused by BLH. By the computer simulation with a real electric load dataset, in two households case, authors show that when l_{th} is set to 1 cent, the scheme can achieve almost the same information leakage with SF-LS2 as well as control monetary loss less than five cents in the US currency. In the multiple households case, the paper shows that the mutual information linearly increases with the number of households. With a 1 kWh battery for BLH, the scheme can execute BLH for 8 households with preserving monetary fairness.

Acknowledgements

This work is partly supported by the Grant in Aid for Scientific Research (No. 26420369) from Ministry of Education, Sport, Science and Technology, Japan.

References

[1] E. L. Quinn, "Smart metering and privacy: existng law and competing policies", Report for the Colorado Public Utilities Commission, University of Colorado Law School, Boulder, Colorado, 2009.

[2] G. W. Hart, "Residential energy monitoring and computerized surveillance via utility power flows", *IEEE Technol. Soc. Mag.*, vol. 8, no. 2, pp. 12–16, 1989.

[3] M. A. Lisovich, D. K. Mulligan, and S. B. Wicker, "Inferring personal information from demand response systems", *IEEE Secur. & Priv. Mag.*, vol. 8, no. 1, pp. 11–20, 2010.

[4] A. Molina-Markham, P. Shenoy, K. Fu, E. Cecchet, and D. Irwin, "Private memoirs of a smart meter", in *ACM Worksh. Embedded Sensing Syst. for Energy-Efficien. in Build. BuildSys 2010*, Zurich, Switzerland, pp. 61–66.

[5] G. Hart, "Nonintrusive appliance load monitoring", *Proc. of the IEEE*, vol. 80, no. 12, pp. 1870–1891, 1992.

[6] M. Marceau and R. Zmeureanu, "Nonintrusive load disaggregation computer program to estimate the energy consumption of major end uses in residential buildings", *Energy Convers. & Manag.*, vol. 41, no. 13, pp. 1389–1403, 2000.

[7] C. Laughman, R. Cox, S. Shaw, S. Leeb, L. Norford, and P. Armstrong, "Power signature analysis", *IEEE Power & Energy Mag.*, vol. 1, no. 2, pp. 56–63, 2003.

[8] N. Batra, J. Kelly, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, and M. Srivastava, "NILMTK: an open source toolkit for non-intrusive load monitoring, in Proc. *50th Int. Conf. Future Energy Syst. ACM e-Energy 2014*, Cambridge, UK, 2014, pp. 265–276.

[9] S. McLaughlin, D. Podkuiko, S. Miadzvezhanka, A. Delozier, and P. McDaniel, "Multi-vendor penetration testing in the advanced metering infrastructure", in *Proc. 26th Ann. Comp. Secur. Appl. Con. Austin ACSAC'10*, Texas, USA, 2010, pp. 107–116.

[10] G. Kalogridis, C. Efthymiou, S. Z. Denic, T. A. Lewis, and R. Cepeda, "Privacy for smart meters: towards undetectable appliance load signatures", in *Proc. 1st IEEE Int. Conf. Smart Grid Commun. SmartGridComm 2010*, Gaithersburg, Maryland, USA, 2010, pp. 232–237.

[11] S. McLaughlin, P. McDaniel, and W. Aiello, "Protecting consumer privacy from electric load monitoring", in *Proc. 18th ACM Conf. Com. Commun. Secur. CCS 2011*, Chicago, IL, USA, 2011, pp. 87–98.

[12] W. Yang, N. Li, Y. Qi, W. Qardaji, S. McLaughlin, and P. McDaniel, "Minimizing private data disclosures in the smart grid", in *Proc. 19th ACM Conf. Com. Commun. Secur. CCS 2012*, Raleigh, NC, USA, 2012, pp. 415–427.

[13] J. Gomez-Vilardebo and D. Gündüz, "Smart meter privacy for multiple users in the presence of an alternative energy source", in *IEEE Trans. on Inform. Forensics & Secur.*, vol. 10, pp. 132–141, 2014.

[14] L. Yang, X. Chen, J. Zhang, and H. Poor, "Optimal privacy-preserving energy management for smart meters", in Proc. IEEE Conf. Comp. Commun. IEEE INFOCOM 2014, Toronto, Ontario, Canada, 2014, pp. 513–521.

[15] L. Alejandro *et al.*, "Global market for smart electricity meters: Government policies driving strong growth", Working Paper, US International Trade Commission, 2014 [Online]. Available: https://www.usitc.gov/publications/ 332/ id-037smart_meters_?nal.pdf

[16] F. Geth, D. Six, J. Tant, T. De Rybel, Peter Tant, and J. Driesen, "Techno-economical and life expectancy modeling of battery energy storage systems", in *Proc. 21st Int. Conf. and Exhibition on Electricity Distrib. CIRED 2011*, Frankfurt, Germany 2011, pp. 1–4.

[17] Ryoju estate develops "ene-self" power supply system for apartment bldgs., adopting PV modules, emergency generator and lithium-ion battery, Mitsubishi Heavy Industries, Ltd., Mar. 2013 [Online]. Available: https://www.mhi.co.jp/en/m/news/story/1303071630.html (accessed: 01.10.2016).

[18] Pecan Street Dataport – a universe of data, available around the world [Online]. Available: https://dataport.pecanstreet.org/

[19] T. M. Cover and J. A. Thomas, *Elements of information theory*. Wiley, 2012.

[20] P. E. Meyer, "Infotheo: Information-theoretic measures", R package version 1.2.0, 2014 [Online]. Available: https://cran.r-project.org/ web/packages/infotheo/index.html

[21] O. Parson, G. Fisher, A. Hersey, N. Batra, J. Kelly, A. Singh, W. Knottenbelt, and A. Rogers, "Dataport and NILMTK: A building data set designed for non-intrusive load monitoring", in *Proc. 1st Int. Symp. on Sig. Process. Appl. in Smart Build. at 3rd IEEE Global Conf. Sig. & Inform. Process. GlobalSIP 2015*, Orlando, USA, 2015.

[22] "Pacific Gas and Electric Company: electric schedule A-1" [Online]. Available: http://www.pge.com/

Ryota Negishi received his B.S. degree from Keio University in 2014. He is a Master student at Keio University. His research interest is security and privacy for smart grid and IoT. He is a member of IEEE.

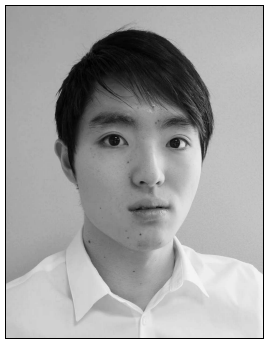
E-mail: negishi@sasase.ics.keio.ac.jp
Department of Information and Computer Science
Keio University
3-14-1 Hiyoshi, Kohoku, Yokohama
Kanagawa 223-8522, Japan

Shuichiro Haruta received his B.Sc. degree from Keio University in 2015. He is a Master student at Keio University. His research interest is security and privacy for social networking systems and IoT. He is a member of IEICE and IEEE.

E-mail: haruta@sasase.ics.keio.ac.jp
Department of Information and Computer Science
Keio University
3-14-1 Hiyoshi, Kohoku, Yokohama
Kanagawa 223-8522, Japan

Chihiro Inamura received his B.Sc. and M.Sc. degrees from Keio University in 2013 and 2015, respectively. His research interest includes privacy issues in smart grid system.

E-mail: inamura@sasase.ics.keio.ac.jp
Department of Information and Computer Science
Keio University
3-14-1 Hiyoshi, Kohoku, Yokohama
Kanagawa 223-8522, Japan



Kentaroh Toyoda received his M.E. degree from Keio University in 2013. He is a Ph.D. student at Keio University. His research interest is security and privacy for RFID, IoT, and Cyber Physical Systems. He was a research associate at Keio University from 2013 to 2015. He received a Telecom System Technology Encouragement Award in 2015 and IEICE communication society encouragement awards in 2012 and 2015. He is a member of IEEE, IEICE, and IPSJ.

E-mail: toyoda@sasase.ics.keio.ac.jp
Department of Information and Computer Science
Keio University
3-14-1 Hiyoshi, Kohoku, Yokohama
Kanagawa 223-8522, Japan

Iwao Sasase received the B.E., M.E., and D.Eng. degrees in Electrical Engineering from Keio University, Yokohama, Japan, in 1979, 1981 and 1984, respectively. From 1984 to 1986, he was a Post Doctoral Fellow and Lecturer of Electrical Engineering at the University of Ottawa, ON, Canada. He is currently a Professor of Information and Computer Science at Keio University, Yokohama, Japan. His research interests include modulation and coding, broadband mobile and wireless communications, optical communications, communication networks and information theory. He has authored more than 280 journal papers and 415 international conference papers. He granted 44 Ph.D. degrees to his students in above field. He serves as Vice President of IEICE in 2014-2016. He is Fellow of IEICE, and Senior Member of IEEE, Member of the Information Processing Society of Japan.

E-mail: sasase@ics.keio.ac.jp
Department of Information and Computer Science
Keio University
3-14-1 Hiyoshi, Kohoku, Yokohama
Kanagawa 223-8522, Japan

Information for Authors

Journal of Telecommunications and Information Technology **(JTIT)** is published quarterly. It comprises original contributions, dealing with a wide range of topics related to telecommunications and information technology. **All papers are subject to peer review**. Topics presented in the JTIT report primary and/or experimental research results, which advance the base of scientific and technological knowledge about telecommunications and information technology.

JTIT is dedicated to publishing research results which advance the level of current research or add to the understanding of problems related to modulation and signal design, wireless communications, optical communications and photonic systems, voice communications devices, image and signal processing, transmission systems, network architecture, coding and communication theory, as well as information technology.

Suitable research-related papers should hold the potential to advance the technological base of telecommunications and information technology. Tutorial and review papers are published only by invitation.

Manuscript. TEX and LATEX are preferable, standard Microsoft Word format (.doc) is acceptable. The author's JTIT LATEX style file is available:
http://www.nit.eu/for-authors

Papers published should contain up to 10 printed pages in LATEX author's style (Word processor one printed page corresponds approximately to 6000 characters).

The manuscript should include an abstract about 150–200 words long and the relevant keywords. The abstract should contain statement of the problem, assumptions and methodology, results and conclusion or discussion on the importance of the results. Abstracts must not include mathematical expressions or bibliographic references.

Keywords should not repeat the title of the manuscript. About four keywords or phrases in alphabetical order should be used, separated by commas.

The original files accompanied with pdf file should be submitted by e-mail: redakcja@itl.waw.pl

Figures, tables and photographs. Original figures should be submitted. Drawings in Corel Draw and PostScript formats are preferred. Figure captions should be placed below the figures and can not be included as a part of the figure. Each figure should be submitted as a separated graphic file, in .cdr, .eps, .ps, .png or .tif format. Tables and figures should be numbered consecutively with Arabic numerals.

Each photograph with minimum 300 dpi resolution should be delivered in electronic formats (TIFF, JPG or PNG) as a separated file.

References. All references should be marked in the text by Arabic numerals in square brackets and listed at the end of the paper in order of their appearance in the text, including exclusively publications cited inside. Samples of correct formats for various types of references are presented below:

[1] Y. Namihira, "Relationship between nonlinear effective area and mode field diameter for dispersion shifted fibres", *Electron. Lett*., vol. 30, no. 3, pp. 262–264, 1994.

[2] C. Kittel, *Introduction to Solid State Physics*. New York: Wiley, 1986.

[3] S. Demri and E. Orłowska, "Informational representability: Abstract models versus concrete models", in *Fuzzy Sets, Logics and Knowledge-Based Reasoning*, D. Dubois and H. Prade, Eds. Dordrecht: Kluwer, 1999, pp. 301–314.

Biographies and photographs of authors. A brief professional author's biography of up to 200 words and a photo of each author should be included with the manuscript.

Galley proofs. Authors should return proofs as a list of corrections as soon as possible. In other cases, the article will be proof-read against manuscript by the editor and printed without the author's corrections. Remarks to the errata should be provided within one week after receiving the offprint.

A copy of the JTIT is provided to each author of paper published.

Journal of Telecommunications and Information Technology has entered into an electronic licencing relationship with EBSCO Publishing, the world's most prolific aggregator of full text journals, magazines and other sources. The text of *Journal of Telecommunications and Information Technology* can be found on EBSCO Publishing's databases. For more information on EBSCO Publishing, please visit www.epnet.com.

(*Contents Continued from Front Cover*)

INSTYTUT ŁĄCZNOŚCI
PAŃSTWOWY INSTYTUT BADAWCZY

Editorial Office

National Institute
of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

tel. +48 22 512 81 83
fax: +48 22 512 84 00
e-mail: redakcja@itl.waw.pl
http://www.nit.eu

JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY 1/2016