JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

3/2011

(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

This issue of *Journal of Telecommunications and Information Technology* includes selected papers devoted to research in the ICT area carried out within the frame of Polish National Project PBZ MNiSW-02-11/2007: *Next Generation Services and Networks – technical, application and market aspects*.

This project was coordinated by the National Institute of Telecommunications and implemented in cooperation with 8 scientific centers: Warsaw University of Technology, Gdańsk University of Technology, Poznań University of Technology, Wrocław University of Technology, AGH University of Science and Technology (Cracow), Research and Academic Computer Network (NASK) and Military Institute of Telecommunications.

It covered diversified spectrum of research in 10 areas: **network architectures and protocols**, **wireless and mobile systems and their security**, **network development planning**, **traffic management – IT QoS**, **digital radio broadcasting networks**, **methods and tools for their design and trials**, **electromagnetic compatibility**, **measurements and monitoring**, **systems aiding regulatory decisions – knowledge mining**, **multimedia services and models for trading network transport resources**.

Examples of work include identification and analysis of evolution directions of NGN architectures and protocols, their design and development methods, traffic management mechanisms, creation of new algorithms, technologies and tools for implementing broadly defined multimedia services, development of technical means supporting introduction of products and services related to wireless/mobile systems and their security, design methods optimizing coverage of single digital (DRM) transmitters and broadcast networks, transmission properties and interference susceptibility of advanced wireless systems and networks in real electromagnetic environment, monitoring and diagnostics of time signals and development of group time standard, creation of innovative mechanisms for trading transport resources of telecommunication networks, in particular on auctions and exchanges to improve efficiency of resource utilization and competition for network resources, and development of tools supporting decision-making for regulation of telecom services markets.

The first five papers are related with **traffic management – IT QoS**. The first one tilted *The IP QoS System* written by Wojciech Burakowski, Jarosław Śliwiński, Halina Tarasiuk, Andrzej Bęben, Ewa Szynkiewicz, Piotr Pyda, and Jordi Mongay Batalla describes the IP QoS system that support a number of, so called, classes of services in the Internet. It is assumed that a user/an application requests from the network a specified service corresponding to

quality of packet transfer, and for doing it, the network allocates an adequate amount of resources if it has. In order to achieve this the network functionalities should be extended comparing to the best effort Internet. These new functionalities are related to signaling, resource provisioning, QoS mechanisms at packet, connection and dimensioning levels. The presented IP QoS system is based on the next generation networks (NGN) and differentiated services (DiffServ) architectures. In the next paper, *Performance Evaluation of Signalling in the IP QoS System* written by Halina Tarasiuk, Jarosław Śliwiński, Piotr Arabas, Przemysław Jaskóła, and Witold Góralski the trial results of the proposed signaling system of the IP QoS system based on NGN and DiffServ architectures, which allows sending a request from a user to the system for establishing new connection with predefined quality of service assurance, are presented. The experiments were performed to measure setup delay utilizing artificial call generator/analyzer. The different distributions of interarrival and call holding times based on the literature were assumed. The results show that the setup delay strongly depends on access time to network devices, however also on the assumed call.

The paper *On Dimensioning and Routing in the IP QoS System* by Witold Góralski, Piotr Pyda, Tomasz Dalecki, Jordi Mongay Batalla, Jarosław Śliwiński, and Waldemar Latoszek presents dimensioning and routing solutions in designed IP QoS system. The functional architecture as well as the description of the functions and methods implemented in the system are discussed.

The quality evaluation of the telecommunication services: VoIP (representing the RT interactive class) and VoD (representing the MM streaming class) is analyzed in the next paper titled *QoS Conditions for VoIP and VoD* by Przemysław Dymarski, Sławomir Kula, and Thanh Nguyen Huy. The objective methods and tools for perceived quality measurement are compared.

The platform for research on auction mechanisms being a distributed simulation framework providing means to carry out research on resource allocation efficiency mechanisms and user strategies is discussed in the paper *A Software Platform for Research on Auction Mechanisms* written by Mariusz Kamola, Ewa Niewiadomska-Szynkiewicz, Krzysztof Malinowski, Wojciech Stańczuk, and Piotr Pałka. Both kinds of algorithms examined are completely user-defined. In the presented approach interaction of algorithms is recorded and pre-defined measures for the final resource allocation are calculated. Moreover, underlying database design provides for efficient results lookup and comparison across different experiments, thus enabling research groupwork. A recognized, open and flexible information model is employed for experiment descriptions.

The next three papers deal with **network architectures and protocols**. The first one titled *The Realization of NGN Architecture for ASON/GMPLS Network* and written by Sylwester Kaczmarek, Magdalena Młynarczuk, Marcin Narloch, and Maciej Sac concerns ASON/GMPLS optical network as a NGN transport layer. In particular, the ASON/GMPLS architecture and its relation to the proposed ITU-T NGN architecture are described. The concept, functional structure and communication among architecture elements as well as the implementation of laboratory testbed are presented. The results of functional tests confirming proper software and testbed operation are stated. In the next one, *Multi Queue Approach for Network Services Implemented for Multi Core CPUs*, the usage of general purpose CPU providing network core functionality is analyzed by Marcin Hasse, Krzysztof Nowicki, and Józef Woźniak. For this purpose parameterized system model has been created, which represents general core networking needs. The problem of the testing in telecommunications and software engineering, is discussed in the next paper titled *Active – Passive: On Preconceptions of Testing* written by Krzysztof M. Brzeziński.

In the next paper related with wireless and mobile systems and their security, *Optimization of Call Admission Control for UTRAN*, Michał Wągrowski, Wiesław Ludwin paper addresses the traffic's grade of service indicators: call blocking and dropping rates as well as the optimization of their mutual relation, corresponding to the call admission control procedure configuration. The results of simulations presented opportunities for the CAC load threshold adaptation according to the traffic volume and user mobility changes observed in the mobile radio network.

The following five papers address some selected topics of multimedia services. In the first one, *Network-on-Multi-Chip (NoMC) with Monitoring and Debugging Support* written by Adam Łuczak, Marta Stępniewska, Jakub Siast, Marek Domanski, Olgierd Stankiewicz,

Maciej Kurc, and Jacek Konieczny, recent research on network-on-multi-chip are summarized. The proposed network architecture supports hierarchical addressing and multicast transition mode. Such an approach provides new debugging functionality hardly attainable in classical hardware testing methodology. In the next paper, *The Design of an Objective Metric and Construction of a Prototype System for Monitoring Perceived Quality (QoE) of Video Sequences*, Lucjan Janowski, Mikołaj Leszczuk, Zdzisław Papir, and Piotr Romaniak present different no reference (NR) objective metrics addressing the most important artefacts for raw (source) video sequences (noise, blur, exposure) and those introduced by compression (blocking, flickering) which can be used for assessing quality of experience. The validity of all metrics was verified under subjective tests. In the next paper, *Communication Platform for Evaluation of Transmitted Speech Quality* written by Andrzej Ciarkowski and Andrzej Czyżewski, a voice communication system designed and implemented is described. The purpose of the presented platform was to enable a series of experiments related to the quality assessment of algorithms used in the coding and transmitting of speech. The system is equipped with tools for recording signals at each stage of processing, making it possible to subject them to subjective assessments by listening tests or, objective evaluation employing PESQ or PSQM algorithms. The framework for testing video streaming techniques is presented in *Video Streaming Framework* by Andrzej Buchowicz, and Grzegorz Galiński. Short review of error resilience and concealments tools available for the H.264/AVC standard is given. The video streaming protocols and the H.264 payload format as well as experimental results are described.

The next paper *The Learning System by the Least Squares Support Vector Machine Method and its Application in Medicine* written by Paweł Szewczyk and Mikołaj Baszun, presents the possibility of using the Least Squares Support Vector Machine to the initial diagnosis of patients is presented. In order to find some optimal parameters making the work of the algorithm more detailed, the following techniques have been used: K-fold Cross Validation, Grid-Search, Particle Swarm Optimization. The result of the classification has been checked by some labels assigned by an expert.

Michał Karpowicz in his paper *Designing Auctions: A Historical Perspective* presents selected results carried out within the framework of **models for trading network transport resources**. In particular, some aspects of auction design is discussed.

Cezary Chudzian, Janusz Granat, Edward Klimasara, Jarosław Sobieszek, and Andrzej P. Wierzbicki in the paper related with **systems aiding regulatory decisions – knowledge mining** and titled *Personalized Knowledge Mining in Large Text Sets* discuss the concept of knowledge engineering, in particular ontological engineering. They present assumptions accepted as a basis for a group research on a radically personalized system of ontological knowledge mining, relying on the perspective of human centered computing and combining ontological concepts of the user with an ontology resulting from an automatic classification of a given set of textual data. Moreover, a pilot system PrOnto that supports research work in two aspects: searching for information interesting for a user according to her/his personalized ontological profile, and supporting research cooperation in a group of users (Virtual Research Community) according, e.g., to a comparison of such personalized ontological profiles is presented.

The next paper tiltled *New SEAMCAT Propagation Models: Irregular Terrain Model and ITU-R P. 1546-4* concerns **electromagnetic compatibility**. The authors, Dariusz Więcek and Dariusz Wypiór present in it implementation of the ITU-R P.1546-4 and ITM propagation models for SEAMCAT prepared and developed in the National Institute of Telecommunications, Poland. Results of their research encompasses methodology, implementation and verification of plug-ins into the SEAMCAT software.

Finally, in the next paper related with **digital radio broadcasting networks**, **methods and tools for their design and trials** and titled *Technical Aspects Outline for the Strategy of Launching Digital Broadcasting in Poland on Wave Bands Below 30 MHz* Andrzej Dusiński and Jacek Jarkowski discus the state of art knowledge concerning the introduction of DRM in the world and prospects for its further development. It presents the possibility of introducing this system in Poland.

Paweł Szczepański
Editor-in Chief

Paper

The IP QoS System

Wojciech Burakowski[a], Jarosław Śliwiński[a], Halina Tarasiuk[a], Andrzej Bęben[a], Ewa Niewiadomska-Szynkiewicz[b,c], Piotr Pyda[d], and Jordi Mongay Batalla[a]

[a] *Institute of Telecommunications, Warsaw University of Technology, Warsaw, Poland*
[b] *Institute of Control and Computation Engineering, Warsaw University of Technology, Warsaw, Poland*
[c] *Research and Academic Computer Network (NASK), Warsaw, Poland*
[d] *Military Communication Institute, Zegrze, Poland*

Abstract—This paper shortly describes the IP QoS System which offers strict quality of service (QoS) guarantees in IP-based networks and supports a number of, so called, classes of services. Such solution requires to implement in the network a set of QoS mechanisms and algorithm working on packet, connection request and provisioning levels. Furthermore, we require signaling system for informing the network about new connection request and network resource allocation capabilities for providing required resources to given connection. The IP QoS System is based on the next generation networks (NGN) and differentiated services (DiffServ) architectures and, at least for now, it is designed for single domain only.

Keywords—classes of service, DiffServ, multi-service networks, NGN, quality of service.

1. Introduction

The current Internet is working under TCP/IP protocol stack and is based on two main fundamentals, which are: the network offers only one class of service named best effort service, and the network resources are overprovisioned as possible in order to minimize packet losses and packet delays. As a consequence, the Internet providers aimed at providing to the users as fast as possible packet transfer but they are far from guaranteeing, so called, strict quality of service (QoS) that is measured by the maximum allowed values of such parameters as IP packet transfer delay (IPTD), IP packet transfer delay variation (IPDV) and IP packet loss ratio (IPLR).

On the other hand, the network capabilities of packet transfer determine the range of applications the users may use with appropriate satisfaction. The lack of guaranteeing strict QoS for packet transfer constitutes the main barrier in introducing, e.g., streaming applications as video on demand (VoD), voice over IP (VoIP), video teleconference (VTC) or e-health teleconsultations. In addition, the network operators may get additional profit if they are able to offer strict QoS instead best effort connections. Concluding, the QoS in the Internet is strongly required for its further evolution.

The recognized approach for guaranteeing strict QoS in IP-based network is the DiffServ architecture [1], [2], [3], [4]. The activities corresponding to this architecture started about 10 years ago and some prototypes were developed, e.g., by European projects. A good example is the AQUILA project [5], [6], [7], [8], which prototyped and tested the system based on DiffServ architecture. The IP QoS System that was recently prototyped and tested in Poland follows the solution from AQUILA project and enhanced it by using the elements from next generation network (NGN) architecture.

The attractiveness of the DiffServ architecture is mainly caused by:

- it allows to provide a number of classes of service differing in handling of traffic profiles as well as in QoS guarantees,
- each classes of service is designed for handling traffic generated by some types of applications,
- per flow handling is necessary only in the border routers while the core routers see only aggregated flows,
- it is a good example of scalable architecture.

In fact, the DiffServ architecture was designed for a single domain but we can observe the activities for extending this architecture for the whole network, as e.g. in EuQoS project [9], [10], [11].

The organization of the paper is the following. In Section 2 we describe the mechanism we need to introduce in the network in order to guarantee strict QoS for packet transfer. The IP QoS System is presented in Section 3. Section 4 concludes the paper.

2. Mechanisms and Algorithms Required to Guarantee Strict QoS in the Network

In order to guarantee a quality for transfer of packets emitted by an application to the network, we need to apply a set of mechanisms, named QoS mechanisms, and algorithms that operate at different levels in the network. These mechanisms and algorithms we can classify to the following categories:

- for handling packets in the routers (time scale – miliseconds),

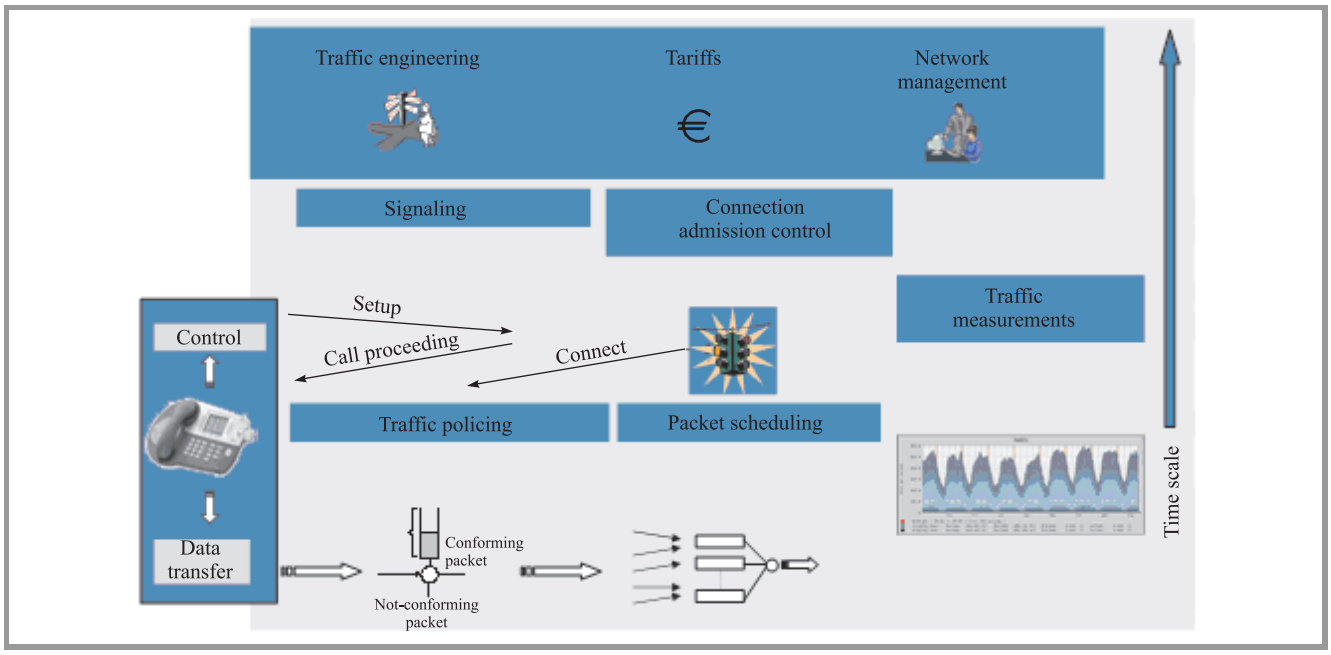


Fig. 1. Required mechanisms and algorithms in the network for providing QoS.

– for establishing/releasing the connections (time scale – seconds or minutes),

– for network dimensioning (time scale – hours or days).

These new set of mechanisms are shown in Fig. 1.
In this section, we briefly describe each of the above group of mechanisms and algorithms.

2.1. QoS Mechanisms at the Packet Level

In order to guarantee strict QoS for a given packet stream, we need to assure its adequate handling in routers. A set of available QoS mechanisms at the packet level is named as per hop behavior (PHB) mechanisms. This set contains such mechanisms as:

– classifier for distinguishing between packets belonging to different classes of service and for sending packets to appropriate path of handling,

– policer for monitoring contracted traffic profile,

– optionally, marker for indicating not conforming packets (they may be discard or send if allowed link capacity, depending on applied algorithm),

– scheduler for managing access to the link when more than one packet in the queues,

– shaper for shaping traffic, if it is needed.

Thanks to the above mechanisms, we may send a packet before the another ones even if this packet arrived later to the system.

2.2. QoS Mechanisms at the Call Level

When a new connection request is sent to the network, first of all we need is to check if we have enough spare network resources for establishing new connection with assuring adequate QoS. The new request is submitted to a given class of service, for which we have earlier, during provisioning phase, allocated an amount of resources. The resources dedicated to a class of service are the buffer size and the link capacity in each output link in the routers. So, we check the availability of spare resources using connection admission control (CAC) function. In general, CAC is a function of such parameters as number of running connections, already accepted volume of traffic (declared or measured), network resources allocated to the class of service and the traffic declarations of new request.
It is worth to mention that performing CAC function is the fundamental for assuring strict QoS. It allows us to control volume of traffic in the network and to avoid network overloading. Unfortunately, this means that some of new requests may be rejected. In addition, for performing CAC we require to implement a signaling system in the network.

2.3. Resource Provisioning (Traffic Engineering)

In a classical approach, before performing CAC function we need to allocate network resources (link capacities, buffers) for all supported classes of service. In addition, we need to specify the nodes in the network, in which we perform the CAC. It would be not practical case to perform CAC in all routers on the path between source and destination since in the case of Internet we have too many connections running in parallel and, as a consequence, the signaling traffic is too high. So, the reasonable solution is to select the routers when the CAC is performed (the best is minimize

the number of these routers) and to overprovision the rest of the network.

3. IP QoS System

In this section we present some details about the IP QoS System that we have recently prototyped and tested in Poland.

The IP QoS System is a proposal for assuring strict QoS guarantees in a single domain network. It follows the DiffServ and NGN architectures [12], [13] and [14]. The architecture of the system is depicted in Fig. 2. It assumes

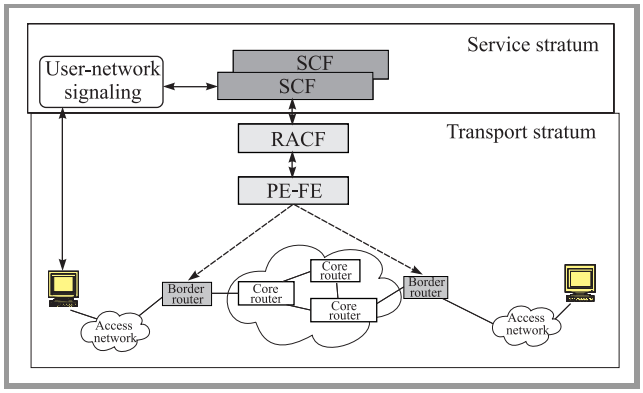


Fig. 2. The architecture of the IP QoS System.

two meta-layers that are: service stratum responsible for service management, and transport stratum responsible for packet transfer in the network. The functions performed by service stratum are called as service control functions (SCF) while the functions performed by transport stratum are resource and admission control function (RACF) as well as policy enforcement functional entity (PE-FE) for setting PHB mechanisms in the border routers.

Figure 3 shows the scenario for establishing connection in the IP QoS System. For establishing the connection, the user/the application sends its request to the network (message "1"). This request is handled by the application server. Next, this request is further send to the server responsible of resource management (message "2"), which checks if the required resources are available. When we have sufficient volume of resources, then it sends the messages to the border routers (messages "3" and "4") for the purpose of tuning

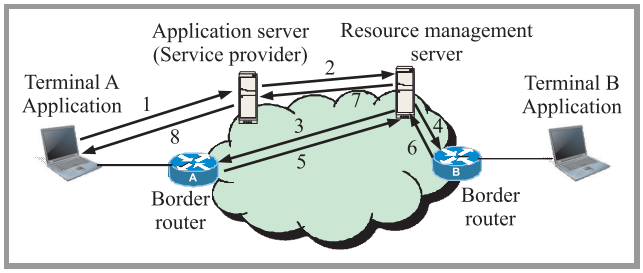


Fig. 3. Scenario for establishing connection in the IP QoS System.

the PHB mechanisms (classifier, policer). After the positive answers from the border router are received (messages "5" and "6"), then the resource management server sends the acknowledge message (message "7") to the service server. Next, this server sends the information to the users/the application of setting the required connection (message "8"). It is essential for the DiffServ architecture that the per-flow operations are performed in the border routers only while in the core routers the operations are performed per aggregated flows.

The details about the control access to the network resources one can find in [14].

3.1. Traffic Management in the IP QoS System

In order to guarantee strict QoS we establish a number of specialized classes of service [8], [15] in the IP QoS System. The term "class of service" expresses the network capabilities to transfer traffic according to a priori specified conditions with respect to maximum allowed values of parameters IPTD, IPDV and IPLR. The IP QoS System supports the classes of service in a single domain network between each pair of the border routers. A given type of application submits its packet stream to a predefined class of service. The classes of service are regarded as globally well known. Since in the IP QoS System the classes of service are supported only in a single domain, we define them in the context of the "end to end" classes of service as specified for multi-domain network and described in [4], [15]. In particular, in the area of a single domain, in one class of service we can merge a number of "end to end" classes of service with similar QoS guarantees. Table 1 shows the list of the classes of service implemented in the IP QoS System with its characteristics of QoS guarantees that are expressed by the maximum allowed values for IPTD, IPDV and IPLR.

Let us recall that in order to establish a given class of service in the network we need:

- to set the values of parameters of the PHB mechanisms that is necessary for assuring adequate handling of submitted traffic and isolation between traffic belonging to different classes of service,
- to allocate an amount of network resources for this class,
- to apply adequate CAC algorithm to control volume of submitted traffic.

In the IP QoS System we perform CAC function only in the ingress border routers while the core network is overprovisioned as it is shown in Fig. 4. It means that the packet delays and the packet losses in the core should be significantly less comparing to the packet delays and losses in the ingress border routers. The above is true only when traffic carried by the network is closed to this allowed by the CAC function. If submitted traffic is rather low then the packets crossing the ingress border routers also experience low

Table 1
Mapping between types of applications jointly with "end to end" classes of service and classes of service in the IP QoS System, QoS guarantees and traffic profiles

Type of application	Classes of service "end to end"	Classes of service in the IP QoS System	QoS requirements			Traffic profile
			IPLR	IPTD (mean value)	IPDV	
VoIP	Telephony	Real time (RT)	10^{-3}	100 ms	50 ms	(PBR, PBRT)*
Interactive games	RT interactive					
Video on demand	MM streaming	MM streaming	10^{-3}	1 s (not critical)	Not critical	(PBR, PBRT)
File transfer protocol (FTP)	High throughput data	High throughput data (HTD)	10^{-3}	1 s (not critical)	Not critical	(PBR, PBRT)
	Standard	Standard (STD)	Not critical	Not critical	Not critical	Arbitrary

* peak bit rate (PBR), peak bit rate tolerance (PBRT), parameters of the token bucket mechanism.

losses and delays comparing with assumed QoS guarantees. Notice that the assumption about core overprovisioning is not critical since in the core we do not perform CAC and, what is also important, usually the link capacities of the core links are rather higher comparing to the link capacities in the access. Furthermore, such overprovisioning we do not have to do for standard class of service.

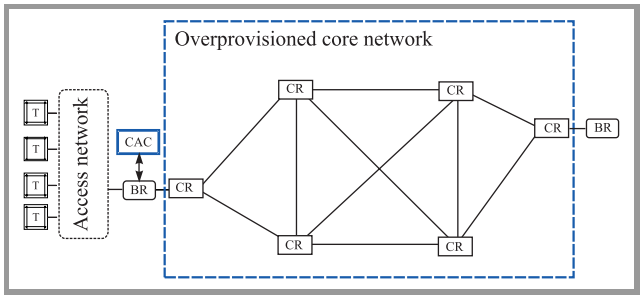


Fig. 4. Traffic management in the IP QoS System: BR – border router, CR – core router, T – terminal CAC – function responsible for admitting/rejection of new connection request.

Now, we explain the rules we have assumed for the network dimensioning. For the sake of simplicity, let we take into account network when traffic offered to classes of services guaranteeing QoS (all classes except STD one), in the further part of the text called as QoS classes of service (or QoS traffic), for all relations ingress-egress border routers is the same and all attached border routers are connected to the core with the links of the same capacity, say C. So, in order to assure core overprovisioning, for QoS traffic we can take part of capacity C, named C_{QoS} ($C_{QoS} < C$) as it is illustrated in Fig. 5. Furthermore, we need to decide which types of connections we have in the system. We can consider two alternative solutions. The first solution is to maintain "point to point" connections between a pair of ingress-egress border routers with allocated link capacities between them. Unfortunately, such approach leads to

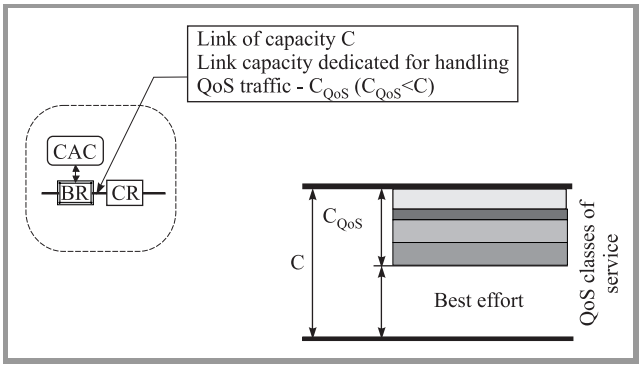


Fig. 5. The partitioning of the link capacity between the border and core routers among the classes of service and STD class of service.

partitioning of the link capacity connecting given ingress border router with core (the link of capacity C_{QoS}) between the directions to the rest of the egress border routers. As a consequence, in the case of temporal QoS traffic fluctuations with respect to which egress border router traffic is submitted, we can expect high level of new connection request losses. Apart this, when we distribute the link capacity between too many directions we lost multiplexing gain. The alternative solution is to maintain the connections "point to any". In this case, we allocate the whole capacity C_{QoS} to handle QoS traffic submitted to a given ingress router without distinguishing the target egress border routers. Such approach is applied in the IP QoS System as illustrated in Fig. 6.

Figure 7 shows a simple example with two ingress and two egress border routers illustrating the applied rule for overprovisioning the core. If we allocate C_{QoS} capacity on the link connecting given ingress border router with the core, then we need to allocate the C_{QoS} capacity on each path connecting this ingress border router to all egress border routers. Of course, such approach leads to the overpro-

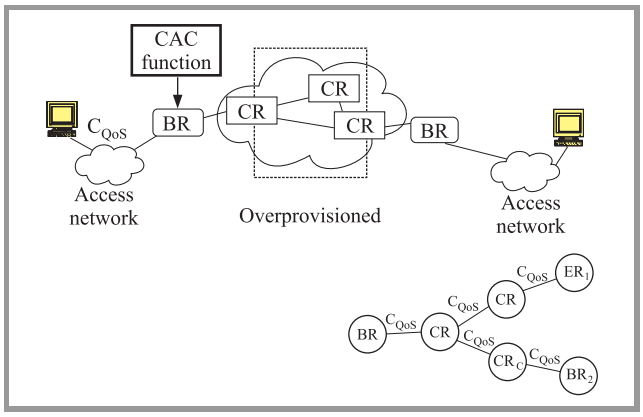


Fig. 6. The concept of the overprovisioning of the core.

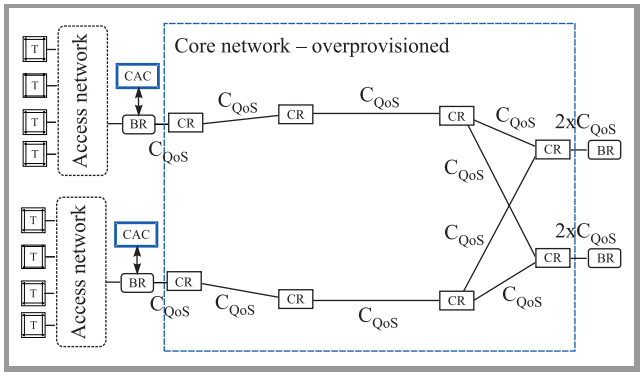


Fig. 7. Example of core overprovisioning in the case with 2 ingress border routers and 2 egress border routers.

visioning of the core. On the other hand, for the incoming traffic to a given egress border router we need to have $(N-1)$ C_{QoS} link capacity, when N border routers are connected to the core. In order to increase this capacity, we need to apply the CAC also in the egress border routers.

4. Summary

The paper provided an overview of the IP QoS System, its architecture and applied approach for traffic control and traffic engineering. The IP QoS System provides strict QoS guarantees by supporting a number of QoS classes of service. Comparing to best effort network, it requires to implement new mechanisms and algorithms at the packet, call request and network provision levels.

The System IP QoS is currently prototyped and tested. Its application to the network depends on the network operators.

Acknowledgement

We would like send by special thanks to all the partners involved in the project for their support as well as for their work on developing the IP QoS System.

References

[1] S. Blake *et al.*, "An Architecture for Differentiated Services", Internet RFC 2475, December 1998.

[2] D. Grossman, "New Terminology and Clarifications for Diffserv", Internet RFC 3260, April 2002.

[3] Y. Bernet *et al.*, "An Informal Management Model for Diffserv Routers", Internet RFC 3290, May 2002.

[4] J. Babiarz, K. Chang, and F. Baker, "Configuration Guidelines for DiffServ Service Classes", IETF RFC 4594, August 2006.

[5] C. Brandauer *et al.*, "AC algorithms in Aquila QoS IP network", *European Transaction on Telecommunications*, Wiley, vol. 16, no. 3, pp. 225–232, May-June 2005.

[6] B. F. Koch and H. Hussmann, "Overview of the project AQUILA (IST-1999-10077)", in Proc. Art-QoS 2003 Workshop, Warsaw, Poland, in *Architectures for Quality of Service in the Internet*, W. Burakowski, B. F. Koch, and A. Bęben, Eds., LNCS 2698, Springer, 2003, pp. 154–164.

[7] A. Bąk, W. Burakowski, F. Ricciato, S. Salsano, and H. Tarasiuk, "A framework for providing differentiated QoS guarantees in IP-based network", *Computer Communications*, vol. 26, Elsevier, pp. 327–337, 2003.

[8] W. Burakowski and M. Dąbrowski, "Wielousługowa sieć IP QoS: architektura i praktyczna weryfikacja w sieci pilotowej", *Przegląd Telekomunikacyjny*, no. 5, pp. 300–309, 2002 (in Polish).

[9] E. Mingozzi *et al.*, "EuQoS: End-to-end quality of service over heterogeneous networks", *Computer Communications*, vol. 32, issue 12, Elsevier, July 2009.

[10] X. Masip-Bruin *et al.*, "The EuQoS system: A solution for QoS routing in heterogeneous networks", *IEEE Commun. Mag.*, vol. 45, no. 2, 2007.

[11] W. Burakowski, A. Bęben, H. Tarasiuk, and J. Śliwiński. "Zapewnienie jakości przekazu "od końca do końca" w sieci Internet: 6.PR IST EuQoS", *Przegląd Telekomunikacyjny*, no. 9, pp. 236–241, 2006 (in Polish).

[12] ES 282 001 ver. 1.1.1 "Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); NGN Functional Architecture Release 1".

[13] "Resource and Admission Control Functions in Next Generation Networks, ITU-T Recommendation Y.2111, 2006.

[14] "Architectural framework for the Q.33xx series of Recommendations", ITU-T Recommendation Q.3300, 2008.

[15] W. Burakowski, A. Bęben, H. Tarasiuk, J. Śliwiński, R. Janowski, J. Mongay Batalla, and P. Krawiec, "Provision of end-to-end QoS in heterogeneous multi-domain networks", in Annals of Telecommunications – Annales des Télécommunications, Springer, vol. 63, issue 11, p. 559, 2008.

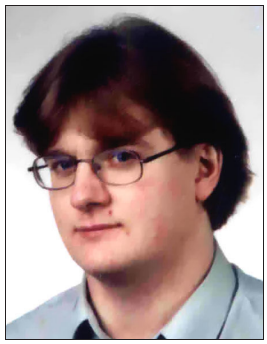
Wojciech Burakowski received his M.Sc., Ph.D. and D.Sc. degrees in Telecommunications from Warsaw University of Technology in 1975, 1982 and 1992, respectively. Now he works as Full Professor at the Institute of Telecommunications, Warsaw University of Technology and at the National Institute of Telecommunications, Warsaw as an R&D Director. He leads the Telecommunication Network Technologies research group.

Since 1990 he has been involved in several COST and EU Framework Projects (AQUILA, EuQoS, MoME, COMET). He is a member of Telecommunication Section of the Polish Academy of Sciences and an expert in 7 FR Programme. He was a chairman and a member of many technical programme committees of national and international conferences. He is author or co-author of about 180 papers published in books, international and national journals and conference proceedings and about 70 technical reports. His research areas include new networks techniques, ATM, IP, heterogeneous networks (fixed and wireless), network architecture, traffic engineering, simulation techniques, network mechanisms and algorithms. Recently, he is working on Future Internet and leads national project "Future Internet Engineering" collecting more than 120 researchers from 9 leading research organizations in Poland.
E-mail: wojtek@tele.pw.edu.pl
Institute of Telecommunications
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Jarosław Śliwiński received M.Sc. and Ph.D. degrees from Warsaw University of Technology in 2003 and 2008, respectively. His research area consists management and control systems in telecommunication, implementation aspects and laboratory networks.

E-mail: j.sliwinski@tele.pw.edu.pl
Warsaw University of Technology
Nowowiejska 15/19
00-665 Warsaw, Poland

Halina Tarasiuk received the M.Sc. degree in Computer Science from the Szczecin University of Technology, Poland, in 1996 and Ph.D. degree in Telecommunications from the Warsaw University of Technology, in 2004. From 1998 she is with Telecommunication Network Technologies Group at the Institute of Telecommunications, Warsaw University of Technology. From 2004 she is an Assistant Professor at the Warsaw University of Technology. From 1999 to 2003 she was collaborated with Polish Telecom R&D Centre. She participated in several European and national projects (2000–2011). Her research interests focus on Future Internet, NGN and NWGN architectures, node and network virtualization, signaling system performance, admission control and resource allocation methods and queueing mechanisms.
E-mail: halina@tele.pw.edu.pl
Institute of Telecommunications
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland



Andrzej Bęben received M.Sc. and Ph.D. degrees in Telecommunications from Warsaw University of Technology (WUT), Poland, in 1998 and 2001, respectively. Since 2002, he has been assistant professor with the Institute of Telecommunications at Warsaw University of Technology, where he is a member of the Telecommunication Network Technologies research group (tnt.tele.pw.edu.pl). He was involved in many European projects, like COST 257 (1996-2000), FP5 IST-AQUILA (2000–2003), COST 279 (2001–2005), FP6 IST-EuQoS (2004–2007). Currently, he is involved as the member of the Management Committees in projects COST IC0703 (2008–2012), FP7 ICT COMET (2010–2012) and Future Internet Engineering (2010–2012) He is author or co-author of about 70 papers published in books, international and national journals and conference proceedings. His research areas include IP networks (fixed and wireless), Future Internet networks, Content Centric Networks, traffic engineering, QoS routing, traffic control, simulation techniques, measurement methods, and test beds.
E-mail: abeben@tele.pw.edu.pl
Institute of Telecommunications
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Ewa Niewiadomska-Szynkiewicz D.Sc., Ph.D., MEng., Professor of optimization and simulation at Warsaw University of Technology, Head of the Control and Optimization of Complex Systems group. She participated in a number of research projects including three European projects within the TEMPUS programme and in the QOSIPS project (5th FP), coordinated a number of the group activities, managed the organization of a number of national conferences. Her interests are in computer simulation, optimization, and network modeling. She is the author of 100 papers, co-author and author of four books.

She also holds the position of associate professor at NASK. She is a member of the IEEE.
E-mail: e-n-s@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

E-mail: ewan@nask.pl
Research and Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland

Piotr Pyda was born in 1972. He received M.Sc. and Ph.D. degrees from the Military University of Technology, Warsaw, Poland, in 1996 and 2003, respectively, both in Telecommunication Engineering. He is now senior researcher in Military Communication Institute, Zegrze. He is engaged in research concerned of communications and information systems. His research interest include QoS and performance evaluation of modern packet networks.

E-mail: p.pyda@wil.waw.pl
Military Communication Institute
Warszawska st 22A
05-130 Zegrze Płd., Poland

Jordi Mongay Batalla was born in 1975. He graduate The Universitat Politecnica de Valencia in 2000 and Ph.D. degree of Warsaw University of Technology in 2009. He is now with Warsaw University of Technology (Poland). His research interest focus mainly on quality of service in Diffserv networks.

E-mail: jordim@tele.pw.edu.pl
Warsaw University of Technology
Nowowiejska 15/19
00-665 Warsaw, Poland

Paper

Performance Evaluation of Signaling in the IP QoS System

Halina Tarasiuk[a], Jarosław Śliwiński[a], Piotr Arabas[b,c], Przemysław Jaskóła[b,c], and Witold Góralski[a]

[a] *Institute of Telecommunications, Warsaw University of Technology, Warsaw, Poland*
[b] *Research and Academic Computer Network (NASK), Warsaw, Poland*
[c] *Institute of Control and Computation Engineering, Warsaw University of Technology, Warsaw, Poland*

Abstract—The IP QoS System is based on next generation networks (NGN) and differentiated services (DiffServ) architectures. Its main part is a signaling system, which allows to send a request from a user to the system for establishing new connection with predefined quality of service assurance. In this paper we present trial results of the proposed signalling system. The experiments were performed to measure setup delay utilizing artificial call generator/analyzer. To obtain results we assumed different distributions of interarrival and call holding times based on the literature. The results show that the setup delay strongly depends on access time to network devices, however also on the assumed call holding time models.

Keywords—IP QoS System, Quality of Service, signaling system.

1. Introduction

The aim of the paper[1] is to present the performance evaluation of signaling in the IP QoS System. The system offers functionalities to assure quality of service (QoS) guarantees for selected flows. We assume that the system is based on differentiated services (DiffServ) [1] and next generation network (NGN) [2] architectures.

To assure QoS for selected flows the system supports five end-to-end QoS classes of service (CoSes) and one best effort CoS. The end-to-end CoSes for data transfer are as follows: *telephony* for voice over IP applications, *RT interactive* for interactive games, *MM streaming* for video on demand, *high throughput data* for handling traffic generated by greedy TCP sources, and *standard* CoS (best effort). Moreover, in the network nodes we map end-to-end *telephony* and *RT interactive* CoS to one aggregated *real time* CoS. For the purpose of defining end-to-end CoSes concept, we follow [3] and [4].

We developed the system for a single IP domain. In this system, Internet service provider (ISP) can provision resources (that is link and buffer capacities) of the domain for each CoS in each network node. In each DiffServ IP router of the domain, an appropriate classifier classifies packets of selected flows to CoSes based on DSCP field of the IP packet header. In addition, edge and core routers offer an appropriate set of mechanisms to support QoS for packet transfer.

In the IP QoS System, we distinguish three types of processes, which operate in different time scales. These processes are:

- management, which manages routing and network provisioning,
- call setup/release,
- packet transfer.

In this paper we focus on call setup and release processes. These processes are handled by so called *signaling system*, which is the main part of the considered IP QoS System. The implemented and next evaluated signaling system consists of signaling entities and protocols developed for transport stratum of the NGN architecture. For handling signaling messages exchanged between functional entities of the architecture, during call setup or release process we utilize a dedicated CoS. We name this CoS as *signaling*. The performance of *signaling* CoS as well as performances of signaling protocols and entities impact on call setup/release delay and as a consequence they impact on user quality of experience (QoE). Following [5], recommended target values of acceptable by user call setup times for national IP network under normal load conditions are: mean delay = 5 s or for 95% of calls setup delay should be not greater than 8 s. For international IP network, the target values are 8 s an 11 s, respectively.

In our approach, we tested signaling system assuming artificial call generator with different call arrival and call holding models. For choosing adequate analytical models[2] to test the *signaling system*, we selected from the literature a number of analytical assumptions, which are based on measurement results obtained in pilot or real networks. However, it is worth to mention that it is a lack of maturity of the *signaling system* solutions developed by the operators [6]. Therefore, the call arrival and holding time models

[1]This work is partially funded by Polish Ministry of Science and Higher Education, under contract number PBZ-MNiSW-02-II/2007 "Next Generation Services and Networks – technical, application and market aspects".

[2]This work is partially funded by Polish Ministry of Science and Higher Education, under contract number N N517 385838 "Modeling and analysis of signaling in QoS Internet".

follow only some pilot trials [7] or measurements from service operators in best effort networks. First approach for modeling call arrivals in IP QoS network is Poisson process [8] as for PSTN network. However, based on [9] we conclude that user behaviors differ in multi-service network comparing with PSTN. For example, [7] presents fractal analysis and modeling of voice over IP traffic in stationary dedicated IP network with 800 users. Based on the obtained results authors conclude that more adequate analytical model of call holding time is based on generalized Pareto distribution (GPD) than exponential distribution. In [10] and [11], authors focus on call analysis of streaming media, e.g., reality show and live news, and sport TV or access to e-teach, video, or audio servers. The call arrivals and call holding times are essentially different than those assumed for PSTN networks. For sake of simplicity, we tested the proposed signaling system for analytical models considered in [7] for voice over IP application. In particular, we compared the system performance for call holding models based on exponential distribution and GPD.

In this paper, we continue our previous work on signaling systems as presented in [12], [13] and [14]. Comparing with that work, we show results for the signaling system based on NGN architecture. It is worth to mention that the system presented in [12]–[14] was developed simultaneously with NGN architecture details. Moreover, we enhance our trials with new analytical models for call arrival and holding times for voice over IP application.

The paper is organized as follows. Section 2 presents details of signaling in IP QoS System. Section 3 describes trial topology. Section 4 presents the obtained trial results for performance evaluation of the signaling system. Section 5 concludes the paper.

2. Signaling in the IP QoS System

The IP QoS System follows the functional decomposition that is similar to the one defined in [15] for NGN architecture. The system focuses strongly on management and control of resources, therefore most of its operations reside in the resource and admission control functions (RACF). In particular, from RACF we selected policy decision functional entity (PD-FE) and transport resource control functional entity (TRC-FE) as essential entities for QoS control. In the subsections below, first we provide an overview of the architecture and then we show typical signaling scenarios when system handles new call.

2.1. Architecture Overview

The architecture of the IP QoS System follows the isolation of functions into service and transport strata. In Fig. 1, we show all considered functional entities mapped over the particular strata. Moreover, the figure indicates the interfaces between particular entities and provides the recommendations defining those interfaces.

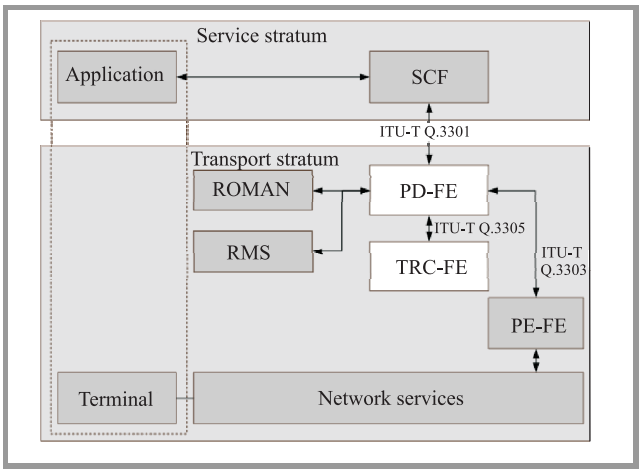


Fig. 1. Functional architecture of the IP QoS System.

In the **service stratum**, there are 2 entities:

- **Application** is available to the end use. We assume that it operates in 2 processes:
  - first it uses application signaling to collect all necessary information about interested parties,
  - then it preforms communication by sending and receiving the user's data.
- **Service control functions** (**SCF**) is a set of services, which are essential to establish a session for the application, e.g., they could cover user registry, authentication and accounting. Nevertheless, for successful integration with the IP QoS System, the SCF must communicate with elements available in the transport stratum.

In the **transport stratum**, we distinguish following entities:

- **Policy decision functional entity** (**PD-FE**) plays the role of the connecting hub between different elements of the RACF. Especially, it is a gateway from service stratum for SCF. This entity routes all messages and performs final decision about acceptance or rejection of the requests.
- **Transport resource control functional entity** (**TRC-FE**) is responsible for abstract representation of the resources available in the network. Moreover, it performs the connection admission control and it also maintains the database of accepted connections.
- **Policy enforcement functional entity** (**PE-FE**) directly interacts with network devices. It is able to translate requests into a set of instructions known by the device. For example, it is able to introduce traffic conditioning into the edge routers or packet scheduling configuration over network interfaces.
- **Routing management** (**ROMAN**) establishes paths in the network between pairs of access networks. Moreover, for each path it assigns the amount of resources which are available for traffic with QoS re-

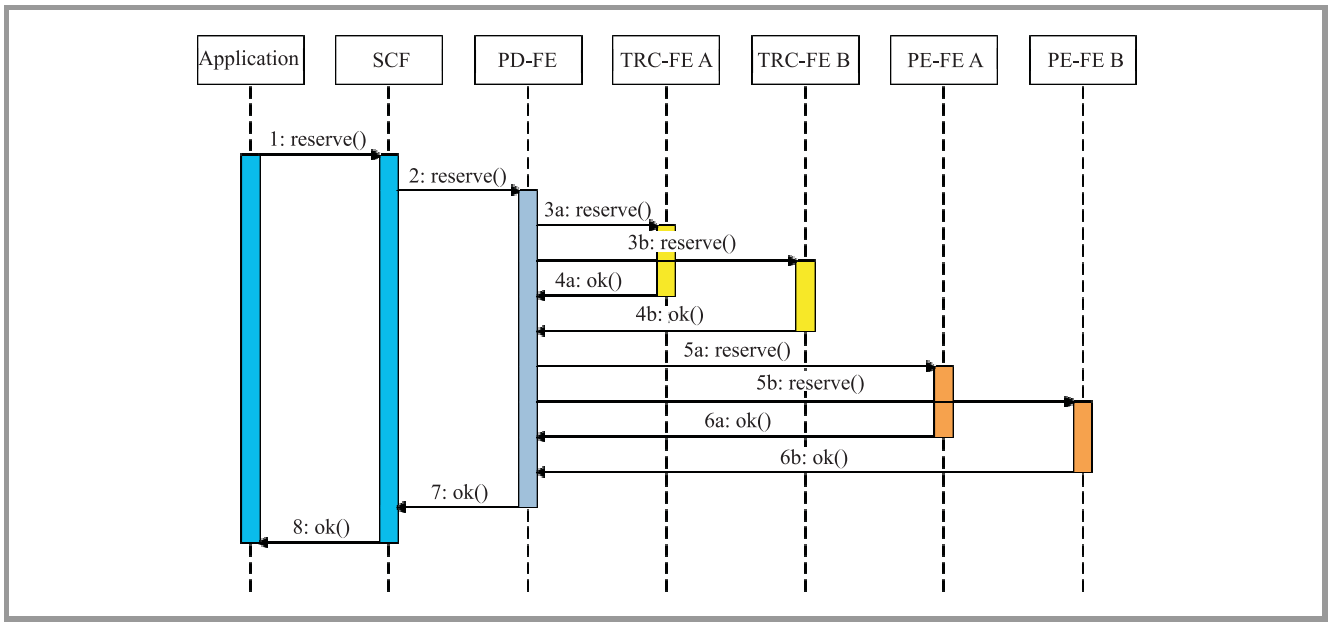


Fig. 2. Message sequence diagram for establishing a session.

quirements. Notice that this entity operates in the long time scale mainly performing management operations.

- **Resource management subsystem** (**RMS**) is responsible for mapping of end-to-end QoS requirements into the resources available in the network. In particular, these are weights assigned to the schedulers and buffer sizes. Similar to the ROMAN, this entity performs management operations, which are independent of call setups and call releases.

In the proposed architecture, we assume that the network consists only of a single *DiffServ* autonomous system. This means that the resources are split among a set of CoSes. Moreover, the traffic introduced into the network is conditioned on the edges, while new connections are controlled by an admission control function.
Notice that the Q-series interfaces that are defined for the NGN architecture in most cases use DIAMETER protocol for transferring signaling messages. However DIAMETER is well defined and standardized, for the purpose of simpler implementation we mapped the structures into the specification language for ICE (SLICE). Those structures were used for implementation of particular signaling nodes, which were built upon the ICE middleware.

2.2. Call Handling Scenarios

The IP QoS System handles 2 main signaling scenarios:

– establishing a session (call setup),

– closing a session (call release).

Below, we present the exemplary message sequence diagrams for both scenarios for the following topology:

- The session is established between two users A and B. Each user is located in different access network, which are connected to the core network using edge routers. Consequently, the traffic between users always crosses 2 edge routers and the routers in the core network.

- The deployment of the system assumes that each access network is handled by one TRC-FE and one PE-FE server. For example, operations for access network A are performed by TRC-FE A server and PE-FE A server.

Establishing a session. Figure 2 depicts the message exchange between entities in the IP QoS System for establishing a session. We distinguish the following steps.

1. Initially, the user A decides to establish a session to the user B. This request is translated by the application into a *reserve* message which is sent to the SCF.

2. The SCF locates user B and performs application level negotiation, e.g., it establishes a codec supported by both sides. Moreover, it must resolve all information required to classify the data streams in the network, i.e., IP addresses, transport protocol type (UDP/TCP) and port numbers. When this process is complete, the SCF sends *reserve* message to the PD-FE. Note that the signature of the *reserve* message may be different in each step. Even though they share the common name, they correspond to different interfaces.

3. The PD-FE verifies that description of the session if the message is complete. Then it decomposes a session into multiple connection structures; one connec-

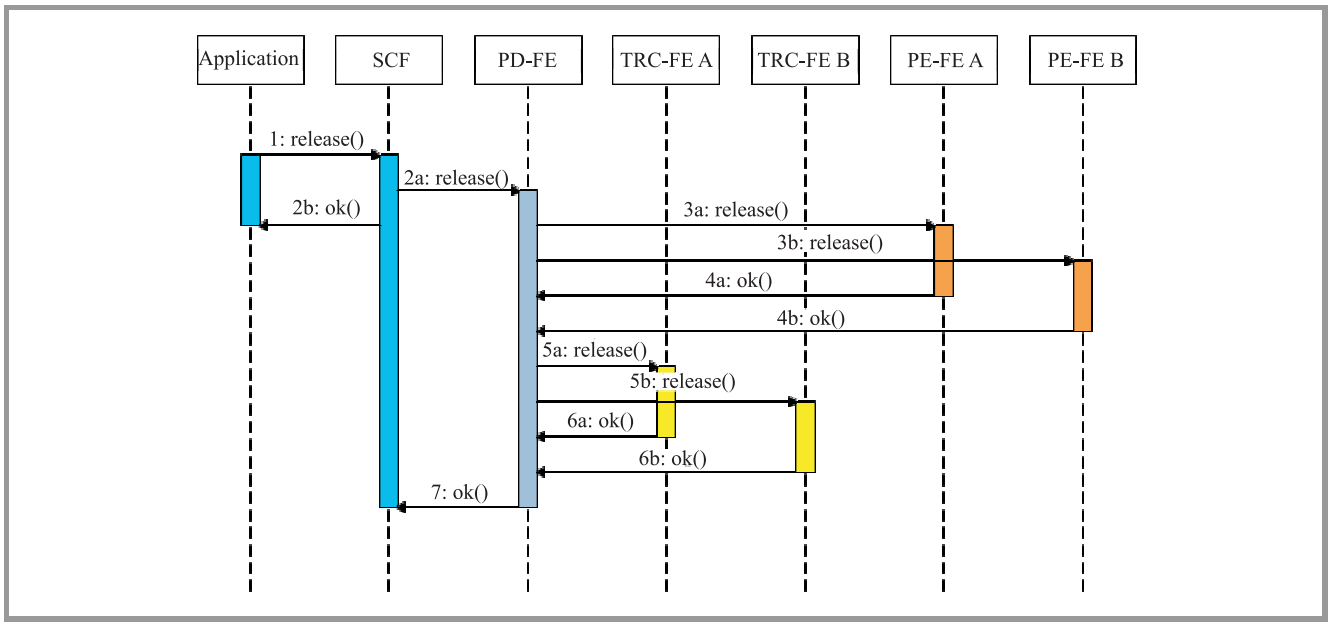


Fig. 3. Message sequence diagram for closing a session.

tion structure represents a single data stream to be established in the network. Before the handling of the connection starts, the PD-FE must map them into appropriate points of control. In case of the IP QoS System, the point of control is mapped directly to the ingress point of the network, i.e., all traffic generated by users belonging to single edge router share the same control point. In our example, we have 2 users attached to different edge routers. Therefore, the scenario will have 2 sets of control points: {TRC-FE A, PE-FE A} and {TRC-FE B, PE-FE B}. Steps number 3a and 3b cover the sending of reserve messages to the TRC FE elements.

4. TRC-FE elements verify that connections are unique and then they perform admission control function. If the result is positive, then they perform reservation by reducing the amount of available resources for given control point. Moreover, they store the connection information as in the IP QoS System, state of the connection is managed by TRC-FE elements. In our example, we assume that the result is positive and that they return the *ok* message to the PD-FE.

5. The PD-FE performs a synchronization upon the result of TRC-FE operation. When all responses are received, it performs the second part of the connection handling, i.e., it sends the *reserve* message to the PE-FE elements.

6. PE-FE elements communicate with edge routers and they enable traffic conditioning mechanisms for accepted connections. They configure policing, shaping and marking mechanisms, as necessary for particular connections. The result is returned to the PD-FE; in our example, the *ok* messages are sent.

7. At this point the resources in the network are reserved. The PD-FE synchronizes the responses for particular connections from PE-FE elements, then merges back connection into a session. This results in a response with *ok* message to the SCF.

8. The SCF forwards the response to the application.

Note that the SCF is responsible for refreshing the soft state of the session by periodically sending the *refresh* messages to the PD-FE. We do not show this exchange in our example as it seems to be straightforward.

Closing a session. Figure 3 depicts the message exchange between entities in the IP QoS System for closing a session. The steps for closing a session are similar as in case of establishing a session, which are:

1. User A decides to close a session with user B. In this case, the application sends a *release* message to the SCF.

2. The SCF verifies that given session is indeed active in the network. Recall that SCF maintains soft state of the sessions, therefore it has a knowledge whether the session is active or not. The SCF removes the session from the list of active sessions and sends confirmation to the application: *ok* message. At the same time it sends *release* message to the PD-FE. Note that the signature of the *release* message may be different in each step. Even though they share the common name, they correspond to different interfaces.

3. The PD-FE verifies that description of the session in the message is complete. Then it decomposes a session into multiple connections; one connection represents a single data stream to be established in

the network. Each connection is mapped into a set of control points, e.g., {TRC-FE A, PE-FE A} and {TRC-FE B, PE-FE B}. Steps number 3a and 3b cover the sending of *release* messages to the PE-FE elements.

4. PE-FE elements communicate with edge routers and they disable traffic conditioning mechanisms which were prepared previously for particular connections. The result is returned to the PD FE: *ok* messages are sent.
5. The PD-FE performs a synchronization upon the result of PE-FE operation. When all responses are received, it performs the second part of the connection handling, i.e., it sends the *release* message to the TRC-FE elements.
6. TRC-FE elements release the resources. Moreover, they remove the connection information from local storage. We assume that the result is positive (there are no errors); they return the *ok* message to the PD-FE.
7. At this point the resources in the network are released. The PD-FE synchronizes the responses for particular connections from TRC-FE elements, then merges back connections into a session. This results in a response with *ok* message to the SCF.

Note that in case of closing a session, neither application or SCF should wait for receiving a response. In fact, the return messages are just informative. For application the session is closed almost immediately, while SCF removes the session information upon reception of first release message.

3. Trial Environment

In this section we present details about call generator/analyzer utilised in the trials and trial topology.

3.1. Call Generator/Analyzer

To allow automatic tests of signaling subsystem performance call generator/analyzer was implemented. It consists of three programs: one for off-line preparation of experiment scenarios, next carrying out experiment itself, i.e., sending requests to the system and collecting data, and the last for postprocessing of results. The reason for performing most operations off-line, was to minimize a delay in the call generation process. As the aim of experiments was to test the signaling system, only call setup and release requests were sent to the system and no data between application hosts were transmitted. The architecture of the call generator is presented in the Fig. 4.

Various characteristics of different services were modeled by the set of distributions used for generating call inter-arrival times and durations. There are two groups of models implemented in the generator. The first one involves simple models in which inter-arrivals and durations are

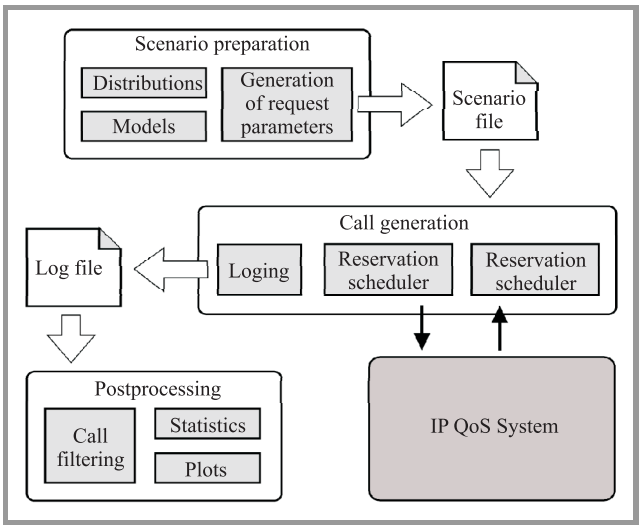


Fig. 4. Call generator architecture.

generated independently using, e.g., exponential or GPD. The second group is represented by a generalized Markov modulated Poisson process (MPPP), brought into play to prepare scenarios with correlations between events. While using Pareto distribution helps to introduce some burstiness as observed in [7], [16], and [17] the last model helps in preparation of more complex scenarios as suggested in [18] and [9].

The call generating program implements two schedulers serving reservation and release requests via independent polls of threads to avoid blocking and reveal full performance of the system. For the same reason logging is postponed to the end of program operation. The communication with the IP QoS System (precisely SCF) is provided by standard user interface using ICE middleware.

The postprocessing of previously collected logs allows to filter data and prepare statistics and various types of plots (time-plots, histogram etc.) without influencing the call generating process.

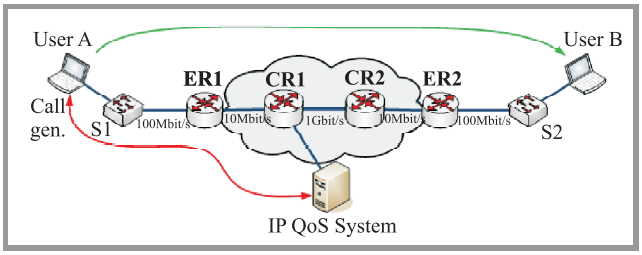


Fig. 5. Trial network topology with call generator and IP QoS System.

Figure 5 shows the topology of the trial network in which presented tests were performed. The trial network consists of two core routers (CR1, CR2 that are Cisco 7201) and two edge routers (ER1 – Cisco 2811 and ER2 – Cisco 1801). In addition, the network uses two Cisco Catalyst 2960G switches (S1 and S2) connected to the edge routers ER1 and ER2. The IP QoS System is connected to the trial core

network. It is responsible for resource provisioning as well as call handling processes.
Call generator/analyzer connects through the network to the IP QoS System in order to generate calls and to make measurements of the setup time in the signaling system. The server with the IP QoS System is running an Fedora 12 operating system on a computer with 3 GHz Intel® Pentium® 4 CPU.

4. Performance Evaluation of the Signaling System

The aim of experiments was to test performance of the IP QoS System in conditions similar to those occurring in operational network and to identify which elements introduce highest delays. Two sets of experiments were performed. In the first one, all system modules except routers were employed. In the second experiment, fully configured system composed of all software modules and network equipment. In this way, it was possible to assess the time necessary to configure edge routers, which turned out to be the main component of total call setup delay.
The scenarios were prepared for end-to-end telephony CoS based on measurement results of VoIP call load, from [7], scaled to the range which allows to estimate the performance limit. The process of call generation used exponential distribution for interarrival times and exponential distribution or GPD for call holding times. No data traffic was generated during experiments as only signaling system was tested, however all signaling system functionalities, including admission control function, were operating. Consequently all requests were processed, stored in database, etc. and sufficient resources (i.e. bandwidth) for telephony CoS were provided. The parameters of experiments are presented in Table 1 and Table 2.

Table 1
Parameters of experiments with exponential call holding times

Variant of the experiment	Call intensity range [1/s]	Mean holding time [s]	Call request [kbit/s]
Without routers configuration	1–10	114.27	8
Full – with routers configuration	0.25–1.25	114.27	8

Table 2
Parameters of experiments with exponential call interarrival time

Variant of the experiment	Call intensity range [1/s]	Holding time – GPD parameters			Call request [kbit/s]
		shape (k)	scale (s)	resulting mean [s]	
Without routers configuration	1–10	–0.39	69.33	49.88	8
Full – with routers configuration	0.25–1.25	–0.39	69.33	49.88	8

Table 1 presents parameters of experiments with Poissonian interarrival time and exponential call holding times. Table 2 presents parameters of experiments with Poissonian call interarrival time and GPD call interarrival time.
It is worth to mention that according to [7] parameters presented in Table 2 better approximate measurement results than those presented in Table 1. As we mentioned above, the aim of trial is also check an impact of the assumed model on the signaling system performance.
The referenced work [7] provided study of VoIP calls for corporate network of approximately 800 subscribers which generated 0.164 call/s. The scenarios used during tests have call intensity significantly up-scaled to test the system performance, while holding times are generated in a way conformant to observations of authors in order to retain typical holding time characteristics.
For each call the system was requested for resources for single VoIP connection, the resources in the IP QoS System were provisioned to allow submit all calls. No calls were reject.
To gather amount of data sufficient for analyzis calls were generated during 32 minutes.

4.1. Experiments without Routers

Two series of ten experiments were performed. As it was previously mentioned their aim was to test performance of the system alone without communication with routers which we expect is so time consuming that may hinder behavior of the software. Such a procedure allows to validate correctness of the implementation and to find the limit of the system performance.

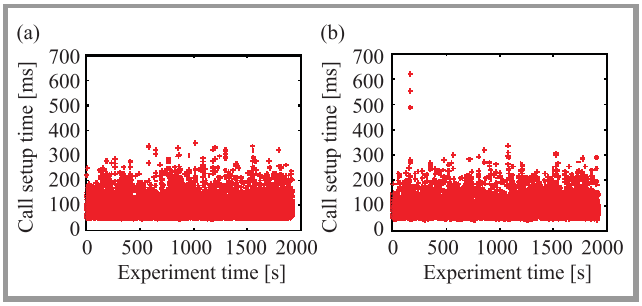


Fig. 6. Call setup times, exponential interarrival times, call intensity 4 call/s, holding times exponential (a) and GPD (b).

Characteristics in Fig. 6 show call setup times measured during experiments with moderate intensity (4 call/s) and exponential holding times (a graph) and GPD holding times (b graph).
Next, two characteristics in Fig. 7 show call setup times measured during experiments with high intensity (9 call/s) and holding times generated according to exponential (a graph) and GPD (b graph) distributions.
In the case when calls are generated with moderate intensity some variation of the service times may be observed in both (exponential and GPD holding time) cases, however

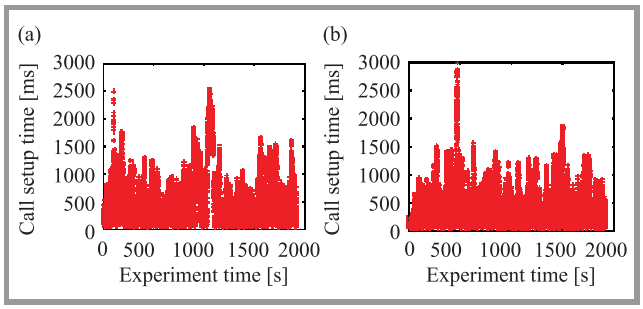


Fig. 7. Call setup times, exponential interarrival times, call intensity 9 call/s, holding times exponential (a) and GPD (b).

the overall system performance may be considered sufficient. When calls are generated with higher rate the graph becomes more rugged and call setup time increases significantly to the level of seconds. Such increase may be attributed mostly to queuing request at database and rate of 9 calls/s may be considered as maximum for the system. The argument for this may be also in characteristics in Fig. 8 presenting performance of the system in scenario

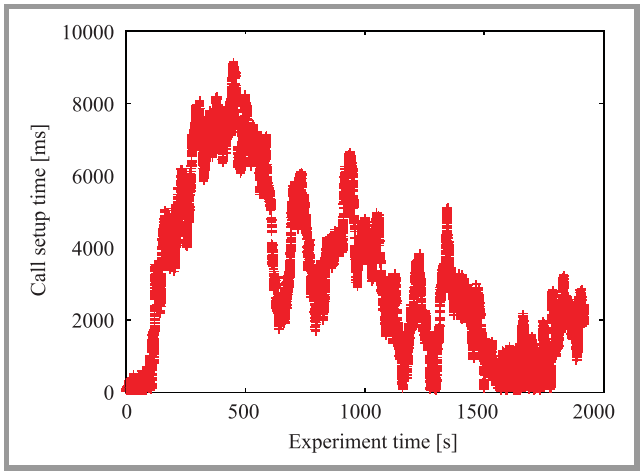


Fig. 8. Call setup times, exponential interarrival times, call intensity 10 call/s, GPD holding times – case of massive congestion.

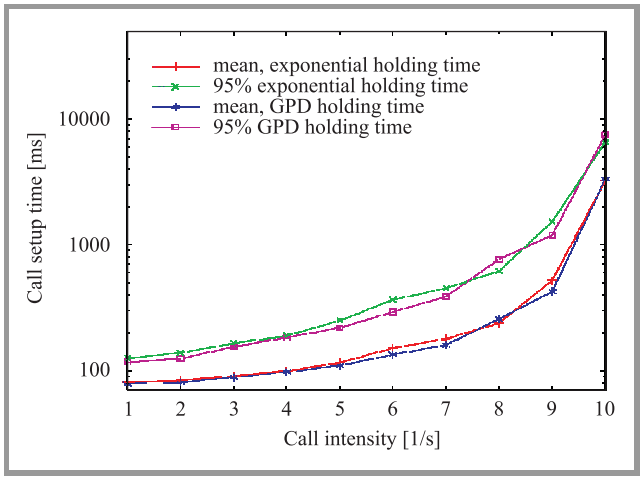


Fig. 9. Call setup times versus call intensity.

with rate of 10 call/s. The significant rise not only in maximum but also minimum call setup time suggests massive queuing occurring in one of the IP QoS System elements, possibly database holding reservation list.

To summarize results of experiments a graph showing call setup time versus call intensity (Fig. 9) was prepared. Two lines for each distribution are presented – one for mean service time and another for 95% – the value much better describing user perception of the system performance.

As we observe the characteristics for both models of holding times are very similar and they are below the target values of setup times.

4.2. Experiments with Router Configuration

Finally similar set of experiments was performed in fully configured signalling system, i.e., with configuring of the routers. The intensities were scaled down as communication time was taken into account, the rest of parameters and procedure following these described previously. Characteristics in Fig. 10 show call setup times for experiments with moderate rate which in this case is 0.25 call/s.

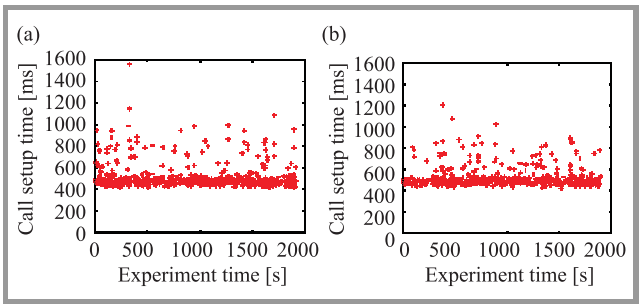


Fig. 10. Call setup times, exponential interarrival times, call intensity 0.25 call/s, holding times exponential (a) and GPD (b).

Next set of graphs (Fig. 11) represents situation of higher load, close to maximum which can be served in acceptable time.

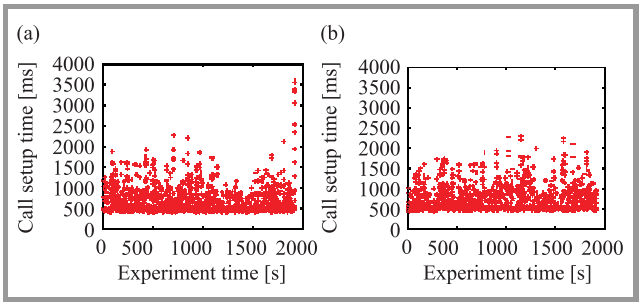


Fig. 11. Call setup times, exponential interarrival times, call intensity 0.75 call/s, holding times exponential (a) and GPD (b).

In these examples some call setup times exceed 2s which is close to the value typically perceived as acceptable for users, so the system limit lays approximately between 0.75 and 1 call/s. As an argument for this another graph,

showing heavy load condition at intensity of 1.25 call/s, is presented in Fig. 12.

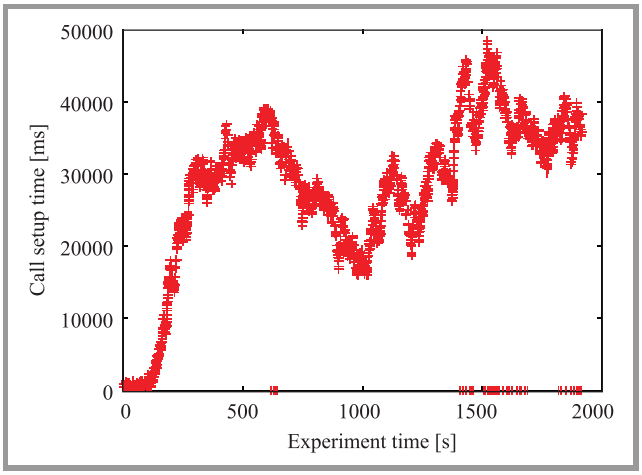


Fig. 12. Call setup times, exponential interarrival times, call intensity 1.25 call/s, GPD holding times.

Rapid grow of service time beyond values observed in examples lacking router configuration suggests that requests are queued in the router access module due to the overload of the control software in the router. It is important to remind that performance limit of the system when router communication was excluded was approximately 9 times higher so the limitation visible in Fig. 12 can be attributed only to the routers. To systematize this findings, characteristics showing average and 95% of call setup time versus call intensity are presented in Fig. 13.

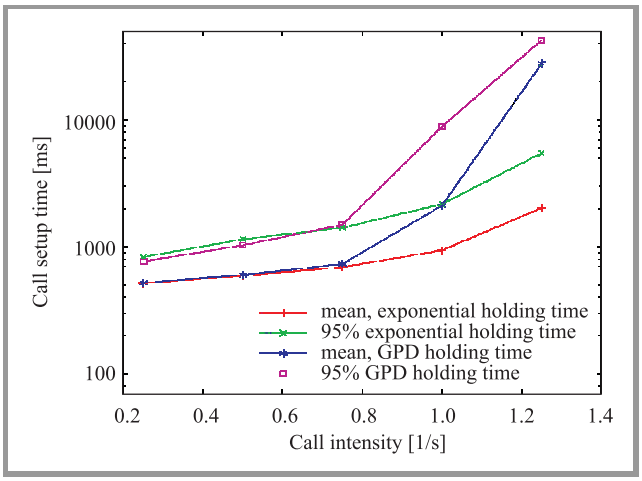


Fig. 13. Call setup times versus call intensity.

As we can see an impact of holding times models play an essential role on the obtained results in this scenario. Moreover, for call intensity above 1 call/s the characteristics for GPD holding time model are above the target values. On the other hand for exponential holding time model are in the acceptance area. So, we observe an impact of the assumed call holding time model on the obtained characteristics.

5. Conclusions

The experiments allowed to evaluate the system performance and to identify elements contributing the most to the experienced delay. The overall performance of fully configured system (0.75 call/s) may be considered low, however it must be stated that it is significantly higher than necessary to serve requests generated by 800 subscribers in [3]. Considering that each edge router can be configured in parallel, the system capacity may be scaled up by partitioning each large access network into a number of smaller ones (analogous as in radio access of the mobile networks). Furthermore, the main delay is connected with configuring routers (around 500 ms under moderate load), so finding more efficient equipment is necessary for real life application. Another source of delay (due to queuing) is database, which may be optimized with help of well-known techniques. The results show also the impact of the assumed call holding time models on setup delay in some scenarios.

Acknowledgement

We would like especially thank all the partners involved in the project for their support as well as for their work on developing the IP QoS System.

References

[1] Y. Bernet *et al.*, "An Informal Management Model for DiffServ Routers", Internet RFC 3290, May 2002.

[2] "Functional Requirements And Architecture Of Next Generation Networks", ITU-T Rec. Y.2012, 04/2010.

[3] J. Babiarz *et al.*, "Configuration Guidelines for DiffServ Service Classes", Internet RFC 4594, Aug. 2006.

[4] K. Chan, J. Babiarz, and F. Baker, "Aggregation of Diffserv Service Classes", Internet RFC 5127, Feb.2008.

[5] "Network Post-selection Delay in PSTN/ISDN Networks Using Internet Telephony for a Portion of the Connection", ITU-T Rec. E.671, March 2000.

[6] T. Aoyama, "A new generation network: Beyond the Internet and NGN", *ITU-T Kaleidoscope*, *IEEE Commun. Magazine*, vol. 47, no. 9, pp. 82–87, 2009.

[7] T. D. Dang, B. Sonkoly, and S. Molnar, "Fractal analysis and modeling of VoIP traffic", in *Proc. 11th Int. Telecommun. Netw. Strategy Planning Symp. NETWORKS 2004*, Vienna, Austria, 2004, pp. 123–130.

[8] J. W. Roberts, "Traffic theory and the Internet", *IEEE Commun. Mag.*, Jan. 2001, pp. 94–99.

[9] W. Chen *et al.*, "Modeling VoIP call holding times for telecommunications", *IEEE Network*, Nov/Dec, pp. 22–28, 2007.

[10] C. Costa *et al.*, "Analyzing client interactivity in streaming media", in *Proc. WWW 2004*, New York, USA, 2004, pp. 534–543.

[11] E. Veloso *et al.*, "A hierarchical characterization of a live streaming media workload", *IEEE/ACM Trans. Networking*, vol. 14, no. 1, pp. 133–146, 2006.

[12] H. Tarasiuk *et al.*, "Designing the simulative evaluation of an architecture for supporting QoS on a large scale", in *Proc. QoSim 2008*, Marseille, France, 2008.

[13] J. Mongay Batalla, J. Śliwiński, H. Tarasiuk, and W. Burakowski, "Impact of signaling system performance on QoE in next generation networks", *J. Telecommun. Inform. Technol.*, no. 4, 2009.

[14] E. Mingozzi *et al*., "EuQoS: end-to-end quality of service over heterogeneous networks", *Computer Commun.*, vol. 32, iss. 12, Elsevier, 2009.

[15] "Resource and Admission Control Functions in Next Generation Networks", ITU-T Rec. Y.2111, Nov. 2008.

[16] A. Brampton *et al*., "Characterising user interactivity for sports video-on-demand", in *Proc. 17th Int. Worksh. Netw. Oper. Sys. Sup. Dig. Audio Video, Urbana-Champaign NOSSDAV 2007*, IL, USA, ACM, 2007.

[17] Sh. Jin and A. Bestavros, "Generating internet streaming media objects and workloads", in *Web Content Delivery*, S. T. Chanson, X. Tang, and J. Xu, Eds. Springer, 2005.

[18] T. Qiu *et al*., "Modelling user activities in a large IPTV system", in *Proc. IMC'09*, Chicago, Illinois, USA, ACM, 2009, pp. 430–442.

Piotr Arabas received his Ph.D. in Computer Science from the Warsaw University of Technology, Poland, in 2004. Currently he is an Assistant Professor at Institute of Control and Computation Engineering at the Warsaw University of Technology. Since 2002 with Research and Academic Computer Network (NASK). His research area focuses on modeling computer networks, predictive control and hierarchical systems.
E-mail: parabas@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

E-mail: Piotr.Arabas@nask.pl
Research and Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland

Przemysław Jaskóła received his M.Sc. in Computer Science from the Warsaw University of Technology, Poland, in 1999. Currently he is a Ph.D. student in the Institute of Control and Computation Engineering at the Warsaw University of Technology. Since 2005 with Research and Academic Computer Network (NASK). His research area focuses on hierarchical optimization and computer networks.
E-mail: Przemyslaw.Jaskola@nask.pl
Research and Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland

Witold Góralski was born in 1985. He graduate The Faculty of Electronics and Information Technology (2009). Since 2009 he is Ph.D. student on The Faculty of Electronics and Information Technology. His research interest focus mainly on QoS, testbeds and queueing mechanisms.

E-mail: w.goralski@tele.pw.edu.pl
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Halina Tarasiuk, Jarosław Śliwiński – for biographies, see this issue, p. 10.

Paper

On Dimensioning and Routing in the IP QoS System

Witold Góralski[a], Piotr Pyda[b], Tomasz Dalecki[b], Jordi Mongay Batalla[a], Jarosław Śliwiński[a], Waldemar Latoszek[c], and Henryk Gut[c]

[a] *Warsaw University of Technology, Warsaw, Poland*
[b] *Military Communication Institute, Zegrze, Poland*
[c] *National Institute of Telecommunication, Warsaw, Poland*

Abstract—This article presents dimensioning and routing solutions in IP QoS System designed during the implementation of the PBZ project: "Next Generation Services and Networks – technical, application and market aspects: Traffic management – IP QoS System". The paper presents the functional architecture together to the description of the functions and methods implemented in the system.

Keywords—IP QoS System, resource dimensioning, routing.

1. Introduction

The architecture of the IP QoS System and traffic control mechanisms have been specified during the PBZ project[1] and the implemented prototype has been tested in the testbed network. The proposed architecture of the IP QoS System is compatible with the next generation network architecture (NGN). Moreover, in terms of quality of service (QoS) assuring, this implementation is compatible with the differentiated services architecture (DiffServ) [1]. Figure 1 shows the IP QoS System architecture with implemented functional modules.

The proposed solution relates to resource management layer, which main objective is to separate the traffic submitted to the four classes of service (CoS): real time, multimedia streaming, high throughput data and standard. The resource management implemented in IP QoS distinguishes three basic processes directed to prepare the network for assuring guaranteed service for the new requests:

- resource dimensioning process between edge routers,
- routing process on the basis of QoS requirements, i.e., QoS-aware routing,
- resource reservation process for new call requests that takes into account quality of service requirements.

These processes we implemented by the following modules: routing management (ROMAN), resource management subsystem (RMS), policy decision – physical entity (PD-PE), transport resource control (TRC) and policy enforcement – physical entity (PE-PE). ROMAN module is responsible for routing in the network. It should be noted that the testbed network implements multiprotocol label switching traffic engineering (MPLS TE) tunnels for carrying traffic of given CoS. ROMAN module sets the path and creates TE tunnels between each pair of edge routers. In the project framework we implemented and tested different routing algorithms for TE tunnels configuration. Besides standard algorithms additional extended Dijkstra's algorithms have been implemented. ROMAN forces the TE tunnel in the routing algorithm by entering the command: *ip explicit-path* with specified intermediate nodes. After configuring tunnel, the module writes the information in a database and provides the information to the RMS module with the list of tunnels. The RMS module performs resource allocation and resource dimensioning, which depend on available resources and matrices of traffic demands, respectively. The primary task of the RMS module is to determine the link capacity and buffer size for each class of services inside a single domain. The resulting capacity and buffer size are set in the edge router and are the parameters used by the call admission control (CAC) function.

The article describes the implementation of modules responsible for proper router configuration within the testbed network. The main objective of the paper is to describe the specification as well as implemented algorithms of the modules responsible for dimensioning process and configuration of MPLS paths. Moreover, we describe how the different modules cooperate with each other and exchange data. The theoretical description comes with exemplary configuration results taken from the testbed network of the IP QoS System. In the conclusion, we summarize the achievements of the implementation of the system by describing the presentation of IP QoS System testbed on national exhibition, and propose system extensions for further development.

2. Functions of IP QoS System Modules

2.1. ROMAN Module

ROMAN module is responsible for the implementation of the routing process in testbed network that is compatible with the DiffServ architecture [1]. In DiffServ networks,

[1]This work is partially funded by Polish Ministry of Science and Higher Education, under contract number PBZMNiSW-02-II/2007 "Next Generation Services and Networks – technical, application and market aspects".

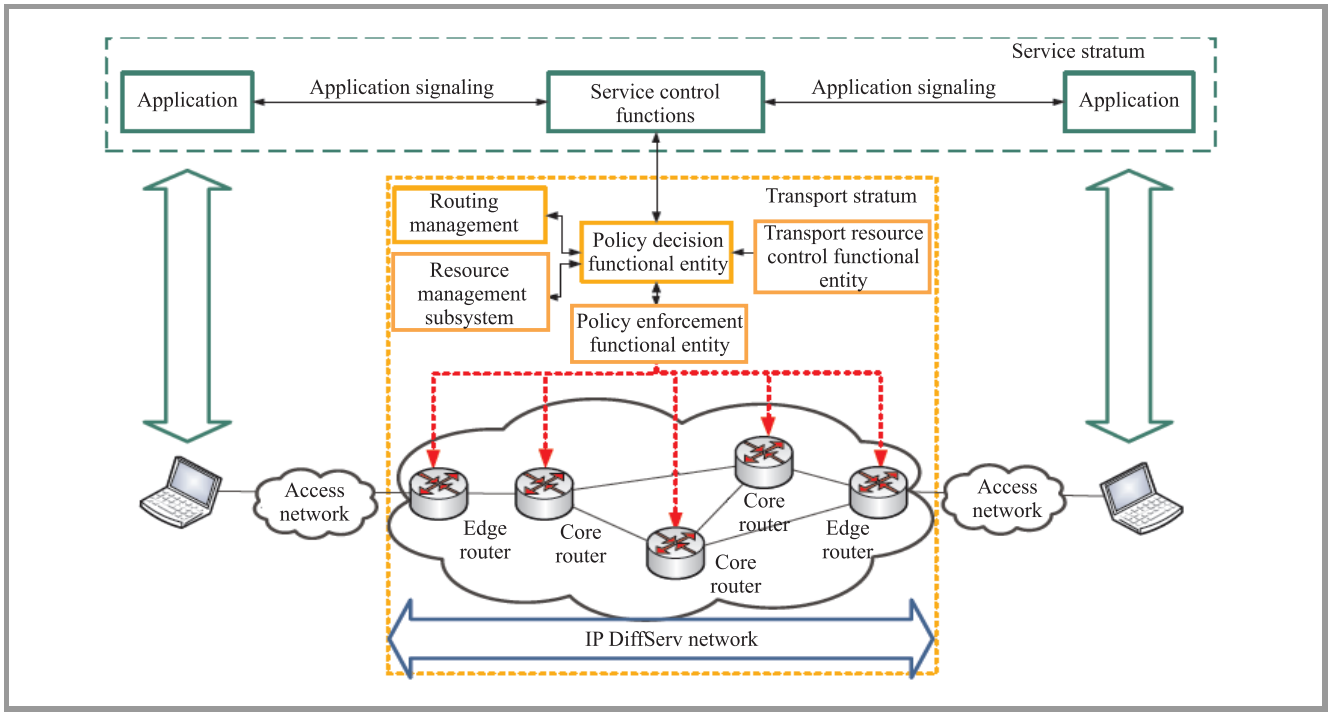


Fig. 1. Architecture of the IP QoS System.

edge routers support functionalities for single streams and core routers are aware only of aggregated traffic in proper CoSes. Testbed network implements MPLS TE tunnels for routing packets belonging to different CoS. Edge routers add and remove MPLS labels for incoming and outgoing packets, respectively. MPLS TE tunnels, or briefly TE tunnels, are defined by the so-called label switched path (LSP). To configure the LSP it is required to specify subsequent nodes from source to destination router. Routers in the MPLS network make forwarding decisions based on their label forwarding instance base (LFIB) tables. These tables contain labels which corresponding input and output interfaces. The paths on which are established the tunnels are determined using a routing algorithm implemented in ROMAN module. ROMAN task is to determine paths and create TE tunnels between each pair of edge routers for traffic classified into proper CoSes. DiffServ architecture assumes that the individual streams of packets sent by applications in the backbone are aggregated into streams of particular CoSes. In DiffServ architecture routers analyze DSCP field in IP headers, and on this basis are handled with appropriate CoS. Packets belonging to the CoSes in the MPLS network are distinguished on the basis of the value of EXP field in MPLS header (see Fig. 2).

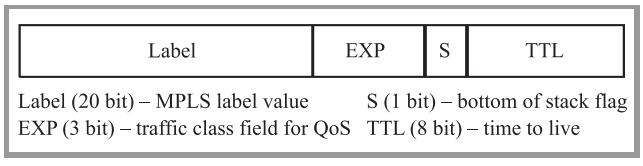


Fig. 2. MPLS header (4 Bytes length).

For this reason we defined mapping between DSCP code values and EXP values of the MPLS header. Method of mapping DSCP codes for aggregated CoSes into MPLS EXP codes is shown in Table 1 [2]. The table is filled in accordance with the EXP-inferred-PSC LSP (E-LSP) model proposed by the IETF [3]. Each router configures per hop behavior (PHB) rules for packets with different EXP field's values. It is possible to do static mapping in the domain but it should be the same in the whole network.

Table 1
Mapping between DSCP in IP QoS System and PMLS EXP field [2]

Type of application	End-to-end class of service	Class of service in IP QoS project	DSCP	MPLS EXP
Signaling	Signaling	Signaling	101000	101
VoIP	Telephony	Real time (RT)	101110	100
Interactive games	RT interactive		100000	
Video on demand	MM streaming	MM streaming	011010 011100 011110	011
FTP	High throughput data	High throughput data	001010 001100* 001110*	010 001*
	Standard	Standard (STD)	000000	000

* DSCP codes and MPLS EXP field used for HTD class of service.

In the project we assumed that the capacity of all the links in the core network is divided into different classes of service. This division is determined by the RMS module, according to the maximum allocation model (MAM) method

described in [4]. The main advantage of the model is its simplicity and, in turn, the model ensures the achievement of isolation between traffic.
ROMAN module performs the following functions:

- retrieves information about the network topology,
- sets required capacity (C_{QoS}) for proper classes of service,
- establishes MPLS TE tunnels,
- provides information about the topology and TE tunnels to the RMS module.

We assume that the input data for ROMAN module is information about current network topology, link capacity and required capacity for CoSes stored in the database. Module retrieves information from one of the router of the network, since all the routers in the network use the OSPF routing protocol [5] and gather information about network topology. In our implementation it is possible to load the status of the network from extensible markup language (XML) configuration file. Due to resource dimensioning model, the value of required capacity for proper CoS cannot be greater than constraint (1):

$$C_{QoS} \leq \frac{C_{\min}}{L_{RD}-1}, \tag{1}$$

where:
$C_{\min}$ – minimal access link capacity,
L_{RD} – number of edge routers.

CISCO routers use PCALC algorithm for MPLS TE tunnels establishment. This algorithm discovers the paths in the network and provides data for explicit route object (ERO) field used in RSVP-TE signaling structure. The proposed solution implemented and tested different routing algorithms for TE tunnels configuration, therefore replacement of PCALC algorithm was mandatory. In addition to standard algorithms such as Dijkstra and Kruskal, additional algorithms have been implemented like extended Dijkstra's described in [6] and self-adaptive multiple constraints routing algorithm (SAMCRA) described in [7], [8]. ROMAN configures TE tunnels by entering the command: *ip explicit-path* with specified intermediate nodes. After configuring the tunnel, the module writes information in the database and provides list of tunnels to the RMS module. Additional implementation details are presented in Section 3.

2.2. RMS Module

RMS module is responsible for implementing the algorithm for resource dimensioning and allocation, depending on available resources and demands in the domain. Performing these tasks requires communication with the ROMAN and PD-PE modules. In the prototype we implemented a simplified static version of resource allocation algorithm.

The primary task for RMS module is to determine the capacity allocated for each class of service in every edge router (ER) of the domain. This capacity is used by call admission control function implemented in the TRC-FE module. In addition, the RMS module sets the buffer size in appropriate port for respective classes of service in accordance to [2].
RMS module receives notification from ROMAN module about changes in network topology. The notification itself does not provide information about new network topology and is the responsibility of the RMS to achieve this information in a pull mode from the ROMAN. Additionally RMS module can be notified that paths has been changed in network by ROMAN module. Like the previous notification, it does not provide additional data structures, therefore the RMS module retrieves structure with up-to-date paths in the network from ROMAN module.
RMS module allocates bandwidth for all classes of service in a chosen path (MPLS TE tunnel). Additionally module allocates bandwidth for the MPLS tunnels that pass through particular link. In addition, buffer sizes are set in each output port and for each class of service (ports through which at least one path passes). The results obtained from the resource allocation algorithm are used in the admission control algorithm.
To describe the algorithm we introduce the following variables and constants:

$l = 1 \ldots L$, link number ($L \equiv$ number of links),
$s = 1 \ldots S$, path number ($S \equiv$ number of paths),
$k = 1 \ldots K$, CoS number ($K \equiv$ number of CoS),
$C_l \equiv$ capacity of link l,
$\delta_{ls} = 1$ if path number s contains link number l, otherwise $\delta_{ls} = 0$,
$M[k,s] \equiv$ matrix of demands including demands for CoS number k in path s.

In particular, the algorithm allocates resources for the classes of service and convert relative input matrix of demands into absolute matrix of demands describing the exact bandwidth for each path and class of service. When for all classes and paths are the same demands then input elements of the matrix are equal to 1.
After running the algorithm, the output matrix contains the bandwidth requirement for each path and class of service. Then, we calculate the value of bandwidth C_{kl} for router output port of link l and class of service k according to the formula:

$$C_{kl} = \sum_{s} M[k,s]\ \delta_{ls}\,. \tag{2}$$

After calculating the bandwidth values for different classes of service we calculate buffer sizes. The length of the buffers allocated for each class depends on the used routers. For example, Cisco routers used in the prototype allowed to work with no more than 64 packets total buffer size for all classes of service. The results presented below take into account this limitation.

Signaling class: Resource dimensioning for signaling traffic is quite complicated. Recent studies showed that a single procedure call statement in the exemplary architecture of

next generation networks needs around 30 kbit/s. If we have reserved bandwidth C_{SIG} for signaling traffic, then we can allow C_{SIG}/30 connection set-up procedures. One procedure connection request, sends simultaneously N messages to the network. In order not to loose signaling packets, the buffer size for signaling CoS should be calculated according to the following formula:

$$B_{SIG} = \frac{C_{SIG}\ [\text{kbit/s}]}{30\ [\text{kbit/s}]}\ N\ [\text{packets}]\,. \tag{3}$$

RT Interactive class: Buffer size for this class must take into account maximum values of delay variation (IPDV) according to the following formula:

$$IPDV_{RT}\ [s] = \frac{B_{RT} \times d_{RT}\ [\text{B/packet}]\ 8\ hop}{C_{RT}\ [\text{kbit/s}]}\,, \tag{4}$$

where:

d_{RT} – the largest packet length of all RT streams,
hop – the number of hops on the longest path,
B_{RT} – buffer size for RT class,
C_{RT} – allocated bandwidth for RT class.

MM streaming and HTD classes of service require small packet loss [2]. Therefore, the buffer size should be large enough to minimize the number of lost packets belonging to these classes of service. On the other hand, while there is no requirement for delay variation the requirement for packet transfer delay (IPTD) is 0.5 s for end-to-end delay [2].
The delay and packet loss levels depend on the load (ρ). Only, we can calculate the maximum buffer size for the delay when the buffer is always full ($\rho \to 1$). Then, the buffer can be calculated using the following formula:

$$IPDV_{MMS/HTD}\,[s] = \frac{(B_{MMS/HTD}+1)\,d_{MMS/HTD}\ [\text{B/packet}]\,8\,hop}{C_{MMS/HTD}\,[\text{kbit/s}]}\,, \tag{5}$$

where:

$d_{MMS/HTD}$ – maximum packet size of all streams MMS or HTD,
hop – the number of hops on the longest path,
$B_{MMS/HTD}$ – buffer size for MMS or HTD,
$C_{MMS/HTD}$ – allocated bandwidth for MMS or HTD.

For the values based on IPTD, the length of the buffer is over-dimensioned ($\rho < 1$). If the buffer size is too small, then we choose a larger buffer size in our implementation (100 packets – as a minimum value).
The input data related to network topology and routing are passed from ROMAN module to RMS module. RMS node can be configured with the appropriate entries in the configuration files. Configuration of QoS requirements is stored in the file "qos.txt", while the value of demands are stored in the file "demands.txt". This file configures the demand for specific paths and specific classes of service: signaling, real time, MM streaming, high throughput data and standard. Demands are expressed as the weights, and allocated capacity is directly proportional to the stated weight. In another configuration file are stored quality of service requirements for different classes of service (Table 2). The file for the relevant class defines the following metrics: IP packet loss ratio (IPLR), IP packet transfer delay (IPTD), IP packet delay variation (IPDV) and packet length d. Additional implementation details are presented in Section 3. Table 2 presents the necessary data of CoS for configuring RMS.

Table 2
The requirements on the quality of service [2]

Number	Class	IPLR	IPTD	IPDV	d
1	Signaling	X	X	X	–
2	Real time	X	X	X	X
3	MM streaming	X	X	–	–
4	High throughput data	X	X	–	–
5	Standard	–	–	–	–

2.3. PD-PE Module

PD-PE module participates in routing process in the domain in the following way. It receives from the ROMAN module information of customers attached to given routers. This function is performed by configureAccessNetworks() method in interface RomanToPd. Current list of customers addresses belonging to the routers overwrites the old one. Function requestAccessNetworks() in PdToRoman interface request list of customer addresses in routers.
PD-PE module provides information for the call admission control function and resource allocation process to the routers, in order to configure the interfaces. Both functions are triggered by the RMS by using configureResources() and configurePorts() functions in RmsToPd interface. During the network resources monitoring process, PD-PE module transmits reports through reportResourceState() function in PdToRms interface.

2.4. PE-PE Module

PE-PE module manages network devices' configuration, which is the final element for configuring the router output interface. For this purpose the interface PdToPe has configurePorts() function that provides router interface configuration. This configuration includes resource allocation (bandwidth and buffer size) for particular class of service. PE-PE module is the last element of the signaling chain for resource allocation. Module communicates with the network devices. Anyway, this communication is not standardized and is specific to each network which actually acts as a driver for the IP QoS System.
For appropriate work of PE-PE module, a database is created containing the following tables: routers, interfaces, class_configurations, pe_points, access_lists, policers, shapers and session. The first four tables provide information about topology of testbed network. It should be noted

that the contents of both interfaces and class_configurations tables will vary depending on the dimensioning of network resources. Subsequent tables access_lists, policers, shapers and session will be filled during admission control process. In routers table is stored basic information about routers in testbed network. Since in the testbed network we implement 2 border routers and 4 core routers, then the routers table contains 6 records, one for each router. Each of these routers have an identifier that is used as a foreign key for accessing the table. The fields username, password and password_enable contain the data necessary to establish a telnet session with the routers. Field ip contains the IP address of virtual loopback interface, which is defined on each router and used as an identifier of the router in the testbed network.

Class_configuration table contains data about classes of service provided in the system. This table contains the following fields: name – the name of the CoS, bitrate – bit rate dedicated to the class on the output link [bit/s] and queue_limit – the queue size for the class. Moreover, the interface_id is a foreign key, which represents the identifier of the network interface. For each interface are defined five classes of service: real time, MM streaming, high throughput data, standard and signaling.

Table Pe_points contains data of all edge nodes in the network. Moreover, the database module of PE-PE contains the tables: policers, shapers i session, which are filled during the per-flow operation of IP QoS System.

2.5. TRC-PE Module

TRC-PE node runs the call admission control algorithm. For this purpose in PdToTrc interface we distinguish the method configureResources() that provide description of resource allocation for each class of service. After starting this method, the following actions are performed:

- TRC-PE module finds in the database points that realize call admission control. In case the point is not found, the method returns a negative result.
- Resource configuration take into account bandwidth, buffer size, packet loss ratio, packet transfer delay and packet delay variation. Based on these parameters is provided maximum load value, which is determined by an call admission control algorithm. This value is stored in the database. If the call admission control algorithm is not supported, the method returns a negative result.
- After the proper run the algorithm returns a positive result.

During configuration of the TRC-PE module was set up a database named pbz_trc. This database stores data about available resources for each class of service on IP QoS System and for all edge routers in the testbed network. Table trc_points contains information for all edge routers (routers ER1 and ER2). Moreover trc_resources table stores information about the classes of service for all the routers specified in the table trc_points. While system is working these tables are filled with records that store data about the current sessions and flows in the testbed network.

3. Implementation of IP QoS System Modules

ROMAN module has been implemented in C# and runs on a virtual MONO platform on Suse Linux. Communication with other modules is provided by the interface using the ICE library. Figure 3 depicts ROMAN module divided in functional blocks.

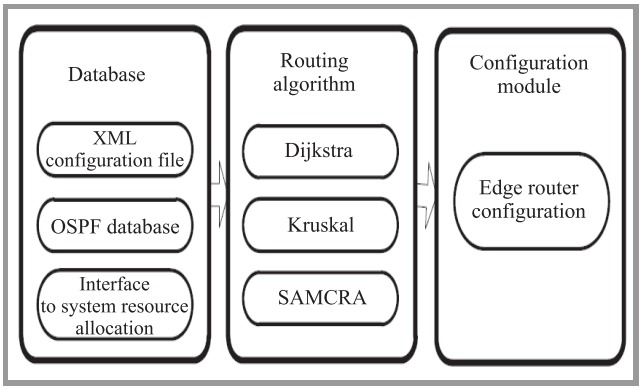


Fig. 3. ROMAN module divided in functional blocks.

The main functionality is included in the library of routing algorithms. This module determines paths between each pair of edge routers for each Class of Service. To find a path between two routers we implemented Dijkstra, Kruskal and SAMCRA algorithms.

To determine the routing topology of the network ROMAN reads OSPF table from one router of the testbed network. This is achieved by using the database library. For testing purposes we can load the network topology from the XML file. In addition, the implemented module provides information to other modules about network topology and established TE tunnels. This functionality of the ROMAN module is used by RMS and PD-PE modules. The interfaces with other modules have been defined using the SLICE language. Figure 4 shows the data structure used by the interfaces.

Another function of ROMAN is to force routing in the network by properly configuring the routers. This is done by sending commands to the edge routers forming the MPLS TE tunnel, giving an ordered list of routers through which the path passes.

RMS module has been implemented in C++. The implementation uses the standard C++ libraries and the ICE library version 3.3.0 for C++. Figure 5 shows the sequence of messages between ROMAN, RMS, and PD-PE modules during the update of paths and network topology.

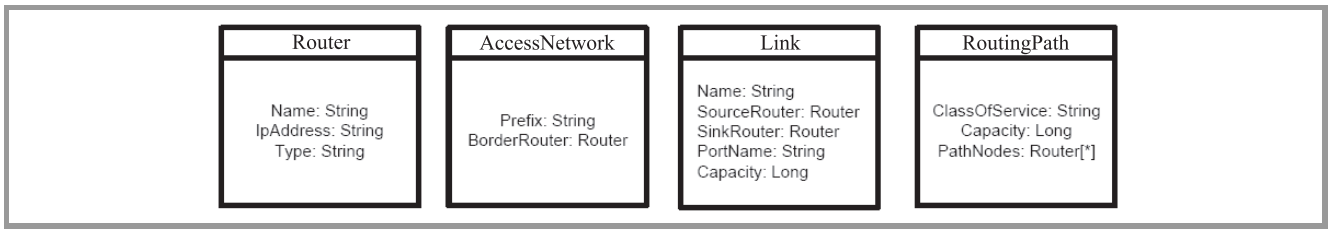


Fig. 4. Data structures used by ROMAN module.

The RomanToRms interface defines the methods topologyHasChanged() and routingPathsHaveChanged() which provide information from ROMAN to RMS module to update network topology and paths, respectively. Moreover, the RmsToRoman interface defines the methods requestTopology() and requestPath() providing download of current network topology and paths.

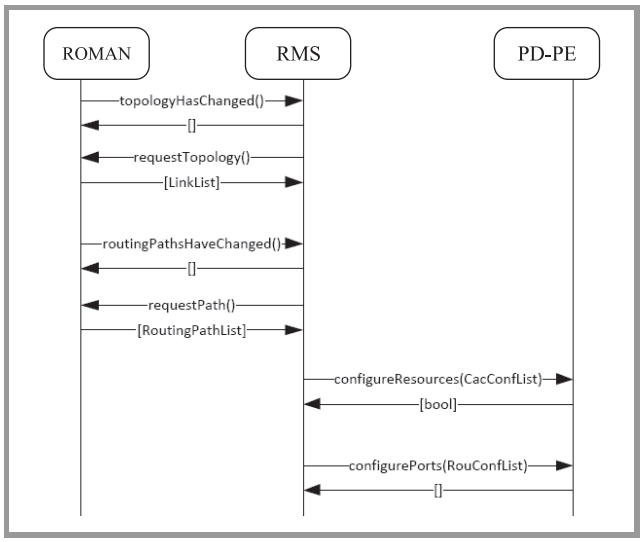


Fig. 5. Sequence of messages – update of paths and topology.

After RMS module receives the new data, it starts call admission control function that calculates the new limits and sets up router output ports. Subsequently, new data are sent to the PD-PE node using the methods configureResources(CacConfList) and configurePorts(RouConfList). CacConfList structure contains the configuration of edge routers for the CAC function (calculated on the basis of the capacity). RouConfList structure contains the configuration used for output ports of the routers.

The RmsToPd interface defines the methods configureResources(CacConfList) and configurePorts(RouConfList) which enable the configuration message from RMS to PD-PE module; specifically, CacConfList contains data for CAC configuration and RouConfList for output ports configuration.

PD-PE, TRC-PP and PE-PE modules were written in Python language. For implementation was used ICE library version 3.3.0 in Python library: readpool, SQLAlchemy, psycopg2, database PostrgreSQL version 8.3.

4. Testbed Network

The described modules of the IP QoS System have been installed in the testbed network. Figure 6 shows the topology of the testbed network in which laboratory tests are performed. Preliminary tests showed that the implemented modules work together correctly and properly configure the network mechanisms for providing DiffServ architecture.

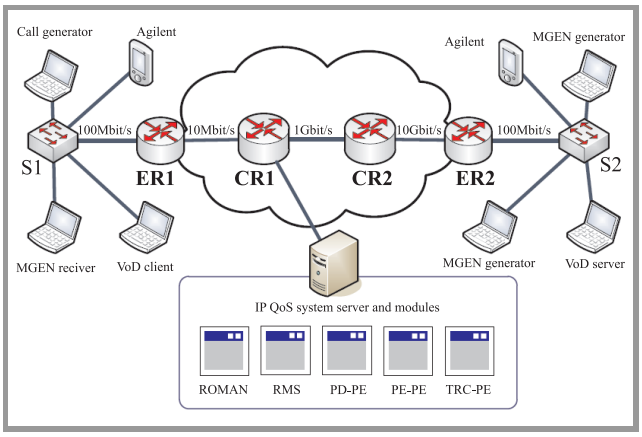


Fig. 6. Testbed network with implemented modules.

The implemented IP QoS System correctly configures the testbed network, which is able to guarantee the QoS parameters set for the proper classes of service. In the following text we expound some issues related to network configuration by the PE-PE and ROMAN modules. Please note that only the ROMAN and PE-PE modules configure the network mechanisms in the network devices. The other modules in the IP QoS System allow proper operation of the whole system.

4.1. MPLS Path Configuration in Testbed Network by ROMAN Module

The implemented software of IP QoS System is tested and demonstrated in the laboratory network as described in [9], [10]. ROMAN requires that edge and core routers

had been pre-configured to send MPLS traffic. In particular, it should be possible to implement OSPF routing and MPLS on the physical interfaces:

 # interface < configured interface name e.g. "GigabitEthernet0/1" >
 # mpls ip
 # mpls traffic-eng tunnels
 # ip rsvp bandwidth < interface bandwidth > < flow bandwidth >

The next step is to create a tunnel between a pair of boundary routers:

 # interface Tunnel1
 # description tunel1
 # tunnel destination < loopback interface ip address of router terminating the tunnel >
 # tunnel mpls traffic-eng path-option 1 explicit name <path name>
 # tunnel mpls traffic-eng record-route
 # no routing dynamic

Each router in the laboratory network has IP addresses associated with physical interfaces and logical address of the router, which is unique throughout the network (Loopback0 logical interface). ROMAN uses Loopback0 interface addresses for configuring MPLS tunnels.
In the laboratory ROMAN application runs in the initial phase of network configuration. This module detects network topology, discovers paths from routing and sets MPLS tunnels. Configured topology with MPLS paths are sent to the RMS module.

4.2. Configuration of the Monitoring Mechanism in Testbed Network by PE-PE Module

Monitoring mechanisms in testbed network are implemented by a single token bucket for real time class of service. To configure this mechanism we should set peak rate value with burst size value [11]. Conforming packets are marked with proper DSCP value, whereas exceeding ones are rejected. Sample commands for configuring Cisco routers mechanism for version 12.1 are listed below:

 # configure terminal
 # ip access-list extended <rule name>
 # permit UDP host <source IP address> eq <port number> host <sink IP address> eq <port number>
 # exit
 # interface <router interface that will be configured>
 # rate-limit output access-group <rule name> <peak rate> <burst size> <burst size> conform-action set-dscp-transmit <number DSCP> exceed-action drop

As we can see from the above list, first we define the group to which we assign the UDP stream between two routers and ports in the network. Then we configure the router interface properly to indicate conforming packets (compatible with token bucket mechanism) with proper DSCP value.

4.3. Configuration of the Scheduling Mechanism in the Testbed Network by PE-PE Module

The first step is the definition of the classes of service. For the exemplary real time CoS, the instructions would be as follows:

 # class-map PbzRealTime
 # match dscp ef cs4

Next, the router must be configured (example for real time class configuration) [11]:

 # policy-map <policy name>
 # class PbzRealTime
 # priority < bit rate allocated to the class of service >
 # exit
 # exit
 # interface < router interface that will be configured >
 # service-policy output <policy name>
 # hold-queue < total buffer size allocated to router interface > out

The mechanisms used in the testbed correspond to these ones which are accessible by Cisco routers. The presented commands are intended to illustrate how we use router mechanisms in the testbed network. It should be noted that implemented IP QoS System can automatically configure a testbed network in accordance with the requirements of guaranteed quality of service.

5. Summary

During the project we carried out preliminary tests of functionality and cooperation of IP QoS System modules. These tests confirmed the correct implementation and cooperation of the modules described in this article. During the tests, we examined not only network configuration, but also the performance of implemented system.
IP QoS System was presented at the conference Krajowe Sympozjum Telekomunikacji i Teleinformatyki 2010 – KSTiT 2010. The exhibition demonstrated the performance of implemented IP QoS System with all modules, calls generator and network analyzer. The exhibition presented a test VoD application with QoE Telchemy analyzer and test VoIP application with MOS Agilent analyzer.
The described implementation of IP QoS System shows that the proposed solution could be used by network operators. However, it should be noted that the implemented system is designed for research and not for commercial purposes. For this, a solution for commercial purposes should be written from the beginning, in order to ensure not only correct work but, especially, effective performance of the system.

At last, let us remark that the proposed solution is limited to single domain network and it does not take into account different network access technologies. The proposed system in future studies could be extended by a further element like access networks such as, e.g., wireless network 802.11 standard.

References

[1] "An Architecture for Differentiated Services", RFC 2475.

[2] H. Tarasiuk, W. Burakowski, A. Jajszczyk, and R. Stankiewicz, "Specyfikacja algorytmów i mechanizmów sterowania ruchem na poziomie pakietów w sieci IP QoS", Raport PBZ-MNiSW-02-II/2007/WUT/B.2/B.6, 2009.

[3] "Multi-Protocol Label Switching (MPLS) Support of Differentiated Services", RFC 3270.

[4] "Maximum Allocation Bandwidth Constraints Model for Diffserv-aware Traffic Engineering", RFC 4125.

[5] "OSPF Version 2", RFC 2328.

[6] P. Pyda, T. Dalecki, "Realizacja rutingu w sieci IP QoS", *Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne*, zeszyt 8–9, s. 1922–1923, 2009.

[7] P. Van Mieghem , H. De Neve, and F. A. Kuipers, "Hop-by-hop Quality of Service routing", *Comput. Netw.*, vol. 37. no. 3-4, pp. 407–423, 2001.

[8] F. A. Kuipers, "Quality of service routing in the Internet: Theory, complexity and algorithms", Ph.D. thesis, Delft University Press, The Netherlands, 2004.

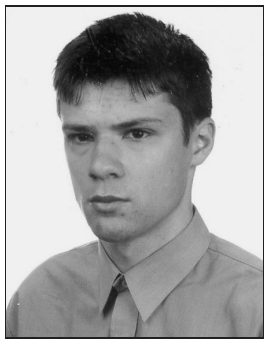
[9] H. Gut, W. Latoszek, M. Gajewski, J. Saniewski, E. Niewiadomska-Szynkiewicz, T. Wiśniewski, P. Arabas, and M. Rotnicki, "Specyfikacja i instalacja sieci laboratoryjnej IP QoS", Raport PBZ-MNiSW-02-II/2007/WUT/B.8, 2009.

[10] H. Gut, W. Burakowski, W. Latoszek, P. Gielmuda, J. Śliwiński, H. Tarasiuk, W. Góralski, A. Bęben, and P. Krawiec, "Integracja oprogramowania i instalacja w sieci laboratoryjnej IP QoS – część II", Raport PBZ-MNiSW-02-II/2007/WUT/D.6, 2010.

[11] W. Burakowski, J. Śliwiński, A. Bęben, H. Tarasiuk, P. Krawiec, "Implementacja modułu do wspomagania konfiguracji sieci", Raport PBZ-MNiSW-02-II/2007/WUT/D.4, 2010.

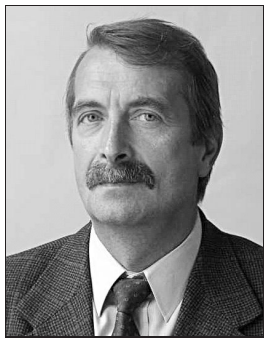
Tomasz Dalecki received the M.Sc.Eng. degree from the Warsaw University of Technology in 2002. Since 2003 he works in Military Communication Institute. His research interests cover management systems design and implementation and security in IP-based systems.

E-mail: t.dalecki@wil.waw.pl
Military Communication Institute
Warszawska st 22A
05-130 Zegrze Płd., Poland



Waldemar Latoszek was born in Otwock, Poland, in 1981. He received engeneer degree from Warsaw University of Technology in 2005. Since 2005 he works in National Institute of Telecomunication. His research interests cover network architectures, testbeds, traffic control.

E-mail: w.latoszek@itl.waw.pl
National Institute of Telecommunication
Szachowa 1
04-894 Warsaw, Poland



Henryk Gut was born in village Zub-Suche, not far from Zakopane, in 1951. He received M.Sc. degrees in Telecommunications from Warsaw University of Technology in 1975. Directly after graduation he began working for National Institute of Telecommunication, where he is employed until now. During working for NIT he was dealing with following topics: technical structure and topology optimization of data transmission network; methods and systems for bit, frame and clock synchronization; modeling of primary bit error processes in binary data channels; designing of terminal and network equipments for national data network SYNCOM and paging network POLPAGER. In period 1998–2006, his research and development activity was focused on utilization power lines and in-home electrical installations as an transmission medium for broadband access and in-home data networks. Recently he is participating in realization two national grate projects: "Next Generation Services and Networks – technical, application and market aspects. Network Traffic Management – IP network with full QoS guarantee" and "Future Internet Engineering". He is author or co-author of about 35 papers published in national journals and conference proceedings and is co-author of a few dozen internal technical reports of NIT.
E-mail: h.gut@itl.waw.pl
National Institute of Telecommunication
Szachowa 1
04-894 Warsaw, Poland

Witold Góralski – for biography, see this issue, p. 20.

Piotr Pyda, Jordi Mongay Batalla – for biographies, see this issue, p. 11.

Jarosław Śliwiński – for biography, see this issue, p. 10.

Paper

QoS Conditions for VoIP and VoD

Przemysław Dymarski, Sławomir Kula, and Thanh Nguyen Huy

Institute of Telecommunications, Warsaw University of Technology, Warsaw, Poland

Abstract—This paper concerns quality evaluation of the telecommunication services: VoIP (representing the RT interactive class) and VoD (representing the MM streaming class). Subjective and objective methods and tools for perceived quality measurement are analyzed and compared. Subjective tests are performed for selected video sequences using the Double-Stimulus Impairment Scale (DSIS) method. Thus the objective algorithms (VQM and VQmon) are calibrated. Speech quality is measured using the objective methods: PESQ and POLQA. Threshold values for network parameters (packet loss rate, delay jitter) are set, that guarantee the acceptable service quality.

Keywords—delay jitter, packet loss rate, PESQ, POLQA, quality of service, VoD, VoIP, VQM, VQmon.

1. Introduction

Quality of telecommunication services grows in importance not only from user point of view. The services providers and operators take into account quality as an element of competition. Services can be delivered using different networks, protocols, devices etc. Quality of service depends on many factors, like the network transmission parameters such as throughput (bandwidth), bit error rate (BER), packet loss rate (PLR), delay, and delay jitter. Influence of these parameters on quality depends on a service. Thus, different end-to-end classes of service were introduced [1]. The examples of such classes are the real time (RT) services like VoIP or videoconference, the multimedia streaming like VoD or IPTV, the high throughput data like FTP and the standard services like email. The requirements concerning network parameters for each class of service were specified in the ITU-T and ETSI recommendations [2], [3], [4]. In DiffServ network [5] the threshold values of network parameters have to be defined, which guarantee the acceptable quality perceived by the end user. Differentiated services enable a scalable service discrimination and the potential users who may violate these threshold values are rejected.

In this paper two examples of telecommunication services are considered, namely the VoIP (representing the class of real time applications) and VoD (representing the class of multimedia streaming) – both accessed by the IP network. Acceptable quality of these services can be achieved by setting appropriate threshold values of network transmission parameters. According to our observations ([6], [7]) the requirements specified in [2], [4] do not always reflect user's preferences. Therefore we start with an analysis of norms and tools for video signal and speech signal quality measurement (Subsections 2.1, 3.1). Subsections 2.3 and 3.1 are dedicated to a calibration of the selected video quality metrics, and setting up credibility conditions for speech quality metrics. Setup of the laboratory stands and methodology of testing the influence of the IP network transmission parameters on speech and video quality are presented in subsections 2.4 and 3.2. Results of these tests as well as the proposed threshold values of transmission parameters are discussed in Section 4.

2. QoS in Multimedia Streaming

2.1. Recommendations and Tools for Video Quality Evaluation

The metrics of video signal quality should reflect the opinion of the end user. Therefore the subjective tests are more credible than the objective quality measures. On the other hand, the subjective tests are more difficult to conduct, they involve a group of participants, require the special acoustic conditions, they are more costly and time-consuming. The subjective tests are described, e.g., in the ITU-T Recommendation P.910 [8] and the ITU-R Recommendation BT.500-12 [9]. In our work they are used to calibrate the selected objective measures.

We have applied the Double-Stimulus Impairment Scale (DSIS), due to relative simplicity of this test. Each participant observes first the reference video sequence and then the sequence to be assessed. Then he/she evaluates the loss of quality according to the mean opinion score (MOS) scale from 1 to 5 (where 5 – no degradation, 1 – unacceptable level of distortions).

The objective quality measures may use the reference video sequence (media based methods), the incoming packets (on line quality evaluation) or the information concerning network structure, codecs etc. (parametric methods). The media based methods may be divided into the following groups.

- The full reference methods (also called the intrusive methods) use the whole reference video sequence and compare it with the tested sequence.
- The reduced reference methods use only some parameters of the original video sequence.
- The no-reference methods (also called the non-intrusive methods) have no access to the original sequence.

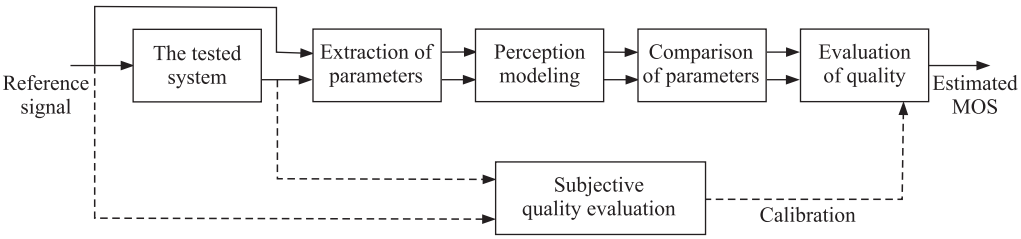


Fig. 1. General scheme of the objective full reference quality evaluation with calibration based on subjective tests.

The full reference quality evaluation is the most credible one – the selected algorithms of this kind are recommended by the ITU-T [10], [11]. The ITU-T Recommendation J.144 [10] presents a series of quality evaluation algorithms without pointing the best one. All the algorithms of this kind may have a similar structure shown in Fig. 1.

The algorithms described in this recommendation may be used for testing the video signals of a relatively high quality, e.g., the cable TV at the bit rates from 768 kbit/s to 5 Mbit/s. These algorithms were not thoroughly tested in presence of the channel errors (e.g., lost packets), therefore they are not recommended to quality evaluation of video sequences transmitted through the channels of low quality. Nevertheless, we have applied one of the J.144 algorithms, namely the Video Quality Metric (VQM), to test the VoD quality. This was possible due to the calibration described in the Subsection 2.3.

The VQM has been proposed by the Institute for Telecommunication Science (ITS), collaborating with the National Telecommunications and Information Administration (NTIA) [12]. It is based on the simplified human visual system model, particularly the spacial and temporal contrast perception.

In order to improve the accuracy and widen the application range of the objective video quality evaluation algorithms, the ITU-T started a new competition, in which the following institutions have taken part: NTT, OPTICOM, Psytechnics, Yonsei University and SwissQual. Finally, the ITU-T proposed:

1. As the full reference methods, recommend four algorithms: NTT, OPTICOM, Psytechnics and Yonsei University. These algorithms are described in the norm J.247 [11].

2. As the reduced reference method, recommend the Yonsei University algorithm. It is described in the norm J.246 [13].

3. Not recommend any of the no-reference methods despite of the relatively good results obtained by SwissQual.

The above mentioned algorithms may be used for evaluation of quality of video signals transmitted through channels of a low quality (packet loss, delay jitter etc.). They have sophisticated synchronization tools for alignment of both video sequences: the reference one and the tested one. The spacial alignment makes it possible to compare the cropped images with the full size images. After the temporal and spacial alignment a series of parameters is extracted from both sequences: they concern luminance, chrominance, edges, block effects etc. The human visual system models are used to compare these parameters in order to calculate the final quality measure using the MOS scale. The algorithm proposed by the Yonsei University is mainly based on the edges processing, therefore it does not require the full reference video sequence, only some information about edges (1 kbit/s do 128 kbit/s, depending on the video sequence). That is why it has been recommended as a reduced reference algorithm.

Unfortunately we had no access to the J.247 and J.246 algorithms, so we have decided to calibrate the VQM, being a part of the J.144 norm.

For the on line quality control, the no-reference methods are useful, particularly the methods based on the IP packets analysis. These algorithms use information concerning the lost packets (some of them identify the coder and analyze the influence of the lost packets on the image quality), the corrupted packets, delay jitter etc. An example of such algorithm is the VQmon/HD distributed by the Telchemy Inc. [14]. The VQmon/HD is used for monitoring of the IPTV, videoconference and VoD quality. It supports the RTP and UDP protocols and many video and audio coding schemes. Each packet is identified as the audio or video I, B or P packet and its influence on the audio/video quality is estimated. The following quality measures are calculated: MOS-A (audio), MOS-V (video) and MOS-AV (aggregated audio and video). The video quality metrics (MOS-V) are evaluated in a relative (mainly the transmission quality is considered) and absolute (not only the transmission but also codec parameters are considered) form. Moreover the instantaneous and average MOS values are delivered. Further comments concerning the VQmon/HD quality metrics will be presented in Subsection 2.3.

For the network planning purposes the parametric quality evaluation algorithms are used. They consider the codec parameters (coding scheme, bit rate) and the parameters of the communication link (bandwidth, packet loss rate, delay, delay jitter) and do not require any measurements. For telephony the E-model (ITU-T Recommendation G.107) and for multimedia the ITU-T Recommendation G.1070 is used.

2.2. Test Procedure to Determine Threshold Values of Transmission Parameters

Our purpose was to determine threshold values of transmission parameters adequate for acceptable perceptual quality of multimedia streaming. It could be done using existing networks and subjective methods of quality evaluation. Unfortunately, such procedure is complicated, very laborious, and time consuming. Because of this we have used a network emulator and objective quality evaluation methods. To test the influence of transmission parameters on perceptual quality of multimedia streaming we have decided to proceed as follows:

1. Calibrate of selected metrics for objective measurements by using the subjective tests.
2. Emulate network and perform multimedia streaming.
3. Using objective methods evaluate perceptual quality of perceived multimedia (video sequence with accompanying audio).
4. Determine threshold values of transmission parameters yielding the acceptable quality.

These steps will be described in the following subsections.

2.3. Calibration of Metrics for Objective Measurements

Because of unavailability of the attested software of the J.247 algorithms [11] we decided to use two objective tests, namely PSNR and VQM (the latter being a part of the J.144 norm). However, the calibration procedure was necessary, in order to express the quality estimates in a MOS scale. The calibration was performed in the following steps:

1. Selection of video material.
2. Using the network emulator (Netem) and streaming application (VLC) for preparation of the distorted video sequences.
3. Performing of the objective tests using the PSNR and VQM quality metrics.
4. Installing the appropriate video display software (MSU video quality measurement tool [15]) on six PCs and preparation of the quality evaluation tasks.
5. Selection of subjects (viewers who evaluate quality of video sequences).
6. Performing of the tests using the DSIS method.
7. Conversion of the objective quality measures to the MOS scale using the results of the objective tests.

As a test material five sequences, MPEG-2 – coded, of size 640×480, and bit rates form about 700 kbit/s to 7000 kbit/s were selected. Laboratory stand for video on demand quality monitoring consisted of two workstations with VLC application to stream and receive video sequences. VLC media player [16] is a free of charge application, which supports various audio and video codecs (MPEG-1, MPEG-2, MPEG-4, DiviX, MP3, OGG Vorbis etc.) and transmission protocols like UDP and RTP. On the receiver workstation the Telchemy's VQmon [14] application (for on-line video evaluation) and video quality measurement tools [15] included VQM and PSNR metrics were installed.
Network and its parameters were emulated by Netem [17] which is a part of Linux system. In our work Netem was used to change the following transmission parameters: bit error rate, packet loss rate, bit rate, and delay jitter.
Subjects were 60 students of the Electronic and Information Technology Faculty (Warsaw University of Technology). Short instruction was given to subjects on arrival for their first visit in the laboratory. The subjects were informed on ideas of the subjective method and the objective methods, and on the measurement procedure.
The subjective method based on DSIS [8], [9] was used. In DSIS subjects watch two video sequences, the original one and the transmitted one. Subjects evaluate quality using the MOS scale recommended by ITU. The scale is given in Table 1.

Table 1
Viewing quality scale for DSIS method

MOS	Quality loss
5	Imperceptible
4	Perceptible, but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

Conversion of PSNR objective metric to MOS is based on finding a proper approximation function. It is relatively easy to find such function under assumption that the function is linear with the minimum value 1 and the maximum value 5. In Fig. 2 the subjective MOS values versus

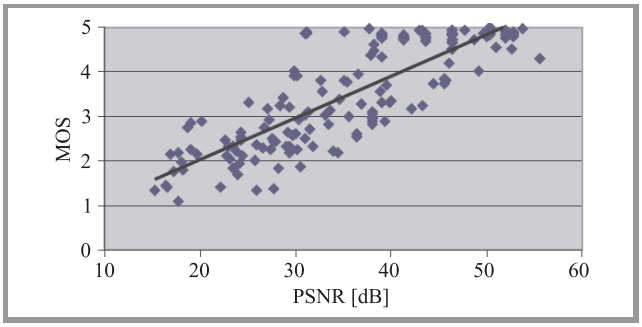


Fig. 2. Subjective MOS versus PSNR.

the measured PSNR [dB] values are given. Note that any point in this figure is an average of scores obtained by many subjects. Indeed, data presented in Fig. 2 suggest the linear

approximation. Using the least mean squares approach the following conversion function is obtained:

$$\text{PSNR\_MOS} = 0.0935\ \text{PSNR} + 0.152. \quad (1)$$

In the case of VQM finding of an approximation function is more difficult because of nonlinearity. In Fig. 3 the subjective MOS versus the measured VQM values are given. Data presented in Fig. 3 suggest the logarithmic approximation curve. Using the least mean squares approach the following conversion function is obtained:

$$\text{VQM\_MOS} = -0.8634\ \ln(\text{VQM}) + 3.4854, \quad (2)$$

where ln – natural logarithm.

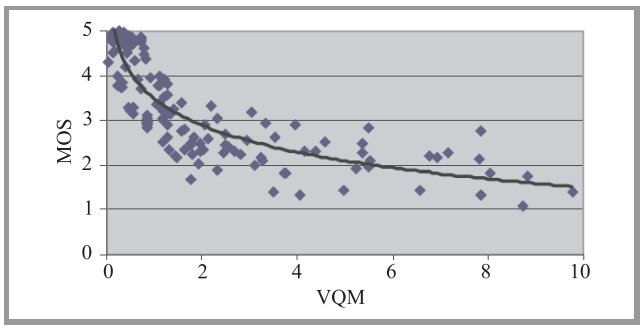


Fig. 3. Subjective MOS versus VQM.

Which estimate, PSNR_MOS or VQM_MOS, is better? In Fig. 4 and Fig. 5 are presented comparisons of the objective estimates: PSNR_MOS and VQM_MOS with the subjective MOS. The linear mapping ($y = x$) is also shown. To answer this question Pearson's correlation was calculated.

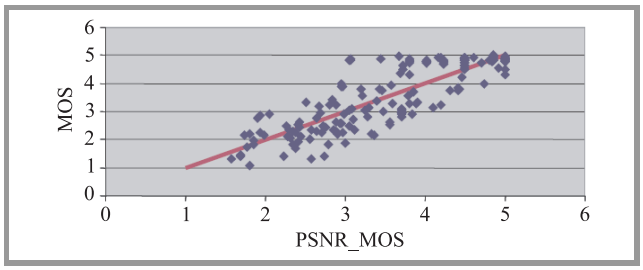


Fig. 4. Subjective MOS versus PSNR_MOS.

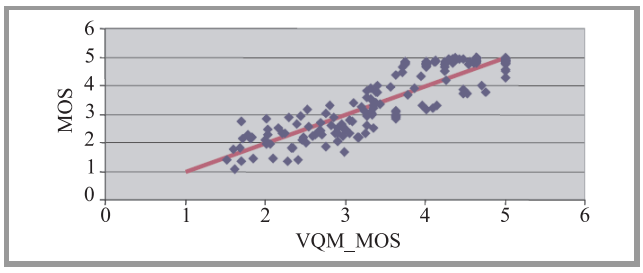


Fig. 5. Subjective MOS versus VQM_MOS.

Calculation of Pearson's correlation needs centering of the sets x_i (e.g., PSNR_MOS values) and y_i (subjective MOS values) – thus the centered data $\hat{x}_i$ and $\hat{y}_i$ are obtained. Then, the correlation is calculated:

$$R_{xy} = \frac{\sum_i \hat{x}_i \hat{y}_i}{\sqrt{\sum_i (\hat{x}_i)^2 \ \sum_i (\hat{y}_i)^2}}. \quad (3)$$

The Pearson's correlation values for PSNR_MOS and VQM_MOS equal 0.849 and 0.883, respectively, so better MOS approximation is obtained with the VQM_MOS.

For the on-line quality assessment, we find the VQmon distributed by the Telchemy Inc. [14] very useful. However, some measures must be taken, to obtain stable and credible results. VQmon delivers packets of results regularly, at a time interval set up by the user. In order to obtain stable results at low packet loss rates, longer measurement intervals should be used. Despite of this, there is sometimes an initial unstable phase, i.e., the first packets of results show lower MOS-V values than the subsequent ones. This concerns not only the instantaneous MOS-V values, but also the averaged values. We have used the measurement intervals of 30 s and we have ignored the initial packets of results, so we have obtained the credible results in most cases.

If a delay jitter causes a drop of quality, the MOS-V values may not reflect the image quality, because it depends on the size of the receiving buffer. VQmon is not informed about the buffer size, because in analyzes the incoming packets before buffering.

In case of corrupted (but not lost) packets, the quality drops, but MOS-V values are high, suggesting a good quality. This is probably because the VQmon analyzes mainly the packet headers, and is not sensitive to corruption od data.

2.4. Influence of IP Network Parameters on Video Signal Quality

For testing the influence of the network parameters on video signal quality, the same laboratory stand as for calibration of metrics was used, i.e., two workstations with VLC application and MSU video quality measurement tools [15] with VQM metric and a server with Netem network simulator. However the number, variety and length of video sequences were much higher. The tested sequences were divided into six categories. Their features are given in Table 2.

Each sequence was transmitted by network emulator. Transmission parameters (BER, PLR, channel throughput, delay jitter) were changed gradually. Quality of video sequence was measured using VQM for each transmission parameter separately. Results were converted to VQM_MOS and averaged for all the sequences – the example of is shown in Fig. 6. Negative influences on video were not being observed for the threshold values of transmission parameters. In Table 3 the obtained threshold values are given. These thresholds are the "worst case" values – they do not stem directly from the average results (like those presented in Fig. 6), but guarantee (according to our tests) lack of distortions in all of the observed video sequences.

According to our results a service provider should guarantee that the transmission parameters are below (BER,

Table 2
Categories of tested video sequences

Cat.	Description	Content	Audio
A	Speaking people	Small and slow changes in picture	Speech
B	Publicity-graphic	Animated cartoons	Music and speech
C	Landscape	Slow movement of camera	Different
D	Pop music	Video clips	Song and music
E	News-reports	Speaker and short videos	Speech, background music
F	Sport	Dynamic changes of picture	Speech, background noise

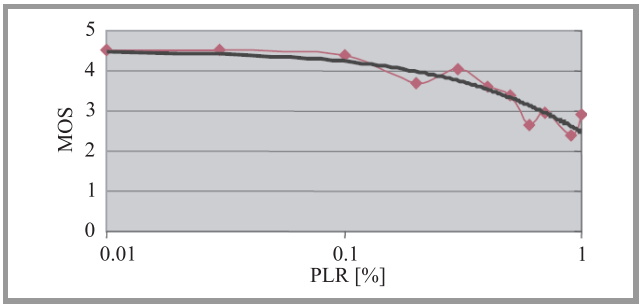


Fig. 6. The average VQM_MOS versus PLR – measured and approximated values.

Table 3
Threshold values for transmission parameters

Parameter	Threshold
BER	0.03%
PLR	0.06%
Throughput	max bit rate*
Delay jitter	0.05 ms
* max. instantaneous video sequence bit rate.	

PLR, jitter) or above (throughput) threshold values given in Table 3. Note that the ITU-T and ETSI Recommendations [1], [2] set the PLR threshold at 1%, which, in our opinion, is too high. However, the Recommendation Y.1541 [3] defines the provisional QoS classes demanding PLR $< 10^{-5}$, but the base threshold is still at 1%. The Recommendation J.241 defines quality levels for videostreaming services. If the PLR is greater than 0.02% quality is referred as poor, the excellent quality is obtained for PLR $<$ 0.001%. Our result (PLR $<$ 0.06%) is somewhat less strict, but we agree that some margin should be used, in order to guarantee very good quality of videostreaming services. For the xDSL networks much more strict conditions are formulated, e.g., PLR $< 10^{-6}$ for SD and PLR $< 10^{-7}$ for HD video transmission [18].

3. QoS in Voice over IP

3.1. Analysis of the Objective Measures of Speech Quality

For quality assessment of the telecommunication services based on speech transmission the media based methods are mainly used. The most popular full reference algorithm, Perceptual Evaluation Of Speech Quality (PESQ) is described in the ITU-T Recommendation P.862 [19]. The algorithm has access to the original speech phrase and the processed one. At the first stage time-domain synchronization of both phrases is accomplished. Then a series of speech parameters, which influence the human perception, are extracted from both signals. These parameters are defined in frequency domain (human ear is not sensitive to phase of the audio signal), using nonuniform scale (thus modeling the basilar membrane in the ear). Then the psychoacoustic representations of both signals are compared, using a human perception model (psychoacoustic model). Mainly the masking phenomena in time and frequency domain are considered in such model. The aggregated result of this comparison, called the Raw MOS, takes values from –0.5 (a big difference of both signals, suggesting a completely unacceptable quality of the tested phrase) to 4.5 (no perceptible difference between both phrases). At last, the Raw MOS is converted to the listening quality MOS (MOS-LQO) which takes values from 1.02 to 4.56 and maximizes correlation with the results of subjective tests.

In order to increase credibility of PESQ MOS-LQO values, ITU has specified conditions in which the measurements have to be performed. These conditions are described in the Recommendation P.862.3 [20]:

- Recommended phrase duration is 8 – 12 s, accepted 3.2 – 30 s, in any case it should not exceed million samples.
- In order to reduce the influence of speaker on the quality assessment results, phrases from 2 feminine and 2 masculine speakers should be used.
- Pure speech signal should take 40% – 80% of the whole phrase (the rest contains initial, inter-word, and final silence), there should be at least 3.2 s of active speech.
- The initial and final silence should last from 0.5 to 2 s. In both phrases being compared, difference of duration of corresponding silences should not exceed 25%.

In the Institute of Telecommunications, Warsaw University of Technology, a series of experiments were performed, in

order to confront the PESQ MOS-LQO values with subjectively assessed quality [7], [21]. The greatest discrepancies were observed if a voice activity detector (VAD) was simulated, which substituted zero-valued samples for silent intervals of the phrase (Fig. 7). Despite of a slight cropping of the initial or final consonants of some words, speech quality was almost unchanged, according to informal listening tests. However, the PESQ_MOS values dropped considerably, almost achieving 2, thus suggesting annoying distortions. We concluded, that PESQ results are not credible if the VAD is applied.

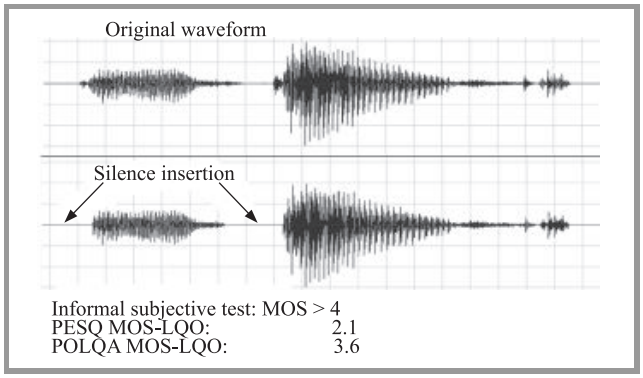


Fig. 7. Influence of the voice activity detector on the PESQ_MOS and POLQA_MOS [21].

The influence of the speaker and the phrase on the PESQ_MOS values is considerable: the results obtained for the same speech coder may differ by a unit on the MOS scale – see Fig. 8. Therefore it is necessary to increase number of speakers and phrases (in comparison with those recommended in [20]), in order to obtain credible average results. In our tests we have used 4 phrases and 4 speakers (2 men and 2 women) – in total 16 phrases.

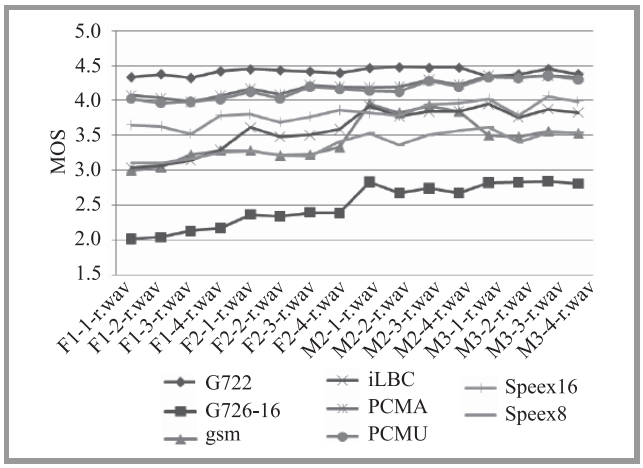


Fig. 8. Influence of the phrase and speaker on the PESQ_MOS [7].

In case of quality testing at low BER or PLR values, the number of phrases and their duration should be increased, because of the scatter of PESQ_MOS values due to random bit and packet loss process. This is illustrated in Fig. 9, where two series of tests were conducted, using the same coder (G.711 PCM) and PLR = 1%. This confirms our decision to use 16 phrases in our tests.

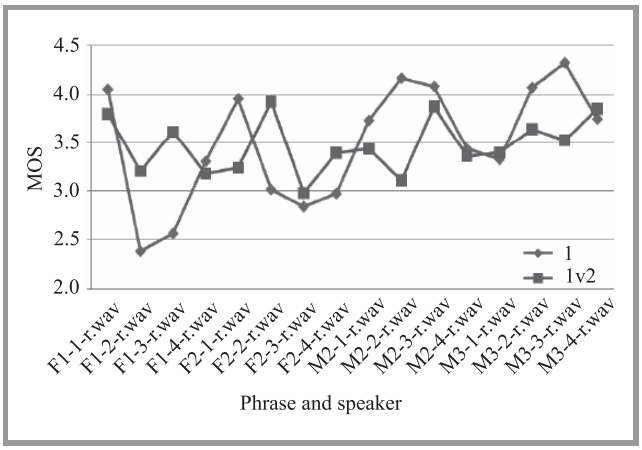


Fig. 9. PESQ_MOS for short speech phrases – PLR=1% [7].

We have also observed some synchronization problems: by increasing or decreasing the inter-word silent intervals the PESQ_MOS values changed despite of no change in subjectively assessed quality [7], [21].

The new algorithm for the objective speech quality evaluation, Perceptual Objective Listening Quality Analysis (POLQA) [22] is an improved version of PESQ. It may be used for quality measurement of speech signals of the bandwidth 4 kHz, 8 (or 7) kHz and 16 kHz. This method has an improved synchronization system and, unless the PESQ algorithm, may be used for the enhanced variable rate coders (EVRC) applied in CDMA systems. We have obtained a one-month license for the POLQA software from the Telchemy, Inc., and we observed, that the POLQA_MOS values exhibit greater correlation with the subjectively evaluated quality than the PESQ_MOS values. In particular, the quality assessment of phrases passed through the VAD simulator were much more realistic (POLQA_MOS = 3.6 while PESQ_MOS = 2.1 – see Fig. 7). Therefore we conclude that POLQA should be used instead of PESQ as a full reference algorithm.

For the on-line speech quality testing, ITU-T has recommended the 3SQM method [23]. It is a non-intrusive method, which does not require the original speech phrase. Speech quality assessment is based on the analysis of the processed phrase: the time-domain discontinuities, increased noise level, non-speech spectra are detected and then the dominant distortion source is found (the listener evaluates the speech quality out of this dominant distortion, usually ignoring the less annoying ones). Then the 3SQM_MOS value is calculated. Despite of relatively high correlation of 3SQM_MOS and PESQ_MOS, the intrusive methods, like PESQ and POLQA yield better accuracy of quality estimation. Therefore, these methods will be used for tests reported in Subsection 3.2.

3.2. Influence of IP Network Parameters on Speech Signal Quality

For testing the influence of the network parameters on speech signal quality, the server with Netem [17] network simulator and two workstations with Ekiga soft-phone applications were used. Ekiga [24] is a tool for VoIP and video-conference communication using SIP and H.323 protocols. It supports many speech coders, like G.711 PCM, Speex, iLBC, GSM-EFR and G.726 ADPCM (the latter with bit rates of 16, 24, 32 and 40 kbit/s). The wideband (speech bandwidth 7 kHz) speech coders Speex and G.722 are also supported.

The original phrase is read from the .wav file and sent directly (in digital form) to Ekiga. It is accomplished due to the application of the virtual audio cable (VAC) [25]. In a similar way, using the VAC, the received speech signal is written directly in the .wav file. Elimination of the A/D and D/A conversions is very important, because these operations cause a drop of the measured PESQ_MOS and POLQA_MOS values.

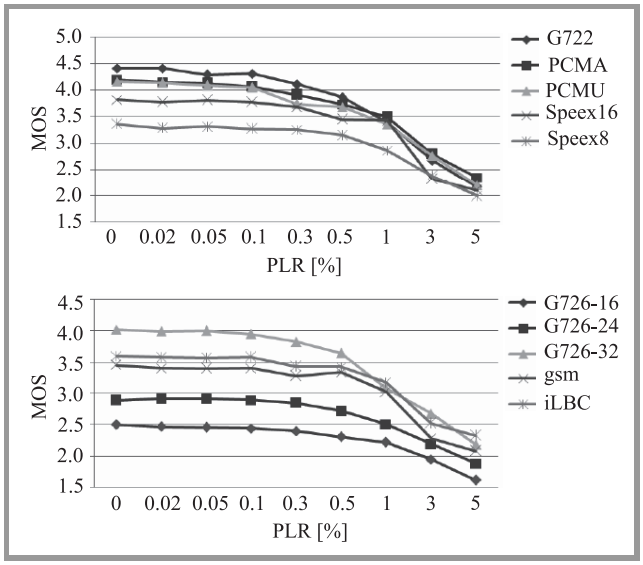


Fig. 10. PESQ_MOS versus PLR for different speech coders [7].

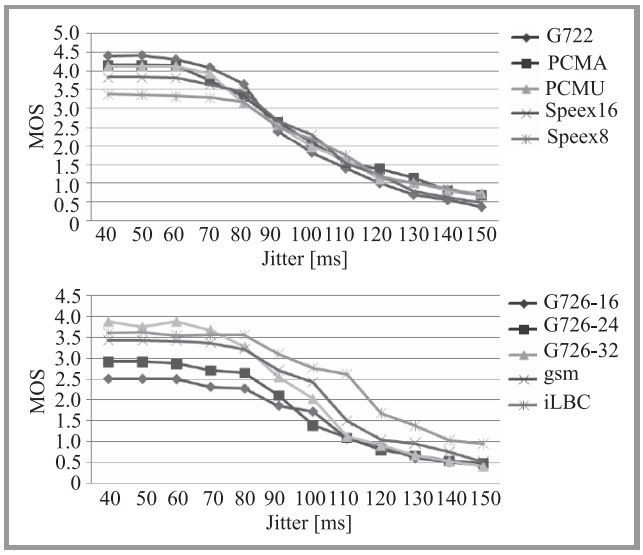


Fig. 11. PESQ_MOS versus delay jitter for different speech coders [21].

In our tests we have used 4 phrases and 4 speakers (2 men and 2 women) – in total 16 phrases concatenated in a single .wav file. In Figs. 10 and 11 the results of tests are shown.

According to our tests, the impact of the packet loss on the speech quality is negligible if PLR $< 0.2\%$. It is much more restrictive condition than the threshold value specified for the conversational voice services in ITU-T Recommendation G.1010 [1], i.e., PLR $< 3\%$. However, in ETSI document [2] similar tests are reported, suggesting that PLR should be less than $0.2\% - 0.5\%$ (depending on speech coder), if MOS > 4 is to be maintained. The ITU-T Recommendation Y.1541 specifies more restrictive threshold value: PLR $< 0.1\%$. Our tests confirm this value.

The delay jitter may cause a drop of speech quality if it is greater than about 60 ms (Fig. 11). In ITU-T Recommendation G.1010 [1] a very restrictive threshold value is specified, namely 1 ms. However in ITU-T Recommendation Y.1541 delay jitter threshold is set at 50 ms, which is also confirmed by our results.

4. QoS Conditions for Selected Communication Services

In this paper two kind of problems are considered: credibility of tools for speech and video quality evaluation and QoS conditions for services based on speech and video signal transmission through the IP networks.

As a tool for the objective full reference speech quality evaluation the PESQ algorithm (ITU-T Recommendation P.862) [19] was examined in detail. According to our tests the credibility conditions specified for this algorithm in Recommendation P.862.3 [20] are not sufficient. In particular, PESQ delivers far too low quality estimation marks (PESQ_MOS values) if a voice activity detector (VAD) is applied. Moreover, the time domain synchronization of two phrases being compared is not perfect, which again yields too low PESQ_MOS values. The number of phrases should be greater than that specified in [20] – instead of 4 phrases we used 16 ones (4 phrases pronounced by 4 speakers). Our comparison of the PESQ algorithm and the newly introduced POLQA (ITU-T Recommendation P.863) [22] reveals that POLQA has better synchronization system and is not so sensitive to modifications introduced by VAD. Experiments illustrated with Fig. 7 show that erroneous drop of MOS is not so considerable as that of PESQ. Therefore we conclude that the POLQA_MOS is a more credible speech quality metric than PESQ_MOS. Our observations concerning the number of phrases and speakers still hold for POLQA algorithm.

PESQ as well as POLQA are intrusive algorithms – selected phrases, known at the receiver's side, must be transmitted through the network. For the on-line speech quality testing,

the 3SQM method [23] is recommended. As a no-reference algorithm, 3SQM is less accurate than PESQ, but a calibration procedure may increase the correlation between the 3SQM_MOS and PESQ_MOS values [26].

For testing the quality of video sequences, we have used a calibrated VQM metric. The calibration process was described in Subsection 2.3. The calibrated VQM values (VQM_MOS) exhibit relatively high correlation with the subjectively evaluated MOS values (Pearson's correlation 0.88). Because of unavailability of the software of the J.247 algorithms [11] we could not make any comparisons using this newly introduced recommendation.

For the on-line video quality evaluation, VQmon of the Telchemy Inc. [14] may be used. In our opinion however, additional conditions should be fulfilled in order to obtain credible results. Stable results are not always obtained at the beginning of the test, therefore it is better to increase the test duration – see Subsection 2.3. The results obtained in presence of packet delay jitter and corruption of the packets' content are not always accurate.

The QoS conditions for telecommunications services based on streaming od video signals were analyzed in terms of the BER, PLR, channel bandwidth (throughput) and delay jitter. The proposed thresholds for BER and PLR (Table 3) are more restrictive that these proposed in ITU Recommendation G.1010 [1], but slightly less demanding as those specified in the ITU-T Recommendation J.241 [4]. According to our tests, video transmission is very sensitive to packet delay jitter. This is due to the UDP protocol which performs no packet numbering and permutation of packets may occur. For the RTP protocol the corresponding threshold would be much higher. It should be considered, that the receiving buffer size may also influence the sensibility of transmission system to delay jitter. So as to the channel bandwidth, we have observed, that any value below the maximum bit rate of a video sequence (in the case of the variable bit rate coding) may cause visible distortions. Therefore we support the opinion, that the channel bandwidth should be greater than the maximum instantaneous bit rate of the transmitted video sequence.

The real time (RT) interactive services based on speech transmission (like VoIP) are less sensitive to BER and PLR than the services based on video transmission. According to our results PLR $< 0.2\%$ enables good speech quality, which is close to threshold value specified in the ITU-T Recommendation Y.1541 (PLR $< 0.1\%$). The quality of the RT interactive services depends on the transmission delay, but in our tests we have skipped this parameter, because the acceptable delay values have been specified in ITU-T Recommendation G.114 [27] and they seem to be credible. According to this recommendation, the one-way delay should not exceed 150 ms. Values less than 250 ms may be accepted, but some users may perceive them as irritating. Our tests have shown, that the delay jitter should be less than 60 ms which is close to the value specified in the ITU-T Recommendation Y.1541 [3]. Note that this value is much greater as the corresponding threshold for video transmission, but this is due to protocols which prevent the permutation of packets.

Comparison of speech coders (Figs. 10 and 11) shows the advantage of the G.722 coder, but it is paid with the relatively high bit rate (64 kbit/s). A good choice for the VoIP service would be the iLBC coder, yielding quite a good quality at bit rates 13 or 15 kbit/s. Note that at the PLR = 1% speech quality of most of the tested coders is similar (MOS about 3.5).

The results presented in this paper were obtained for typical VoIP and VoD hardware and software configurations and currently available tools for speech and image quality evaluation. The telecommunication services as well as tools for quality evaluation are still in phase of development [28]. Therefore the QoS conditions are still being reformulated and specified with greater accuracy.

Acknowledgement

The authors are deeply indebted to all co-authors of the papers quoted, especially A. Sadowska, G. Szmyd, M. Golański and M. Gora, for conducting series of tests. This work was supported by the PBZ-MNiSW-02-II/2007 "Next Generation Services and Networks – technical, application and market aspects".

References

[1] "End-user Multimedia QoS Categories", ITU-T Recommendation G.1010, 2001.

[2] "Speech and Multimedia Transmission Quality (STQ); Audiovisual QoS for Communication over IP Networks", ETSI ES 202 667, 2009.

[3] "Network Performance Objectives for IP-based Services", ITU-T Recommendation Y.1541, 2006.

[4] "Quality of Service ranking and measurement methods for digital video services delivered over broadband IP Networks", ITU-T Recommendation J.241, 2005.

[5] W. Burakowski, J. Śliwiński, H. Tarasiuk, A. Beben, P. Krawiec, S. Kula, P. Dymarski, S. Kaczmarek, M. Narloch, H. Gut-Mostowy, W. Latoszek, P. Pyda, T. Dalecki, E. Niewiadomska-Szynkiewicz, P. Arabas, M. Rotnicki, and T. Wiśniewski, "Specyfikacja Systemu IP QoS opartego na architekturze DiffServ" (in Polish), in *Proc. KSTiT, Przegl. Telekom.*, vol. 8-9, pp. 966–972, 2009.

[6] S. Kula, P. Dymarski and G. Szmyd, "Wpływ parametrów sieci na jakość sygnałów wideo (in Polish)", in *KSTiT, Przegl. Telekom.*, vol. 8-9, pp. 796–799, 2009.

[7] P. Dymarski, S. Kula and A. Sadowska, "PESQ jako narzędzie do oceny jakości sygnału VoIP (in Polish)", *KSTiT, Przegl. Telekom.*, vol. 8-9, pp. 1299–1308, 2010.

[8] "Subjective Video Quality Assessment for Multimedia Applications", ITU-T Recommendation P.910, 1999.

[9] "Methodology for the Subjective Assessment of the Quality of Television Pictures", ITU-R Recommendation BT.500-12, 2009.

[10] "Objective Perceptual Video Quality Measurement Techniques for Digital Cable Television in the Presence of a Full Reference", ITU-T Recommendation J.144, 2004.

[11] "Objective Perceptual Multimedia Video Quality Measurement in the Presence of a full Reference", ITU-T Recommendation J.247, 2008.

[12] Y. Wang, "Survey of Objective Video Quality Measurements", Department of Computer Science, Worcester Polytechnic Institute, June 2006.

[13] "Perceptual Visual Quality Measurement Techniques for Multimedia Services over Digital Cable Television Networks in the Presence of a Reduced Bandwidth Reference", ITU-T Recommendation J.246, 2008.

[14] "TVQM video quality metrics", *Understanding IP Video Performance*, Feb. 2008, Telchemy application note.

[15] "MSU Video Quality Measurement tools", MSU Graphics and Media Lab (Video Group), May 2011 [Online]. Available: http://compression.ru/video/quality_measure/index_en.html

[16] "VLC Media Player", VideoLan Organization, May 2011 [Online]. Available: www.videolan.org/vlc

[17] "Netem", The Linux Foundation, May 2011 [Online]. Available: www.linuxfoundation.org/collaborate/workgroups/networking/netem

[18] DSL Forum, "Triple-play Services Quality of Experience (QoE) Requirements", Technical Report TR-126, 2006.

[19] "Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs", ITU-T Recommendation P.862, 2002.

[20] "Application Guide for Objective Quality Measurement Based on Recommendations P.862, P.862.1 and P.862.2", ITU-T Recommendation P.862.3, 2007.

[21] A. Sadowska, "Algorytm PESQ jako narzędzie do oceny jakości sygnału mowy (in Polish)", master's degree dissertation, Inst. of Telecommunications, Warsaw University of Technology, 2011.

[22] "POLQA – Perceptual Objective Listening Quality Analysis", ITU-T Recommendation P.863, 2010.

[23] "Single-ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications", ITU-T Recommendation P.563, 2004.

[24] "Ekiga", http://ekiga.org/

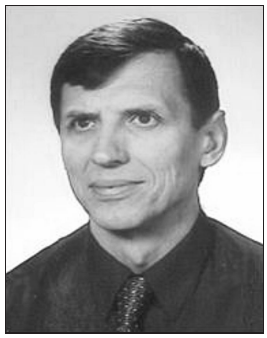
[25] E. Muzychenko, "Virtual audio cable" [Online]. Available: http://software.muzychenko.net/eng/vac.html

[26] Ł. Apiecionek, "Metoda oceny jakości transmisji głosowej w telefonii VoIP", Ph.D. dissertation, IPPT PAN, 2010.

[27] "International Telephone Connections and Circuits – General Recommendations on the Transmission Quality for an Entire International Telephone Connection – One Way Transmission Time", ITU-T Recommendation G.114, 2003.

[28] P. Dymarski and S. Kula, "Metody i standardy badania postrzeganej jakości sygnałów audio i wideo (in Polish)", in *proc. KSTiT, Przegl. Telekom.*, vol. 8-9, pp. 775–781, 2009.

Przemysław Dymarski received the M.Sc. and Ph.D. degrees from the Wrocław University of Technology, Poland, in 1974 and 1983, respectively, both in Electrical Engineering. In 2004 he received the D.Sc. degree in Telecommunications from the Faculty of Electronics and Information Technology of the Warsaw University of Technology. Now he is with the Institute of Telecommunications, Warsaw University of Technology. His research includes various aspects of digital signal processing, particularly speech and audio compression for telecommunications and multimedia, audio watermarking and applications of Hidden Markov Models.
E-mail: dymarski@tele.pw.edu.pl
Institute of Telecommunications
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland



Sławomir Kula received the M.Sc. and Ph.D. degrees from the Faculty of Electronics, Warsaw University of Technology, in 1977 and 1982, respectively. In periods 1999–2002 and 2005–2008 he was Vice Dean of Faculty of Electronics and Information Technology. Since 2008 he is Deputy Director for Education at Institute of Telecommunications, Member of IEEE Communication Society (chairman of ComSoc Warsaw), Member of SIT (chairman of SIT-WUT). He is an author of books (in Polish): *Systemy teletransmisyjne*, *Przewodowe systemy dostępowe xDSL*, and *Systemy i sieci SDH* (also editor).
E-mail: skula@tele.pw.edu.pl
Institute of Telecommunications
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Thanh Nguyen Huy was born in Hanoi in 1986. Since 2005 he was a student at the Faculty of Electronics and Information Technology of the Warsaw University of Technology. In the years 2009/10 he participated in the LLP Erasmus Program, continuing his studies at St. Pölten University of Applied Science in Austria. His master's degree dissertation, *The impact of network parameters on perceived video quality* concerned the quality of selected telecommunications services, like VoD and IPTV. In 2011 he obtained master's degree in Telecommunications.
E-mail: T.NguyenHuy@stud.elka.pw.edu.pl
Institute of Telecommunications
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Paper

A Software Platform for Research on Auction Mechanisms

Mariusz Kamola[a,b], Ewa Niewiadomska-Szynkiewicz[a,b], Krzysztof Malinowski[a,b], Wojciech Stańczuk[a], and Piotr Pałka[a]

[a] *Institute of Control and Computation Engineering, Warsaw University of Technology, Warsaw, Poland*
[b] *Research and Academic Computer Network (NASK), Warsaw, Poland*

Abstract—The platform for research on auction mechanisms is a distributed simulation framework providing means to carry out research on resource allocation efficiency mechanisms and user strategies. Both kinds of algorithms examined are completely user-defined. Interaction of algorithms is recorded and pre-defined measures for the final resource allocation are calculated. Underlying database design provides for efficient results lookup and comparison across different experiments, thus enabling research groupwork. A recognised, open and flexible information model is employed for experiment descriptions.

Keywords—auctions, market simulation, multi-commodity markets.

1. Introduction

The rules for interchange of goods are the legal setting determining market operation. Design of such rules is a fascinating, demanding and important task, often preconditioning efficient operation of economies. When present on a market driven by a set of rules, an entity always implements its own best strategy, developed subject to those rules. However, the entity's initial decision to participate depends on the market attractiveness, comprising its legal framework.

Developing rules for market operation such that desired aims are reached, or trading mechanism design, was one of subjects in collaborative research project "Next-Generation Services and Data Networks — technology, application and market aspects", supported by Polish Ministry of Science and Higher Education. The structure of activities for thematic group "Trading models for transmission services marketplace" is presented in Fig. 1. Note that market clearing, bidding and resource allocation strategies have been considered there as parallel tasks. Examining how they interact when put together is rarely available with analytical models, especially that they are developed by many research teams, and thus with various approach.

This is where the developed platform for research on auction mechanisms (PRAM) comes in, providing those research groups the common language and information model to express the settings of the market, the common entity-market interaction scheme, and the common repository of searchable results. PRAM is, chronologically, the finalisation of the project research activities, making it possible to carry out simulation-driven analysis of strategies developed. It provides for verification of strategies while their assumptions are partially not met or the information about market state is incomplete.

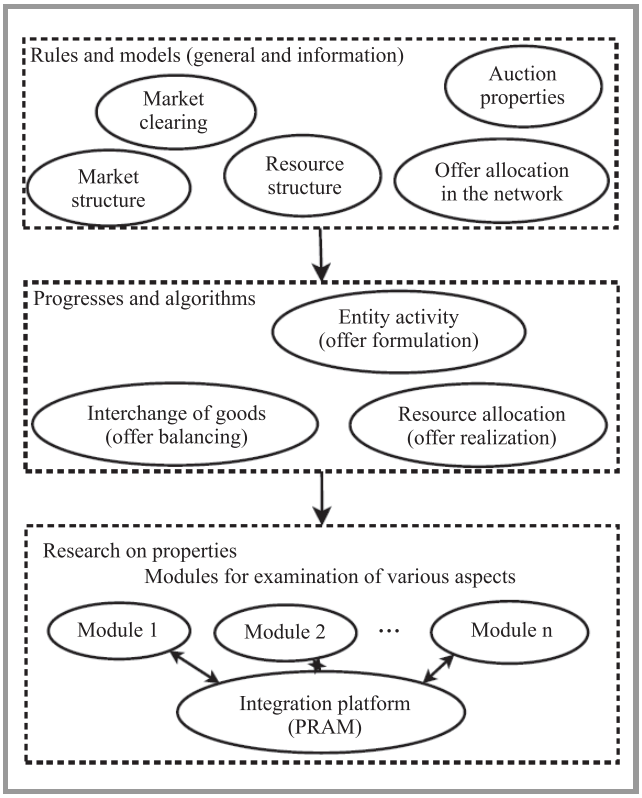


Fig. 1. The position of PRAM creation task within "Trading models for transmission services marketplace" thematic group.

The structure of this document is as follows. Section 2 presents existent and mature trading platforms, while PRAM architecture and functionality is explained in Section 3. This is followed by discussion on comparison criteria for bandwidth trading mechanisms in Section 4. Conclusions are given in Section 5.

2. Existent Frameworks for Interchange of Goods

Information, functional and physical architectures of exemplary trading platforms are presented below. This overview

is an improved version of the material presented in [1]. Examples supporting multilateral trade, e.g. where many participants place sell and/or buy offers at the same time, have been selected. The examples come from various branches and support interchange of different kinds of goods. Great majority of presented examples are fully operational in the business, but some concepts still in research phase are presented as well.

2.1. FCC's Integrated Spectrum Auction System

The automated auction system (AAS) was the first Federal Communications Commission's system used to support frequencies auctioning. AAS required from bidders to use a dedicated software and to use dialup connections to FCC's call centre. Because of growing Internet popularity, AAS was decided to be upgraded to an online web application.

The current new integrated spectrum auction system (ISAS) has replaced the former AAS and Form 175 systems (the latter serving filling up frequency request forms). When compared to its predecessors, ISAS offers extended functionalities including request data validation, advanced data query, integration with other FCC forms, ergonomc interface, improved bid placement [2]. The system is now open to every Internet user.

ISAS system provides for simultaneous auctions with many rounds and the possibility to bid for a bunch of licenses.

2.2. WARSET – Warsaw Stock Exchange Trading System

The quoting at Warsaw stock exchange (WSE) is done via WARSET transaction system [3]. WARSET supports fully automated offers processing and transaction making. It is easily accessible and provides complex information about the market. Moreover, WARSET is now integrated with brokerage, thus facilitating bidding for broker's customers. The principal WARSET contractor, acting within a consortium, has provided the necessary hardware and the client application, making it possible for 37 brokerages collocated with WSE to connect instantly, and for the rest to connect via wide area network. Auxiliary transfer agents have been developed, like the one to interact with the Polish National Depository for Securities.

2.3. Polish Power Exchange

Polish power exchange (PPE) has been founded as a central component of the Polish energy market undergoing liberalization. Since its very beginning, PPE was in the fore while deploying novel solutions for energy trading. Within six months of PPE's operation the spot energy market has started, with its prices being the reference point in bilateral contracts. In 2003 PPE has been licensed by Polish financial supervision authority to operate an electricity marketplace.

In 2008 PPE has started commodity derivatives market. Derivatives for energy quoted there make it possible to calculate longer-term electricity prices, which enables big market players to forecast and optimize buy or sell prices. PPE runs on a state-of-the-art trading platform, provided by NASDAQ OMX – the biggest manufacturer of trading platforms in the world [4]. PPE is technically capable to serve whole Polish energy market.

2.4. MERKATO – Bandwidth Marketplace

New York-based bandwidth trading platform MERKATO did not count much on the market, but it was a remarkable enterprise. MERKATO was designed as an open and scalable platform for real-time bandwidth acquisition in Internet. The auction algorithm serving requests, invented and patented by The Invisible Hand Inc., was based on a progressive second-price auction. MERKATO had adopted a distributed approach: a microauction was organized for each network resource, as bandwidth. Therefore, bids for resource bundles had to be submitted and processed independently [5].

Market clearing was done every 5 minutes, by collecting bids, calculating equilibrium prices and allocating throughput to winners. Interestingly, the whole process was fully automated: once the winners got selected, the operators reconfigured their networks and access points accordingly. Moreover, MERKATO offered derivatives market where futures were traded in broad time-scale.

Unfortunately, MERKATO is not operational since 2007, and without an apparent successor or a competitor.

2.5. PeerMart – a Distributed P2P Auction System

PeerMart is a technology for auction-based resource interchange in peer-to-peer (P2P) networks [6]. P2P networks growth is driven by an idea of sharing own and using others' resources worldwide, by means of agents running on home PCs, without centralized management of any sort. However, P2P users often act egoistically, e.g. by not providing any resources and switching their PCs on only when they need to use others' resources.

PeerMart was designed to solve such problems by introducing incentives to share own resources. It utilizes double auctions in distributed setting, thus rewarding valuable content. Furthermore, redundancy mechanisms are applied to ensure system robustness in presence of non-cooperating agents.

Every resource is sold or bought in PeerMart via a double auction carried out by dispersed broker-peers (auctioneers).

2.6. Storage Exchange

Storage Exchange is another double-auction based trading platform [7]. The goods being traded is storage space: the sellers are storage providers or any businesses possessing free disk space, and the buyers are institutions in need of virtual disks.

2.7. Band-X – Architecture for Bandwidth Trade in IP Networks

The band-X system is not the operational bandwidth trading platform; it rather a mature concept of such system, developed at Drexel University [8]. Quality of service provided through DiffServ and IntServ technologies are the system technical foundations. Multilateral agreements are supported, as well as spot and derivative trading. The work focused on organizational aspects of the platform operation.

3. The Platform Architecture and Functionality

The platform for research on auction mechanisms is a tool supporting examination of behavior of models developed in the course of the project. There are two kinds of models: representing auctioning and resource allocation process, and representing market entities activities. The basic scope of PRAM application includes

- verification of theoretical model properties through simulation,
- estimation of model sensitivity,
- assessment of observed model properties in scenarios where selected models interact.

Specifically, PRAM helps in searching for Nash equilibria in games induced by models interaction, in analysis of those equilibria properties, and in estimating sensitivity of results for constraints defining market rules. Consequently, the platform makes it easy to compare market clearing mechanisms, and to infer about their practical efficacy. Specific use scenarios are examining robustness of those mechanisms when players exhibit unusual behavior (e.g., speculate) or estimating the maximum disproportion in players' market strength when a trading mechanism still remains efficient.
The main architectural PRAM assumption is ergonomics for mechanisms testing and ranking. It is also important to extend existing repository of models easily, by implementing new resource allocation mechanisms and new agents simulating user behavior.

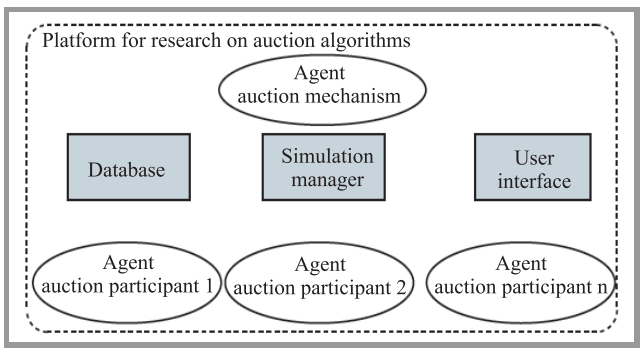


Fig. 2. The main PRAM modules.

PRAM architecture is a modular one, cf. Fig. 2. The information exchanged vertically by the modules are conforming to multicommodity market data model (M^3), developed earlier by project participants [9]. The modules of auction mechanism and auction participants are replaceable, while the middle layer modules constitute PBMA core. Simulation manager is responsible for starting and setting up links to agents, performing the simulation, and the cleanup. Database module provides persistence to experiment configuration data as well as intermediate and final results. User interface module defines forms for experiments creation, configuration, running and processing.

3.1. Multicommodity Market Data Model – M^3

M^3 is a method and format for a formal description of a market where trade of resources takes place. It has been initially developed to describe offer structure in the energy market in Poland. For its generality, it has been next used to model IP network bandwidth trade [10]–[13]. Used in PRAM it effectively describes properties and dependencies between goods being traded.
M^3 defines the following basic entities and relations between them:

- network nodes and arcs, describing the topology of the network where capacity trade takes place,
- market entities (users, providers) that buy or sell resources (capacity),
- resources being offered, with their proper attributes,
- offers, i.e. bindings of market entities and resources, offered or demanded at a specific price.

It is also possible to define compound resources, i.e. containing simple resources and other compound resources. Analogously, one can define simple and compound offers and market entities, exploiting the model generality and flexibility. However, it can also be applied without knowledge of advanced features, like aggregation facilities. It is possible to declare only key values: *offeredPrice*, *min/maxValue* and *shareFactor* (1 for sell, –1 for buy offers), leaving other unset. Fields *acceptedVolume*, along with *sell/buyPrice* parameters of commodity structure contain results of the market clearing process.

3.2. Functional Requirements

Functionality of PRAM is determined by three major assumptions:

- The platform will principally be used in research context, and applied for varying set of models interacting.
- Ease of historical results retrieval and comparison is central.

- Design patterns, data models and communication mechanisms applied must make PRAM a valuable proof of concept of a commercial trading system.

3.3. Platform Users and Resources Being Subject to Interchange

It is assumed that the main PRAM user is a researcher or designer of trading mechanisms, i.e. algorithms for resource allocation and quoting. PRAM simulation framework provides means to examine interaction of mechanisms and market entities, both being represented by software agents. Research via simulation aims to discover phenomena difficult to analyze and predict, like in scenario where all or part of market participants are human.
Within the project, PRAM is applied to bandwidth allocation in data networks, but it can easily be used in other, quite distant, application domains, like parallel problem solving (cf. [14]). The platform is a framework for market simulation; it implements its fixed components (the database, simulation manager and user interface) and exemplary replaceable components (user and auction mechanisms agents). Two auction mechanisms have been implemented: balancing communication bandwidth trade (BCBT – see [12]) and effective bandwidth auction mechanism (EBAM – see [13]). They constitute a good starting point for eventual further development and research.

3.4. The Architecture

PRAM design follows an open system concept, i.e. it facilitates rapid new agent prototyping. System architecture is multigrained, composed of federations of simulation components, like user agents that act synchronously, and the market mechanism, coupled via the simulation manager. Data persistence is provided by the underlying relational database module, and PRAM web interface is managed by the interface module. A single, universal application programming interface and communication protocol between the management module and agents have been designed. It makes possible to treat all agents uniformly, wherever possible, as black boxes. On the other hand, simulation manager operation is, to much extent, transparent for the agents. Functionalities of the five types of PRAM modules are described below.

Auction mechanism agent. It is a functional module responsible for market clearing, resource allocation and quoting. It runs, in principle, by performing optimization tasks, which can be done by external solvers (e.g. CPLEX, LPSolve) or built-in custom optimization routines. This module must implement the following operations:

1. *Auction initiation.* This operation is executed once at each auction start. It is invoked by the simulation manager, thus informing the auction mechanism that the auction has just started. The simulation manager passes information about the system (i.e. network topology, list of market entities, list of resources being traded, auction-specific parameters) to the auction mechanism. Most of the data are stored in M^3 format.
2. *Resource allocation.* This operation may be executed more than once, depending on the type of auction. The simulation manager invokes this operation with buy/sell offers as arguments. Operation results are resource allocations and prices set by the auction mechanism. The operation arguments and results are expressed in M^3 format.
3. *Simulation termination.* This operation is executed once for each auction. It is invoked by the simulation manager to communicate the auction mechanism that it is going to be destroyed because the simulation has just ended. The auction mechanism agent is given a chance to perform cleanup activities, like disconnecting from a remote solver.

User agents. It is a functional module implementing market entity behavior. Many types of user agents can take in a single simulation experiment, their emergent collective behavior being often impossible to be expressed analytically. The main results of user modules operation are trading buy and/or sell offers that, after being merged by the simulation manager, get presented to the auction mechanism. These modules must implement the following operations:

1. *Auction initiation.* Like for the auction mechanism agent, this operation is executed once at each auction start. This operation gives also an opportunity to pass any extra parameters to agents that parametrize their working, including, e.g. parameters of probability distribution, IDs of other agents that are going to form a cartel etc.
2. *Offer preparation.* On operation invocation actual resource allocation and prices are passed to agents. In response, agents return their new offers to the auction mechanism. For iterative mechanisms this operation is executed many times; for one-step mechanisms this operation is called twice (on the first call, there are no allocations yet; on the second call the auction is over and no user response is expected).
3. *Simulation termination*. Like for the auction mechanism, this operation is called only once, on simulation end.

Simulation manager. It is the PRAM central module, responsible for running experiments, i.e. resource allocation sessions. Its functionality covers both pure managerial activities and in-depth analysis of the data being exchanged. The simulation manager:

- manages spawning, cleanup and synchronization between agents;

- forwards data between user agents and the mechanism;
- merges separate offers into one M^3 model, presented to the mechanism;
- calculates predefined simulation outcome statistics.

The fundamental requirement is that multiple heterogeneous agents must be managed timely and reliably results in multithreaded design of the simulation manager. Communication state with each agent is handled in a separate thread, and the communication technology is Java Messaging Service (JMS). Commands spawning the agents are delegated from Java to the underlying operating system.

Database module. The database module manages persistence of M^3 structures and PRAM-specific data structures into a relational database. The data module is the only interface between other PRAM modules and the database, providing efficient mapping of Java objects into data tables, and the database is the only repository of any PRAM data, which does not preclude database direct access from other applications. Any M^3 data, before being stored into the database, need to be converted into plain old Java objects (POJO), using the converters generated automatically from M^3 XML schemas (XSD). PRAM database contains therefore:

- agents configuration (startup parameters, mapping between software agents and market entities),
- scenario definitions (selection of M^3 models, agent types, general scenario attributes and descriptions),
- M^3 intermediate and final scenario results (offers, prices, allocations),
- solution statistics calculated by PRAM.

Graphical user interface. PRAM user web interface makes it possible for a user to define and run test scenarios, and to filter, analyze and visualize individual and aggregated results. Using a diagram, graph or table form, a user can observe data that are:

- simulation-oriented – all output data (final and intermediate allocations, bids and prices) are shown for a selected scenario;
- resource-oriented – selected resource allocations and prices are shown for various testing scenarios;
- user-oriented – selected user bids and allocations are shown for various testing scenarios.

3.5. Data Flows between PRAM Modules

Figure 3 illustrates the data exchanged between PRAM modules. It must be emphasized that the central role of the simulation manager is evident as soon as the user requests to carry out the simulation, while the simulation environment is being initialized. This process, and the simulation itself, is executed in the following steps:

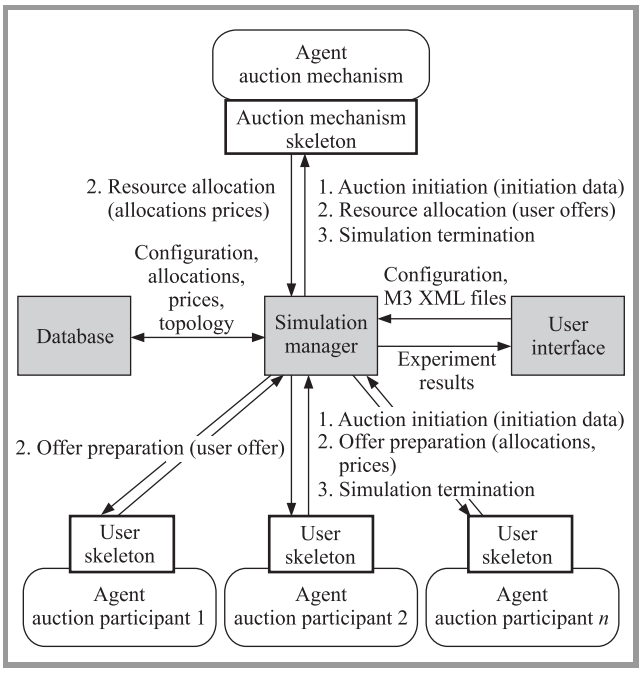


Fig. 3. Data flows between PRAM modules.

1. Configuration for simulation is read by the simulation manager from the database: network topology, market entities and, product and offer definitions, agents and simulation parameters are loaded.
2. Simulation manager spawns the agents (or waits for those spawned externally) and waits until they report they may start simulation.
3. The agents get initialized by the manager.
4. The simulation runs by alternately collecting user agents offers, merging them and forwarding to the auction mechanism agent. Mechanism reply, containing prices and allocations, is passed back to agents, and the procedure is repeated. Intermediate results are stored in the database.
5. The simulation is broken on mechanism request, or when the maximum number of iterations is reached. Finally, all agents are requested to decommit resources and quit.

PRAM has been implemented in Java language, using selected Java Enterprise Edition components as JMS and Java Persistence API (JPA – Hibernate implementation). Communication between platform components during simulation is presented in Fig. 4. All experiments data that can be stored using M^3, are stored in both plain XML text files, and as persistent POJO structures. For communication with agents, only plain XML is used, and any extra non-M^3 parameters are passed using Java native serialization.

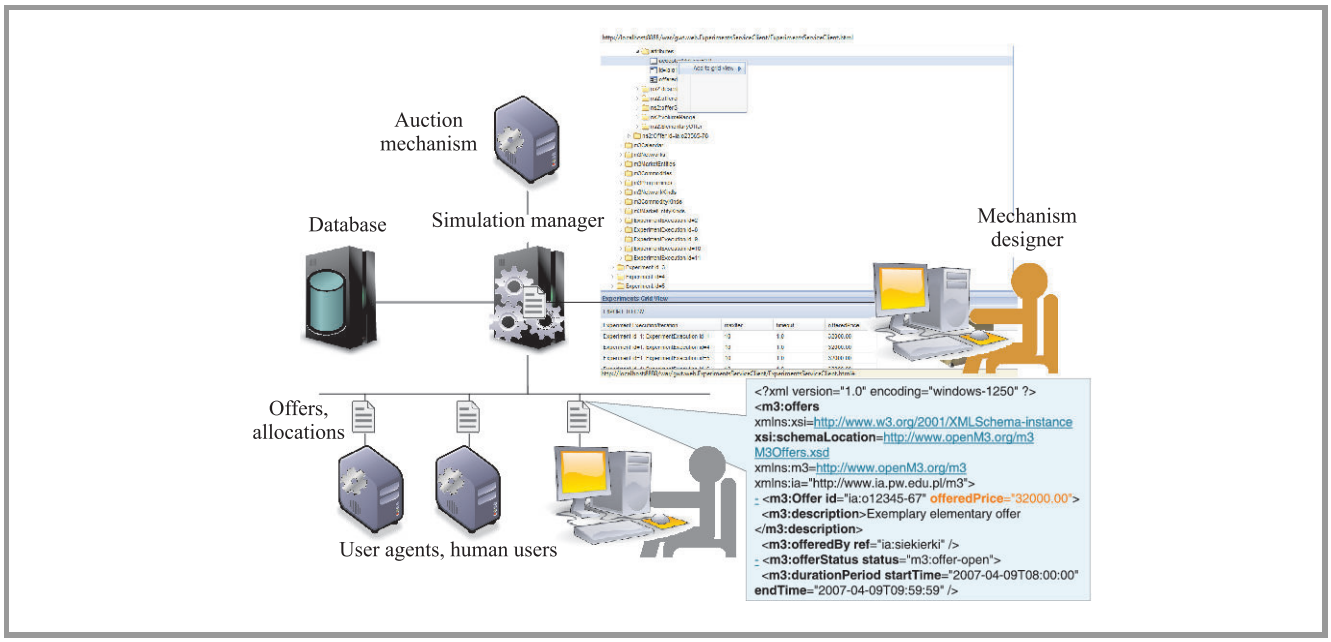


Fig. 4. Communication between modules during simulation.

Graphical user interface acts as a control station for the whole process of simulation and data analysis. GUI component, running on the server side as a web application, is loosely coupled with other PRAM components via Spring framework [15]. While operating PRAM, a user can utilize existent external software supporting network topology and calendar edition. The software is dedicated to use M^3 model, and to operate on M^3 files. It operates locally, and the resulting M^3 files can be uploaded to PRAM afterwards.

4. Comparison Criteria for Bandwidth Trading Mechanisms

The theory of mechanisms defines a number of mechanism properties. Many of those properties are desirable for any market mechanism being under construction. The properties can be perceived as criteria for mechanism ranking. Mechanism engineer, knowing about market peculiarities (e.g., legal layout, kinds and number of entities, kinds and structure of resources) may indicate mechanism properties that are considered important in a given situation. In a liberal market system the criteria important for individual entity are usually disjoint from the global criteria, i.e. important for the society as a whole. Apart from mechanism evaluation according to the two above viewpoints, one may consider the third approach: mechanism technical efficiency.

4.1. Global criteria

Economic efficiency. Social welfare is considered the principal measure of a mechanism efficiency. Social welfare is the total of real economic benefits from the trade of commodities. If the user best strategy in a market mechanism is to bid according to his valuation, then the social welfare can be calculated using the prices offered by market participants. Otherwise, social welfare can be approximated by so called economic benefit, i.e. the difference of the total value of goods being bought and the total value of goods being sold, using transaction prices instead of the valuations.

Incentive compatibility. A mechanism is incentive compatible if a user best strategy is to announce all his private valuation information, i.e. if a user has no incentive to bid untruthfully. A mechanism is incentive compatible if users' truthful strategies are their best strategies, and therefore are the market game equilibria. Incentive compatibility prevents any individual or collective actions diverging from the optimal strategy. A good measure of such prevention is a so-called allocation inefficiency.

Budgetary balance. A market is in budgetary balance if the money flow from goods acquisition is equal to the flow from goods sales. This means that a market driven by a mechanism enforcing budgetary balance does not require any subsidy, neither it generates any surplus.

Market concentration. Entities that have considerable market share may influence the clearing process and jeopardize assumed mechanism properties. Herfindahl-Hirschman index (HHI) is used to measure market concentration; it is calculated by summing squared percentage market shares for all market entities.

Pareto efficiency. The game outcome is Pareto-efficient when it is possible to increase profits of one market entity

only at the cost of the profit loss by some other entity or entities. In other words, the outcome of the game is not dominated in Pareto sense by any other outcome. It is particularly instructive to apply the term of Pareto efficiency to resource allocation problem. The resource allocation problem solution in market economy is a detailed register or description of what resource has been assigned to whom. Solution space is defined by the current state of technology and the amount of available resources in the economy. The final allocation solution depends always on customers' preferences. Therefore, for given preferences, technology and resources if an allocation is Pareto-efficient, it is impossible to find another allocation improving somebody's profits without spoiling someone else's profits.

4.2. Individual Criteria

Individual profit maximization. Every single market participant is interested that the market mechanism makes it possible to maximize participant's profit. On incentive compatible markets individual profit maximization do really takes place. However, society expectations are often that market prices should stay as low as possible, for the benefit of customers. Under such demand one can still design a mechanism for individual manufacturer profit maximization. It requires the original problem to be reformulated so that society expectations are considered superior to profit maximization – they can be, for example, treated as constraints to mechanism outcome.

Absolute fairness – individual rationality. A mechanism is considered to be absolutely fair when none of the players will incur individual loss, i.e. the player profit will be positive. In fact, this simple criterion preconditions the player participation in the market.

Individual relative fairness. A mechanism is considered to be relatively fair from one's point of view if the other competitive offers are not favored at his/her costs. This broad term encompasses more specific criteria:

- Anonymity. Market players are treated anonymously if the order of their numbering does not influence the outcome.
- Symmetry. Any two players characterised by equal parameter values, i.e. players having the same preferences (in the sense of their utility functions) and capabilities (in the sense of quality, amount and geographical availability of services offered) should be given equal allocations outcome.
- Price uniformity. A mechanism is fair if the price of a service is equal for all customers.

4.3. Mechanism Technical Efficiency

The most important criterion of mechanism technical efficiency is the possibility of the mechanism to be deployed and successfully used. Successful implementation of a mechanism depends on complexity of the underlying algorithms and on algorithms robustness. Typical measures characterizing mechanism technical efficiency are:

- market clearing time,
- total market clearing time (for iterative mechanisms),
- number of exchanged messages (average per user),
- number of lost messages (i.e. messages that did not count in the process of market clearing).

5. Conclusion

The platform for research on auction mechanisms in its current state of development should be perceived as a groupwork environment used for design and simulation verification of resource allocation mechanisms. Genericity of PRAM algorithms for experiments data selection and evaluation have implied use of somewhat prolix M^3 data format and clumsiness of graphical user interface. The next step in PRAM development is to implement user agents being operated by a human. This will make possible to run simulation scenarios where some market users will be played by e.g. students, being confronted with each other as well as with automated software agents.
In the long run, PRAM commercialization can be considered. Although the platform has already adopted a number of concepts and technologies used in enterprise applications (tiered architecture, JMS, Hibernate [16], GWT [17]), making it a fully-fledged business application requires adding many new functionalities. They include: authentication and authorization, scalability, repository protection, SLA guarantees etc. The architectural solutions implemented so far in PRAM have been selected deliberately to facilitate such transition.

Acknowledgement

The research presented in this paper was partially supported by Polish Ministry of Science and Higher Education grant PBZ-MNiSW-02/II/2007-LUB.

References

[1] W. Stańczuk, P. Pałka, J. Lubacz, and E. Toczyłowski, "A framework for evaluation of communication bandwidth market models", *J. Telecom. Inform. Technol.*, no. 2, pp. 52–60, 2010.

[2] "ISAS briefing sheet", Federal Communications Commission, 7th March 2011 [Online]. Available: http://wireless.fcc.gov/auctions/conferences/combin2003/papers/ISASbriefingsheet.pdf

[3] "Gpw.pl – Trading system", 7th March 2011 [Online]. Available: http://www.gpw.pl/trading_system

[4] "IT System – Towarowa Giełda Energii", 7th March 2011 [Online]. Available: http://www.polpx.pl/en/20/it-system

[5] G. Giammarino, J.-F. Huard, and N. Semret, "Merkato: a platform for market-based resource allocation", in *Proc. Networking 2000, Intelligent Agents for Telecommunications Environments*, Paris, France, 2000.

[6] "PeerMart: A Decentralized Auction-based P2P Market", 7th March 2011 [Online]. Available: http://www.peermart.net

[7] M. Placek and R. Buyya, "Storage exchange: a global trading platform for storage services", in *Euro-Par 2006*, LNCS 4128, W. E. Nagel et al., Eds. Springer, 2006, pp. 425–436.

[8] D. M. Turner, V. Prevelakis, and A. D. Keromytis, "The bandwidth exchange architecture", Tech. Rep. DU-CS-04-07, Drexel University, 2004 [Online]. Available: https://www.cs.drexel.edu/files/ts467/DU-CS-04-07.pdf

[9] P. Kacprzak, M. Kaleta, P. Pałka, K. Smolira, E. Toczyłowski, and T. Traczyk, "Communication model for M3 – open multicommodity market data model", in *Proc. TPD'2007 Conf.*, Poznań, Poland, 2007.

[10] P. Kacprzak, M. Kaleta, P. Pałka, K. Smolira, E. Toczyłowski, and T. Traczyk, "M3: Open multicommodity market data model for network systems", in *Proc. XVI Int. Conf. Sys. Sci.*, Wrocław, Poland, 2007.

[11] P. Kacprzak, M. Kaleta, P. Pałka, K. Smolira, E. Toczyłowski, and T. Traczyk, "Application of open multicommodity market data model on the communication bandwidth market", *J. Telecom. Inform. Technol.*, no. 4, pp. 45–50, 2007.

[12] P. Pałka, K. Kołtyś, E. Toczyłowski, and I. Żółtowska, "Model for balanced aggregated communication bandwidth resources", *J. Telecom. Inform. Technol.*, no. 3, pp. 43–49, 2009.

[13] M. Karpowicz and K. Malinowski, "Network Flow Optimization with Rational Agents", NASK internal rep., 2009.

[14] M. Kamola, "Software environment for market balancing mechanisms development, and its application to solving more general problems in parallel way", in *Proc. PARA 2010 Conf.*, Reykjavik, Iceland, 2010.

[15] C. Walls and R. Breidenbach, *Spring in Action*. Manning Publications Co., 2005.

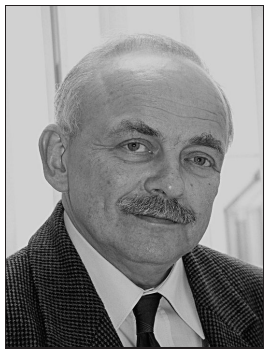
[16] Ch. Bauer and G. King, *Hibernate in Action*. Manning Publications Co., 2005.

[17] R. Cooper and Ch. Collins, *GWT in Practice*. Manning Publications Co., 2008.

Mariusz Kamola received his Ph.D. in Computer Science from the Warsaw University of Technology in 2004. Currently he is Associate Professor at Institute of Control and Computation Engineering at the Warsaw University of Technology. Since 2002 with Research and Academic Computer Network (NASK). His research area focuses on economics of computer networks and large scale systems.

E-mail: mkamola@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

E-mail: Mariusz.Kamola@nask.pl
Research and Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland



Krzysztof Malinowski Prof. of techn. sciences, D.Sc., Ph.D., MEng., Professor of control and information engineering at Warsaw University of Technology, Head of the Control and Systems Division. Once holding the position of Director for Research of NASK, and next the position of NASK CEO. Author or co-author of four books and over 150 journal and conference papers. For many years he was involved in research on hierarchical control and management methods. He was a visiting professor at the University of Minnesota; next he served as a consultant to the Decision Technologies Group of UMIST in Manchester (UK). Prof. K. Malinowski is also a member of the Polish Academy of Sciences.
E-mail: K.Malinowski@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

E-mail: Krzysztof.Malinowski@nask.pl
Research and Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland

Wojciech Stańczuk received his M.Sc. in Telecommunications from the Warsaw University of Technology in 2001. Now he is Ph.D. student and research assistant in the Institute of Telecommunications at Warsaw University of Technology. His scientific interests cover techno-economic aspects of telecommunication networks operation, including resource allocation and pricing as well as strategies for infrastructure investments.
E-mail: w.stanczuk@tele.pw.edu.pl
Institute of Telecommunications
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Piotr Pałka is an Assistant Professor of the Operations and Systems Research Division in the Institute of Control and Computation Engineering at the Warsaw University of Technology, Poland. He received the M.Sc. and Ph.D. degrees in computer science in 2005 and 2009, respectively, both from Warsaw University of Technology. His research interest focus on the market mechanisms, especially the incentive compatibility on the infrastructure markets and on the multi-agent systems. His current research is focused on application of multicommodity turnover models.

E-mail: P.Palka@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Ewa Niewiadomska-Szynkiewicz – for biography, see this issue, p. 10.

Paper

The Realization of NGN Architecture for ASON/GMPLS Network

Sylwester Kaczmarek, Magdalena Młynarczuk, Marcin Narloch, and Maciej Sac

Department of Teleinformation Networks, Gdańsk University of Technology, Gdańsk, Poland

Abstract—For the last decades huge efforts of telecommunication, Internet and media organizations have been focusing on creating standards and implementing one common network delivering multimedia services – Next Generation Network. One of the technologies which are very likely to be used in NGN transport layer is ASON/GMPLS optical network. The implementation of ASON/GMPLS technology using open source software and its results are the subject of this paper. The ASON/GMPLS architecture and its relation to the proposed ITU-T NGN architecture are described. The concept, functional structure and communication among architecture elements as well as the implementation of laboratory testbed are presented. The results of functional tests confirming proper software and testbed operation are stated.

Keywords—ASON, Connection Control Server, Diameter, GMPLS, IP QoS, NGN, RSVP.

1. Introduction

The changes that take place in the area of modern community indicate the great value of information. The information has various areas of applications and forms of presentation. Constant information growth and the need for fast availability to the whole public in direct or processed form generate necessity of new telecommunication network architecture proposition. For this reason networks have to be developed to meet the needs of new requirements. The next generation network (NGN) is a proposition of architecture which has a chance to fulfill society requirements.

Standardization of the NGN dates back to the NGN workshop held in 2003 by ITU-T (International Telecommunication Union – Telecommunications) standardization group. The Y.2000 series of recommendations has been given for the NGN specification and requirements. Functional requirements and architecture of next generation networks are described in Recommendation Y.2012 [1]. Conceptually, the NGN architecture consists of service stratum and transport stratum. The transport stratum provides the IP connectivity services to NGN users. The service stratum provides the service control and content delivery functions. From economic aspect, the conversion of current networks to NGN architecture has to proceed evolutionally because of the high costs associated with realization of this architecture.

The ITU-T automatically switched optical network (ASON) [2] concept and generalized multi-protocol label switching (GMPLS) [3] Internet Engineering Task Force (IETF) solution were combined by Optical Internetworking Forum (OIF) into ASON/GMPLS optical network [4], which is one of the most promising solutions for NGN transport layer. Key issue for the ASON/GMPLS proposition is the provision of effective network control servers. An implementation of ASON/GMPLS connection control layer is presented in this paper.

The aim of our work was a realization of selected functionality of ASON/GMPLS network [5]. We decided to carry out the tasks in two stages. The first step was to write software for network elements with respect to ASON/GMPLS standardization and the latest trends in ITU-T NGN architecture. Secondly, implemented software had to be tested in laboratory testbed.

The paper is organized as follows. In Section 2 ASON/GMPLS control plane in context of the NGN architecture is presented. An ASON/GMPLS control plane implementation concept is introduced in Section 3. The realization of ASON/GMPLS architecture, functional architecture of connection control server (CCS) and service control server (SCS) as well as resource terminal (RT) representing transport resources are described in Section 4. The results of functionality and operation tests are reported in Section 5. Conclusions and outlook to future are presented in Section 6.

2. ASON/GMPLS Architecture

The automatically switched optical network (ASON) was proposed by ITU-T and described in Recommendation G.8080 [2]. In fact ASON is only a concept of architecture. It does not specify all protocol details necessary to implement the control plane solution. Optical Internetworking Forum (OIF), a group of international network service providers, made an effort to apply GMPLS protocols to ASON architecture [4]. This solution, an ASON control plane built on GMPLS protocols is known as ASON/GMPLS.

Due to the fact that ASON/GMPLS is a highly complicated architecture, in this section the main aspects regarding the ASON/GMPLS control plane functionality are presented. The main task of the ASON/GMPLS control plane

is to facilitate fast and efficient configuration of connections within the transport layer network to support both switched and soft permanent connections. It consists of different components providing specific functions (including routing and signalling). The interactions between and within domains are defined in terms of reference points: UNI, E-NNI, I-NNI [2].

ITU-T standardization group recommended control plane components like: routing controller (RC), protocol controller (PC), connection controller (CC), link resource manager (LRM), termination and adaptation performer (TAP), calling/called party call controller (CCC), network call controller (NCC). These components can be combined in different ways depending on the required functionality.

The architecture of ASON/GMPLS network is presented in Fig. 1. The connection controller is responsible for coordination among the link resource manager, routing controller and other connection controllers for the purpose of setup, release and modification of connection parameters [2]. For this reason the CC components utilize a connection controller interface (CCI) to the transport plane. As stated in [2] the routing controller is an abstract entity that provides routing functions. The link resource manager maintains the network topology. The role of protocol controller is to map the operation of the components in the control plane into messages that are carried by communication protocols between interfaces in the control plane. The termination and adaptation performer holds the identifiers of resources that can be managed using the control plane interfaces. Call components are concerned with call service and implemented in service control server. The main role of the calling/called party call controller is generation of outgoing call request and acceptance or rejection of incoming call request.

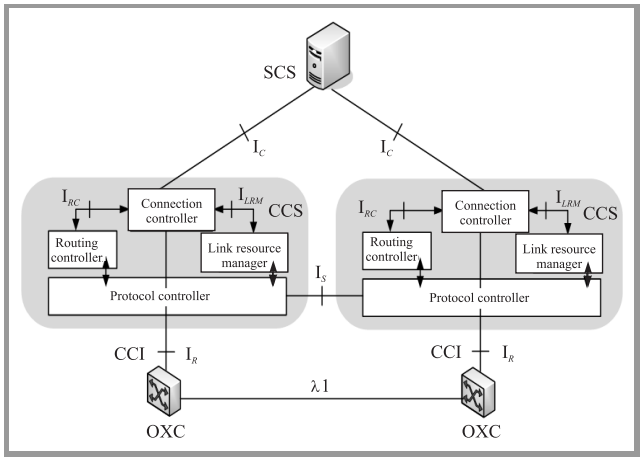


Fig. 1. ASON/GMPLS network architecture. SCS – service control server, CCS – connection control server, OXC – optical cross-connect.

As it has been already mentioned, the ASON/GMPLS architecture is one of the solutions considered as NGN transport layer. ITU-T NGN resource and admission control function (RACF) [1] performs operations similar to ASON/GMPLS connection control server consisting of connection controller, routing controller, LRM and protocol controller elements. Selected functionality of ITU NGN service control functions is performed by service control server in the ASON/GMPLS architecture. NGN transport functions correspond to resource elements depicted as OXC in Fig. 1. The concept of ASON/GMPLS architecture implementation is presented in the next section.

3. Concept of Implementation

The concept of the proposed ASON/GMPLS implementation is presented in Fig. 2. According to the three-layer architecture depicted in Fig. 1, the implementation includes functionality of service control layer, connection control layer as well as optical resource layer represented by resource terminals (RTs) emulating optical cross-connect operation. Corresponding layers communicate over dedicated interfaces using communication protocols.

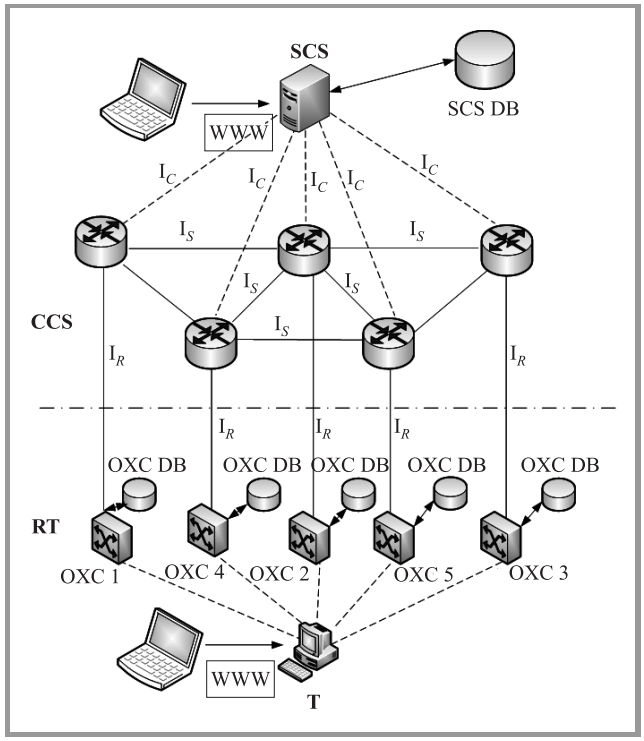


Fig. 2. Concept of ASON/GMPLS implementation. RT – resource terminal, T – terminal.

Service control server (SCS) is responsible for handling of user call and call termination request. Call requests are furtherly transformed into connection requests in the connection control layer. SCS provides www interface and stores all necessary information in local database. The database in service control server consists of the following tables: CTRL_TABLE, CALL_STATE and CALL_STAT. CTRL_TABLE table maps addresses from resource layer into addresses of corresponding connection control servers (CCSs). The CALL_STATE table stores the state of processed requests. The CALL_STAT table gather statisti-

cal data regarding performance of handling requests in the system. Particularly, duration of operations regarding call setup and termination requests in the system is registered. Connection in optical layer is established only if there exist enough free resources to allocate. Each resource terminal informs corresponding connection control server about the result of optical resource allocation.

Each connection control server is in charge of dynamic management of optical resources in transport layer by processing requests for establishing (setting) and releasing (deleting) paths. Connection control servers utilize RSVP [6] protocol extended to transport objects regarding resource reservation in optical layer. The design of CCS functionality was based on the following assumptions:

– mapping of elements from control layer to transport layer is one-to-one,

– single reservation session results in reservation of one or more transport units, depending on the bandwidth demand request,

– identifiers of resource layer are transported using mechanisms of LMP protocol [7],

– fixed filter (FF) reservations style is applied [6].

Resource terminals emulate optical resources. For this reason they maintain information regarding state of the emulated device in a local database. Terminal T is used for configuration and verification of reservation state in emulated OXCs. In order to perform these operations dedicated www interface which presents database content of resource terminals is provided.

Communication between cooperating layers is performed over well defined interfaces. Service control layer and connection control layer communicate over I_C interface. Information between connection control server and resource terminal is exchanged over I_R interface. In both cases Diameter protocol [8] is used. In the next section realization of the testbed architecture based on the presented concept is described.

4. Architecture Realization

In this section realization of the proposed ASON/GMPLS architecture concept (Fig. 2) is described. The implementation of ASON/GMPLS network elements limited to selected functionality was based on Linux platform. The realization of the software was performed in two phases (variants). The first step was variant I regarding implementation of connection control layer with basic functionality and the final step was variant II regarding implementation of connection control layer with ability to communicate with surrounding layers.

Functional structure of the implemented ASON/GMPLS software (variant II) is presented in Fig. 3. Service control server (SCS) handling user requests can be managed through www browser, which communicates with Apache HTTP embedded server. The www server uses PHP hypertext preprocessor compiled to shared object library libphp5.so loaded as a module at the beginning of its initialization. SCS server configuration and description of connection control layer are stored in local Oracle database. Communication between database and www server uses SQL queries. PHP OCI8 packet provides functions to communicate with the database. That allows to update the database content by WWW/PHP scripts of user interface.

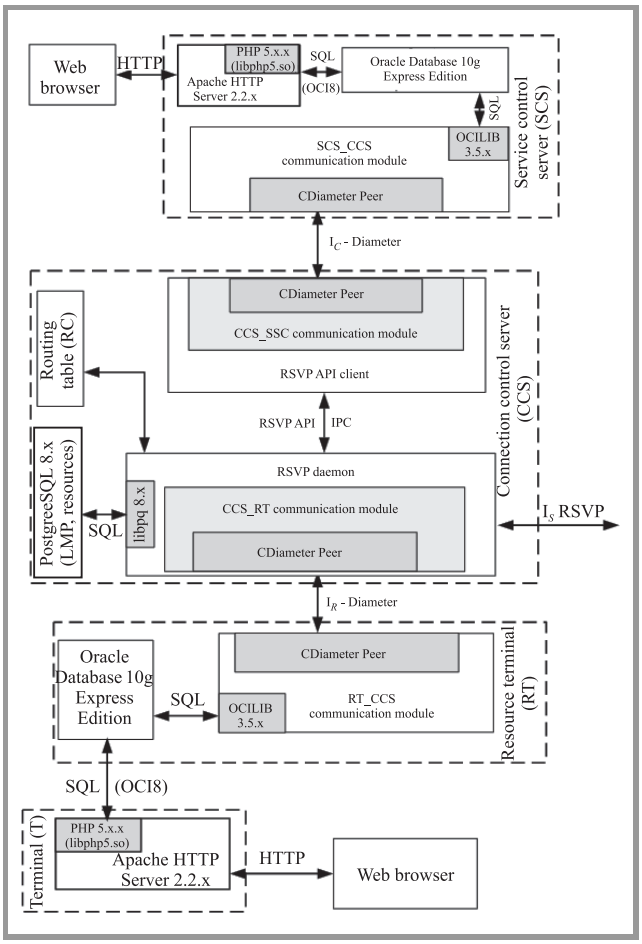


Fig. 3. ASON/GMPLS network realization.

The most important part of service control server (SCS) is the communication module responsible for exchanging information between SCS and CCS. In the communication module OCILIB library is used to provide communication with database including database change notification (DCN) mechanism. DCN allows to asynchronously notify the communication module of the SCS about changes in the database generated by the user through the WWW/PHP interface. Moreover, OCLIB allows to modify the content of the database according to the results of user request processing in connection control servers and resource terminals. Requests results, particularly regarding processing performance in the architecture can be statistically analyzed in SCS and presented by the WWW/PHP interface. Furthermore, service control server communicates with connection control layer using Diameter protocol. In the re-

alization of SCS an open source implementation CDiameter Peer was used. Thus, communication module of the SCS has the functionality of Diameter Peer, which allows to provide connection control layer with the following parameters of user request: Call-ID call identifier [9], source and destination address in connection control and resource layers, bandwidth demand, reservation priority. The above mentioned request parameters are carried in Diameter messages as appropriate attribute value pairs (AVP) elements. Detailed description of Diameter application used in the architecture is described in the last part of the section.

The core component of the proposed and implemented ASON/GMPLS architecture is connection control layer. The following ASON/GMPLS functionalities: connection controller, protocol controller, routing controller and link resource manager were implemented in each connection control server. The functionality of protocol controller and connection controller is performed by the RSVP daemon element, which is the part of KOM RSVP implementation [10] of RSVP protocol. Functionality of KOM RSVP project was appropriately extended to transport information for control of optical network and to communicate with service control layer as well as resource terminal layer [11], [12]. Thus, some limited functionality of RSVP-TE [13] signalling protocol was achieved. Extension of KOM RSVP preserved original structure of KOM RSVP implementation, with division into RSVP daemon and RSVP API client. The role of API client is to invoke API functions provided by RSVP daemon with respect to Diameter messages received from service control server in order to control the process of setting-up and releasing connections in the transport layer. Apart from RSVP API functions, due to distributed nature of CCS processing, POSIX signals were also used as another way of interprocess communication (IPC). The main functions of RSVP daemon are to send, receive and process RSVP messages as well as to allocate and release optical resources emulated by resource terminals. Information necessary to affect RTs are carried by Diameter protocol. CDiameter Peer implementation was used in RSVP daemon as well as in RSVP API client. Each resource terminal has its single representation in the server from the connection control layer. The state of OXC resources under the control of connection control server is maintained in resource database which plays the role of LRM ASON component. To identify physical resources for each RSVP session some additional information was introduced to KOM RSVP code including resource ID in PSB and RSB blocks interchangeably defining physical resource in the transport plane. A functionality of Call-ID object representing call identifier according to [9] was also added. RC ASON element is based on the routing table in the operating system kernel and managed by iproute package. For resources discovery purpose link management protocol (LMP) procedures are implemented. All information necessary for proper operation of CCS server are stored in local PostgreSQL database which is accessible thanks to the mechanisms provided by libpq library.

Resource terminal is responsible for emulation of optical cross-connect (OXC) device. Resources of the emulated OXC are mapped to the content of the Oracle Database 10g Express Edition database. The state of the resources can be checked and each OXC can be configured through www interface at any time. That operation is possible through terminal T with Apache www server and PHP hypertext preprocessor which generates SQL queries to databases in resource terminals. Similarly to service control server PHP OCI8 packet is used. The main part of resource terminal is the communication module exchanging information with connection control server. Like in SCS, in resource terminal OCILIB library provides database communication mechanisms allowing to retrieve and modify the content of the Oracle database. In resource terminal CDiameter Peer is also utilized to provide communication with connection control server using Diameter protocol. Diameter messages at I_R interface transport parameters necessary to establish and release connection in the transport layer.

Service control server as well as resource terminal communicate with connection control server using Diameter protocol [8]. For this reason a new Diameter application PBZ_App and a new Diameter vendor identifier PBZ_vendor_id have been defined. Moreover, new types of Diameter messages and AVPs have been proposed. Diameter messages sent between service control server and connection control server (I_C interface) are listed in Table 1. Messages exchanged between resource terminal and connection control server (I_R interface) are presented in Table 2. Both tables contain full description of used messages and their contents. According to Diameter specification [8], all requests and corresponding answers have the same command codes. The type of the message (request or answer) is determined by "R" flag in header field, which is set to 1 for requests.

Message flow for connection creation scenario in the realized ASON/GMPLS architecture is presented in Fig. 4. Arrows with numbers correspond to consecutive stages of exchanging information. Particular messages have the following meaning:

1. Asynchronous DCN notification of changes in the CALL_STATE table of Oracle database. Changes were caused by generating request for connection creation through www/php interface.
2. Diameter CCR (Connection-Create-Request) message.
3. Execution of RSVP API createSender() function.
4. RSVP PATH message to RSVP daemon of the next connection control server (according to the routing table).
5. RSVP RESV message in case of successful resource reservation in the remaining connection control servers on the path or RSVP PATH ERROR message otherwise. Steps 6 and 7 are skipped when receiving PATH ERROR message.

Table 1
Diameter messages exchanged over I_C interface

Abbr. name	Full command name and meaning	Code	Direction	Carried parameters
CCR	Connection-Create-Request (*Request for creating connection*)	10000	SCS→CCS	– call identifier (Call-ID) – IP address of source connection control server – IP address of destination connection control server – IP address of source resource terminal – IP address of destination resource terminal – bandwidth – priority
CCA	Connection-Create-Answer (*Answer to CCR message*)	10000	CCS→SCS	– call identifier – result of connection creation
CDR	Connection-Delete-Request (*Request for terminating connection*)	10001	SCS→CCS	– call identifier
CDA	Connection-Delete-Answer (*Answer to CDR message*)	10001	CCS→SCS	– call identifier – result of connection termination

Table 2
Diameter messages exchanged over I_R interface

Abbr. name	Full command name and meaning	Code	Direction	Carried parameters
RCR	Resource-Create-Request (*Request for resource allocation – allocation of optical transport units*)	20000	CCS→RT	– call identifier – identifiers of incoming transport units – identifiers of outgoing transport units
RCA	Resource-Create-Answer (*Answer to RCR message*)	20000	RT→CCS	– call identifier – result of resource allocation
RDR	Resource-Delete-Request (*Request for resource release – release of optical transport units*)	20001	CCS→RT	– call identifier – identifiers of incoming transport units – identifiers of outgoing transport units
RDA	Resource-Delete-Answer (*Answer to RDR message*)	20001	RT→CCS	– call identifier – result of resource release

6. Diameter RCR (Resource-Create-Request) message.
7. Diameter RCA (Resource-Create-Answer) message.
8. POSIX signal of number 50 in case of successful resource reservation in all connection control servers (including source connection control server) or POSIX signal of number 51 otherwise.
9. Diameter CCA (Connection-Create-Answer) message.
10. Update of connection state in Oracle database.

Message flow for connection termination scenario in the realized ASON/GMPLS architecture is presented in Fig. 5. Arrows with numbers correspond to consecutive stages of exchanging information. Particular messages have the following meaning:

1. Asynchronous DCN notification of changes in the CALL_STATE table of Oracle database. Changes were caused by generating request for connection termination through www/php interface.
2. Diameter CDR (Connection-Delete-Request) message.
3. Execution of RSVP API releaseSession() function.
4. RSVP PATH TEAR message to RSVP daemon of the next connection control server (according to the routing table). Simultaneously Diameter RDR (Resource-Delete-Request) message to the corresponding resource terminal is sent.

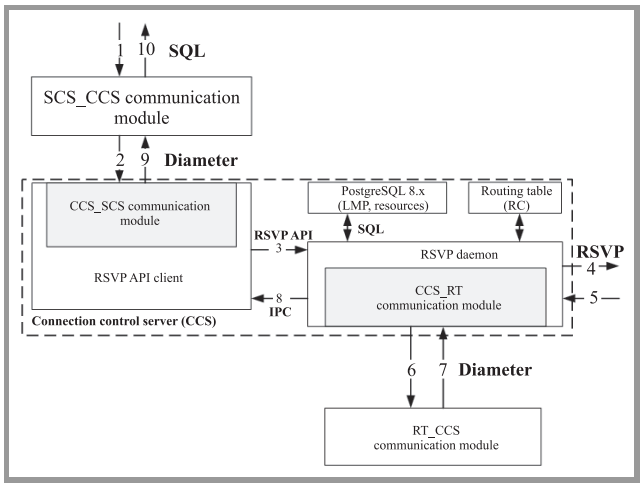


Fig. 4. Connection creation scenario in the realized ASON/GMPLS network.

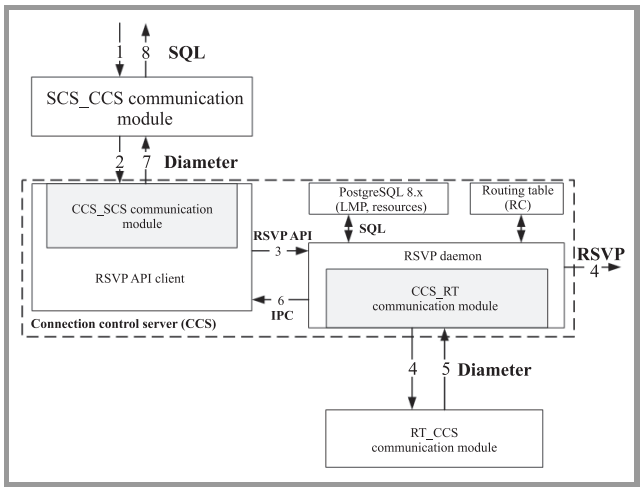


Fig. 5. Connection termination scenario in the realized ASON/GMPLS network.

5. Diameter RDA (Resource-Delete-Answer) message.
6. POSIX signal of number 52.
7. Diameter CDA (Connection-Delete-Answer) message.
8. Update of connection state in Oracle database.

5. Results and Tests

The implemented software for the ASON/GMPLS network elements has been installed and validated in laboratory testbed. Testing system was created to study not only the basic functionality, but also to investigate communication of connection control layer with the whole architecture. The structure and configuration of the testbed is described in Subsection 5.1. The results of performed functional tests are presented in Subsection 5.2.

5.1. Testbed Architecture and Configuration

Network architecture from Fig. 2 has been implemented for the purpose of ASON/GMPLS software testing. Connection control server software has been installed on the NTT TYTAN computers with the following hardware parameters:

– Supermicro X8DTL-3F motherboard,

– Intel XEON E5506 (2,13 GHz) quad core processor,

– 4GB DDR3 ECC R RAM memory,

– 2x500GB SATA HDD.

Resource terminal software has been installed on NTT computers with the following hardware parameters:

– Gigabyte GA G31M-ES2L motherboard,

– Celeron E3300 (2,5GHz) dual core processor,

– 2GB DDR2 DIMM memory,

– 250GB SATA HDD.

The architecture and the configuration of the realized ASON/GMPLS testbed (along with IP addresses of all network equipments) are described in Fig. 6. Execution of

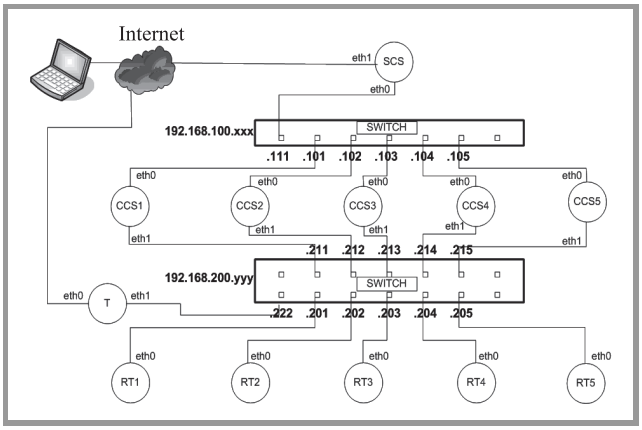


Fig. 6. Architecture and configuration of the implemented ASON/GMPLS testbed.

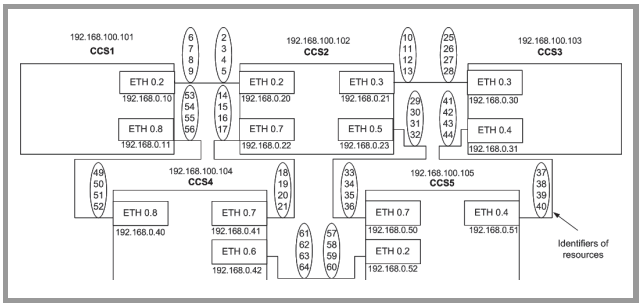


Fig. 7. Architecture of the implemented ASON/GMPLS connection control layer. ETH *x.y* stands for virtual network interface *y* based on physical network interface eth*x*.

all functional tests has been preceded by proper configuration of Debian Linux operating system and implemented software on all computers in the testbed.
Due to limited number of physical network interfaces in hardware platform, implementation of the connection control layer from Fig. 2 required configuration of virtual interfaces. The structure of the used virtual local area networks (VLANs) and the sets of optical resources identifiers assigned to particular interfaces are presented in Fig. 7.

5.2. Functional Tests

In order to verify the implemented and installed ASON/GMPLS software a set of functional test has been executed. Performed test scenarios included validation of:

- handling of single connection creation request, including the situation when there is not sufficient amount of resources to fulfill the request,
- handling of single connection termination request,
- handling of multiple connection creation requests,
- handling of connection termination request in case there are other established connections.

During execution of the scenarios the following aspects of the implemented ASON/GMPLS network operation have been checked:

- communication between service control server and connection control server using Diameter protocol,
- communication between API client and RSVP daemon,
- RSVP daemon operation:
  - PATH message handling and processing of RSVP objects (particularly RSVP_HOP object carrying optical transport unit identifiers as well as Call-ID object carrying call identifier),
  - creating PSB block carrying outgoing optical resource identifiers,
  - creating PHopSB block carrying incoming optical resource identifiers,
  - handling of RESV message with Call-ID object extension,
  - creating RSB block,
  - Diameter protocol communication between RSVP daemon and resource terminal (RT) emulating optical resources,
  - handling of PATH TEAR message with Call-ID object extension,
  - updating local resource database,
  - communication with API client in order to confirm resource allocation/release.

All performed test scenarios confirmed correctness of ASON/GMPLS architecture implementation. Results of the selected basic scenarios are described in the next part of the paper.

Handling of single connection creation request. Single connection creation request was generated using WWW/PHP graphical interface of service control server (SCS) (Fig. 8). Bandwidth of optical transport unit (TRU)

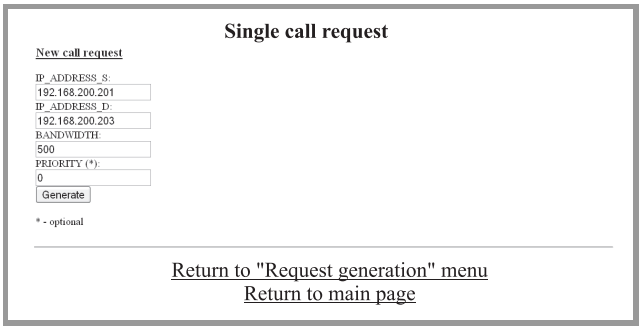


Fig. 8. Generation of connection creation request with 500 Mbit/s bandwidth.

was set to 1000 Mbit/s. The request generated in the test scenario concerned creating connection in the transport layer between resource terminals RT1 (emulating OXC of IP address 192.168.200.201) and RT3 (emulating OXC of IP address 192.168.200.203). According to Fig. 7, this connection involved one relay connection control server (CCS2) as well as one relay resource terminal (RT2). Requested bandwidth was 500 Mbit/s.
After generating the request and receiving the response from connection control server, state of the request stored in the service control server database was updated (Fig. 9). In implemented system allocation of optical TRUs is emu-

CALL_ID	IP_ADDRESS_S	IP_ADDRESS_D	BANDWIDTH	STATE	T_REQ_1	T_CONF_23	T_REQ_4	T_CONF_56
42	192.168.200.201	192.168.200.203	500	2	2010-10-26 15:45:42.649267	2010-10-26 15:45:42.967983		

Fig. 9. Content of the CALL_STATE table in the service control server database after successful connection creation.

ID	ID_OXC	ID_PORT	ID_TRU	STATE
1	1	0	1	0
2	1	2	2	1
3	1	2	3	0
4	1	2	4	0
5	1	2	5	0
6	1	3	10	1
7	1	3	11	0
8	1	3	12	0
9	1	3	13	0
10	1	5	29	0
11	1	5	30	0
12	1	5	31	0
13	1	5	32	0
14	1	7	14	0
15	1	7	15	0
16	1	7	16	0
17	1	7	17	0

Fig. 10. Content of the TRU_OXC table for resource terminal RT2 after successful connection creation between CCS1 and CCS3. TRUs with STATE = 1 are allocated.

lated in resource terminals as changes in TRU_OXC tables of local databases. Contents of TRU_OXC table for resource terminal RT2 after successful test scenario execution is presented in Fig. 10. As RT2 is a relay resource terminal, two optical TRUs are allocated (one incoming and one outgoing).

Handling of single connection termination request. Similarly to connection creation, single connection termination request was generated using WWW/PHP graphical interface of service control server (SCS). Connection termination request generated in the test scenario concerned releasing previously allocated (in subsection: Handling of single connection creation request) optical resources for the connection of unique Call-ID value (Fig. 11).

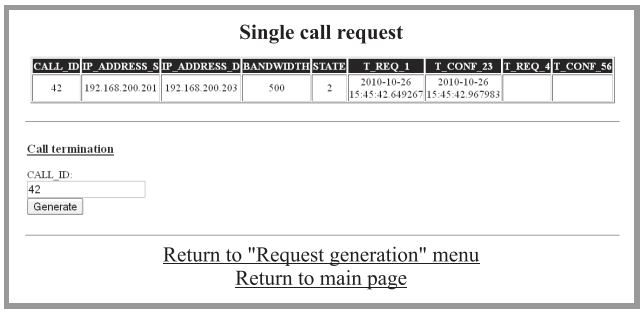


Fig. 11. Generation of connection termination request for the connection with Call-ID = 42.

The service control server software identified parameters required to release resources using the Call-ID value carried by the request generated using user interface. Once the request was processed by service control server it was sent using Diameter protocol to the connection control server which had initiated connection creation. After generating the request and receiving the response from connection control server, state of the request stored in the service control server database is updated (Fig. 12). In implemented system release of optical TRU is emulated in resource terminals as changes in TRU_OXC tables of local database. Contents of TRU_OXC table for resource terminal RT2 after successful test scenario execution is presented in Fig. 13.

CALL_ID	IP_ADDRESS_S	IP_ADDRESS_D	BANDWIDTH	STATE	T_REQ_1	T_CONF_23	T_REQ_4	T_CONF_56
42	192.168.200.201	192.168.200.203	500	2	2010-10-26 15:45:42.649267	2010-10-26 15:45:42.967983	2010-10-26 15:49:07.638234	2010-10-26 15:49:07.735563

Fig. 12. Content of the CALL_STATE table in the service control server database after successful connection termination.

ID	ID_OXC	ID_PORT	ID_TRU	STATE
1	1	0	1	0
2	1	2	2	0
3	1	2	3	0
4	1	2	4	0
5	1	2	5	0
6	1	3	10	0
7	1	3	11	0
8	1	3	12	0
9	1	3	13	0
10	1	5	29	0
11	1	5	30	0
12	1	5	31	0
13	1	5	32	0
14	1	7	14	0
15	1	7	15	0
16	1	7	16	0
17	1	7	17	0

Fig. 13. Content of the TRU_OXC table for resource terminal RT2 after successful connection termination. TRUs with STATE = 0 are free.

6. Conclusions

The aim of the work described in the paper was to implement software and testbed of ASON/GMPLS network, which elements can be mapped into ITU-T NGN functional architecture. Development of initially implemented connection control layer for ASON/GMPLS resulted in creation of the fully operational architecture with basic functionality consisting of service control layer, connection control layer and resource layer – the latter emulates control functions of real devices, optical cross-connects.

In the realization of this goal we used open-source implementations of communication protocols (RSVP and Diameter) and achieved full success. Selecting the most appropriate implementation for both protocols involved comprehensive review and tests of the available code. Finally, we chose KOM RSVP Engine and CDiameter Peer solutions as the most proper and introduced indispensable extensions. In order to use RSVP to control optical transport network consisting of optical cross-connects we had to extend chosen protocol implementation by adding new objects. Moreover, we added new Diameter application as well as messages for the purpose of communication between the particular layers of the architecture.

The implemented functionality of ASON/GMPLS was thoroughly tested in laboratory conditions. The set of tested scenarios included setting-up and releasing connections in various conditions. Performed tests validated correctness of all network elements operation including communication procedures and request processing. Furthermore, for implemented ASON/GMPLS architecture basic performance tests regarding call setup and termination operations durations were conducted. Performance results of request handling in the implemented architecture highly depend on database system used in connection control server. We applied and performed tests with PostgreSQL, Oracle 10g Express Edition databases and our own implementation based on data structures written in C language. Due to space limitation, the results of system performance tests with different types of database system are not included.

Summing up, we managed to implement ASON/GMPLS architecture, which is suitable for transport layer of ITU-T NGN proposition. As developed software is flexible, it could be readily adopted to any network architecture of optical network. Application of our solution in existing and future telecommunication networks requires further work and research concerning performance and reliability issues.

Acknowledgement

This work was partially supported by the Ministry of Science and Higher Education, Poland, under the grant PBZ-MNiSW-02-II/2007.

References

[1] "Functional Requirements and Architecture for Next Generation Networks". ITU-T Recommendation Y.2012, April 2010.

[2] "Architecture for the Automatically Switched Optical Network (ASON)". ITU-T Recommendation G.8080/Y.1304, June 2006.

[3] E. Mannie, "Generalized Multi-Protocol Label Switching (GMPLS) Architecture", IETF RFC 3945, October 2004.

[4] OIF Guideline Document: Signaling Protocol Interworking of ASON/GMPLS Network Domains, June 2008 [Online]. Available: http://www.oiforum.com/public/documents/DIF-G-Sig-IW-01.0

[5] S. Kaczmarek, M. Narloch, M. Młynarczuk, M. Sac, "Next generation services and teleinformation networks – technical, application and market aspects; Network architectures and protocols", PBZ-MNiSW-02-II/2007/GUT/2.6, Gdańsk, December 2010 (in Polish).

[6] R. Braden *et al.*, "Resource ReSerVation Protocol (RSVP) – Version 1 Functional Specification", IETF RFC 2205, September 1997.

[7] A. Fredette *et al.*, "Link Management Protocol (LMP) for Dense Wavelength Division Multiplexing (DWDM) Optical Line Systems", IETF RFC 4209, October 2005.

[8] P. Calhoun *et al.*, "Diameter Base Protocol", 3588 IETF RFC, September 2003.

[9] "Distributed Call and Connection Management: Signalling Mechanism Using GMPLS RSVP-TE". ITU-T Recommendation G.7713.2/Y.1704.2, March 2003.

[10] M. Karsten, "Design and implementation of RSVP based on object relationships", in *Proc. Networkings 2000*, vol. 1815, Springer, LNCS, 2000, pp. 325–336.

[11] S. Kaczmarek, M. Młynarczuk, and M. Waldman, "The realization of ASON/GMPLS Control Plane", in *Information Systems Architecture and Technology, System Analysis Approach to the Design, Control and Decision Support*, J. Świątek *et al.*, Eds. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej, 2010, pp. 313–324.

[12] S. Kaczmarek, M. Młynarczuk, and M. Waldman, "RSVP-TE as a reservation protocol for optical networks", in *Proc. XIV Poznań Telecommun. Worksh. PWT 2010*, Poznań University of Technology, 2010, pp. 28–31.

[13] D. Awduche *et al.*, "RSVP-TE: Extensions to RSVP for LSP Tunnels", IETF RFC 3209, December 2001.

Sylwester Kaczmarek received his M.Sc. in Electronics Engineering, Ph.D. and D.Sc. in Switching and Teletraffic Science from the Gdańsk University of Technology, Poland, in 1972, 1981 and 1994, respectively. His research interests include: IP QoS and GMPLS networks, switching, QoS routing, teletraffic, multimedia services and quality of services. Currently, his research is focused on developing and applicability of VoIP technology. So far he has published more than 200 papers. Now he is Professor and the Head of Teleinformation Networks Department at GUT.
E-mail: kasyl@eti.pg.gda.pl
Department of Teleinformation Networks
Faculty of Electronics, Telecommunications and Informatics
Gdańsk University of Technology
Gabriela Narutowicza st 11/12
80-233 Gdańsk, Poland

Magdalena Młynarczuk was born in Poland in 1980. She received her M.Sc. degree in Telecommunication Systems and Networks from Gdańsk University of Technology in 2004. Since 2008 till March 2011 she has been an assistant at Gdańsk University of Technology, Faculty of Electronics, Telecommunications and Informatics. Now she works as lecturer at GUT. Her research interests include control of optical networks, transmission and switching technology and network design. Her doctoral thesis is entitled: "QoS Routing in multi-domain optical network with hierarchical structure of control plane".
E-mail: magdam@eti.pg.gda.pl
Department of Teleinformation Networks
Faculty of Electronics, Telecommunications and Informatics
Gdańsk University of Technology
Gabriela Narutowicza st 11/12
80-233 Gdańsk, Poland

Marcin Narloch was born in Poland in 1974. He received M.Sc. and Ph.D. in Telecommunications from Gdańsk University of Technology in 1998 and 2006, respectively. Since 2006 he has kept assistant professor position at Gdańsk University of Technology, Faculty of Electronics, Telecommunications and Informatics. His research activities focus on traffic control and service design in IP QoS and GMPLS Next Generation Networks.
E-mail: narloch@eti.pg.gda.pl
Department of Teleinformation Networks
Faculty of Electronics, Telecommunications and Informatics
Gdańsk University of Technology
Gabriela Narutowicza st 11/12
80-233 Gdańsk, Poland

Maciej Sac was born in Poland in 1985. He received his M.Sc. degree in Telecommunications from Gdańsk University of Technology in 2009. Since 2009 he has been a Ph.D. student at Gdańsk University of Technology, Faculty of Electronics, Telecommunications and Informatics. His research interests are focusing on ensuring reliability and Quality of Service in IMS/NGN networks by the means of traffic engineering.

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

Paper

Multi Queue Approach for Network Services Implemented for Multi Core CPUs

Marcin Hasse, Krzysztof Nowicki, and Józef Woźniak

Gdańsk University of Technology, Gdańsk, Poland

Abstract—Multiple core processors have already became the dominant design for general purpose CPUs. Incarnations of this technology are present in solutions dedicated to such areas like computer graphics, signal processing and also computer networking. Since the key functionality of network core components is fast package servicing, multicore technology, due to multi tasking ability, seems useful to support packet processing. Dedicated network processors characterize very good performance but at the same time high cost. General purpose CPUs achieve incredible performance, thanks to task distribution along several available cores and relatively low cost. The idea, analyzed in this paper, is to use general purpose CPU to provide network core functionality. For this purpose parameterized system model has been created, which represents general core networking needs. This model analyze system parameters influence on system performance.

Keywords—generic purpose CPU, multi core, network, queue, network services.

1. Introduction

In the recent years, CPU technology has turned into multi core to brake the clock barrier and improve applications performance. Higher solutions performance require much faster medium to transfer data. Computer networking remains not only a medium, but also network software stack, which is a significant part of the network performance. Once application efficiency is improved by using CPU technology, networking stack software could also be considered as area, were multi core can increase productivity. There are many hardware CPU architectures used for high speed network packet processing. Network processors Intel IXP28XX series [1] introduce hardware micro engines able to process pipeline oriented traffic with significant performance improvement. More recent design from Cavium Networks called OCTEON Multi-Core Processor Family [2] or Freescale [3] are providing hardware driven opportunities to divide multi flow traffic into separated cores. These sophisticated hardware designs characterize high cost of implementation and deployment. Cost of these high performance solutions are significant but still worth their price to satisfy network needs.

From the software point of view there are multiple implementations of network stacks, protocols and solutions provided by the market today. These solutions characterize high performance, modular architecture and relatively easy integration. Due to this they already have strong market position for dedicated network solutions. There are also lots of research going on to optimize hardware support for multiple software threads [4] and to provide possible highest performance for computer network nodes. In personal computers segment multi core CPUs have also strong presence. More and more end user applications taking advantage of this solution and increasing their performance by adding software support for multi core hardware architecture.

In this paper, multi queue approach to network services implemented in multi core general usage CPUs [5] is presented. This approach has been proposed to verify, if this is reasonable to optimize network performance for solutions with standard CPUs. Most of new software architectures, which support multi core environments, are dedicated to data processing but not to the packet processing. It important to identify how much software architecture design can influence on network application performance.

2. Multi Core Architecture Approach

Nehalem is the codename for an Intel processor microarchitecture [6], [7]. The most popular available in end user market is Intel's Core i7 [8] processor. Higher performance is achieved on this CPU by processing multiple data in the same time using parallel cores. Most of operating systems are providing multi core support and assigning tasks to be processed by hardware with possible highest performance. An operating system is not focusing on the type of application beside I/O bound or CPU bound types distinguished by the task scheduler [9]. This approach might not be enough to support high performance networking and low performance management and monitoring activities, running at the same time. In the server area this problem has been solved by introducing visualization on both application and platform levels [10], [11]. Many visualization solutions including hypervisor [12], became standard parts of the operating systems [13] providing opportunities to work with several contests at the same time. The software architecture described in this paper also account visualization approach in order to work with different contests on multiple CPU cores. Such solutions like that are available now - for instance Sun xVM Virtual Box [14], which allows multiple operating systems run the same time on a single PC.

Network applications are responsible for servicing data streams. Network streams are transporting significant amount of data, which needs to be processed and depending on application functionality either consumed (provided to the end user) or transmitted (back to the network). Effective stream processing could have significant impact on networking application performance.

For this research purposes, there has been defined a network stream consisted of a limited sequence of data packets $D^1, D^2, \ldots$. To simplify the mathematical model we assume, that each packet represents different amount of data but has the same, permanent defined size. Such approach is often used to specify application throughput for defined packet size

$$D_k = \{D_k^1, D_k^2, ..., D_k^n\}, n \in N. \tag{1}$$

The network application can get streams from several different sources

$$S = \sum_{k=1}^{K} D_k, k \in N. \tag{2}$$

Each stream, responsible for delivery of potentially different sets of information, is directed to different applications for different sorts of processing. This diversity can be a good approach for the system design. In our model we additionally accept, that some streams require more privilege (priority) service and task delivery to the end user.

Multi core CPUs, ensure better performance by redirecting tasks to be executed in the same time on multiple cores. Each core offers the same execution condition including access to peripherals, memory etc. [15]. Properly configured operating system is able to manage tasks to assure possible best performance. In this paper authors propose to change the task approach for multi core CPUs provided by most modern operating system to the network stream approach. The intention on this research is to verify, whether it is reasonable enough to consider general purpose CPU as hardware platform to systems dedicated for networking.

Standard operating system (like, e.g., Linux kernel 2.6.X) in general treats I/O coming from networking card as interrupts coming from any other computer peripherals. Analyzing different communication devices from the data rate per second point of view one can differ between devices (PCI bus 528 MB and gigabit Ethernet 128 MB), however communication expectation is to service significant amount of data in shorter possible time. The network card driver in the Linux kernel associated with networking stack, is able to provide packet delivery services to dedicated application. In cases of many different streams, which need to be serviced by different applications existing in OS together with another, not strict network related tasks, can meet computer performance critical point. When critical point is crossed, it could manifest long tasks queues and delays in service (see Fig. 1).

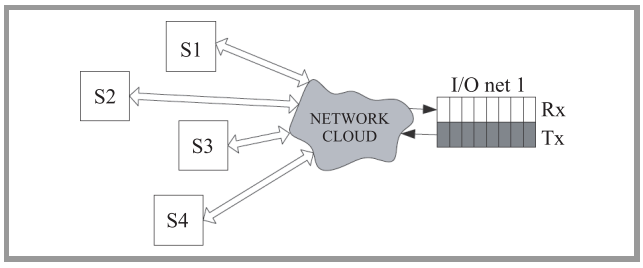


Fig. 1. Single package queue for single physical interface.

Considered system approximation assumes single package queue associated with physical interface. Packages in the queue are arriving from a multiple resources. Each data stream D_k in queue income, specified by amount of data S, should be serviced by different application. Performance of networking system from this approximation can be determine by specifying time t in which dedicated networking application completed service of S

$$P_s = S(t). \tag{3}$$

In the recent days networking technology is based either on networking pipeline [1] or dedicated multi core solutions [2]. These solution offers different approaches for networking applications than standard applications service on generic purpose CPU. In the networking area application can be divided into the following areas:

- Management/Controlling – responsible for controlling network settings, managing network events; type of traffic – control plane able to work with full Linux stack;
- Preprocessing – responsible for redirecting packets to exceptions or forwarding; type of traffic – slow path/fast path able to work with limited L3 stack to classify packages;
- Exceptions – responsible for servicing packages, dedicated to particular node – type of traffic – slow path able to work with full Linux stack;
- Forwarding – responsible for fast forwarding packages to additional network segments; type of traffic – fast path able to work with basic/limited Linux stack.

Each of these areas has different requirements for throughput and different user expectations related to accessibility, stability and manage ability. A general purpose operating system is not providing a different set of system resources for networking application areas, beside the ability to configure different application with different priorities [9]. For example performance of Linux application depends on number tasks executed on CPU in specified amount of time, and scheduler latency can cause unexpected delays for application, which require quick system reaction time.

Multicore CPUs, when hypervisor is used, provides ability to distinguish different system expectation and execute each networking application area on separated core [16]. In this model physical network interface can be assigned directly to OS running with fast and limited Linux stack responsible for preprocessing. This networking application role is to

classify packet and as soon as possible send it to exception or forwarding applications running on additional cores and with dedicated networking stacks. Last core in the system is responsible for dealing with preprocessing, exceptions and forwarding rules. Its management and control interface should be easy available for system administrator though slow port management interface (Fig. 2).

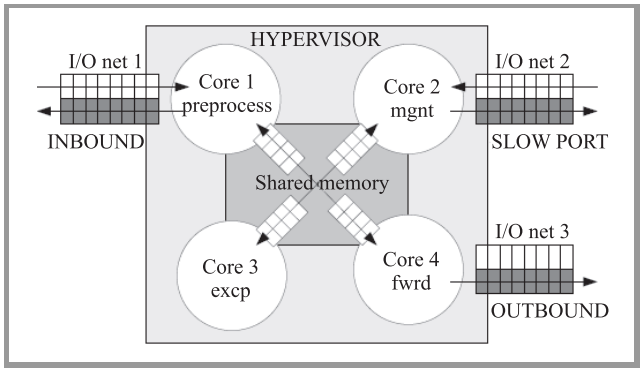


Fig. 2. Network system architecture for general purpose CPU.

Hypervisor role is to physically assign networking interfaces to dedicated cores. Communication between cores is delivered trough shared memory available from each core. This approach allows to install on each core [17] dedicated version of OS (with dedicated networking stack) where physical access to I/Os is pre configured by hypervisor. Interface access latency should be much lower than for solutions with master OS [13], [14].

3. Common Network Traffic Models

Generic approach for traffic modeling is to mathematically describe the physical arrival of packets as a point function of countable values. Points describe packets arrival instances starting from $T_0 = 0$ and is limited only to model assumed time frame: $T_0, T_1, \ldots, T_n, \ldots$. Countable values are usually dependent on two point stochastic processes – one, describing time distance between arrival instances and second, describing actual value/number of delivered packets. Packet arrival time (PAT) $\{PAT_n\}_{n=1}^{\infty}$ process is non negative random sequence:

$$PAT_n = T_n - T_{n-1}\,. \tag{4}$$

The point process is describing each instance of packet arrival:

$$T_n = \sum_{k=1}^{n} PAT_k\,. \tag{5}$$

Number of delivered packages (NDP) $\{NDP_n\}_{n=1}^{\infty}$ is associated with amount of work – workload (WLD), which has to be done to service traffic in a specified time. It can be limited by bandwidth bottleneck and can be described via stochastic function D_n

$$NDP_n = \begin{cases} D_n(T_n, \tau) & \text{if } \tau <= \tau_{max} \\ D_n(T_n, \tau_{max}) & \text{otherwise} \end{cases}\,. \tag{6}$$

It can be useful in traffic modeling to incorporate also function describing workload WLD_n associated with n-th delivered NDP. This workload can be dependent on queue delays in case traffic is redirected between multiple queues.
Both stochastic functions PAT_n and NDP_n are characterized by independent distributions. In some systems also the workload factor can be described by a stochastic distribution.
In related works authors describe different stochastic functions in each component of the traffic model. One of the oldest traffic model, which has been used in many analyzes uses Poisson process [18]. This process assumption random sequence arrive independently from one another and depends on constant value. This model can be successfully used to model different types of traffics like ex. VoIP [19]. There are also studies, which proves that Poisson modeling can fail [20] and using fractal (self similar) model [21] based on results analyze of hundreds of million packets observations in LAN and WAN area, can be more effective. Self-similar model cumulates traffic volume as self-similar process with increments that are strictly stationary to specified shifts in time. Alternative model for Poisson independence in arrival can be Markov process, which introduces dependence into random sequence. Markov chains can be used for example for TCP traffic classification [22], however the most commonly used Markov model is Markov-modulated Poisson process (MMPP) model, which can be used for modeling self-similar traffic [23], [24].

4. Networking System Model

The easiest way to determine possible system performance is to crate parameterized mathematical model and calculate system throughput indicators. There are many factors, which can affect system performance, in the multicore system architecture. Some parameters remain constant as shared memory access latency and some depend on other aspects like number of additional tasks executed on OS-es dedicated to separated cores.
QIRx is an inbound queue associated with I/O net1 in Fig. 2. This interface delivers set of K data streams D_k to the Core 1 OS networking stuck for preprocessing. Accepting burstiness limitation of Poisson process model, traffic can be described through a counting process:

$$D_k\{N(t+\lambda) - N(t) = n\} = \frac{(\tau\lambda)^n e^{-\tau\lambda}}{n!}\,, \tag{7}$$

where $N(t)$ is the number of packet arrivals at time t from single source.

$$S(t) = \sum_{k=1}^{K} D_k(n_k), k \in N \tag{8}$$

For each stream k number of arrival n_k is called stream activity level (SAL) and should be represented by different constant value. This model parameter correspond to NPD_n and could be parametrized by stochastic function.

Preprocessing core is responsible for saving packages from *QIRx* queue in dedicated shared memory queues: *SSME* – exception and *QSMF* – forwarding. This model assumes that in *QIRx* for every 10 packages η is classified as forwarding and $10-\eta$ as exception. Shared memory queue can service p packages in time t, but core ability to read and write to the queue also depends on several other factors. Our model limits these factors to number of tasks in CPU queue. Higher number of tasks in queue can increase CPU average waiting time [25] and writing/reading operation can be delayed. Little's formula describes relationship between L – average number of processes in the queue and W – CPU average waiting time. Parameter α can be associated with average arrival time, which is proportional to traffic rate.

$$W = \frac{L}{\alpha}. \tag{9}$$

System performance is often determined by traffic processing latencies. Sum of whole latencies specified by system model is able to show how long traffic will be processed by the system. In this system model δL_1 is latency specified between *QIRx* and *QSMF*/*QSME*

$$\delta L_1(t,\tau,L) = \frac{L_p}{S(t)}. \tag{10}$$

Latency δL_2 is specified between *SSME*, *QSMF* and its target interface. For exception core target interface would be $QITx$ and for forwarding it would be $QFTx$.

$$\delta L_2(t,\tau,L,\eta) = \frac{\frac{\eta}{10}L_f}{S(t)} + \frac{(1-\frac{\eta}{10})L_f}{S(t)} \tag{11}$$

Assuming, that traffic exception service suppose to notify packet sender about exception condition, there need to be considered additional process in the CPU preprocessing, responsible for dealing with back to sender traffic.

$$\delta L_2(t,\tau,L,\eta) = \frac{\frac{\eta}{10}L_f}{S(t)} + \frac{(1-\frac{\eta}{10})(L_f+L_p+1)}{S(t)} \tag{12}$$

Total system latency should also consider hypervisor latency. For this model this is constant value δL_h.

$$\delta L(t,\tau,L,\eta) = \delta L_1(t,\tau,L) + \delta L_2(t,\tau,L,\eta) + \delta L_h \tag{13}$$

For the sake of simplicity the model has been limited to consider only significant factors, which can influence the system performance. It can be extended to provide more detailed data, if this accuracy level is not satisfied enough to determine its value against standard general purpose CPU system. The authors' intention was to show how fluctuations of commonly accepted factors can affect the modeled system performance, e.g., influence the packet service latency.

5. Model Parameterization Results

Networking system model performance and scalability depends on several configurable system parameters, starts from incoming traffic, through different latencies and finish at operating system process utilization ability. Value of model presented in this paper is ability to verify, how system model parameters can influence whole system performance.

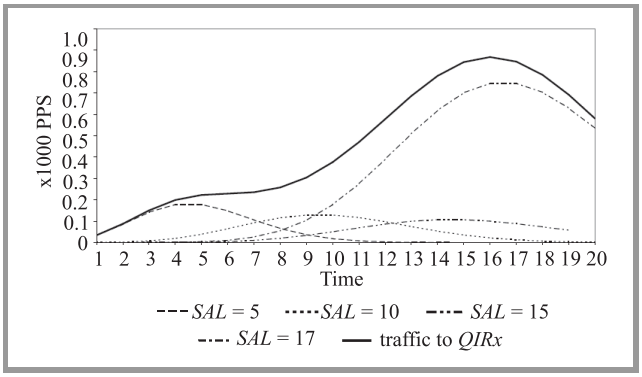


Fig. 3. Data stream distribution associated with QIRx.

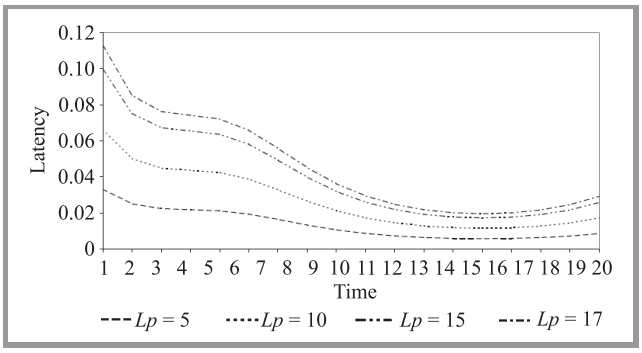


Fig. 4. Latency between QIRx and QSMF/QSM.

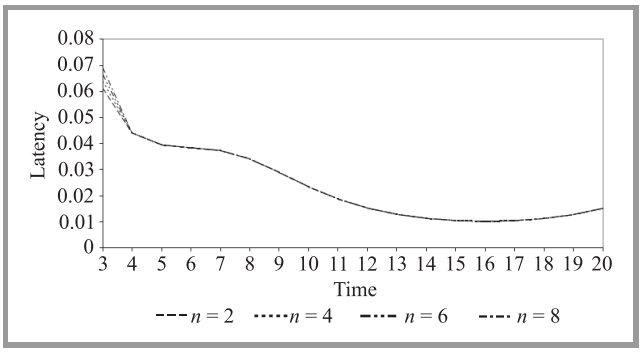


Fig. 5. Latency between SSME, QSMF and its target interface, $L_p = 5$, $L_f = 5$ – approach 1.

Sample charts presented on Figs. 3–8 provides overview of system model parameters influence on measured latency. Multi queue approach assumes several parameters affecting performance value in more or less significant way. For example, number of task, which need to be serviced by OS in presented system model plays marginal role. Another parameters can affect time, in which packages are serviced, in more significant way. Data stream distribution represented by stochastic model and parametrized by system activity

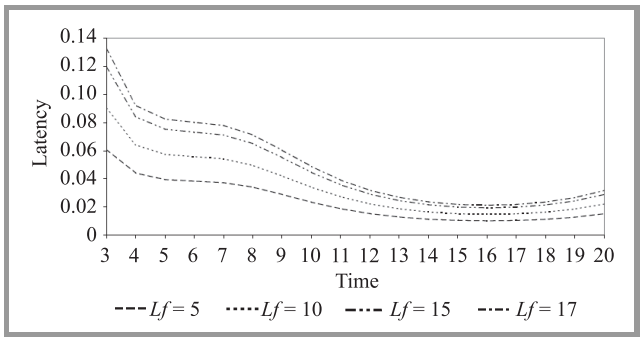


Fig. 6. Latency between SSME, QSMF and its target interface, $n = 2$, $L_p = 5$ – approach 2.

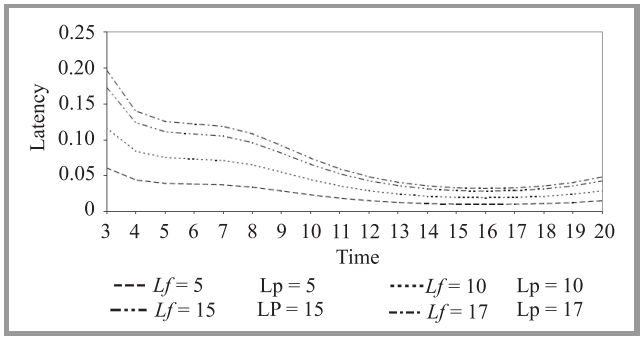


Fig. 7. Latency between SSME, QSMF and its target interface, $n = 2$ – approach 3.

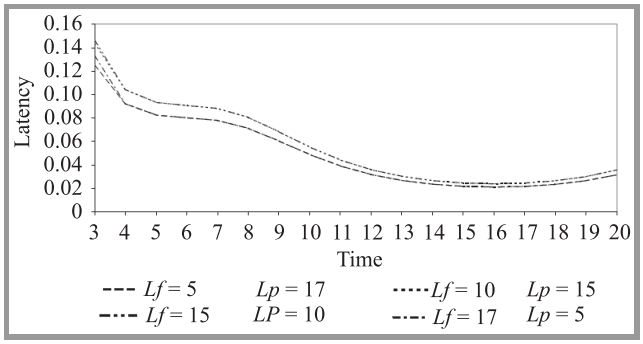


Fig. 8. Latency between SSME, QSMF and its target interface, $n = 2$ – approach 4.

level can directly influence exception and forwarding queue latencies. Naturally higher number of transmitted packets (average arrival time) affect latencies and can cause system delays. Size of calculated lag seems to be reasonable small and its effect to whole system for most of the cases should be negligible, however system designers/architects should be aware, that in case of adverse stream distribution unexpected delays can happen.

6. Summary and Conclusions

The need for flexibility and performance and cost reduction in the networking systems make general purpose CPUs worth to be considered as valuable alternative for expensive solutions designed for packet processing. If system architecture agrees for general purpose computing limitations (like ex. us speed), there can be considered many system designs, which would make general purpose personal computer dedicated networking solutions. Fast development of CPU technologies allows to assumes, that CPUs dedicated to common marked are more capable to play valuable roles in dedicated (not generic) solutions. Good example can be presented in this paper usage of hypervisor technology in designing dedicated system model. The only limitation in the possible system architecture designs can be imagination of system architects. System model can help verify usability of dedicated solution assuming hardware/software limitation of model parameters. Networking system presented in this paper can be a good reference for further work, which could bring more detailed model providing more complex analyze of system performance indicator like queues delays, latencies as well as more self similar delivery distribution to better present ex. Ethernet characteristic. It has been proved that idea with specialized core functions (forwarding, exception), due to relatively small latencies caused by between core communication, could open easy way for generic purpose CPU usability in niches like eg. computer networking. Solutions based on generic purpose, multicore CPUs could be truly considered in complex, functionality oriented system designs.

Acknowledgment

The work was partially supported by the Polish National Center for Research and Development under the PBZ grant MNiSW-02/II/2007.

References

[1] "Intel IXP4XX product line of network processors", http://www.intel.com

[2] "OCTEON Multi-Core Processor Family", http://www.cavium.com/OCTEON_MIPS64.html

[3] "Semiconductors overview", http://www.freescale.com/

[4] "Improving network performance in multi-core systems", in Intel Corporation White Paper, http://www.intel.com

[5] "Intel CPU documentation", in Intel CPU, http://www.intel.com/design

[6] "Press release", in Intel Corporation, March 2007, http://www.intel.com/pressroom

[7] L. Shimpi, "AnandTech", June 2008, in The Nehalem Preview: Intel Does It Again, http://www.intel.com

[8] "Intel Core i7" in The Nehalem Preview: Intel Does It Again, http://www.intel.com

[9] M. Hasse and K. Nowicki, *Linux Scheduler Improvement for Time Demanding Network Applications, Running on Communication Platform Systems*. Gdańsk, Polska: Politechnika Gdańska, 2011.

[10] P. Barham *et al*., "Xen and the art of virtualization", in *Proc. ACM Symp. Operat. Sys. Principles*, New York, USA, 2003.

[11] A. Gavrilovska *et al*., "High-performance hypervisor architectures: virtualization in HPC systems", in *Proc. 1st Worksh. System-level Virtu. High Perform. Comput. HPCVirt 2007*, Lisbon, Portugal, 2007.

[12] "A Performance comparison of hypervisors", in *VMWare Performance Study*, Technical paper VMWare.

[13] "Guide to virtualization on Red Hat enterprise Linux", in *Virtualization Guide*, http://docs.redhat.com

[14] "Virtual box reference", in *Sun xVM Virtual Box*, http://www.virtualbox.org

[15] C. Pitter and M. Schoeber, "Time predictable CPU and DMA shared memory access", in *Proc. FPL 2007*, Amsterdam, The Netherland, 2007, pp. 317–322.

[16] K. Kolyshkin, "Virtualization in Linux", September 2006, http://www.pdfmob.com

[17] R. Ennals, R. Sharp, and A. Mycroft, "Task partitioning for multi-core network processors", in *Proc. Eur. Symp. Programming ESOP*, LNCS, 2005, vol. 3443, SpringerLink.

[18] L. Moddelmog and P. Johnson, "Poisson distribution", February 2006 [Online]. Available: http://pj.freefaculty.org/stat/Distributions/Exponential_v2.lyx

[19] I. Al Ajarmeh, J. Yu, and M. Amezzine, "Framework of applying a non-homogeneous Poisson process to model VoIP traffic on tandem networks", in *Proc. 10th WSEAS Int. Conf. Applied Informatics and Communications AIC 2010*, Taipei, Taiwan, 2010, pp. 164–169.

[20] V. Paxson and S. Floyd, "Wide-area traffic: the failure of Poisson modeling", *IEEE/ACM Trans. Netw.*, vol. 3, no. 3, pp. 226–244, 1995.

[21] W. Leleand, M. Taqqu, W. Willinger, and D. Wilson, "On the self similar nature of Ethernet traffic", *IEEE/ACM Trans. Netw.*, vol. 2, no. 1, pp. 1–15, 1994.

[22] G. Munz, H. Dai, L. Braun, and G. Carle, "TCP traffic classification using Markov models", in *Proc. Traffic Monitoring and Analysis Workshop TMA 2010*, Zurich, Switzerland, 2010, pp. 127–140.

[23] A. Nogueira, P. Salvador, R. Valadas, and A. Pacheco, "Modeling self-similar traffc through Markov modulated Poisson processes over multiple time scales", *Telecommun. Sys.*, vol. 17, no. 1–2, pp. 185–211, 2001.

[24] S. Scott and P. Smyth, *The Markov Modulated Poisson Process and Markov Poisson Cascade with Applications to Web Traffic Modeling.* Oxford University Press 2003.

[25] J. F. Brady, "Virtualization and CPU wait times in a Linux guest environmnet", January 2008.

Marcin Hasse received the M.Sc. degree in Telecommunication from the Gdańsk University of Technology, Poland in 2005. Currently he is working for embedded computing leading company providing solutions for telecommunication market. His research interest and current work are related to operating system improvements for networking/telecommunication usage scenarios. He is author of several publications in computer networking mechanisms improvements for end user services.
E-mail: marcin@hasse.pl
Gdańsk University of Technology
G. Narutowicza st 11/12
80-952 Gdańsk, Poland

Krzysztof Nowicki received his M.Sc. and Ph.D. degrees in Electronics and Telecommunications from the Faculty of Electronics, Gdańsk University of Technology (GUT), Poland, in 1979 and 1988, respectively. He is the author or co-author of more than 150 scientific papers and author and co-author of five books, e.g., "LAN, MAN, WAN – Computer Networks and Communication Protocols" (1998), "Wired and Wireless LANs" (2002) (both books were awarded the Ministry of National Education Prize, in 1999 and 2003, respectively), "Protocol IPv6" (2003), "Ethernet-Networks" (2006), Ethernet End-to-End. Eine universelle Netzwerktechnologie (2008). His scientific and research interests include network architectures, analysis of communication systems, network security problems, modeling and performance analysis of cable and wireless communication systems, analysis and design of protocols for high speed LANs.
E-mail: krzysztof.nowicki@eti.pg.gda.pl
Gdańsk University of Technology
G. Narutowicza st 11/12
80-952 Gdańsk, Poland

Józef Woźniak is a Full Professor in the Faculty of Electronics, Telecommunications and Computer Science at Gdańsk University of Technology. He received his Ph.D. and D.Sc. degrees in Telecommunications from Gdańsk University of Technology in 1976 and 1991, respectively. He is the author or co-author of more than 250 journal and conference papers. He has also co-authored 4 books on data communications, computer networks and communication protocols. In the past he participated in research and teaching activities at Politecnico di Milano, Vrije Universiteit Brussel and Aalborg University, Denmark. In 2006 he was Visiting Erskine Fellow at the Canterbury University in Christchurch, New Zealand. He has served in technical committees of numerous national and international conferences, chairing or co-chairing several of them. He is a member of IEEE and IFIP, being the vice chair of the WG 6.8 (Wireless Communications Group) IFIP TC6 and. For many years he chaired the IEEE Computer Society Chapter at Gdańsk University of Technology. His current research interests include modeling and performance evaluation of communication systems with the special interest in wireless and mobile networks.
E-mail: jozef.wozniak@eti.pg.gda.pl
Gdańsk University of Technology
G. Narutowicza st 11/12
80-952 Gdańsk, Poland

Paper

Active – Passive: On Preconceptions of Testing

Krzysztof M. Brzeziński

Institute of Telecommunications, Warsaw University of Technology, Warsaw, Poland

Abstract—In telecommunications and software engineering, testing is normally understood to be essentially *active*: a tester is said to stimulate, control, and enforce. Passive testing does not fit this paradigm and thus remains the niche research subject, which bears on the scope and depth of the obtained results. It is argued that such limited understanding of testing is one of its many community-bound preconceptions. It may be acceptable in the current engineering approach to testing, but can and should be challenged in order to converge on the core concepts of the proposed science of testing ("testology"). This methodological work aims at establishing that there are no fundamental reasons for admitting the dominant role of the active element in testing. To show this, external (also extra-technical) areas are consulted for insight, direct observations, and metaphors. The troublesome distinction between (passive) testing and monitoring, as well as unclear relations between testing and measurements, are also addressed.

Keywords—behavior, development, metrology, monitoring, passive testing, reactive systems, Scientific Method.

1. Introduction

Testing is intertwined with the development (creation, construction, and further use) of artifacts – intentionally designed objects. Artifacts of a certain complexity are "mechanically" referred to as artificial *systems*. We understand testing as the umbrella term for a particular set of concepts, methods, and techniques of verification and validation (V&V), i.e., assessing whether a system is *correct* w.r.t. a given notion of correctness. This assessment leads, pragmatically, to deciding whether a system is *acceptable*.

Testing cannot be replaced by other, "non-testing" V&V techniques. Placed in a loop of development activities, testing is a vital element in achieving and maintaining correctness (and thus quality) of systems. The complexity of testing, however, is known to grow *exponentially* in the complexity of tested systems. Accordingly, despite the undisputable improvements in testing concepts and techniques, spectacular system failures (including those that entail loss of life), attributed to inadequate testing, still happen. In order to sustain the pace of development of complex systems (including telecommunications systems), it is necessary to seek improvements in testing *beyond* its current, relatively steady development. The aim of this work is to contribute towards this end. Its underlying assumptions and theses are briefly presented below (see [1]–[5] for discussion).

Testing is currently researched and practiced mostly by specialized groups, or schools, within separate communities concerned with particular classes of systems to be tested. The immediate context of this work are systems characteristic of information and communications technology (ICT) – a field defined by convergence of traditional *telecommunications* and informatics (*software engineering* and *computer science*). The convergence of concepts and approaches to the development (and thus – also to testing) of ICT systems is far from complete. It thus makes sense to refer, within ICT, to separate communities of *software* testing, *protocol* testing, *circuit* testing, etc. There are also groups concerned with testing outside ICT (*chemical* testing, *material* testing, etc.), with their own, important insight.

Testing communities speak particular languages (or, to quote Wittgenstein, they play different *language-games*). They are reluctant to borrow the concepts and terms from peer groups. Consequently, any *preconceptions* they may have on testing cannot be easily confronted with other patterns of understanding, and are very hard to uproot (even if they clearly form a crippling self-restriction). This phenomenon of conceptual and linguistic (terminological) "lock-in" has been noticed, e.g., by Lamport [private communication, 2010], who called it a "Whorfian syndrome".

Testing schools tend to follow the *engineering* approach, with its *apprentice* tradition of vocational study. The testing concepts and terms are defined stipulatively, to mean what a given community *wants* them to mean. This particular understanding, as well as skills for its practical use, are taught, and then checked during exams that lead to obtaining professional titles of a "certified tester" or the like. Accordingly, there are sources of community-bound "standardized knowledge" of testing [6]–[11]. There is, however, no common definition of testing that would be accepted *across* the testing communities. Lack of such definition indicates a serious problem with testing, as "*a definition influences future perceptions – a too narrow or misleading one may block future investigations for a long time*" [12].

In order to gain new perspectives, impetus, and funding, testing needs to properly address the issues identified above. To do so, it should transcend the limitations imposed by the apprentice model, and establish itself as a *science*, with academic recognition. This seems necessary not only for immediate professional application of testing, but also for its proper *teaching*, in a way that avoids seeding and perpetuating the existing preconceptions in the new generation of researchers – a concern that is not unique to testing [4]. This path has already been taken by *metrology* – the "*sci-*

ence of measurement and its application" [13, 2.2]. For the postulated new testing science, of the *design science* kind, we propose the name: "Testology". It would be the producer and bearer of *first principles* and *core concepts* of testing, regardless of its area and context of application, and would also allow forming the "applied testing" subsets and specializations, with meaningful relations to each other. Just as any other (design) science, testology is free to seek linguistic and conceptual *metaphors* [14] without any *a priori* restriction of the range of possible "donors", and to look into languages currently spoken by particular testing schools, in the hope that "*a core theory... can be synthesized from writings across a number of disparate fields*" [15]. Also the general understanding of "testing", encoded in everyday language and reflected in dictionary entries, should not be neglected. In this context, the existing sources of vocational knowledge on testing, including the definition(s) of testing contained there, are only one of the inputs available for consideration, and not *the authoritative* body of the concepts of testology.

To illustrate the postulated approach to testing, in the sequel we focus on a single idea that currently prevails in virtually all ICT-related testing schools, namely, that *testing is active*. We argue that it is a community-bound preconception – a conventional disciplinary modification of the concept of testing, which is not substantiated by any "deeper roots" of emerging testology. We further argue that sticking to this preconception is an unnecessary handicap for applied testology – valuable research on *passive testing* is currently conducted away from the mainstream, in a *niche* research area, which bears on the breath, depth, and consistency of obtained results. We claim that abandoning the "active" preconception should bring more consistency to the mainstream of testing research, by allowing the uniform treatment of both active and passive testing, which in turn may contribute to the postulated "nonlinear" improvement in testing.

2. Status of Active and Passive Testing

In telecommunications, software engineering, and most other technical disciplines, the prevailing intuition of testing is reflected in its operational characterization as an activity (stating what is being *done* while testing), in which a tester:

- generates and *applies* (sends) stimuli, or "test input data", in order to *control* a system under test (**Sut**) – to provoke and guide phenomena (in our case – mainly behavior) to be investigated by testing;
- *observes* phenomena as they appear under the influence of applied stimuli;
- *analyzes* the relation between observed phenomena and some reference (such as a pre-computed, intended behavior);
- *decides* on a suitable verdict, which expresses the assessment made.

This operational characterization is often taken as the *operational definition* of testing: all the enumerated operations are quite tangible, and their joint presence is said to constitute what shall (and, by complement, what should not) be regarded as testing. This characterization is then implicitly employed in the role of the definition of testing (as in [16, pp. 14–16], where, on 600+ pages, no other *explicit* definition of testing is given), or is suitably rephrased, as in [17]: "*The principle of testing is to apply inputs... and to compare the observed outputs to expected outputs*". Similar definitions[1], in various wording, prevail in "official", vocational compendia, dictionaries of terms, and meta-standards, and are also cited in research papers. Their common element is that they stipulate the *active* character of testing – a tester controls, solicits, and enforces. Active testing constitutes the mainstream of testing.

On the other hand, since the early 1980s there has been ongoing interest in *passive testing*, technically defined as a testing activity in which a tester does not influence (stimulate) a **Sut** in any way – it does not apply any test stimuli. Two typical approaches to such testing may be identified.

One party claims that the active character of testing reflects its *essence*, and thus cannot be surrendered. It is natural for this party to maintain that "passive testing" simply does not respond to the concept of testing – that it is a spurious interpretation, a mere *façon de parler*, or the case of confusion of tongues. Indeed, passive testing has not been identified as a dimension of the discourse space of testing, nor even alluded to, in the annotated bibliography of formal testing [22], the proceedings of the prestigious Dagstuhl Seminar on testing [23], taxonomies developed to get insight into the notion of testing [24], [25], standardized glossaries of terms pertaining to testing [6], [7] or broader software engineering activities [19]. It is also, apparently, not covered by the new, forthcoming international software testing standard ISO 29119. The telecommunications-oriented methodology of *conformance testing* [26] openly excludes passive testing from its scope. The standardized test language TTCN-3 [27] was meant to express active tests, and there have been very few proposals to re-use it also for passive testing [28].

The other party investigates passive testing basing on its arbitrarily adopted technical definition, without any deeper concern for methodological harmonization with active testing. This is how the majority of valuable results on passive testing have been achieved so far. In order to avoid the "politically incorrect" term, various euphemisms [3] are used: *observer*, *trace checker*, *the oracle*, *passive monitor*, *arbiter*, *supervisor*. Another indication of the present niche character of research on passive testing is its apparent discontinuity: frequent "restarts" and "re-inventions" of its key elements – a phenomenon not unknown in science, but in

[1] "*Implementation... is exercised with selected sequences of inputs*" [18]. "*... the process of operating a system or component under specified conditions* [as explained elsewhere – understood to be imposed by a tester]..." [19]. "*... testing always implies executing the program on (valued) inputs*" [20, ch. 5]. "*Software testing involves... systematically executing the software, while stimulating it with test inputs...*" [21].

this case it is particularly easy to be ignorant of previous work on passive testing (see [3] for examples).

Between these two approaches, there is an apparent gap: very little has been written on the fundamental methodological issue of whether passive testing "should" be admitted as *bona fide* testing. Serious attempts at starting a discussion at the meta-level, *about* passive testing, are known to have been vigorously rebuffed, as unnecessary, idle, and – allegedly – showing disrespect for "established and accepted truths". This stance is understandable in the vocational, *engineering* tradition, with its apprentice model, but questioning the present state of a conceptual framework is natural, healthy, and indispensable for any *science*, and should not be confused with the "know-better" attitude. This is yet another justification for testology.

Apart from the intellectual challenge of establishing a place for passive testing within testology, it can be shown that there is the growing *need* for it. Recently, a stream of reservations has been raised, by different authors, concerning the ability of testing, *as it is traditionally understood*, to respond to evolving needs, as briefly surveyed below.

Among the new tendencies and postulated further developments of formal model-based testing, [23] identifies:

- integration of test techniques, in order to be able to choose for every task their best combination,
- accepting that a product, however thoroughly tested, evolves and changes.

Both postulates may be re-cast in terms of passive testing, in the following way. Passive testing may be considered as a particular set of combinations of constituent elements, or "modules", of the general testing methodology; active testing would then be *another*, different set of such combinations. This conceptual and technological modularity was proposed by this author already in 1996 [29], and it has been researched since then under the name of protocol multimeter (PMM) [30]. One of the hypotheses tacitly adopted for active testing is that a system under test does not change *during* the tests, and that it is still meaningful to refer to test verdicts *after* the testing is over [31]. In fact, however, *all* real-world implementations do change, in unexpected ways and moments in time. This makes active testing, performed in finite sessions, inherently inconsistent with its hypotheses[2], while the "campaign-less" passive testing is not affected.

In his unpublished keynote speech at the recent software testing conference (ICST, Berlin, 2011), Ian Sommerville, the authority on the design and testing of ICT systems, put to doubt the universal validity of very foundations of testing, as it is currently researched and practiced within ICT. He identified these foundations as deriving from Hume's *reductionism* – reducing complex systems into manageable parts, simple enough to be understood, and interpreting the whole system in terms of interactions of these parts. This approach is conspicuous in the succession of activities in software testing: *unit*, *integration*, *system*, and *acceptance* testing. It is based on strong assumptions: that system boundaries, boundaries of its parts, and the detailed specifications and correctness notions for these parts can always be established, and that there is *control* over both the process of decomposition and putting together, and the operation of the parts (the latter being directly tied to active testing). In *systems of systems* (including global telecommunications and information technology systems), these assumptions simply do not hold: a system is multi-purpose (and these purposes are not established a priori), it exhibits emergent behavior ("*we put it together and strange things happen*"; ibid.), it is not built at once, it is unlimited in size and time scope, it is dynamically changing, it is not clear what constitutes its parts, the boundaries of its stipulated parts are constantly re-negotiated, and there is no *single* notion of its failure. The consequences and recommendations for testing include: basing the testing of such systems on "*actual system operation, not mythical specifications*" (ibid.), and accepting that there is no *single*, pragmatically meaningful result of testing (obtained by executing a particular test suite). Although this was not explicitly stated, passive testing clearly addresses both concerns.

In the sequel, in order to question the distinguished role of the "active" part of the concept of testing, we turn to external, arguably – more generic ideas, including those that had been established much earlier, before any current, disciplinary connotations had any chance to set in.

3. Towards the Generic Concept of Testing

It is possible to characterize testing very generally [3], in a way that avoids preconceptions, as:

- an activity with at least some empirical, *experimental* elements, the results of which can only be established *a posteriori*;
- where experiments are conducted on a particular object – *thing under test* (**Tut**);
- in order to evaluate a certain entity that partakes in testing – the *object of assessment* (**Ooa**);
- conducted with a certain *aim*. The quasi-equivalent formulations of this aim, adopted by different schools of thought, are:
  - to establish whether a given *relation*, normally – an equivalence or preorder holds between a **Tut** and a given *reference* (**Ref**) as often adopted within the formal testing community;
  - to establish whether a given *hypothesis*, which also means – all its necessary consequences, or "requirements" concerning a **Tut**, can be regarded as true (this is the essence of the logical approach to testing, as exemplified by the "industry-oriented" testing framework [26], and also that of the *Scientific Method*);

[2]Completeness of testing is usually defined as a *relative* notion, based on the assumption that hypotheses [32] *are* true.

– to obtain *knowledge* as to whether **Tut** corresponds to a **Ref** in a specific way (where the need for testing may be restated as the need to *know* if a system is correct – this is the language of *epistemology*).

A *test result*, or *test outcome*, always pertains to a **Tut**[3] – it records how a **Tut** behaved during a test. A *test verdict*, normally in {`Pass`,`Fail`,`Inconclusive`}, pertains to the object of assessment, which may, or may not be a **Tut**. The (non)identity of **Tut** and **Ooa** is subject to some debate. In engineering (and thus also in traditional ICT testing), a **Tut** is indeed taken to *be* the object of assessment (and thus, also the object of the ensuing corrective actions, if necessary), which is often reflected in various definitions of testing. In natural sciences, however, the object of assessment is normally a **Ref** – a hypothesis that explains and predicts facts about **Tut** – it is this hypothesis, not "the world", which may be found by testing to be defective. Redirecting the assessment is also possible, e.g., for *reverse engineering* [33].

The relations between **Tut**, **Ref**, and **Ooa** are just one dimension in a *matrix of choices* for sensible combinations of elements in the conceptual space of testing. Some of these combinations are actually claimed and occupied by different research schools, and other – are still to be "discovered" and tried out. In [1] it has been shown how much flexibility is to be gained by surrendering some *habitual* choices. Herein it is claimed that insisting on testing being *active* is one of such choices, the origin of which is no longer clear. Nowhere in the proposed "generic" exposition of testing "being active" appears as a necessary property of testing. In Aristotelian theory of predication, such property might be *essential* – included in a definition (as it is currently presented), or might be a *proprium* (*idion*) – still necessary, and derivable from a definition, but not explicitly present in this definition. In this author's opinion, "being active" appears rather to be of the third kind of predication – an *accident* of testing.

4. Testing and the Scientific Method

Testing, as a concept, did not emerge with technical systems. The important pre-technical, philosophical (epistemological) aspects of testing are present in the *Scientific Method* (SM) – a group of paradigms of sound scientific enquiry; in particular, we refer to one member of this group, attributed to W. Whewell, J. S. Mill, and K. Popper. It is primarily applicable to natural sciences, which does not preclude it from being a viable source of insight in a more technical context. As illustrated in Fig. 1, the application of SM consists in taking a series of steps:

– identifying a problem (i.e., a set of phenomena);

– stating a hypothesis that explains this problem – a statement p about "the world", preferably presented as a logical formula;

– deducing a set of the necessary logical consequences of the hypothesis: $\{p \rightarrow q_1, p \rightarrow q_2, \ldots\}$, where q_k *must* hold if p is indeed true;

– expressing the selected consequences in terms of their individual "empirical content" – predicted phenomena f, in principle amenable to empirical observation, such that q is true iff f "exists" (i.e., depending on its nature, *occurs*, *holds*, is *present* or *absent*);

– *testing the hypothesis* – performing *experiments* aimed specifically at confirming or denying the existence of predicted phenomena.

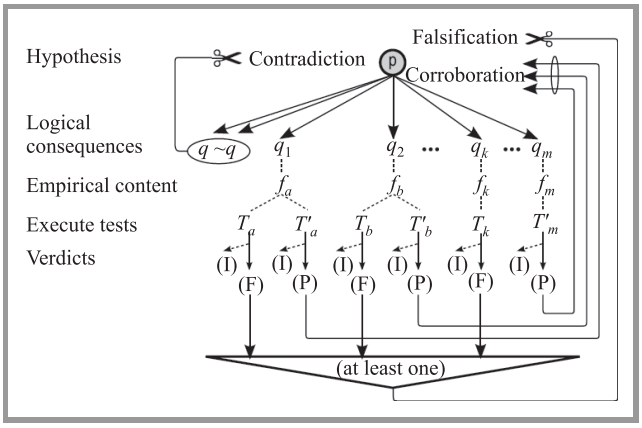


Fig. 1. Tests in the Scientific Method ([3]).

It is possible to check by purely formal means if the derived consequences of a hypothesis are non-contradictory. This non-empirical element appears in testing theories and practices as a "static phase" of testing; e.g., as *static conformance review* [26]. Being *a priori*, it does not count as testing proper, and is not presented as such in SM. In general, however, even the fundamental *a posteriori* character of testing is not universally admitted. Two quite opposite views on this matter, both voiced in "official" publications of the testing community, are: "*unlike dynamic testing... static testing techniques rely on...*" [8], and "*Different from testing, and complementary to it, are static techniques...*" [20]. This shows, again, how *arbitrary* the conceptual foundations of testing are, and further legitimatizes raising and investigating doubts about these foundations.

Experiments entail empirical observation. The same method and means of such observation could be used in different *way*s, with different *aim*s, e.g., in the initial phase

[3]Admittedly, "**Tut**" is not an established term, but other, more conventional terms such as **Sut** (*System under Test*), **Iut** (*Implementation under Test*), **Eut** (*Equipment Under Test*) are too specific, being related to a particular test architecture or kind of tests.

of the application of the method – to "charge" one's intuition as to the phenomena about which one is to propose an explanation. This activity, although empirical, does not qualify as *testing* – it should rather be called *monitoring*.

Two extreme views on how to approach the testing of a hypothesis, or *directions of testing*, are (disregarding nuances): *verificationism*, according to which a hypothesis, to be accepted as true, must be convincingly confirmed or *corroborated*, (also – *verified*, in the sense: shown to be true), and *falsificationism* (attributed to Karl Popper), which holds that it is essentially not possible to empirically verify a hypothesis, and the only sensible (meaningful) direction is to try to *falsify* (refute) it. This very influential Popperian stance [34], taken to the ground of ICT, re-emerged in the well known observation by Edsgar Dijkstra that "*testing can only show the presence of bugs* [i.e., can falsify the claim of correctness] *but never their absence* [i.e., cannot verify that all the system's properties are as predicted]". Pure approaches, however, are extremely rare – practical applications of the scientific method almost always *combine* the elements of verification and falsification, in varying proportions (as was also explicitly postulated for technical validation activities in [35]). Popper admitted that corroboration *does* count scientifically, if obtained for genuinely risky predictions. In this sense, the Dijkstra's observation seems surprisingly shallow and misleading. It overlooks the very principles of *model-based testing* [24] with its accompanying assumptions (or *test hypotheses* [32]), under which it is perfectly possible to (conditionally) *prove* correctness.

The role of experiments is to confirm or deny the existence of phenomena, regardless of how "existence" and "experiment" are technically defined. The outcome of executed experiments is thus associated with verdicts: `P` (pass) if an experiment *confirms* the existence a predicted phenomenon, `F` (fail) if it *denies* this existence, and `I` (inconclusive) if neither holds. In general, `P` is not the converse of `F`, although in particular testing theories this may be the case. According to the idea of the "non-orthodox" Scientific Method, as shown in Fig. 1, tests-experiments for each phenomenon are divided into two groups: $\{T\}$ – tests aiming at falsification (so only able to issue a `F` or an `I`), and $\{T'\}$ – tests aiming at corroboration (so only able to issue `P` or `I`). For some predicted phenomena (like f_k and f_m), only one of these groups of tests may be present. It is also conceivable that a falsification and corroboration test be combined in a single experimentation unit, so that its verdict is in $\{\texttt{P},\texttt{F},\texttt{I}\}$. This is the basic form of tests considered in [26]. It has direct representation in the linguistic devices of the TTCN-3 test language [27]. It is, however, by no means generic. Confirmation and refutation, not being the converse of each other, may need entirely different experimental approaches, and these are likely to translate into unrelated, orthogonal test programs, the composition of which may be unnecessarily complex and purely artificial.

Let us now combine the Scientific Method with the general view on tests presented in the preceding section. It can be seen that a hypothesis p is, at the same time, a **Ref** and an **Ooa**, while a **Tut** is a fragment of "the world", in which predicted phenomena occur. This setting can be brought closer to what is customary in ICT-related testing, by stating "**Tut** is correct" as p, and accepting, as the necessary logical consequences $\{q_1, q_2, \ldots\}$ of this hypothesis, the *requirements* (if a testing theory is cast in logic [36]) or particular features of a *behavioral model* (for a process-oriented approach). In this case, the consequences are obtained in a different way – they are not really *derived* from p, as very little can be derived, by pure logic, from "**Tut** is correct"[4]. Instead, some subset of consequences would contain the explicitly stated, *essential* requirements that define a correct **Tut**, and another, possibly *very large* set would contain its *propria*, derivable from the defining requirements, but not explicitly stated in a definition of **Tut** (and so, formally, not counted as essential). The place of *accidental* consequences in this picture is most dubious, as it is in philosophy in general. Accidents are not really instrumental in *distinguishing* things (correct implementations of A from correct implementations of A', or correct and incorrect implementations of A). Stating the "proper" set of q that would be subject to testing is one of the primary problems in testing theories. In natural sciences, new q are produced and subjected to testing continually. In technical testing, this problem is recast as *test generation and selection*. Associating particular q with different general *kinds* of properties (e.g., according to the Aristotelian concepts) is also tacitly practised, which transpires from, e.g., the telecommunications-oriented concept of "essential requirements" (as opposed to non-essential requirements, the testing of which might be skipped).

Within SM, "experiment" and "test" have, for all practical purposes, the same sense. What *is* this sense, has been investigated by philosophy of science [37], but nowhere within the context of the Scientific Method an experiment is described as necessarily *active*, i.e., that in which influence is purposefully exerted upon investigated phenomena. The distinction between *passive* (or "natural", or quasi-experiments) and *active* (or "controlled") tests/experiments has been, however, noticed and discussed. J. S. Mill calls them, respectively, *pure observation* and *artificial experiments* [38], and finds a place for both in scientific enquiry (and thus – in the Scientific Method). According to Mill, their essence is, respectively, to "*find an instance in nature suited to our purposes, or, by an artificial arrangement of circumstances, make one*". At a sufficiently high level of abstraction there seems to be "*no difference in kind, no real logical distinction, between the two processes of investigation... as the uses of money are the same whether it is inherited or acquired*". Mill does acknowledge the "great disadvantage" of pure observation, such as the apparent inability to ascertain causal relations and "*to produce a much greater number of variations in the circumstances than nature spontaneously offers*". He also identifies circumstances in which pure observation is advantageous, and his argu-

[4] If we ignore some quite fundamental, but still disputed philosophical consequences, such as "**Tut** exists" or "*something* is correct".

ments resemble the current expositions of the distinguishing features and applications of passive testing.

One Mill's remark, if taken literally, may provide the understanding of passive testing that *directly* maps onto active testing: "*Instead of being able to choose what the concomitant circumstances* shall be*, we now have to discover what they* are" (ibid.). The "concomitant circumstances" map to a particular *test purpose* as pursued by means of a particular *test preamble* (both being the elements of active tests, as stipulated in [26]). Instead of actively executing a test preamble, a passive tester recognizes it *if* and *when* it happens to occur. According to this interpretation, it is, in principle, possible to use the same test suite, however generated, for active *and* passive testing. The idea of recognizing a sequence of events embedded in a trace of behavior has been explored within the *pattern-matching* approach to passive testing [39], but, to this author's knowledge, it has not been developed so far as to suggest that these patterns may directly correspond to preambles taken from an active test suite.

5. Connotations of Passive Testing

As already stated, the prevailing operational characterization of testing derives from Mill's *controlled* experiments. Similarly, passive testing may be said to be based on *quasi-experiments* – the observation and assessment of phenomena that are not invoked (provoked, stimulated, influenced) by a tester. Pragmatically, this lack of influence may be intended or required for the following reasons.

The nature of a phenomenon may *not allow* for such influence (e.g., as in the investigation of the radiation spectrum of a distant star). In testing applied to technical systems, this translates to the absence of input port(s) – their genuine, physical absence, their administratively imposed inaccessibility for testing, or (as may be common for systems of systems) lack of information on whereabouts of these ports. Proposing that a **Tut** should provide the "testing ports" is a part of the *design for testability* framework [40]; one of its ideas postulates equipping a system with *additional* devices (interfaces and special functional properties), *specifically* for the purposes of its prospective, eventual testing. This approach has currency, e.g., in electronic circuit design, but is not advocated (or is even "prohibited") in most testing contexts in telecommunications.

A phenomenon may be "intensive enough". As an analogy, consider the dictionary meaning of "test" in chemistry: it is defined as a process of identifying the presence or the nature of a substance, *commonly by the addition of a reagent*. A reagent (and also a *catalyst*, which may be used for similar purposes) is analogous to a focused stimulus, which makes a phenomenon or substance *reveal itself*.

External stimulation (although feasible) might change or distort a phenomenon. There remains, however, a philosophical question as to whether passive observation really solves this problem. The *observer effect* (not to be conflated with the Heisenberg's *uncertainty principle*) is a posited principle, according to which a mere act of observation necessarily changes the phenomenon being observed. On the macro scale, in the setting of complex ICT systems, this concern may be safely dismissed[5].

A **Sut** may be a larger system whose *integrity, safety, and performance* critically depends on non-interference with its internal parts and processes. Normally, active tests would include incorrect, invalid, unexpected (inopportune [26]) stimuli that would have an *a priori* unknown effect upon a system[6], which may well be catastrophic to the system's mission – as in the U.S. network-wide failure of 1990 [42], and also in the Chernobyl's disaster, both caused by a stimulus that was not even incorrect or unexpected *per se*.

Finally, it may be *too cumbersome or costly* to build and operate the "sending" channel of a tester, through which stimuli would be administered.

One may claim that, regardless of any technical, local definitions, it is "common knowledge" that testing is active, as (supposedly) codified in the language and reflected in the common use patterns of the term, recorded in dictionary entries[7]. The passive nature of testing is stipulated, or at least not rejected, in the following dictionary entries for "test":

- *the means by which the presence… of anything is determined* (this closely resembles tests for the presence of a phenomenon in SM, which, as already indicated, do not have to be active);
- *trial – the examination before a judicial tribunal of the facts… in a case* (where the tribunal has no power to influence the course of the past events to re-enact alternative scenarios).

The active elements are emphasized in the following entries:

- *trial; to try out* (in order to be tested, something must be actually *used*, which connotes both-way interactions with this entity);
- *a set of standardized questions, problems, or tasks designed to elicit responses for use in measuring the traits, capacities, or achievements of an individual* (to elicit responses is the explicit role of stimuli in active tests).

Altogether, basing on [44] it may be concluded that the active and passive connotations of "test" are well balanced.

Distinctions made in the purely technical context are also much less clear-cut than it is usually admitted. One

[5]Although there have been insightful discussions on the behaviour of automata, in which both principles have been used (at least metaphorically) under the name of "complementarity" [41].

[6]If this effect were known *a priori*, testing would not be needed at all.

[7]This is a genuine and acknowledged problem. For example, in [43] it is noted that *"the field of IS* [Information Systems] *development is severely hampered by the limitation of meaning derived from the everyday use of some representative words…*".

of the early attempts at harmonizing active and passive testing [45] has been to employ two testers operating in parallel: an *active* tester for confirmation tests leading to the P verdicts, and a *passive* tester, originally called a *trace analyser*, for refutation tests leading to the F verdicts. This solution was based on the earlier idea of separating the problem of choosing and applying *test inputs* (stimuli – the "pure" active part of testing) from the problem of assessing the behavior of a **Tut** for *these* or *any other* stimuli that may have been provided, by any means [46]. In both approaches, the passive testing functionality may form a part of a compound (effectively – active) tester, or may, in the limit, form the *whole* of a tester – a self-contained *passive tester*. The conditions for such transformations were discussed in [3]. In these early stages, a passive tester was apparently treated as a *bona fide* entity, and not as a metaphor. Interestingly, one of the first direct uses of the term "passive testing" (instead of various euphemisms) was made in [47], when the initial acceptance for passive testing as (a kind of) testing seemed to evaporate.

6. Testing versus Monitoring

It is surprisingly difficult to precisely state the difference between *monitoring* and *testing* a system. The common tendency so far has been to conflate monitoring with passive testing[8]. The prevailing intuition is that monitoring is *not* testing, so "passive testing is not testing" as well. We argue that it is both necessary and possible to keep the two notions apart – they have different *sense*, even if, in the limit, they may refer to technically "the same" activity.

Let us assume that at least some level of technical instrumentation is necessary, and that the suitable technical instruments: a *monitor* and a *passive tester*, respectively, are present. In the following, we treat monitoring as *using* a *monitor*, resp. testing as *using* a *tester*, and we look at each of the constituent parts of the decomposed concepts separately. Using a thing (an apparatus) presupposes the existence of a *user* – some external entity that is *not* a monitor (resp. a passive tester). Clearly, not *every* use of technical instruments lies within the scope of the respective notions – using a monitor to hammer down nails would certainly not count as monitoring. The pertinent question is *what* use of a monitor (tester) makes for monitoring (testing), and how this "proper" use is related to the functionality of the instrument.

We first consult the basic dictionary meanings of "monitoring" [44], noting the recurring use of two key terms: *looking* (or *watching*) and *seeing*:

1. *Listening to transmitted signals in order to check the quality of the transmission*. Monitoring is thus performed *in order to* check (some properties), but checks themselves are left to the user. Consider a medical monitor (e.g., an electrocardiograph). The output of a monitoring system is a stream of data, suitably (e.g., graphically) *presented* so that it can be conveniently *interpreted*. The interpretation itself rests with the doctor, who on different occasions can *look* at (the same) data from different perspectives, in order to *see* if there is any activity of the heart, or if the heart-beat is regular, etc.

2. *Observing, recording, or detecting (an operation or condition) with instruments that have no effect upon the operation or condition*. This definition stresses the *passive* and *technical* character of the operation. "Detecting" suggests the higher-level functionality that will be later assigned to an *extended* monitor.

3. *Keeping track of, checking continually.* This stipulates a "campaign-less", possibly infinite process.

Points (2) and (3) correspond to the joint characteristics of monitoring and passive testing, while point (1) seems useful for differentiating the two. It relies on differences between *looking* (watching, listening) and *seeing* (hearing). According to [44], to *look* means: to direct one's glance, attention, consideration (to *watch* – to keep under attentive view or observation, as in order to see something); to *see* means: to perceive (things) mentally, to discern, to understand, to recognize.

The first approach to differentiating between monitoring and passively testing a *thing under investigation* – **Tui** (it is not known yet whether it is *thing under monitoring* – **Tum**, or **Tut**) is based of different *levels* of interpretations (see Fig. 2). *Monitoring* is a technical counterpart of watching a particular aspect of a system. Monitoring provides, in a *passive* way, a *continued* stream of processed data on the behavior of a **Tum**. These data are intended to be interpreted by an external process (in particular, but not exclusively, by a human operator), where the underlying phenomena or properties of a **Tum** can be *seen*. Unlike monitoring, *passive testing* involves both, the (syntactic) process of *watching*, and the (semantic) interpretative process of *seeing*. The latter may still be performed by a human test operator (as is often stipulated in approaches to testing characteristic of software engineering). The output

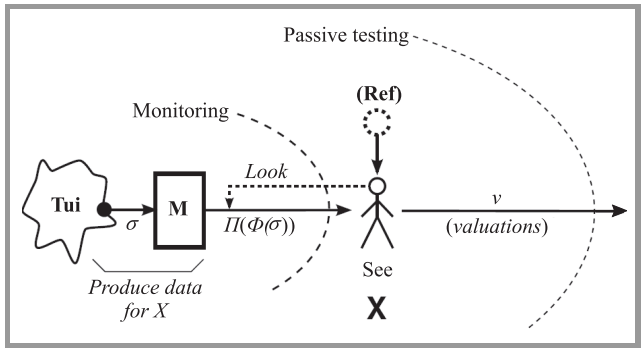


Fig. 2. Monitoring and passive testing.

[8] "*Monitoring is ... called passive testing...*" [48].

of the passive testing process is a stream of *interpretations*, or *valuations* (v) of monitored data.

A basic technical apparatus for monitoring, a *simple* (or *plain*) *monitor* (**M** in Fig. 2) is thus assigned functions for: syntactically transforming (filtering, projecting) a stream of "raw data" about the behavior of **Tum** into a stream of processed data: $\sigma \mapsto \sigma' = \Phi(\sigma)$; and suitably presenting (formatting) the processed data: $\sigma' \mapsto \Pi(\sigma')$, so that the stream of formatted data can be *conveniently* interpreted. The presentation function is not mere aesthetic decoration; it is an important part of the notion of monitoring. Both syntactic processing and pragmatic presentation mode depend on the intended external interpretation process – the understanding is that of "monitoring *for*..." or "monitoring *in order to*...". Monitoring is thus *not* purpose-agnostic.

Experience shows that it is quite common to shift some semantic interpretation functions (e.g., raising an alarm if a threshold is exceeded) from an external process to the monitoring process itself, which will now be referred to as *extended monitoring*. A monitor enhanced with the interpretation function **I** is an *extended* monitor (**Me** in Fig. 3a). Such a device would be able to, e.g., raise an alarm when certain threshold values are exceeded (as in a class of medical monitors), while still being referred to as a monitor, and not a tester. A single layer of inbuilt interpretations yields what might be called *tier 1* of *extended monitoring of* **S**. There may be many consecutive tiers of extended monitoring.

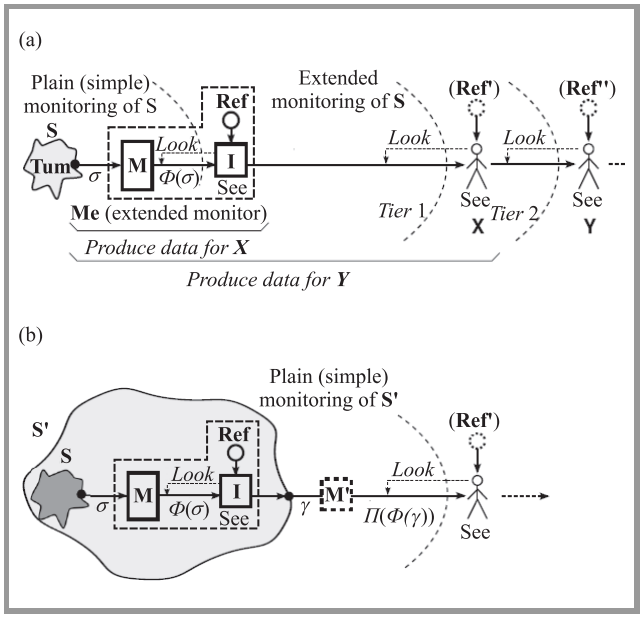


Fig. 3. Plain and extended monitoring: (a) tiers of extended monitoring; (b) shifting the boundary of the monitored system.

One way to conceptually dispose of the notion of extended monitoring (and to stay with "just" monitoring), is to re-position the boundary of a system under monitoring – the stream of interpretations is now regarded as raw data γ on the behavior of *another* system **S'** (Fig. 3b). The original system **S** is now *embedded* in a context, which makes it clear that its behavior can only be investigated indirectly[9].

When interpretations become a part of the monitoring process itself, the distinction between monitoring and passive testing, as proposed in Fig. 2, seems to collapse – the output of *both* processes is now a stream of interpretations. Additionally, with the growing number of tiers of extended monitoring (Fig. 3a), there is no clear point at which monitoring would "magically" change into (passive) testing. It thus becomes apparent that another, additional criterion for distinguishing monitoring and passive testing is necessary. We take this additional criterion to be the *kind* of interpretations that are carried out within the respective processes, as already hinted in [49]. Let us consider a pair: $< B, C >$ consisting of a particular *behavior*, and circumstances (*conditions*) in which this behavior is exhibited. We claim that monitoring and (passive) testing differ in the pragmatically meaningful, logical ordering of the elements of this pair. In (extended) monitoring, interpretation (valuation) serves to infer, from the observed behavior, the conditions (circumstances), or the general *mode of operation* of a **Tum**, such as "being overloaded (congested)", "being down", "being stable", "being under attack" (in the context of intrusion detection [1]), or "being dead" (in the medical context). This is consistent with the view that "*monitoring consists of measuring properties of the network, and of inferring an aggregate predicate from these measurements*" [50]. In testing, interpretation is related to the defined circumstances (conditions), called in this context *test purposes*. In active testing, a test system, steered by a test program, *establishes* (forces) these conditions, while in passive testing a test system *recognizes* them. Note that the same understanding was also arrived at earlier, although in a different way.

It follows from the foregoing discussion that monitoring can be located as a lower-layer functionality with its results interpreted by testing, or as a higher-layer functionality acting on a stream of lower-layer test verdicts. According to this view, both monitoring and passive testing are "full", but *different* functionalities, with no fixed subordination relation between them. We conclude that passive testing *can* be distinguished from monitoring, and thus can be freed from one of its strongest "non-testing" connotations.

7. Testing versus Measurements

Intuitively, testing and measuring are closely related (as in: test *and* measurement), but distinct concepts. This intuition makes metrology an interesting source of concepts and mechanisms to be directly imported, and also general insight and analogies. Some links between the two domains have already been briefly identified in [3]. To be a measurement, determining/assigning a value must be based on empirical observation of a real, existing object – similarly

[9]This setting can be re-cast as *verification-in-context* [35]. This is also the classical telecommunications setting, where **S**, called **Iut** (*Implementation under test*), is embedded in, and only indirectly accessible through, other parts of a **Sut**.

for testing. According to [51], the necessary conditions for calling an evaluation a measurement are: a well-defined, external *reference*, and a well-defined measurement *operation* which can be carried out independently of any specific measurer. These two postulates have always been the cornerstones of formal testing: the former is at the very core of *model-based* testing, and the latter was given due consideration, e.g., in [26] as "Conformance Assessment Process".

Surprisingly, the in-depth, direct, non-metaphorical discussion of the relations between measurements and testing is lacking. The measurement community at least declares interest in suitably extending and adjusting their methodology so as to accommodate testing, but any reciprocal effort from the testing community has not manifested itself. The metrological interest in testing is mainly due to the conceptual troubles with applying traditional concepts of measurement (implicitly focused on *physical* quantities[10]) to information technology artifacts, with their *logical* properties [53]. Despite this declared interest, the current effects of harmonization are modest. Conspicuously, "measurement" and "testing" are defined not side-by-side, but in different metrological documents: "Measurement – *process of experimentally obtaining one or more quantity values that can reasonably be attributed to a quantity*" [13, 2.1] vs. "Testing: *determination of one or more characteristics of a given object of assessment according to a procedure*" [54]. One of the rare explicit explanations of the distinction is "*sometimes made by considering testing to be a measurement or measurements together with a comparison to a specification*" (ibid.). This explanation is, however, flawed in that "comparison to a specification", in a broad sense, is also the *internal* element of most methods of measurement.

The main obstacle to directly applying metrological concepts to testing seems to be the core concept of metrology – the *measurand*, which is controversial in itself, and subject to internal, metrological debate [55]. A measurand is a *quantity*, i.e., a property that has a magnitude that can be expressed using numbers [13, 1.1] (more generally – expressed symbolically). The minimum requirement seems to be that the objects of measurement can be *ordered* w.r.t. the magnitude of a quantity in question. The mainstream concepts of metrology pertain to such quantities that meaningful algebraic operations on their values (expressions of magnitude) can be defined, so that the results of these operations reflect the empirical relations between quantities of respective objects. Metrology also explicitly admits *ordinal* quantities, which can be (numerically) expressed and which enter into (empirical) ordering relations, but with no corresponding algebraic operations on the expressions of their values (e.g., garment sizes: $\{\mathtt{XS},\mathtt{S},\mathtt{M},\mathtt{L},\mathtt{XL}\}$). The values of these quantities can be obtained by a conventional measurement procedure. There are also properties that have been specifically excluded from the scope of the concept of "quantity", and thus also from the scope of "measurement" – *nominal* properties that have no magnitude [13, 1.30], although can be assigned a (symbolic) value. Sex and colour have been given as an example (ibid.), although this is debatable – ordering (the "value" of) humans by sex or race has often, sadly, been practiced[11], and colour has an obvious "objective" value (wavelength). Despite this somewhat arbitrary exclusion, there have been attempts at the metrological treatment of taxonomic, nominal relations (such as postal codes [56]).

For testing, the "measurand" would be correctness, and the conventional expression of its value is a verdict, in $\{\mathtt{Fail},\mathtt{Inconc},\mathtt{Pass}\}$, or, possibly, in $\left\{0,\frac{1}{2},1\right\}$ (if this should bring more metrological connotations). It may seem to be the nominal, taxonomic property, officially – beyond the scope of metrology. In testing, however, the verdicts (reflecting the "magnitude" of correctness) do introduce ordering on systems – incorrect systems are "less than" correct ones, and there may be different implementations of a standard that are "equally correct". There is also the explicit ordering on verdicts: $\mathtt{P} \rightarrow \mathtt{I} \rightarrow \mathtt{F}$, built into the semantics of the test language TTCN [27]. On the other hand, composing a correct and incorrect system may yield a system that is correct or incorrect, which *a priori* cannot be established by applying any operations to their individual correctness values. This is why, after conformance testing, combined systems are subjected to interoperability tests.

It may be concluded that the *direct* application of metrological concepts and language to testing seems no more controversial than the ongoing debates within metrology itself (including the notion of a measurand). If, however, mutual harmonization is for any reason unacceptable, then metrology may always be used as a source domain for *scientific metaphors* [5] aimed at explicating testology. As in any similar case, the metrological community has no "right" to stop testology from applying such metaphors (or to enforce the observance of all the metrological definitions and agreements to the last detail). The only criterion of the validity of metaphors is their effectiveness. Obviously, the canonical metaphor to try out is: "Testing is measurement".

Having established the applicability, either direct or metaphorical, of metrological concepts to testing, we now briefly return to the "active-passive" dimension. Measurement is popularly believed to be essentially *passive*, as it is intended to assess the object of measurement "as it is". Contrary to this impression, the techniques of measurement, which clearly constitute some part of its essence, are explicitly divided into passive (as in measuring the radiation spectrum of a body) and active. Measuring resistance may be performed actively, by applying a certain voltage (a stimulus) and observing the resulting current. The same quantity may also be measured passively, by observing the

[10] "*. . . the measurement of a well-defined physical quantity — the measurand*" [52, 1.2]

[11] Such pragmatic "ordering" should not be *a priori* rejected, in view of the general shift from regarding measurement as *determination* (of some elusive "true value") towards treating it as *assignment* [51].

relation between the current and the voltage in a circuit, while refraining from actually applying either. Both would be readily called "measurement", with no methodological and linguistic reservations. It is acknowledged that under certain circumstances one technique is preferred over the other, or is exclusively applicable, but no paradigmatic preference is given to either. This symmetry is a feature of metrology that should, by analogy, at least be given due consideration in testology. Should metrology be consulted for insights, testology would not find there any justification for its current asymmetric views on the active nature of testing.

8. Concluding Remarks

Any scientific community, including the testing community, is free do define the scope of its interest, conceptual horizon, and terminological (linguistic) devices. Such choices are, however, not beyond the scope of external scrutiny. They are also often the object of *internal*, intra-disciplinary debate. For example, Gaudel [32] felt that it was necessary to examine the general dictionary entries to recharge the failing intuition of testing. Similarly, within the broad context of information systems there are schools of thought that, dissatisfied with a certain methodological lock-in, try to re-define their discipline in terms of *semiotics*. Also investigations into how people use words have a long tradition in social sciences and philosophy of science.

The presented high-level methodological discussion of testing is not the first of its kind. It is similar in vein, and complementary (but more focused in scope) to [57]. It also builds on [1] and [3], where an attempt is made at identifying and dismissing spurious incompatibilities between the "testing-like" concepts developed by different research communities, and on [5], which surveys the methodological aspects of looking for insight and borrowing concepts.

The aim of this work has been not to arbitrarily fix a terminological "misunderstanding", but to show how testology could be freed from a particular family of preconceptions that seem to impede one direction of its development.

Acknowledgements

This work presents the motivation and results of the fundamental research track of Research Task 7: "Verification and validation of network protocols by passive testing", within the framework of Research Project PBZ-MNiSW-02-II/2007 contracted by the Polish Ministry for Science and Higher Education and financed with the 2007-2010 reseach funds. Fragments of the author's work [3] were re-used in Section 4, in accordance with the publishing agreement.

References

[1] K. M. Brzeziński, "On common meta-linguistic aspects of intrusion detection and testing", *Int. J. Inform Assurance and Secur. (JIAS)*, vol. 2, no. 3, pp. 167–178, 2007.

[2] K. M. Brzeziński, "A joint meta-linguistic taxonomy of intrusion detection and testing/verification", in *Proc. 2nd Int. Worksh. Secure Informa. Syst. SIS'07*, Wisła, Poland, 2007.

[3] K. M. Brzeziński, "Towards the methodological harmonization of passive testing across ICT communities", in *Engineering the Computer Science and IT*, S. Soomro, Ed. In-Tech, 2009, pp. 143–168.

[4] K. M. Brzeziński, "On conceptual struggles over 'testing'", in *Proc. XIV Poznańskie Warsztaty Telekomunikacyjne PWT 2010*, Poznań, Poland, 2010.

[5] K. M. Brzeziński, "Standards are signs", in *Proc. 15th EURAS Ann. Standardization Conf.*, Lausanne, Switzerland, 2010, pp. 43–60.

[6] *Software testing. Vocabulary.* BS 7925-1. British Standards Institution, 1998.

[7] *Standard Glossary of Terms Used in Software Testing, version 2.0*. ISTQB (Glossary Working Party), 2007.

[8] *Certified Tester. Foundation Level Syllabus*. ISTQB, 2007.

[9] *Certified Tester. Advanced Level Syllabus*. ISTQB, 2007.

[10] A. Spillner, T. Linz, and H. Schaefer, *Software Testing Foundations*. Rocky Nook, 2007.

[11] G. Bath and J. McKay, *The Software Test Engineer's Handbook*. Rocky Nook, 2008.

[12] C. F. Tschudin, "On the structuring of computer communications", Ph.D. dissertation, University of Geneve, 1993.

[13] VIM, *International vocabulary of metrology – Basic and general concepts and associated terms*. Joint Committee for Guides in Metrology (JCGM), 2008, vol. JCGM 200.

[14] R. R. Hoffman, *Metaphor in Science*. Lawrence Erlbaum Associates, Inc., 1980, pp. 393–423.

[15] S. Purao, C. Y. Baldwin, A. Hevner, V. C. Storey, J. Pries-Heje, B. Smith, and Y. Zhu, "The sciences of design: observations on an emerging field", Harvard Business School, Working Paper 09-056, 2008.

[16] K. Naik and P. Tripathy, *Software Testing and Quality Assurance: Theory and Practice*. Wiley, 2008.

[17] L. Cacciari and O. Rafiq, "Controllability and observability in distributed testing", *Inform. Software Technol.*, vol. 41, no. 11-12, pp. 767–780, 1999.

[18] C. Sunshine, "Formal techniques for protocol specification and verification", *Computer*, vol. 12, no. 9, pp. 20–27, 1979.

[19] *IEEE Standard Glossary of Software Engineering Terminology*. IEEE Std 610-12. IEEE, 1990.

[20] *Guide to the Software Engineering Body of Knowledge*. SWEBOK, IEEE, 2004.

[21] L. Frantzen and J. Tretmans, "Model-based testing of environmental conformance of components", in *Formal Methods of Components and Objects (FMCO'06)*, LNCS 4709. Springer, 2007, pp. 1–25.

[22] E. Brinksma and J. Tretmans, "Testing transition systems: an annotated bibliography", in *Proc. MOVEP 2000*, Nantes, France, 2000, pp. 187–195.

[23] E. Brinksma, W. Grieskamp, and J. Tretmans, "Summary", in *Perspectives of Model-Based Testing* of *Dagstuhl Seminar Proceedings*, E. Brinksma, W. Grieskamp, and J. Tretmans, Eds., no. 04371. IBFI, 2005.

[24] M. Utting, A. Pretschner, and B. Legeard, "A taxonomy of model-based testing", Univ. of Waikato, Hamilton, New Zealand, Working Paper 04/2006, 2006.

[25] J. Ryser, S. Berner, and M. Glinz, "On the state of the art in requirements-based validation and test of software", Tech. Rep. IFI-98.12, Univ. of Zurich, May 1998.

[26] *Conformance Testing Methodology and Framework*. ISO/IEC 9646. ISO/IEC, n.d., vol. 1–7.

[27] *MTS; The Testing and Test Control Notation version 3*. ETSI ES 201 873, ETSI, n.d.

[28] K. M. Brzeziński, "Intrusion detection as passive testing: linguistic support with TTCN-3", in *DIMVA*, LNCS 4579. Lucerne: Springer, 2007, pp. 79–88.

[29] K. M. Brzeziński, D. Mastalerz, and R. Artych, "Practical support of testing activities: the PMM Family". COST 247 WG3 Internal Report, IT P.W., LTiV Tech. Rep. 965, 1996.

[30] K. M. Brzeziński, "Weryfikacja i testowanie", *Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne*, no. 4, pp. 139–140, 2010 (in Polish).

[31] K. M. Brzeziński, "Testowanie w cyklu życia systemu: nieregularności meta-standaryzacji", in *Krajowe Sympozjum Telekomunikacji i Teleinformatyki (KSTiT)*, Warszawa, 2009 (in Polish).

[32] M.-C. Gaudel, "Formal methods and testing: hypotheses, and correctness approximation", *Formal Methods*, pp. 2–8, 2005.

[33] K. M. Brzeziński, A. Gumieniak, and P. Jankowski, "Passive testing for reverse engineering: specification recovery", in *Proc. IASTED Int. Conf. Paral. Distrib. Comput. Netw. PDCN 2008*, Innsbruck, Austria, 2008, pp. 27–32.

[34] K. Popper, *Conjectures and Refutations: The Growth of Scientific Knowledge*. London: Routledge, 1963.

[35] L. Heerink and E. Brinksma, "Validation in context", in *Proc. 15th IFIP Int. Symp. Protocol Specification, Testing and Verification PSTV 1995*, Warsaw, Poland, 1995, pp. 221–236.

[36] *Framework on Formal Methods in Conformance Testing*. ITU-T Z500. ITU-T, May 1997.

[37] M. Heidelberger, "Experimentation and instrumentation", in *Encyclopedia of Philosophy*, D. M. Borchert, Ed. Thomson Gale, 2006, vol. 10, pp. 12–20.

[38] J. S. Mill, *Of Observation and Experiment*. Routledge and Kegan Paul, 1974.

[39] J. A. Arnedo, A. R. Cavalli, and M. Núñez, "Fast testing of critical properties through passive testing", in *Proc. IFIP Int. Conf. Testing Commun. Syst. TestCom 2003*, Sophia Antipolis, France, 2003, pp. 295–310.

[40] R. Dssouli and R. Fournier, "Communication software testability", in *Protocol Test Systems III*. North-Holland, 1991, pp. 45–55.

[41] K. Svozil, "Extrinsic-Intrinsic Concept and Complementarity", in *Inside Versus Outside*, H. Atmanspacher and G. J. Dalenoort, Eds. Berlin: Springer, 1994, pp. 273–288.

[42] P. G. Neumann, "Cause of AT&T network failure", *Risks Dig.*, vol. 9, no. 62, 1990.

[43] M. Boahene, "Information systems development methodologies: are you being served?" in *Proc. 10th Australasian Conf. Information Syst.*, Wellington, New Zealand, 1999, pp. 88–99.

[44] *Webster's Encyclopedic Unabridged Dictionary of the English Language*. Gramercy Books, 1996.

[45] R. Wvong, "A new methodology for OSI conformance testing based on trace analysis", Master's thesis, University of British Columbia, 1990.

[46] G. von Bochmann and O. B. Bellal, "Test result analysis with respect to formal specifications", in *Proc. 2nd. Int. Worksh. Protocol Test Syst.*, Berlin, Germany, 1989, pp. 272–294.

[47] D. Lee, A. N. Netravali, K. K. Sabnani, B. Sugla, and A. John, "Passive testing and applications to network management", in *Int. Conf. Netw. Protoc. ICNP'97*, Atlanta, USA, 1997, pp. 113–122.

[48] J. Tretmans, "Testing concurrent systems: a formal approach", in *10th Int. Conf. CONCUR'99*, Eindhoven, The Netherlands, 1999, pp. 46–65.

[49] K. M. Brzeziński, "Towards Practical Passive Testing", in *Proc. IASTED Int. Conf. Paral. Distrib. Comput. Netw. PDCN 2005*, Innsbruck, Austria, 2005, pp. 177–183.

[50] M. Dilman and D. Raz, "Efficient reactive monitoring", *IEEE J. Selec. Areas Commun.*, vol. 20, no. 4, pp. 668–676, 2002.

[51] L. Mari, "The role of determination and assignment in measurement", *Measurement*, vol. 21, no. 3, pp. 79–90, 1997.

[52] *Evaluation of Measurement Data – Guide to the Expression of Uncertainty in Measurement*. GUM. Joint Committee for Guides in Metrology (JCGM), 2008, vol. JCGM 100.

[53] "Metrology for Information Technology (IT)". NISTIR 6025. NIST, White paper, 1997.

[54] *Conformity Assessment – Vocabulary and General Principles*. ISO/IEC 17000, ISO/IEC, 2004.

[55] A. C. Baratto, "Measurand: a cornerstone concept in metrology", *Metrologia*, vol. 45, pp. 299–307, 2008.

[56] R. M. Olejnik, "Kod pocztowy jako przykład metrologicznej skali nominalnej", *Pomiary Automatyka Robotyka*, no. 7–8, pp. 186–188, 2004 (in Polish).

[57] T. S. E. Maibaum, "The epistemology of validation and verification testing", in *Proc. 17th IFIP Int. Conf. TestCom 2005*, Montreal, Canada, 2005, pp. 1–8.

Krzysztof M. Brzeziński is Assistant Professor at the Institute of Telecommunications, Warsaw University of Technology, Poland. He obtained his Ph.D. in Telecommunications from the same University in 1995. His research interests concentrate on rigorous, formalized design and testing of distributed (esp. telecommunications) systems, and theory of standardization. He is the author of a book on ISDN technology (also translated into Russian), four book chapters, over 30 conference papers, and over 60 research reports. He is a certified TTCN-3 specialist.

E-mail: kb@tele.pw.edu.pl
Institute of Telecommunications
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Paper

Optimization of Call Admission Control for UTRAN

Michał Wągrowski and Wiesław Ludwin

Department of Telecommunications, AGH University of Science and Technology, Kraków, Poland

Abstract—This paper addresses the traffic's grade of service indicators: call blocking and dropping rates as well as the optimization of their mutual relation, corresponding to the call admission control procedure configuration. In the presented results of simulations authors showed opportunities for the CAC load threshold adaptation according to the traffic volume and user mobility changes observed in the mobile radio network.

Keywords—blocking and dropping of calls, call admission control, measurement based optimization, radio resource management, traffic analysis.

1. Introduction

Basically, there are two major ways to increase efficiency of mobile networks. The first one is to enhance performance of systems (in particular, radio technologies). The main objective of the system development is to better and better cope with problems related to the transmission in a mobile radio channel, which is faded and interfered. The second way is to ensure that the systems are effectively deployed and used. This regards to the issues of network planning, configuration and optimization. Putting into practice new technologies requires gaining knowledge about their efficient usage. Hence, looking for new solutions of adaptation the network to specific conditions of its operation in a given geographic area, taking into account the current state of environment, generated traffic and the radio resources occupation in that particular area, becomes the essential part of mobile communications development.

In this paper we focus on two important indicators of the traffic's grade of service (GoS), namely the call blocking and dropping rates (denoted *BR* and *DR*). They correspond to the operation of the call admission control (CAC) and congestion control (CC) procedures, respectively. We show the impact of the traffic volume and user mobility profile on the abovementioned indicator values as well as their mutual relation. According to the purpose of CAC procedure optimization we show new opportunities and consider limitations of their taking.

2. Dropping and Blocking of Calls

Users' perception of the mobile network performance is based on their experience on its operation. Their feelings result from several factors, like e.g., the availability of services, their quality as well as the frequency of unwelcome events. Providing availability and quality of services is the issue of both network planning and system characteristics. It is up to the operator to design the network layout in the best way to enable the radio transmission with required signal to interference ratio (SIR) and serve users with expected quality, i.e., popularly speaking to provide a "good range" over the network operation area. The service quality and related SIR requirements are characteristics of used technology and provided services.

The WCDMA radio interface load may vary while serving a constant number of transmissions in particular cells due to changes in radio channels [1], [2]. If we assume the system is able to provide users with a guaranteed quality of service (QoS), which for CS domain should be ensured, then the quality of serving traffic can be expressed with the GoS indicators. In UMTS, the CAC and CC procedures take care of keeping the load below a certain threshold to ensure the stability of network operation. These two procedures are responsible for preventive blocking of new calls and dropping of the serving ones in case of congestions, respectively. The CAC procedure estimates the additional load of each new call before it is accepted and based on the total estimated load level decides whether to block it or not. Unfortunately, due to some traffic variations (caused mainly by users mobility) as well as signal fading (e.g., due to shadowing) the threshold of maximum allowed load may be occasionally exceeded. In such a case the CC procedure performs actions to keep the load below this threshold and protect the system against congestions. The order of performed actions starts from higher layers, where at first bit rates for particular transmissions are tried to be reduced. If the quality of served calls cannot be decreased anymore and next no handover to any other cell or radio access technology is possible, then finally, one or more calls have to be dropped. Other common reasons for dropping of calls include not defined cell neighborhoods, faults that may occur during the handover signaling procedure, in which whole the protocol stack is involved as well as not enough transmitter maximum power.

Dropping of calls is perceived very frustrating by users, much more than blocking of new ones and impacts their mean opinion on the network quality more than blocking. Hence, operators pay attention to protect their networks against dropping more than blocking. For this purpose, usually a certain reserve of load is assumed in the CAC procedure, which means that the CAC load threshold for new calls (denoted as η_{max_new}) is set below the CC one (η_{max}).

Both blocking and dropping rates are important measures used for the network quality assessment in terms of the traffic serving efficiency. They can also be used as key performance indicators (KPIs) for the network optimization process. Blocking rate is defined as the ratio of blocked calls to all call attempts and the dropping rate is the ratio of dropped calls to the admitted calls number. Their definitions are based on counters, which indicate numbers of events that occurred during the observation (measurement) time period ΔT.

2.1. The Model for UTRAN Traffic Analysis

For 2G FDMA/TDMA systems, analytical models of serving policies for fresh and handover calls, including various network load and user mobility, were deeply investigated and described, e.g., in [3] and [4]. For UMTS, the state of WCDMA interface load, distinct from 2G systems, may change not only according to the traffic volume, but also to its distribution as well as the propagation environment variations. Thus, the radio resource occupation in particular cells may change in continuous manner. Moreover, it depends nonlinearly on the number of calls being served as well as on the users' distance to their base stations. The analytical approach to such a dynamic process for a network with cell coupling, such as UTRAN, is very complex and problematic. Hence, we used computer simulations to better recognize blocking and dropping occurrence characteristics and appropriate rate indicators.

For the sake of the considered problem nature we built a dynamic model, in which consecutive evaluated system states were correlated in time. Only fresh voice calls were generated in the network model and the handover ones resulted from users' mobility. New calls appeared with exponentially distributed intervals. The same distribution, with an average value of 120 s, characterized the connection duration. Mobile stations could have moved over straight lines in randomly selected directions and with speeds assigned according to the normal distribution characterized by four factors: average, standard deviation, minimum and maximum value.

Based on defined this way traffic volume and density, user mobility as well as all link budgets analysis, the implemented system procedures generated blocking and dropping of calls. The load factor and the maximum transmit power were the only reasons for the performed actions by CAC and CC procedures. To ensure the results accuracy, it was necessary to assure the events occurrence precisely in time. Therefore, all transmissions in the network must have been treated asynchronously, as in real. Hence, all transmitter powers were calculated each time the active set update procedure was performed by any active mobile station. The frequency of network evaluations depended on the number of mobile stations served and the event schedule configured for particular simulation scenarios. Simulations usually examined several hours of the network operation, during which a huge number of events were processed. Note that each mobile station performed its active set update procedure every one second. All that led to the simulations were time consuming and required considerable computing power. Hence, while planning of simulation experiments the issues of their feasibility in terms of a reasonable time to achieve results had to be considered.

The author's concept of dynamic network model assumed simplifications leading to decrease the simulation time. Most of all the highest system time resolution related to the fast power control loop was not considered. The toroidal network structure (presented in Fig. 1) was chosen and limited just to seven cells in which calculations were performed. Users could have moved in the area served by seven base stations and never left it. A user served, e.g., in cell 2, if moved outside the network, would hit the cell 4, 6 or 5, depending on the movement direction. It corresponds to the 4', 6' and 5' cells layout in the model shown in Fig. 1, however, the user's position was automatically shifted to the appropriate cell with the indicator 4, 6 or 5. According to [5] and [6] at least two rings of neighboring cells should be taken into account for the correct external interference calculation for WCDMA interface. Therefore, appropriate copies of cells, marked with ' were used for that purpose. They are just the same only geographically shifted cells, thus they do not need to be evaluated and introduce no additional computational load. The nominal cell radius was assumed to 1 km.

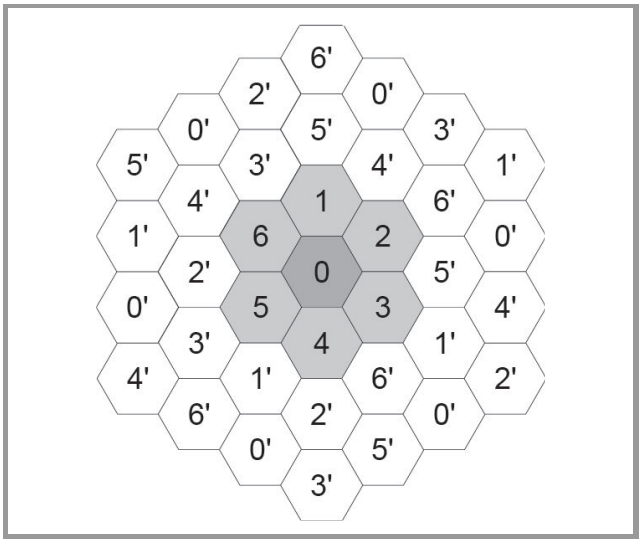


Fig. 1. The network layout model used for simulations.

The most of link parameters for mobile and base stations were set according to the specification [7] and [8] requirements and commonly used values for CS voice service [1], [2]. For the estimation of propagation loss we used the Okumura-Hata model. We assumed the area was flat and the propagation environment homogenous, without shadowing.

In case of a congestion, the most commonly applied strategy for selection of connections to be dropped uses the load or power criterion, which can be met either for uplink or for downlink. We examined the voice service in CS FDD

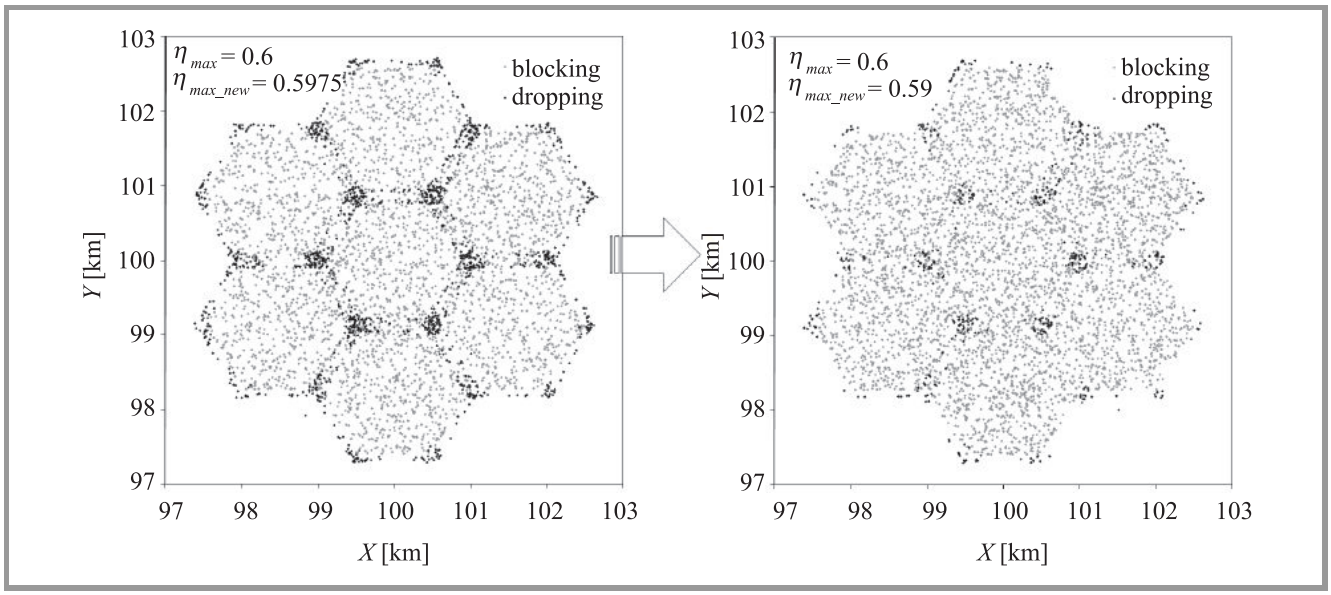


Fig. 2. Positions of blocking and dropping events in the simulated network model for selected CAC configuration cases.

domain, thus at least one congested transmission direction could have caused the CC procedure action.

Assuming equal users priorities, the connection selected to be dropped should be the one, which occupies the largest amount of resources at a given moment in time. In UMTS this means that such a connection introduces the largest interference or uses the largest transmission power among all the existing ones. This approach leads to minimization of unwelcome dropping events. Due to such a strategy, positions in the network of users suffering call dropping are usually close to the cell boarder, if no indoor users are assumed. The occurrence of blocking and dropping events in the regular simulated network model is shown in Fig. 2.

2.2. Call Admission Control

The most important procedure responsible for the *BR* and *DR* values mutual relation is call admission control. It assures a required balance between these indicators, which is always a trade-off. To decrease dropping, the network must accept a smaller number of new calls and thus, decreased *DR* is paid with increased *BR*. Unfortunately, the overall number of both unwelcome events may increase in this case, so, in the effect we can serve less traffic, but with a better (or more desirable) GoS. En example of this trade-off achieved thanks to the CAC procedure configuration is illustrated in Fig. 2.

For all the examined cases the constant value of $\eta_{max} = 0.6$ was used. It was assumed that the CAC procedure worked perfectly. This means that there was no possibility of making wrong decision about the admission of a new call. Thus, there was no possibility of dropping any call due to the new one admission in a cell. If the CAC decided to admit a new call it meant that all transmissions in the particular cell would be maintained directly after that. Moreover, it was assumed that CAC worked immediately, so the decisions were made without any delay.

The cost of achieved dropping improvement for the defined network operation scenario (with a constant traffic volume and user mobility profile) is shown in Fig. 3. Served traffic rate (*SR*) is defined as the ratio of well served calls number to all call attempts, where "well served" means all the ones which were admitted and successfully served.

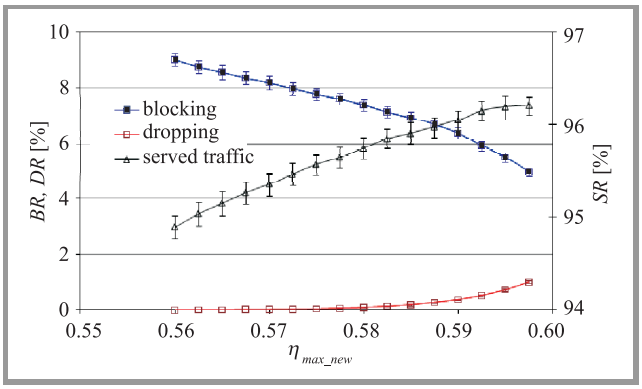


Fig. 3. CAC threshold configuration: reward and costs.

We can observe that the dropping protective CAC configuration (achieved by decreasing the load threshold η_{max_new} while maintaining a constant value for η_{max}) causes an increase in the overall number of unwelcome events as well as a decrease in the served traffic rate. The difference between η_{max} and η_{max_new} implies how much the network is better protected against dropping than blocking. By examining consecutive values of the CAC load threshold we can also estimate a curve showing the *BR* versus *DR* relation that is possible to obtain for the particular network operation case. This curve shows Pareto front for the CAC optimization if only *BR* and *DR* are taken into account. The results

obtained by simulations for consecutive η_{max_new} values assuming a constant value of η_{max} are presented in Fig. 4.

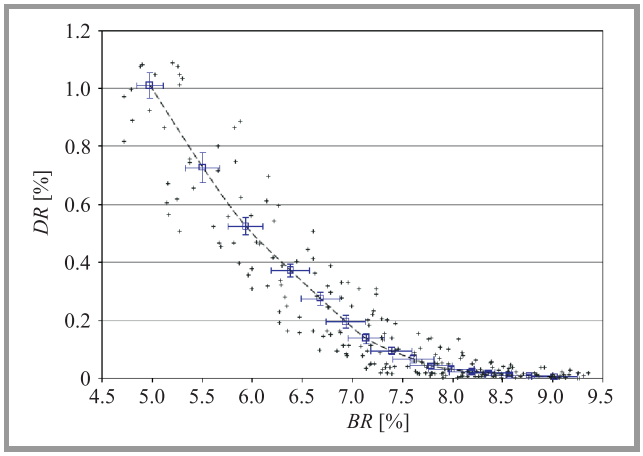


Fig. 4. Blocking-dropping Pareto front estimation for the simulated case of CAC procedure configuration.

Each averaged point corresponds to a different value of the load threshold η_{max_new} set the same in all the simulated base stations. Each configuration case is presented in the figure as a cloud of small crosses measured during the simulation in all the cells as well as their average estimator. Results from all the cells in the simulation model could have been averaged because they were obtained for the same traffic and environment conditions. Moreover, it is important to note that all the simulations covered the same network operation period as well as to enable reliable comparisons, every time the random generator was initiated with the same value. The confidence level was assumed at 95%.

2.3. Traffic Volume and User Mobility Impact on BR and DR

Blocking and dropping rates are function of the traffic volume and the user mobility profile (especially their average speed). Based on the described network model we examined the impact of the offered traffic volume and the mean mobile stations speed of movement on *BR* and *DR* values. The measurement period ΔT was defined to 900 s and the overall simulation time was set to 50 000 s. Figure 5a presents results of performed simulations assuming a constant user mobility profile and different (uniformly distributed) traffic volumes offered to the network. We can observe much bigger increase of blocking rate than the dropping one in case of a heavy network load, which is a direct result of the CAC operation.

In the second case, shown in Fig. 5b, a constant volume of the offered traffic was assumed, but tests were performed for mobile stations moving with different velocities (V_{MS}). When mobile stations moved fast, variations of the interface conditions increased, more handovers were performed and thereby the risk of dropping a call was also increased. That

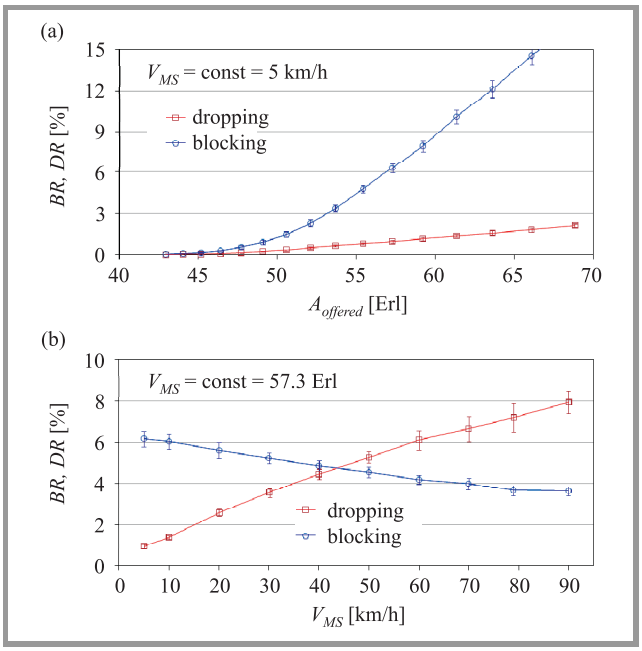


Fig. 5. Impact of (a) traffic volume offered per one cell and (b) users velocity on blocking and dropping rate values for a selected CAC configuration case.

caused, on the other hand, more room for admission of new calls and resulted in decreased *BR* value.

3. Optimization of CAC

The values of the η_{max_new} and η_{max} thresholds impact the *BR* and *DR* indicators. Besides an obvious care about the rates minimizing, a proper relation between them should be assured according to an operator's policy that can be defined by the following objective function:

$$CF = BR + W\ DR, \tag{1}$$

where $W \geq 0$ is a weight factor that was assumed as $W = 4$ (similar as in [5]). The optimization of the CAC procedure can be based on the cost function CF (η_{max_new}) minimization.

The means of estimation of the optimal η_{max_new} value for selected and similar in all cells traffic conditions is illustrated in Fig. 6. Each point on the plot is a result of multiple tests performed for a given simulation scenario and the η_{max_new} threshold value. The cost function approximation CF (η_{max_new}) enables finding its minimum.

The above case assumed stable conditions of traffic and user mobility during all the simulation time. In a real network they change periodically, so, obtained this way result would provide an optimal value of the CAC threshold ($\eta_{max\_new\_opt}$) for averaged traffic conditions.

To examine the dependence of $\eta_{max\_new\_opt}$ on traffic volume offered within a cell ($A_{offered}$) and the mean speed of mobile stations (V_{MS}) we simulated the network model described in subsection 2.1 for many configuration cases.

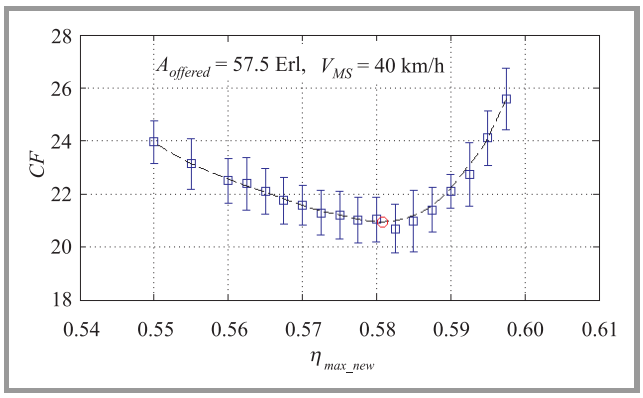


Fig. 6. Optimal η_{max_new} value estimation for the selected traffic volume and mobile stations speed.

Tests related to the traffic volume and users velocity were performed separately, assuming stability of other conditions. To estimate the relation of $\eta_{max\_new\_opt}$ to $A_{offered}$ the network model was examined for different values of mean time interval between consecutive calls. During simulations a constant value of mobile station speed equal to 40 km/h was assumed. Next, the dependence of $\eta_{max\_new\_opt}$ on V_{MS} was analyzed for the constant value of $A_{offered} = 53.6$ Erl.

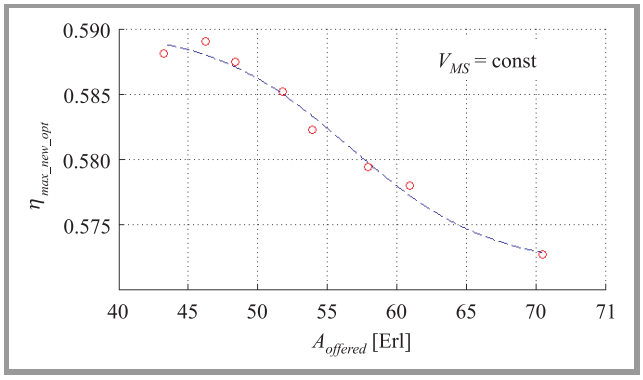


Fig. 7. Optimal η_{max_new} value dependence on the traffic volume offered to one cell for a constant speed of mobile stations.

As shown in Fig. 7, when the traffic volume is bigger the cost function Eq. (1) reaches its minimum for smaller values of η_{max_new}. Thus, the difference between η_{max} and η_{max_new} increases.

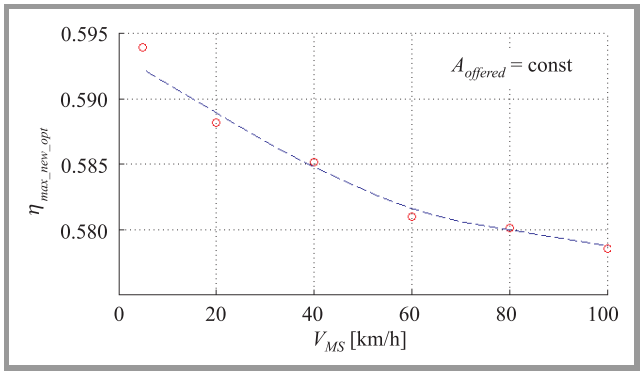


Fig. 8. Optimal η_{max_new} value dependence on the mobile station speed for a constant traffic volume offered to one cell.

Considering the second simulation scenario for constant $A_{offered}$ and variable V_{MS}, when mobile stations are moving faster the optimal η_{max_new} value is smaller (Fig. 8).

Results presented in Figs. 8 and 9 show relations of the optimal CAC configuration in a qualitative manner. Although the estimated curves should be treated as only approximated, the crucial fact is that the optimal CAC configuration depends on the values of $A_{offered}$ and V_{MS}, which in a real network change periodically. En example of traffic volume measurement result is shown in Fig. 9.

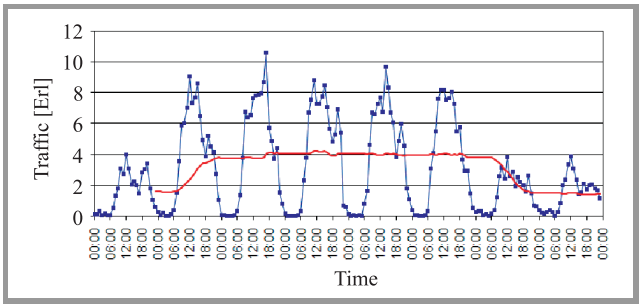


Fig. 9. Traffic volume for a sample cell (8 day period) with a moving 24 h average; real cell measurements, data received from a Polish operator.

We can expect that if we are able to adjust the η_{max_new} parameter according to the traffic variation during a day or a week, we could reach a better traffic GoS in the meaning of the defined objective function. To ensure the optimal relation between blocking and dropping of calls the η_{max_new} threshold value should be decreased when the offered traffic volume increases as well as in the case of increased speed of mobile stations (as shown in Figs. 8 and 9).

The relations presented in Figs. 8 and 9 can help in definition of such a dynamic CAC adaptation process, however there are some important issues that must be considered. If we are going to take advantage of measurements performed online, we must take into account their reliability and feasibility.

Basically, if we want the adaptation process to work effectively we need to perform the η_{max_new} threshold changes on a relatively short time scale. Hence, direct usage of *BR* and *DR* indicators in this process (as proposed, e.g., in [9]) might be problematic. First of all, short periods of measurements result in a poor reliability of the obtained indicators. According to *BR* and *DR* it is crucial, since these indicators are based on counters of events which happen rarely and thus, require long periods of measurement. Moreover, for frequently performed measurements a lot of resources are required for sending reports from all the monitored cells as well as for data processing in RNC [10] which is the entity responsible for gathering measurement reports.

To solve these problems the concept of decentralized (single cell oriented) RRM architecture can be applied. It is assumed to pass over information about cell coupling and close the whole measurement and decision process in a sin-

gle cell. Although it suggests to reorganize the scheme of measurement gathering and processing in UMTS, which would require a certain effort, the reward of opening new possibilities for managing the network seems to be tempting. Moreover, reconfigurations of the η_{max_new} threshold can be based on indicators that are related to *BR* and *DR* but measured with much better reliability during short periods of time.

The offered traffic volume can be better estimated in a short period. It is also based on event counters but such that occur much more often, i.e., incoming new calls. Hence, the measurement of $A_{offered}$ provides much better reliability results than the measurement of *BR* and *DR* during the same period ΔT. The reliability of these three indicators can be compared based on scattering of consecutive samples shown in Fig. 10. The same simulation scenario assuming stationary traffic conditions was examined for different measurement schedules. Note that the $A_{offered}$ standard deviation related to its average estimator value is one order of magnitude smaller that for *BR* and *DR*.

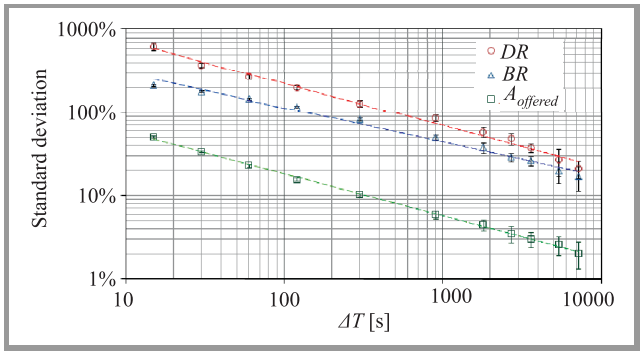


Fig. 10. *BR*, *DR* and $A_{offered}$ scattering for different measurement periods.

Because it is hard to get information about user speeds and movement directions, therefore, for the purpose of dynamic CAC adaptation the mean time of serving calls in a cell ($t_{average}$) can be used. It is directly related to V_{MS} as presented in Fig. 11.

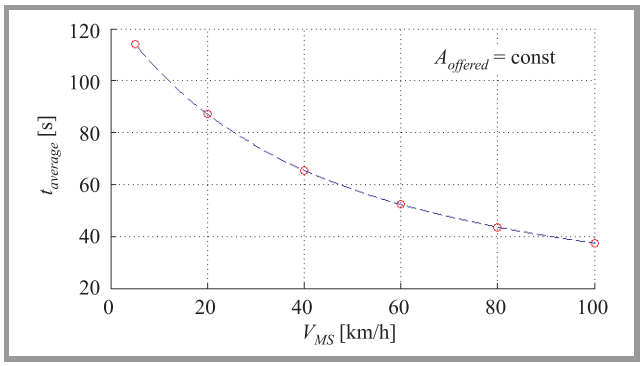


Fig. 11. Average calls duration in a cell dependence on the mobile stations velocity.

Based on relations shown in Figs. 8 and 11 the η_{max_new} dependence on the easy to measure $t_{average}$ was estimated (Fig. 12). If users move faster, there are more handovers in the network and the mean time of holding calls in a single cell decreases (assuming the mean overall connection duration remains the same).

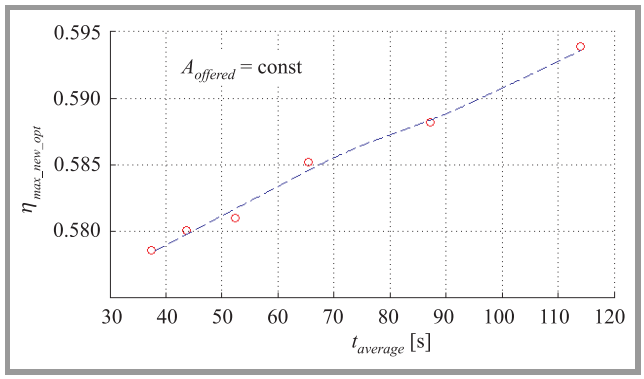


Fig. 12. Optimal CAC load threshold dependence on the average calls duration in a cell.

4. Conclusions and Future Work

Mobile radio network optimization and management methods have to follow the evolution of radio networks and systems. This evolution leads to more dynamic and flexible systems supporting a wider range of services and business areas. In this paper we indicated opportunities for dynamic adaptation of the call admission control procedure. The performed simulations showed that the optimal CAC load threshold depends on the traffic volume offered within a cell and the speed of mobile stations, which is related to the mean time of serving calls in the cell. Based on these relations as well as the indicated limitations a short term CAC threshold adaptation process can be applied. We expect it would enable a better network flexibility according to the traffic's GoS requirements.

References

[1] H. Holma, A. Toskala, *WCDMA for UMTS, Radio Access For Third Generation Mobile Communications*. Wiley, 2004.

[2] *Radio Network Planning and Optimization for UMTS*. J. Laiho, A. Wacker, and T. Novosad Eds. Wiley, 2005.

[3] D. Everitt, "Traffic engineering of the radio interface for cellular mobile networks", in *Proc. IEEE*, vol. 82, no. 9, pp. 1371–1382, September 1994.

[4] W. Ludwin, *Projektowanie sieci komórkowych w aspekcie ruchowym*. Kraków: Uczelniane Wydawnictwa Naukowo-Dydaktyczne AGH, 2003 (in Polish).

[5] *Understanding UMTS Radio Network Modeling, Planning and Automated Optimization: Theory and Practice*. M. J. Nawrocki, M. Dohler, and A. H. Aghvami Eds. Wiley, 2006.

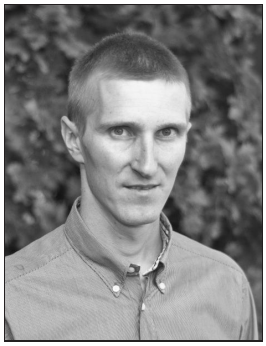
[6] M. Wągrowski, "Analiza interferencji współkanałowych w łączu w górę dla interfejsu WCDMA/FDD", *Krajowa Konferencja Radiokomunikacji, Radiofonii i Telewizji*, Warszawa, 2004, pp. 514–517 (in Polish).

[7] 3GPP TS 25.101, User Equipment radio transmission and reception (FDD).

[8] 3GPP TS 25.104, Base Station radio transmission and reception (FDD).

[9] GANDALF (Monitoring and self-tuning of RRM parameters in a multi-system network), Eureka, Celtic (04.2005–12.2006). Available: www.celtic-gandalf.org

[10] M. Nawrocki, M. Wągrowski, K. Sroka, R. Zdunek, and M. Miernik, *On Input Data for the Mobile Network Online Optimisation Process*. COST 2100 TD(09)747, Braunschweig, Germany, 2009.



Michał Wągrowski is an Assistant Professor at the Department of Telecommunications, AGH University of Science and Technology (AGH-UST), Kraków, Poland. He received his M.Sc. in Electronics and Telecommunications in 2000 and Ph.D. degree in Telecommunications in 2011, both from AGH-UST. His interests include mobile networks planning and optimization. He has been actively working in European IST and Celtic projects as well as in grants supported by Polish Ministry of Science. He was also an active member of COST Action 2100. He is co-author of two books as well as many technical papers and reports. He has served as a reviewer for several IEEE conferences and journals.

E-mail: wagrowski@kt.agh.edu.pl
Department of Telecommunications
AGH University of Science and Technology
Mickiewicza Av. 30
30-059 Kraków, Poland

Wiesław Ludwin received the M.Sc. and Ph.D. degrees in Electronic and Telecommunications Engineering from the Faculty of Electrical Engineering, AGH University of Science and Technology (AGH-UST), Kraków in 1978 and 1983, respectively. In 2005 he received the Dr Hab. degree in Telecommunications and Radiocommunications from the Military University of Technology, in Warsaw. From 1978 to 1986 he was with the Institute of Control Systems Engineering and Telecommunications; since 1986, he has been with Department of Telecommunications AGH-UST, where he currently holds a position of professor. His general research interests are in applied radiocommunications. Particular topics include wireless system design for telecommunications and traffic and mobility modeling in cellular networks. He is the author or co-author of three books on wireless networks and more than 50 research papers. The paper "Is Handoff Traffic Really Poissonian?" published in proceedings of the 4th IEEE International Conference on Universal Personal Communications, ICUPC'95 held in November 1995 in Tokyo, Japan has been widely referenced.

E-mail: ludwin@kt.agh.edu.pl
Department of Telecommunications
AGH University of Science and Technology
Mickiewicza Av. 30
30-059 Kraków, Poland

Paper

Network-on-Multi-Chip (NoMC) with Monitoring and Debugging Support

Adam Łuczak, Marta Stępniewska, Jakub Siast, Marek Domański, Olgierd Stankiewicz, Maciej Kurc, and Jacek Konieczny

Chair of Multimedia Telecommunications and Microelectronics, Poznań University of Technology, Poznań, Poland

Abstract—This paper summarizes recent research on network-on-multi-chip (NoMC) at Poznań University of Technology. The proposed network architecture supports hierarchical addressing and multicast transition mode. Such an approach provides new debugging functionality hardly attainable in classical hardware testing methodology. A multicast transmission also enables real-time packet monitoring. The introduced features of NoC network allow to elaborate a model of hardware video codec that utilizes distributed processing on many FPGAs. Final performance of the designed network was assessed using a model of AVC coder and multi-FPGA platforms. In such a system, the introduced multicast transmission mode yields overall gain of bandwidth up to 30%. Moreover, synthesis results show that the basic network components designed in Verilog language are suitable and easily synthesizable for FPGA devices.

Keywords—debugging, FPGA, multi-chip, NoC, video coding.

1. Introduction

Network-on-chip (NoC) is a relatively new design approach that provides a methodology of implementing Systems on chip (SoC) interconnections. NoC-based systems incorporate a network infrastructure that offers remarkable improvement over conventional communication systems like bus-based or circuits-switching-based [1]. Because basic network components are reused, there is no need to implement network infrastructure from the scratch and thus, the design costs related to communications are reduced. Moreover, scalability of the system is greatly improved because new devices can be added in a structured way. Finally, NoCs provide communication abstraction, which allows independent design of devices [2]. NoC based architecture can be used in both ASICs and FPGAs. Commonly, the first step of system design is to implement the system on FPGAs and the second is to move it to ASIC [3]. In most of cases, the circuit optimized for FPGA is also efficient as ASIC (but not in reverse). Due to this fact, it is more worthy to consider NoC networks for FPGAs. In order to create a useful and efficient NoC architecture, the proposed solutions should meet certain requirements related to transmission bandwidth, communication latency, structure flexibility and many others. The implementation cost and possibility to reuse NoC component is also important. In practice, when complex hardware is designed (such as a video encoder) certain features such as scalability and unified communication interfaces are highly expected.

Hardware implementations of recent video coding standards, as for example advanced video coding AVC/H.264 [4], consist of many compression tools and pre-/post-processing blocks. Additionally, in order to implement a decoder or encoder that works in real time parallel processing has to be applied. It means that the design consists of many processing elements that require high communication bandwidth (especially with memory) and in some cases the whole design requires more than one device (FPGA).

The work has been aimed at development of a communication infrastructure based on the idea of NoC, which allows to dynamically combine multiple integrated circuits and will support the testing and monitoring functionalities, as a result, a new variant of NoC have been proposed. The new network architecture will have none of the drawbacks listed in the section below.

2. Main Network-on-Chip Drawbacks

As it was already said, Network-on-Chip (NoC) is an efficient solution for connecting modules of hardware application but has two main drawbacks:

- There is **no scalability and flexibility** for multi-chip systems. Scalability may be achieved by using a hierarchy in interconnect system. However, not all hierarchical networks are flexibly scalable in terms of multi-chip scalability. Some works, e.g., [5], [6], introduce a hierarchy to improve flow of network traffic and ease resource management but the proposed NoC extensions are not suitable for multichip systems because any change in the structure of the network requires its reconstruction. Another example has been shown in [7]. Despite it is designed for hierarchical arrangement of chip-multi-processors (CMP), based on mesh topology, such structure is inefficient in the case of non-homogeneous tiles. Moreover, mesh topology, as a higher level interconnect system, is hardly scalable. There are more solutions [8], [9] but none of them is appropriate, nor do they meet the aforementioned requirements.
- **Only a unicast transmission is supported** but in multimedia applications, as for example video encoding, many processing cores use the same source data. In the case of the unicast transmission these data need to be sent to their destinations multiple times. Such

unnecessary data transmissions can be significantly reduced by applying the multicast transmission.

3. Network-on-Multi-Chip

The authors propose a variation of NoC for multi-chip systems called network-on-multi-chip (NoMC). The NoMC is a **hierarchical NoC network**. The proposed way of intergroup and interchip connection management enables dynamic linking of multiple chips without a need of redesigning. In general, the NoC structure has been split into two areas: local and global. The global part of NoC has a tree structure with full dynamic of the linking mechanism, but the local one can be implemented as any structure with one gateway to the global part. This solution simplifies system expansion with new functionalities/processing cores.

Additionally, to provide efficient data processing and to improve network performance authors introduce **a multicast transmission**. The idea is simple: more than one destination address in packet header is allowed. Although, the implementation requires proper packet replication in network switches, we get ability to send the same data to several locations, even to several chips. Moreover, because it is possible to add an additional address to any packet, we suddenly get the ability to send all packets not only to primary destination location but also to debag/monitoring location. In this way the authors achieve very useful additional functionality on the NoC level that is not yet described in literature.

4. Scalability and Hierarchical Addressing

We consider scalability in terms of the ability to easy extend the system by new hardware components. Our new scalable architecture of NoMC consists of 3 levels of hierarchy, starting from the lowest level:

- **Local network** – also called a **group** of processing elements (PEs), that contains PEs, network interfaces called endpoints (EPs) and routers. One chip consists of at least one group of PEs.
- **Cluster level** that provides connectivity for a set of groups (local networks) (Fig. 1a). One chip can consist of more than one cluster, but for small projects only a local network may exist without higher level of hierarchy.
- **System level**, which is introduced to interconnect clusters. The higher level of interconnects enables linking multichip boards together. Active elements at the system level are characterized by hot plug support.

We also introduce gateways to the NoC network, which separate all of the hierarchy levels from each other. The main goal of gateways is to parse packets and extract or include information necessary for proper routing. Such an approach allows designing of each hierarchy level individually. The local network architecture is defined with only a set of devices (routers and endpoints, i.e., network interfaces for PE) that can be connected applying any topology. Since routers are expensive in terms of hardware consumption, their number should be as low as possible. In comparison to commonly known network interfaces [10] the functionality of endpoint has been extended to meet the aforementioned requirements. The endpoints are able to perform basic switching operations and may be connected to each other without a need for more sophisticated routers. The detailed description of the hierarchical addressing was presented in previous works [11]–[17]. The addressing scheme is adjusted to hierarchical architecture (Fig. 1b). Each address consists of three parts, each referring to one level of network hierarchy. At a particular level only own part of address is recognized. In order to introduce multicast transmission mode the authors propose to add more than one destination address per packet (see Fig. 3), each address is then checked in every network element (gateways, routers, etc.). The packet is copied if routes for any of the destination addresses are splitting. The proposed solution for external network architecture (cluster and system level) is based on a tree topology. Distinction between cluster and system level has been introduced in order to connect clusters flexibly. Moreover, tree structure allows designing of a simplified routing algorithm and packet handling protocol which yield reduction of hardware consumption.

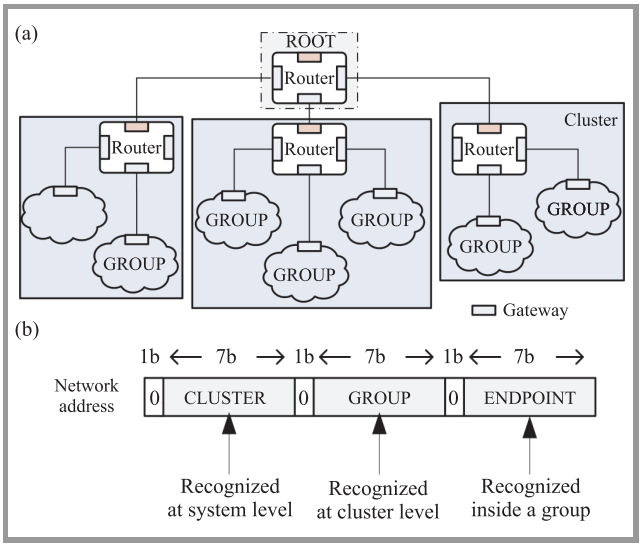


Fig. 1. (a) Hierarchical structure of cluster and system level of a network, (b) network address format.

5. Multicast Transmission

Classical NoC networks support only simulcast transmission, which is enough for most simple applications, but is not sufficient for complex applications and for debug and

monitoring features. Our research indicated that implementation of multicast (similar mode to the Ethernet network) is possible: instead of a single destination address, multiple addresses are assigned to every packet (Fig. 2). The main change includes network routers which must be able to duplicate packets consisting multiple addresses. This means that the main cost of multicast feature implementation is placed in routers. As it has been already said, multicast functionality allows sending of a copy of the packet to any location but in a particular case it may be a monitoring/debugging device. In order to design a router that uses

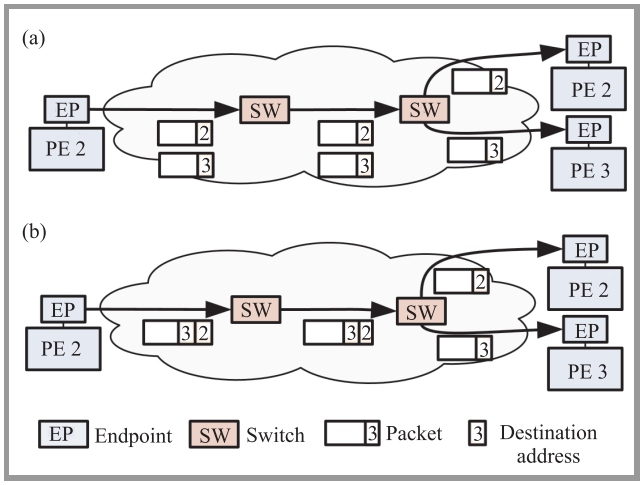


Fig. 2. (a) Unicast and (b) multicast communication with example of packet replication.

a reasonable amount of memory, the network packets size has been limited to 32 words. Such short packets/messages make network traffic more fluent and reduces the cost of packet replication process. Packets always start with the field *Destination address* and end with *EndOfPacket* command, as shown in Fig. 3.

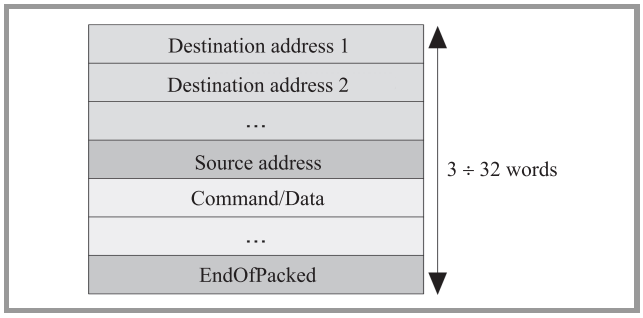


Fig. 3. NoMC packet structure with multiple destination address.

6. Debugging and Monitoring Features

As a result of introducing of the multicast feature we have obtained additional functionality such as debugging and monitoring. The well-known standard for in-circuit test is JTAG [4] protocol, which is intended for system management tasks. It requires two physical components: test access port (TAP) which interprets JTAG protocol, and boundary scan register (BSR). Implementing of those modules in each PE may require large amount of chips resources regarding the scale of current designs. There are several approaches in literature of NoC embeded debugging functionality [5]–[9], but the hardware cost of this functionality is still significant. Moreover, the described proposals are not scalable and mostly based on JTAG standard. None of them offers full and scalable monitoring feature [13], [18]. Multicast based mechanism introduced by the authors include real-time monitoring, management of the devices and system configuration. The debugging is supervised by the so-called remote debugging host (RDH) (Fig. 4). RDH is an off-chip control device or software application on a personal computer, connected to the system with any physical interface. The role of RDH is to provide user interface to the debugging functionality, such as: applying test vectors, gathering debugging data, handling exceptions or emulating hardware devices in software. More about multicast transmittion and debugging can be found in [12]–[17].

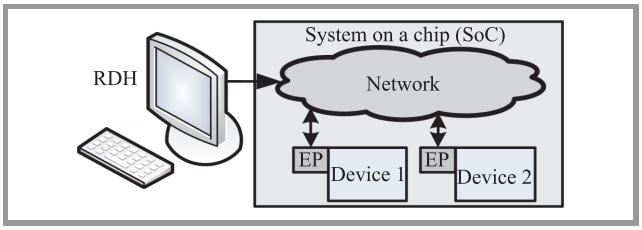


Fig. 4. System with one remote debuging host (RDH) as a root and examoplary system-on-chip.

Debug-mode in the endpoint forces sending of a copy of each packets outgoing to RDH. Endpoints use multicast transmissions and add RDH address to the packets address list. Debug-mode can be switched on and off for each endpoint individually. RDH receives packet duplicates and with the use of specific application is able to recognize and present packets data to the user. Also, correctness of packets and data format can be verified. With sufficient network bandwidth, real-time debugging/monitoring is possible.
The authors have assessed the proposed ideas during design, implementation and testing process of AVC/H.264 video decoder. At that time, many examples of debugging functionality usage were observed, which otherwise would be very difficult to attain. For example, without debugging functionality, it would be required to resynthesize the whole project with additional testing benches in order to test what was wrong: the transmission through the link was corrupted, there was some kind of hazard situation somewhere or it was just a synthesis error.

7. Hardware Platform

In order to verify the proposed solutions a hardware platform has been designed and produced. The test platform made at Poznań University of Technology consists of 2 to 9 FPGA devices. A Xilinx FPGA Virtex-4/5 and

Spartan-3 devices have been used (Figs. 5 and 7). All the NoMC network components were implemented in Verilog hardware description language and synthesized using the ISE design suite. Using such a system, the authors were able to conduct many experiments for various NoMC configurations and for a wide range of parameters.

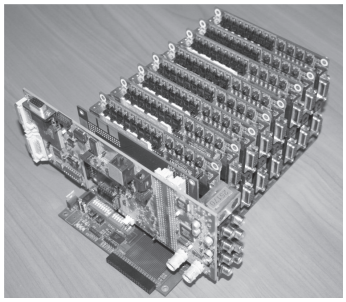


Fig. 5. The multi-board experimental system with FPGA devices and an SDI video grabber.

Table 1
Synthesis results for Spartan6 XC6SLX75-3 FPGA device

Elements	LUT	FlipFlop	CLK [MHz]
Router (4 ports)	1106 (2%)	759 (1%)	278.8
Router (3 ports)	647 (1%)	573 (1%)	277.3
Endpoint	436 (1%)	345 (1%)	315.4
Gateway	515 (1%)	409 (1%)	315.3

In Table 1 the synthesis results of basic network components are shown. As one may see, for 32-bit bus of NoC and Spartan-6 FPGA, it is possible to achieve 1 GB/s of throughput.

8. Conclusions

In this paper, the authors summarize research and development of new NoMC architecture. In the course of development of addressing scheme and packet flow control in the network strong emphasis was put on certain features, such as multichip scalability, debugging and monitoring functionality that was expected. Consequently, the new architecture of interconnect system consists of three levels of hierarchy, each separated with a dedicated device, referenced as a gateway. As it was highly expected, the multicast transmission mode which provided improved network performance and significant reduction of the required bandwidth was successfully introduced.
The main achievements include expansion of network to support the packet remote monitoring and hierarchical addressing for scalability support. An assessment of the proposed debug system on an exemplary real debugging scenario has been made using multi-FPGA boards (Figs. 6 and 7). The authors tested many applications targeted to distributed systems. Among them a H.264/AVC decoder, motion estimation algorithm and several transmission and data broadcast schemes (for example, real-time HD video

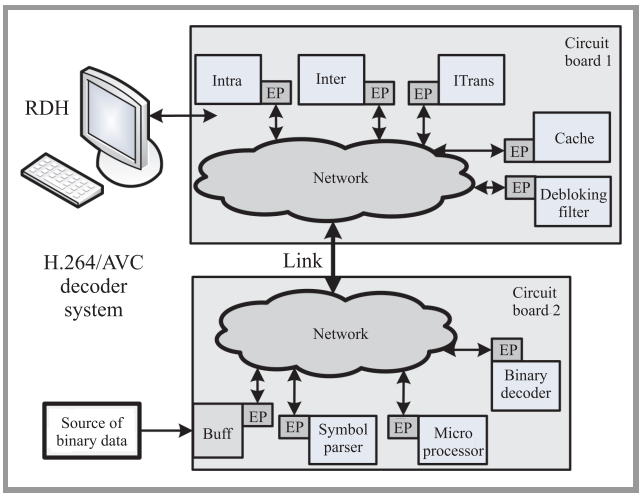


Fig. 6. The implemented H.264/AVC system on two circuit boards with remote debuging host and external source of testing data.

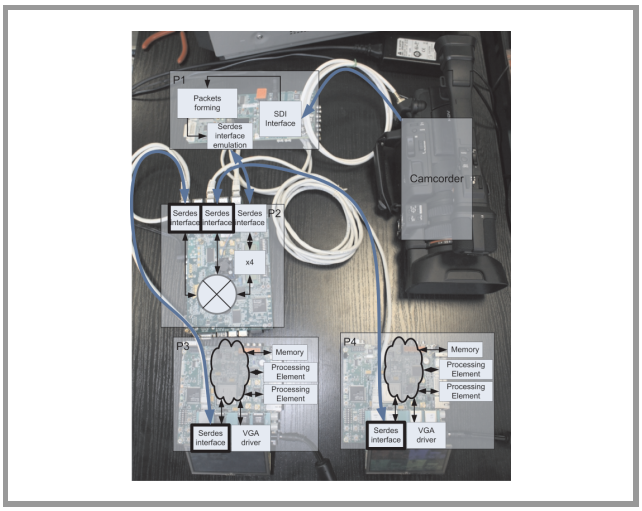


Fig. 7. A video capture and processing system based on two Virtex-4 boards with a video grabber.

sequence capture and video data broadcast to all FPGA devices in system (Fig. 7)). Finally, the conducted research and analysis prove that the designed network-on-multi-chip works correctly and meets all the assumed requirements.

Acknowledgement

The work was supported by public funds as a research project “Next Generation Services and Networks – technical, application and market aspects”, PBZ-MNiSW-02/11/2007.

References

[1] C. Hilton and B. Nelson, “PNoC: a flexible circuit-switched NoC for FPGA-based systems”, *Comput. Digit. Techn., IEEE Proc.*, vol. 153, no. 3, pp. 181–188, May 2006.
[2] J. Henkel, W. Wolf, and S. Chakradhar, “On-chip networks: a scalable, communication-centric embedded system design paradigm”, in *Proc. 17th Int. Conf. VLSI Design*, 2004, pp. 845–851.

[3] P. Subramanian, J. Patil, and M. K. Saxena, "FPGA prototyping of a multi-million gate system-on-chip (SoC) design for wireless USB applications", in *Proc. Int. Conf. Wirel. Commun. Mob. Comput. Connect. World Wirel.*, Leipzig, Germany, 2009.

[4] *Information Technology Coding of Audio-Visual Objects, Part 10: Advanced Video Coding.* ISO/IEC FDIS 14496-10.

[5] A. Lankes, T. Wild, A. Herkersdorf, "Hierarchical NoCs for optimized access to shared memory and IO resources", in *Proc. 12th Euromicro Conf. Digit. Sys. Design DSD 2009*, Patras, Greece, 2009, pp. 255–262.

[6] R. Holsmark, S. Kumar, M. Palesi, and A. Mejia, "HiRA: a methodology for deadlock free routing in hierarchical networks on chip", in *Proc. 3rd ACM/IEEE Int. Symp. Netw.-on-Chip*, La Jolla, USA, 2009, pp. 2–11.

[7] C. Puttmann, J.-C. Niemann, M. Porrmann, and U. Ruckert, "GigaNoC – a hierarchical network-on-chip for scalable chip-multiprocessors", in *Proc. 10th Euromicro Conf. Digit. Sys. Design DSD 2007*, Lubeck, Germany, 2007, pp. 495–502.

[8] X. Leng, N. Xu, F. Dong, and Z. Zhou, "Implementation and simulation of a cluster-based hierarchical NoC architecture for multiprocessor SoC", in *Proc. IEEE Int. Symp. Commun. Inform. Technol. ISCIT 2005*, Beijing, China, 2005, vol. 2, pp. 1203–1206.

[9] *WISHBONE System-on-Chip (SoC) Interconnection Architecture for Portable IP Cores*. Revision: B.3, Sept. 2002.

[10] E. Salminen, A. Kulmala, and T. D. Hmlinen, *Survey of Network-on-chip Proposals*, White Paper, OCP-IP, March 2008.

[11] A. Łuczak, M. Kurc, and J. Siast, "Szeregowy interfejs komunikacyjny dla układów FPGA serii Virtex", *Pomiary Automatyka Kontrola*, vol. 56, no. 7, 2010 (in Polish).

[12] A. Łuczak, M. Kurc, M. Stępniewska, and K. Wegner "Platforma przetwarzania rozproszonego bazująca na sieci NoC", w XII Konf. Naukowa Reprogramowalne Układy Cyfrowe, Szczecin, Polska, maj 2009 (in Polish).

[13] H. Yi, S. Park, and S. Kundu, "A design-for-debug (DfD) for NoC-Based SoC debugging via NoC", in *Proc. 17th Asian Test Symp.*, Sapporo, Japan, 2008, pp. 289–294.

[14] M. Stępniewska, A. Łuczak, and J. Siast, "Network-on-multi-chip (NoMC) for multi-FPGA multimedia systems", in *Proc. 13th Euromicro Conf. Digit. Sy. Design DSD 2010*, Lille, France, 2010.

[15] M. Stępniewska, O. Stankiewicz, A. Łuczak, and J. Siast, "Embedded debugging for NoCs", in *Proc. 17th Int. Conf. Mixed Design of Integr. Circ. Sys.*, Wrocław, Poland, June 2010.

[16] A. Łuczak and J. Siast, " Network-on-chip with multicast transsmition support", to be published.

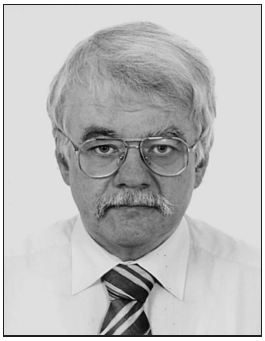
[17] A. Łuczak, M. Stępniewska, and J. Siast, "Hierarchical addressing with hot-plug support in Network-on-Multi-Chip", to be published.

[18] H. Yi, S. Park, and S. Kundu, "On-chip support for NoC-based SoC debugging", *IEEE Trans. Circ. Sys.*, vol. 57, no. 7, pp. 1608–1617, 2010.

Adam Łuczak was born in 1972. He received his M.Sc. and Ph.D. degrees from Poznań University of Technology in 1997 and 2001, respectively. In 1997 he joined the image processing team at Poznań University of Technology. He is Member of of Polish Society Theoretical and Applied Electrical Engineering (PTETiS). His research activities include video coders control, MPEG-4/H.264 systems and hardware implementations of digital signal processing algorithms.

E-mail: aluczak@multimedia.edu.pl
Chair of Multimedia Telecommunications
and Microelectronics
Faculty of Electronics and Telecommunications
Poznań University of Technology
Polanka st 3
60-965 Poznań, Poland

Marta Stępniewska was born in 1981. She received her M.Sc. degree from Poznań University of Technology in 2005. She is a Ph.D. student at the Chair of Multimedia Telecommunications and Microelectronics. She takes part in some projects taking up hardware programming. She is interested in video transmission in internet network, recent history, physics, anthropology and cycling.
E-mail: mstep@multimedia.edu.pl
Chair of Multimedia Telecommunications
and Microelectronics
Faculty of Electronics and Telecommunications
Poznań University of Technology
Polanka st 3
60-965 Poznań, Poland



Marek Domański was born in 1954. He received the M.Sc., Ph.D. and Habilitation degrees from Poznań University of Technology, Poland, in 1978, 1983 and 1990, respectively. He headed many research projects on image and video compression, image and video enhancement and restoration, multidimensional digital filters and telemedicine. Recent activities include industry-oriented research on 3D video, advanced video and audio compression techniques and as well as on video analysis and video surveillance. Prof. M. Domański serves as the head of Polish delegation to MPEG and he actively participates in MPEG standardization activities. He is an author or co-author of over 200 peer-reviewed papers in journals and proceedings of internationally recognized conferences. He has already advised 15 Ph.D. dissertations that have been finished. Currently he is a professor at Poznań University of Technology and he is the head of Chair of Multimedia Telecommunications and Microelectronics at this university.
E-mail: domanski@et.put.poznan.pl
Chair of Multimedia Telecommunications
and Microelectronics
Faculty of Electronics and Telecommunications
Poznań University of Technology
Polanka st 3
60-965 Poznań, Poland

Jakub Siast received the M.Sc. degree in Electronics and Telecomunications from the Poznań University of Technology, Poland, in 2009. He is a Ph.D. student at the Chair of Multimedia Telecommunications and Microelectronics. The main area of his professional activities are video compression, networks on chip and FPGA devices.

E-mail: jsiast@multimedia.edu.pl
Chair of Multimedia Telecommunications
and Microelectronics
Faculty of Electronics and Telecommunications
Poznań University of Technology
Polanka st 3
60-965 Poznań, Poland

Olgierd Stankiewicz was born in 1982. He received the M.Sc. degree from Poznań University of Technology in 2006. In 2005 he won second place in IEEE Computer Society International Design Competition (CSIDC), held in Washington D.C. Currently, he is a Ph.D. student at the Chair of Multimedia Telecommunications and Microelectronics. His professional interests include signal processing, video compression algorithms, computer graphics and hardware solutions.

E-mail: ostank@multimedia.edu.pl
Chair of Multimedia Telecommunications
and Microelectronics
Faculty of Electronics and Telecommunications
Poznań University of Technology
Polanka st 3
60-965 Poznań, Poland

Maciej Kurc was born in 1984. He received his M.Sc. degree from Poznań University of Technology in 2008. He is a Ph.D. student at the Chair of Multimedia Telecommunications and Microelectronics. The main areas of his professional activities are image processing, video compression algorithms and electronic hardware solutions. He is interested in electronic circuit design and programming, signal processing using FPGA, digital photography and cycling.

E-mail: mkurc@multimedia.edu.pl
Chair of Multimedia Telecommunications
and Microelectronics
Faculty of Electronics and Telecommunications
Poznań University of Technology
Polanka st 3
60-965 Poznań, Poland

Jacek Konieczny was born in 1984. He received the M.Sc. degree from Poznań University of Technology in 2008. He is a Ph.D. student at the Chair of Multimedia Telecommunications and Microelectronics. The main area of his professional activities is video compression in multipoint view systems. His interests are image and audio compression algorithms and their implementation on PC and FPGA platforms.

E-mail: jkonieczny@multimedia.edu.pl
Chair of Multimedia Telecommunications
and Microelectronics
Faculty of Electronics and Telecommunications
Poznań University of Technology
Polanka st 3
60-965 Poznań, Poland

Paper

The Design of an Objective Metric and Construction of a Prototype System for Monitoring Perceived Quality (QoE) of Video Sequences

Lucjan Janowski, Mikołaj Leszczuk, Zdzisław Papir, and Piotr Romaniak

Department of Telecommunication, AGH University of Science and Technology, Kraków, Poland

Abstract—The paper presents different no reference (NR) objective metrics addressing the most important artefacts for raw (source) video sequences (noise, blur, exposure) and those introduced by compression (blocking, flickering) which can be used for assessing quality of experience. The validity of all metrics was verified under subjective tests.

Keywords—mean opinion score, no reference metric, objective metric, quality of experience, video artefacts.

1. Introduction and the General Prototype Concept

The importance of "live" streaming, which operates on the basis of wireless networks, has been verified in recent years by the emergence of numerous applications such as mobile TV and IP video monitoring systems in urban areas. Unlike traditional applications such as web browsing, multimedia applications require real-time content transmission mechanisms with a low negative impact on user-perceived quality of video communication [1]. To meet this requirement, increase user satisfaction and, consequently, increase the profits for service suppliers, a system of evaluation/verification of video artefacts must be developed and implemented. This solution should be designed for wireless transmission infrastructures in order to control the pseudo-subjective quality of "live" transmitted video sequences [1]. The term "pseudo-subjective" means control with the use of objective metrics, verified on the basis of subjective assessments.

Limitations of traditional solutions based on the notion of quality of service (QoS) require arrangements such as described in [2], i.e., taking into account the characteristics of the transmission media, human vision (human visual system, HVS), and the level of quality as perceived by the user (quality of experience, QoE). However, most of the currently available QoE assessment systems are designed either for one specific type of visual content and application, or for one specific scenario of a wireless service. In recent years, models without a reference (also known as no reference, NR, models) have gained particular focus. To evaluate the QoE, they do not require access to the reference (undistorted) sequence.

It should be noted that the development of new QoE models working in the NR scenario is still a challenging area of research because of the limitations of current metrics (which must be applicable in a non-laboratory environment), diversification of the evaluation based on the content and user profile, resistance to variety of emerging distortions, and the need to meet the requirements of low computational complexity.

This paper highlights the need for assessment of imaging artefacts for "live" streaming applications in a wireless environment and describes the models implemented in the NR scenario assessment. The proposed solution is verified using the results of psycho–physical experiments. The results obtained show the usefulness of the proposed mechanisms for assessing the quality of streaming applications in a wireless environment, and confirm the high correlation with the feelings of users.

The concept presented in this article is to create techniques and tools that can be implemented by service providers with a view to continuing the monitoring of overall video sequence streaming service quality. The results (technology and tools) are expected to be used (mainly) in the wireless service.

The most innovative and distinctive functionality of the system is the introduction of NR metrics to assess and monitor the QoE. It should be noted that the proposed credible assessment and control of the perceived quality of video sequences, based on the QoE numerical estimations and the accuracy calculation of video reconstruction in the context of specific parameters and playback rate conditions, play a fundamental role in ensuring QoE for services based on video sequence streaming.

As already mentioned, the quality estimation solution, allowing to assess video sequences when there is a lack of available references, remains a challenge. In contrast to all methods based on reference solutions (full reference, FR, and reduced reference, RR), which are limited by shortcomings in the quality of the source video sequence, the NR approach evaluates absolute quality, as seen from the perspective of the user. NR does not require the additional,

ideal channel to transmit data to be used as a reference. In addition, the NR solution allows for "live" transmitted session tracking, allowing the delivery of estimated results in real-time.

For real applications the authors are interested in absolute quality throughout the supply chain of media (known as end-to-end); in other words, from the beginning (the impact of focus, noise, exposure), through the transition stages (the impact of stream bandwidth scaling), to the end (the impact of the presentation and application). NR-type methods of assessing quality are therefore a natural response to the needs of real video sequence streaming scenarios.

It is particularly important to assess the impact of scalable stream bandwidth. Gaining increasing popularity, video sequence streaming services are still faced with the problem of limited access links bandwidth. Although for wired connections bandwidth is generally available in the order of megabits, higher bit rates are not as common for wireless links. Users of wireless links cannot expect a stable high-bandwidth connection.

Therefore the solution to run video sequence streaming services for such access lines is transcoding video streams "on the fly". The transcoding result is bit rate (and quality) scaling to personalize the stream sent to the current parameters of the access link. Scaling the quality of the video sequences is usually in the (often inseparable) domains of compression, space and time. Scaling in the compression domain usually boils down to operating the codec quantization parameter. Scaling in the spatial domain means reducing the effective image resolution, resulting in increased granularity when one tries to restore the original content to screen size. Scaling in the temporal domain amounts to the rejection of frames, i.e., reducing the number of frames transmitted per second (FPS).

The abovementioned scaling methods inevitably lead to a lower perceived quality of end-user experience. Therefore the scaling process should be monitored for QoE levels. This gives the opportunity to not only control but also to maximize QoE levels in real time, depending on the prevailing transmission conditions. In case of failure to achieve a satisfactory level of QoE, an operator may intentionally interrupt the service, which may help in preserving and allocating network resources to other users.

Unfortunately, determining the level of QoE in any case cannot be reduced to a simple maximization of quantitative parameters in each of the three domains. Consumer perception based on HVS, is highly non–linear and depends on many variables (such as visual content). Therefore attempts are made to create models for automatically determining QoE levels through the analysis of visual content as seen by the user [3].

Attempts to determine the impact of scaling in the compression domain on the perceived QoE quality are particularly difficult. The same compression ratio is not a sufficient indicator of perceived quality. In the NR model it is necessary to identify the impact of this manipulation on the effects shown in the image. The most important effects associated with lossy compression are block artefacts and block flickering. To determine the QoE it is necessary to accurately and quantitatively assess the severity of these effects. Numerous models given in literature [3], [4], [5] usually do not achieve a sufficient correlation with actual user ratings.

It is far easier to model the impact of scaling in the temporal domain, because here at least the FPS value is openly available. Attempts to model the impact of scaling in the domain of perceived quality have been made in several studies, including [6]. It is relatively less complicated to evaluate the effect of a reduction in motion picture effective resolution (i.e., increase in granularity) on the visual effects. These effects were studied in [3], [7], although the former work used other applications.

The methodology of studies presented in the section covering the evaluation of scalable video sequences of this paper is based on subjective quality tests on independent influences of the three abovementioned scaling methods. In addition, the study was carried out on developing metrics, evaluating each quality parameter and presenting the results of statistical analysis of results.

The first value-added feature of this research is the provision of an identical environment for the psycho-physical experiment for all test artefacts and for scaling all three domains of quality using an eleven-point quality scale. This provides an opportunity to compare the results obtained for all considered scaling methods, and to build an integrated model (still being refined) that takes into account the simultaneous combination of methods. Another innovative element is the measure of evaluation of lower qualities due to the large value of QP – a measure with a very high correlation with subjective assessments. Another added value is a detailed statistical analysis of the results obtained for correlation with the mean opinion score (MOS) and statistical reliability. It is an often overlooked element in work on QoE modeling. Furthermore, different video sequences have been used and considered in subjective tests as an additional independent variable, in some cases allowing for the statistical analysis of the impact of a given sequence on the accuracy of the resulting measurement.

In summary, the authors present a concept that involves creating and implementing QoE metrics based on user preferences, assessments of subjective and observer characteristics, and the feedback loop formed by the iterative verification of metrics, modifying their parameters on the basis of these subjective assessments.

The remainder of the article is structured as follows. Section 2 deals with the measurement of quality and artefacts (based on video parameters – Subsection 2.1 and network parameters – Subsection 2.2). Section 3 presents the psycho-physical experiment verification environment. Section 4 presents a statistical analysis of the results in terms of measurement and scaling artefacts in the compression, spatial and temporal domain, and information on how to implement the prototype, while Section 5 contains conclusions and plans for further research.

2. Quality Measures

This section includes a description of video quality metrics dedicated to the assessment of certain video artefacts in a no reference mode. The metrics are used to build a QoE monitoring prototype. This section provides a summary of the work on video quality metrics developed in recent years. For detailed descriptions of the metrics please refer to [1], [8], and [9].

2.1. Measures of Quality Based on QoV Parameters

Exposure. An exposure distortion is understood as the overall quality degradation caused by incorrect exposure time. The metric was inspired by the shape of histograms of images taken for different exposure times. A histogram of a correctly exposed image spreads over the whole luminance range. Histograms of over- and under-exposed images are shifted to the bright and the dark side respectively. The higher the exposure distortion, the more significant the shift. In other words, there are no completely black and white regions on over- and under-exposed images respectively. Consequently, the exposure metric is based on histogram range inspection.
The metric is calculated locally for each video frame. In the first step mean luminance is calculated for each macro-block of a given video frame. The average of the three macro-blocks with the lowest and highest luminance represent luminance (histogram) bounds. The exposure metric for a single frame is calculated as:

$$ex = \frac{L_b + L_d}{2}, \tag{1}$$

where L_b and L_d are bright and dark luminance bounds.

The video level metric is calculated by averaging frame metrics over one scene. The proposed methodology assumes that each natural video sequence has at least some bright and dark regions. It is a significantly more accurate approach than a simple histogram average luminance calculation. For instance, it eliminates the problem when images showing black objects with very few bright regions would be classified as under-exposed.

Blur. The most common approach of image blur estimation utilizes the fact that blur makes image edges less sharp. Recent work representing this approach is described in [10] and [11]. The proposed metric is based on an average width of sharp edges only. Sharp edge selection is critical in terms of predicting accuracy since it eliminates strong content dependency. In the first step, strong edges are detected using the Sobel operator. In the second phase, edge width is measured as a number of neighbouring pixels (localized on the left and right in the same horizontal line) that fulfils the following criteria:

- the right-localized pixel intensity values systematically increase/decrease for rising/falling edges,
- ditto for left-localized pixels,
- the edge slope value does not fall below a certain level, defined by the standard deviation of surrounding pixel intensity.

Noise. The idea behind the proposed noise metric is based on research presented by Lee in [12]. According to this work, the most convenient method of estimating noise in remotely sensed images is to identify homogenous image areas and use them to calculate noise statistics. More recent work utilizing this approach is presented by Farias in [10] and Dosselmann in [11].
The presented approach of identifying homogenous regions guarantees the selection of a comparable number of blocks for images ranging from low to high spatial complexity. It outperforms approaches based on a fixed threshold in terms of a visual nuisance prediction performance. In order to eliminate moderate content dependence (a drawback of existing metrics), the spatial masking phenomenon is addressed by weighting frame-level noise values according to the spatial activity of a given frame. This compensates the well-known property that images with low spatial complexity are more exposed to visual distortion caused by noise.

Block artefact. Blockiness artefact measurement is based on the well-known discrete cosine transform (DCT)-based coding. Each blockiness artefact has at least one visible corner. Recent research utilizing this fact is described in [10] and [11]. In the proposed approach the blockiness artefact is calculated locally for each coding block. The absolute difference of pixel luminance is calculated separately for intra-pairs, represented by neighbouring pixels from one coding block, and inter-pairs, represented by pixels from neighbouring blocks. A ratio between the total value of intra- and inter-difference is considered as the blockiness level.

Block flickering. The flickering metric described here was inspired by the work presented by Pandel in [13]. The implementation task was threefold. The first aim was to define the threshold used to decide whether a given macro-block remains in state of no-update. In [13] the threshold was defined as the mean squared difference between the pixels of the current and corresponding macro-blocks, although the exact value was not revealed. The authors calculated the threshold as an average of absolute differences in pixel luminance for each 16×16 macro-block. Second, a different method for spatial pooling was proposed by calculating the frame-level flickering measure as a mean value over a small number of macro-blocks with the highest values (number of transitions between states). Third, the two previous parameters were adjusted in order to optimize prediction performance defined as a correlation with subjective scores. Similarly to the blockiness metric, averaging over a time window is required; the window size was equal to the sequence length for the purposes of the experiment.
In order to maximize the correlation of the flickering metric F with MOS, the authors considered several threshold values (between 0.5% and 2% of luminance change) and several numbers of macro-blocks with the highest num-

ber of transitions between states (between 0.5% and 10% of macro-blocks). The highest correlation with MOS was achieved for the threshold equal to 1% and frame-level flickering averaging over 3% of the total number of macro-blocks.

2.2. Measurement of Video Content Characteristics

For the purpose of the subjective experiment the authors were interested in choosing a good representation of videos to be included in the sequence pool; this means video sequences which would obtain different MOSs for the same compression parameters. On the other hand, sequences which are similar in terms of scene complexity should be avoided because they would not provide any additional information to the experiment.

The key parameters describing any video sequence characteristics are spatial and temporal information, i.e., the number of details and the movement dynamics respectively. In order to make the selection task easier, the authors use a scene complexity measure [14]. Scene complexity is a function of both spatial and temporal information which provides information on how difficult a given video sequence is to encode. It should be noted that it is represented by a single value, therefore the task of sequence selection becomes significantly simpler than for selection based on spatial and temporal information. It is easier to decide which scenes are close to each other and which are not.

The question how to measure all these content characteristics remains open. This paper utilizes a method presented in [14] where *scene complexity* o is defined as

$$o = \log_{10}\left(\underset{n}{\text{mean}}\left[SA(n)\cdot TA(n)\right]\right), \tag{2}$$

where $SA(n)$ is spatial activity computed for the nth frame and given by

$$SA(n) = \underset{space}{\text{rms}}\left[\text{Sobel}\left(F(n)\right)\right] \tag{3}$$

and $TA(n)$ is temporal activity computed on the base of nth and $n-1$th frames given by

$$TA(n) = \underset{space}{\text{rms}}\left[F(n) - F(n-1)\right]. \tag{4}$$

In both Eqs. (3) and (4) $F(n)$ denotes the nth video frame luminance channel. Sobel is the Sobel filter [14] and rms_{space} is the root mean square function over an entire video frame.

3. Verification of Measurement by Subjective Psycho–Physical Experiments

In order to properly model the image quality parameters to the assessment of subjects, an appropriate environment for the psycho-physical experiment was created. The experiments were performed at the AGH University of Science and Technology in Kraków. They were attended by over 100 students. Very similar conditions (LCD monitors and lighting) were provided for all test positions (see Fig. 1), and the experiments themselves followed the Video Quality Experts Group (VQEG) methodologies [15] wherever possible.

Fig. 1. Psycho–physical experiment environment.

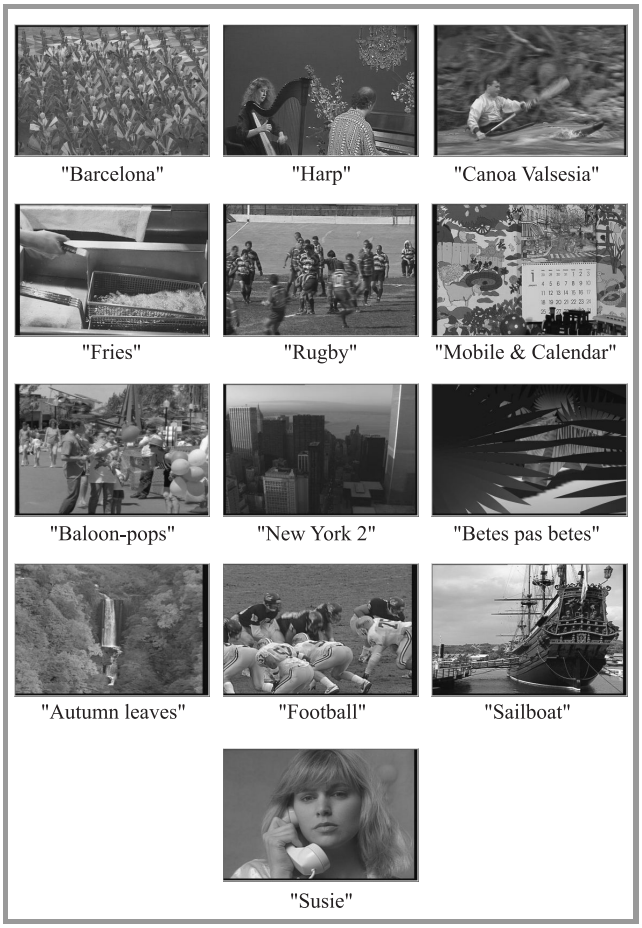


Fig. 2. Thirteen VQEG test sequences.

The experiment used thirteen VQEG test sequences (see Fig. 2) [15], [16], [17]: "Barcelona" (#2), "Harp" (#3), "Canoa Valsesia" (#5), "Fries" (#7), "Rugby" (#9), "Mobile

& Calendar" (#10), "Baloon–pops" (#13), "New York 2" (#14), "Betes pas betes" (#16), "Autumn leaves" (#18), "Football" (#19), "Saitboat" (#20) and "Susie" (#21). These sequences reflect the broad spectrum of two different characteristics of the content (temporal video activity and spatial video activity). Video sequences were encoded using the H.264 codec (X264 implementation), main-profile (Level 40). In accordance with the VQEG recommendations, QP was selected to obtain the order of the average bit rate streams of 5000 kbit/s (Compression Ratio, $CR = 50.38848$), 1000 kbit/s ($CR = 251.9424$), 500 kbit/s ($CR = 503.8848$), 300 kbit/s ($CR = 839.808$), 200 kbit/s ($CR = 1259.712$) and 100 kbit/s ($CR = 2519.424$).

The initial FPS rate was 30 with FPS rates values of 15, 10, 7.5, 6 and 5 also examined.

The effective initial resolution was the SD/D-1 NTSC resolution (720×486). Additionally the HHR 525 (352×480), SIF (352×240), QCIF (176×144) and SQCIF (128×96) resolutions were examined.

The authors used the ACR methodology, as described in ITU-T Recommendation P.910 [18]. This methodology represents the single-stimulus (SS) approach, which means that all video sequences in the test set are presented one after another without the option of comparison with the reference. Reference sequences are included in the test set and evaluated on the same basis as the others. This approach is known as ACR with hidden reference (ACR-HR). An eleven-step, numerical quality scale was used [18].

4. Statistical Analysis of Results of the Evaluation – Implementation of Prototype

This section contains a description of the methodology used to build models which are the components of the prototype used to evaluate the perceived QoE of streaming video sequences. The prototype includes the following components:

- single metrics to evaluate the quality of the source material,
- metric scaling in the time domain,
- metric scaling in the space domain,
- integrated metrics for the evaluation of H.264 compression (scaling in the domain of compression).

Descriptions of individual quality metrics and sequence characteristics of the video sequences are presented in Section 2. For scaling in the time and space domains, the values are taken directly from the sequence parameters, and the FPS number and resolution do not require specific metric algorithms.

The prototype was implemented in MATLAB, using standard libraries for processing images and video sequences. The system is able to analyze video sequences stored in files on a local disk. The parameters passed when the starting script is called allow any individual metrics, integrated metrics, and metrics that will be counted for the analyzed video sequence selection. In addition, there is the option of script setting, which is able to analyze multiple sequences and record the obtained results into the database.

4.1. Methodology of Model Building

Ratings obtained for en eleven-point scale are a significantly better approximation of normal distribution than results obtained for a five-point scale. This is because the eleven-point scale has two extreme answers, which should not be popular choices (responses 10 and 0). This allows to obtain less skewed distribution than for a five-point scale answer distribution. It should be noted that a skewed distribution is distinctly different from the Gaussian distribution. Therefore in order to model the obtained results the authors assumed a Gaussian distribution of results, making it possible to use the classical regression model.

In addition, all sequences were divided into test and learning sets. All models alongside the presented coefficients were obtained for the learning sets. Only after the final acceptance of the model was it confronted with the test set. Such methodology guarantees correct checking of whether the resulting model has the ability to predict subjective quality, and generalizes the results obtained for the learning sequences.

4.2. Scaling in the Time Domain

Metric scaling in the time domain appears to be very simple, because the information on the number of frames displayed each second is known. However, the constructed metric is not able to correctly model the quality perceived by the user. The reason is the lack of information about the sequence content. The model presented here also takes into account another factor, which is the amount of image detail. In addition, statistical analysis showed that the natural logarithm of the FPS number is a better predictor than the FPS number itself.

For the entire collection of analyzed films the obtained R^2 coefficient is lower than that obtained for the test sequences. However, the coefficient $R^2 = 0.88$ is a very good result and testifies the accuracy of the resulting model.

4.3. Scaling in the Spatial Domain

As is the case for the time domain, scaling in the space domain is easy to spot because the resolution of the present sequence is precisely known. Similarly to the time domain scaling the information about the image resolution is found to be inadequate because the content of the presented sequence affects the quality change.

The resolution change model takes into account both the amount of detail (SA) and the dynamics of the sequence (TA). In addition, using the logarithm of the resolution provided better results than the resolution value itself. In this case, both the coefficients of R^2 obtained for the test sequences and all sequences are equal.

4.4. Scaling in the Compression Domain

Creating a quality model for scaling in the domain of compression was a significantly more difficult task. The first and most important reason is compression complexity. Each compression scheme has numerous different parameters that define the encoding method. Therefore there is no obvious parameter which affects the QoE. Nevertheless, the test sequences obtained by the authors had relatively high single R^2 values, ranging from 0.74 to 0.89.

4.5. Packet Loss

Packet losses have a significant impact on the quality perceived by users. It is obvious that for larger losses the obtained quality is worse, but it is not true that a particular packet loss level indicates a particular quality of the sequence.

Detailed analysis shows that it is important to identify the location of the losses, both in the GOP structure and within the frame. In order to take into account these relationships it will be necessary to build a model based on additional information. In further studies the authors aim to rely on two possible scenarios. The first is image analysis similar to that used in the construction of a metric scale model in the fields of time, space and compression. The second solution is far more accurate inspection and detection of packets which form part of the picture, and/or the lost GOP. Research is being carried out on this analyzer as part of AGH's activities in the Joint Effort Group (JEG) forum.

5. Conclusions and Further Research

The paper presents a system for assessing the QoE based on measurements of artefacts present in video sequences. The validity of objective metrics has been verified under subjective tests. Statistical analysis of results demonstrates relatively high correlation coefficients as far as a no reference scenario is concerned.

Experiments reveal that the validity of quality measures is influenced by video content. Future subjective experiments will focus on a diversity of video sequences in terms of their spatial (number of details) and temporal (motion level) activity.

Co-operation with the VQEG JEG project will provide an opportunity to enhance the derived metrics by packet loss in the near future.

The proposed metrics, which have been coded in the MATLAB environment, will be moved to optimized, fast C/C++ libraries. Preliminary tests of blocking and flickering artefacts confirm the acceleration of metric computation which is important for a real time deployment.

Acknowledgement

Experiments have been performed under a grant of Ministry of Higher Education "Next Generation Services and Networks – Multimedia Services" and a EU FP7 Collaborative Project 218086 INDECT ("Intelligent information system supporting observation, searching and detection for security of citizens in urban environment").

References

[1] P. Romaniak, "Towards realization of a framework for integrated video quality of experience assessment", in *Proc. 28th IEEE Int. Conf. Comp. Commun. Worksh. INFOCOM'09*, Piscataway, USA, 2009, pp. 417–418.

[2] H. Derbel, N. Agoulmine, and M. Salaun, "ANEMA: autonomic network management architecture to support self-configuration and self-optimization in IP networks", *Comp. Netw.*, vol. 53, no. 3, pp. 418–430, 2009.

[3] L. Janowski, M. Leszczuk, Z. Papir, and P. Romaniak, "Ocena jakości sekwencji wizyjnych dla aplikacji strumieniowania na żywo w środowisku mobilnym", *Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne*, vol. 82, no. 8–9, pp. 800–804, 2009 (in Polish).

[4] A. Leontaris and A. R. Reibman, "Comparison of blocking and blurring metrics for video compression", in *Proc. IEEE Int. Conf. Acoust. Speech, Sig. Process. ICASSP 2005*, Philadelphia, USA, 2005, vol. 2, pp. 585–588.

[5] S. Tourancheau, P. Le Callet, and D. Barba, "Impact of the resolution on the difference of perceptual video quality between CRT and LCD", in *Proc. IEEE Int. Conf. Image Process. ICIP 2007*, San Antonio, USA, 2007, vol. 3, pp. 441–444.

[6] M. Ries, O. Nemethova, and M. Rupp, "Performance evaluation of mobile video quality estimators", in *Proc. 15th Eur. Signal. Proces. Conf. EUSIPCO*, Poznań, Poland, 2007.

[7] H. Knoche, J. D. McCarthy, and M. A. Sasse, "Can small be beautiful?: assessing image resolution requirements for mobile tv", in *Proc. 13th Ann. ACM Int. Conf. Multim. MULTIMEDIA '05*, New York, USA, 2005, pp. 829–838.

[8] P. Romaniak, M. Mu, A. Mauthe, S. D'Antonio, and M. Leszczuk, "A framework for integrated video quality assessment", *18th ITC Specialist Seminar on Quality of Experience*, May 2008.

[9] P. Romaniak, L. Janowski, M. Leszczuk, and Z. Papir, "A no reference metric for the quality assessment of videos affected by exposure distortion", in *Proc. IEEE Int. Conf. Multime. and Expo*, Barcelona, Spain, 2011.

[10] M. C. Q. Farias and S. K. Mitra, "No-reference video quality metric based on artifact measurements", in *Proc. IEEE Int. Conf. Image Process. ICIP 2005*, Genoa, Italy, 2005.

[11] R. Dosselmann and X. Dong Yang, "A prototype no-reference video quality system", *Proc. 4th Canadian Conf. Comp. Robot Vision CRV 2007*, Montreal, Canada, 2007, pp. 411–417.

[12] J. S. Lee and K. Hoppel, "Noise modeling and estimation of remotely-sensed images", in *Proc. Int. Geosci.Remote Sensin*, Vancouver, Canada, 1989, vo. 2, pp. 1005–1008.

[13] J. Pandel, "Measuring of flickering artifacts in predictive coded video sequences", in *Proc. Ninth Int. Worksh. Image Analys. Multim. Interact. Serv. WIAMIS 2008*, Washington, IEEE Computer Society, pp. 231–234, 2008.

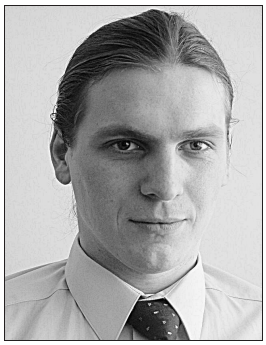
[14] C. Fenimore, J. Libert, and S. Wolf, "Perceptual effects of noise in digital video compression", in *140th SMPTE Techn. Conf.*, Pasadena, USA, 1998, pp. 28–31.

[15] VQEG. *The Video Quality Experts Group* [Online]. Available: http://www.vqeg.org/

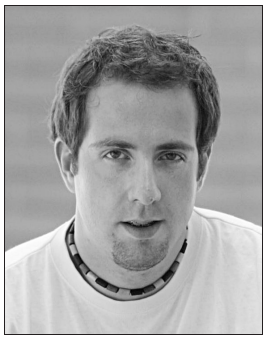
[16] VQEG. *Index VQEG Test Sequences*, 2008 [Online]. Available: http://media.xiph.org/vqeg/TestSeqences/ThumbNails/

[17] A. Webster, *Objective Perceptual Assessment of Video Quality: Full Reference Television*, 2004 [Online]. Available: http://www.itu.int/itu-t/, http://www.itu.int/ITU-T/studygroups/com09/docs/tutorial_opavc.pdf

[18] *Subjective Video Quality Assessment Methods for Multimedia Applications*, 1999, ITU-T.



Lucjan Janowski is an assistant professor at the Department of Telecommunications (AGH University of Science and Technology). He received his M.Sc. degree in Telecommunications in 2002 and Ph.D. degree in Telecommunications in 2006 both from the AGH University of Science and Technology. During 2007 he worked on a post-doc position in Laboratory for Analysis and Architecture of Systems of CNRS in France where prepared both malicious traffic analysis and anomaly detection algorithms. In 2010/2011 he spent half a year on a post-doc position in University of Geneva working on QoE for health applications. His main interests are statistics and probabilistic modeling of subjects and subjective rates used in QoE evaluation. He has participated in several commercial and scientific projects. He is an author of several research papers.
E-mail: janowski@kt.agh.edu.pl
Department of Telecommunication
AGH University of Science and Technology
Mickiewicza Av. 30
30-059 Kraków, Poland



Mikołaj Leszczuk is an assistant professor at the Department of Telecommunications, AGH University of Science and Technology (AGH-UST), Kraków, Poland). He received his M.Sc. in Electronics and Telecommunications in 2000 and Ph.D. degree in Telecommunications in 2006, both from AGH-UST. He is currently teaching Digital Video Libraries, Information Technology and Basics of Telecommunications. In 2000 he visited Universidad Carlos III de Madrid (Madrid, Spain) for a scientific scholarship. During 1997–1999 he was employed by several comarch holding companies as a Manager of Research and Development Department, President of the Management and Manager of the Multimedia Technologies Department. He has participated actively as a steering committee member or researcher in several national and European projects, including: INDECT, BRONCHOVID, GAMA, e-Health ERA, PRO-ACCESS, Krakow Centre of Telemedicine, CONTENT, E-NEXT, OASIS Archive, and BTI. He is a member of the Video Quality Experts Group Board and a co-chair of Quality Assessment for Recognition Tasks Group. His current activities are focused on e-health, multimedia for security purposes, P2P, image/video processing (for general purposes as well as for medicine), the development of digital video libraries, particularly video summarization, indexing, compression and streaming subsystems. He has been a chairman of several conference sessions. He is a member of IEEE Society since 2000. He has served as an expert for the European Framework Programme and the Polish State Foresight Programme as well as a reviewer for several conferences publications and journals.
E-mail: leszczuk@agh.edu.pl
Department of Telecommunication
AGH University of Science and Technology
Mickiewicza Av. 30
30-059 Kraków, Poland

Piotr Romaniak received his MS.c. Eng. degree in Telecommunications (2006), currently is a Ph.D. student in the Department of Telecommunications at the AGH University of Science and Technology (Kraków, Poland). His area of interest includes knowledge about multimedia systems, content-based indexing, perceptual no-reference video and image quality assessment (objective and subjective techniques), video streaming and video delivery technologies (IPTV, VoD), security and watermarking issues. He was/is involved in international projects funded by Culture 2000, eContentPlus, FP6 and FP7, as well as nation projects funded by Polish Ministry of Science and Higher Education. He was also involved in few research activities: for polish top telecommunication services provider (assessment of the 3rd generation fax image quality, subjective and objective evaluation methods) and in cooperation with other educational entities (panoramic images concatenating, subjective tests – MOS). Currently he is working on 3DTV quality of experience assessment and 2D content quality optimization and assessment in surveillance systems (FP7 IP INDECT Project).
E-mail: romaniak@kt.agh.edu.pl
Department of Telecommunication
AGH University of Science and Technology
Mickiewicza Av. 30
30-059 Kraków, Poland

Zdzisław Papir is professor at Department of Telecommunications (AGH University of Science and Technology in Kraków, Poland). He received the M.Sc. degree in Electrical Engineering in 1976 and Ph.D. degree in Computer Networks both from the AGH University of Science and Technology. In 1992 he received the Dr Hab. degree from the Technical University of Gdańsk. He is currently lecturing on Signal Theory, Modulation and Detection Theory, and Modeling of Telecommunication Networks. During 1991–98 he made several visits at universities in Belgium, Germany, Italy, and US working on statistical modeling of telecommunication traffic. During 1994–95 he was serving for the Polish Cable Television as a Network Design Department Manager. Since 1995 he serves also as a consultant in the area of broadband access networks for the Polish telecom operators. He authored five books and about 60 research papers. The book *Telecommunication Traffic and Packet Network Congestion* was awarded by the Ministry of Science and Higher Education. He was involved in organization of several international conferences at home and abroad. Between 1999–2006 he was a guest editor for IEEE Communications Magazine responsible for the Broadband Access Series. He is a member of an editorial board of a bookseries *Global Information Society* (in Polish). He has been participating in several R&D IST European projects being personally responsible for statistical modeling of telecommunication traffic and subjects/users behavior for quality assessment of multimedia communication services.

E-mail: papir@kt.agh.edu.pl
Department of Telecommunication
AGH University of Science and Technology
Mickiewicza Av. 30
30-059 Kraków, Poland

Paper

Communication Platform for Evaluation of Transmitted Speech Quality

Andrzej Ciarkowski and Andrzej Czyżewski

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

Abstract—A voice communication system designed and implemented is described. The purpose of the presented platform was to enable a series of experiments related to the quality assessment of algorithms used in the coding and transmitting of speech. The system is equipped with tools for recording signals at each stage of processing, making it possible to subject them to subjective assessments by listening tests or, objective evaluation employing PESQ or PSQM algorithms. The functionality for the simulation of distortions typical for voice communication over the Internet was implemented, making it possible to obtain reproducible, quantifiable results. An application of the presented platform for evaluation of acoustic echo canceler algorithm based on watermarking techniques, which was developed earlier is presented as an example of an effective deployment of the described technology.

Keywords—acoustic echo cancelation, doubletalk detection, echo-hiding.

1. Introduction

Development process of new algorithms for coding and improving the quality of transmitted speech entails the need to submit the results to assessments, making possible for the author of the algorithm to observe the introduced changes effect on the speech signal quality. An essential element of the assessment procedure is reproducibility of the results, which is often not feasible when working with active, "live" communication system. An obstacle in obtaining reproducible results is typically the element of randomness associated with a variable system load, choice of routes of communication and many other factors whose impact can be minimized or neglected by creating a separate, isolated platform for research and evaluation purposes only. Another need is the ability to simulate certain phenomena that typically occur randomly in communication systems, also assuming certain quantitative or qualitative parameters. This paper summarizes the design and implementation of such a system, which was conducted by the authors.

The implemented platform was practically utilized during the evaluation of the novel acoustic echo canceler (AEC) algorithm based on semi-fragile watermarking techniques, which was introduced by the authors [1], [2]. Obtained results are presented in the final part of this paper as a proof of usability of the developed system.

2. Platform Description

The developed communication platform is based on the use of typical elements, common in Internet telephony (VoIP) implementations, but extends them with some additional tools for collecting measurement data, including recording signals at the various intermediate stages of processing. An important aspect of the developed system is the automation of the results collecting, which is possible by using non-interactive execution mode. This allows to create scripts easily which enable obtaining a series of results depending on the specific parameter values. On the other hand, the interactive mode, allowing the user to manipulate UI elements facilitates the single passage, quick measurements as well as checking the behavior of algorithms while certain parameter changes over time. For this reason, the described system consists of two applications based on common software libraries, but differing with regard to the method of interaction.

Both applications are in fact the VoIP clients (terminals) communicating via standard RTP protocol and incorporating a complete communication stack thereof. The used implementation was conceived entirely at the Gdańsk University of Technology (GUT) and constitutes a fully functional realization of the RTP specification [3], together with a number of additional extensions and profiles. The entire source code associated with the implementation of the system was prepared in C++ programming language, with the support of numerous open source libraries for enhanced portability. Thus, although the primary work environment of the authors is Microsoft Windows OS, the developed libraries and non-interactive applications can be easily run on other operating systems, including Linux and MacOS X.

Figure 1 shows the architectural elements of the system. Four main groups of blocks can be identified, corresponding to consecutive stages of speech signal processing in the path from I/O interface to the transport layer. The input/output stage lists the three possible types of signal path "endpoints". In the case of the online analysis it is possible to use a real audio interface (sound card); the user of the system provides a test signal through the microphone and it gets possible to rehearse the result immediately. The signal can also be read from an audio file, what allows for a series of experiments using the same signal and differ-

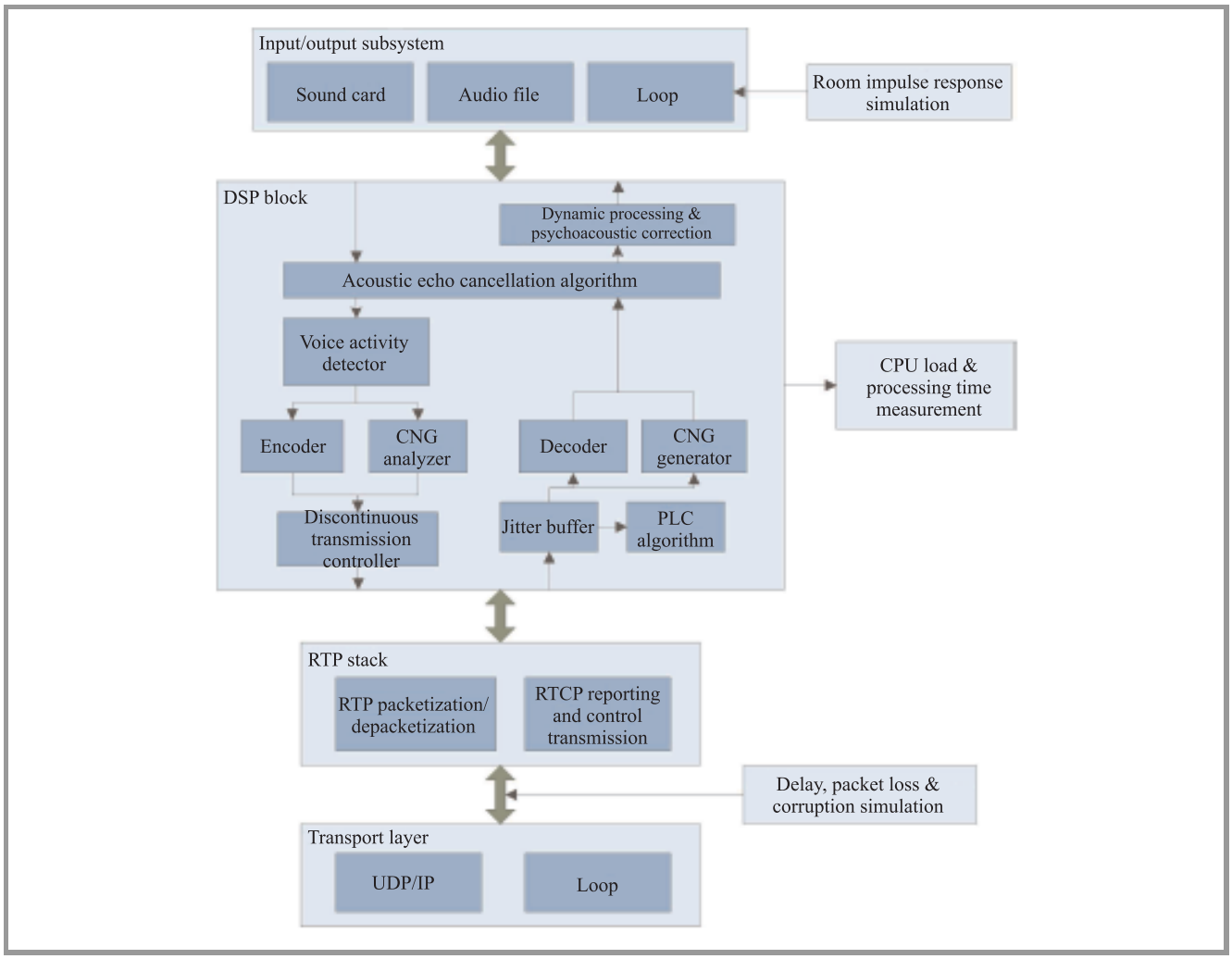


Fig. 1. Elements of communication platform for monitoring and evaluating voice transmission quality.

ent parameters of algorithms. Finally, it is also possible to use the local loop, optionally equipped with a filter that implements a specified transient response, which is particularly applicable in the case of testing algorithms for echo cancelation, because it allows simulating echoes with some specific properties using the remote terminal. It should be noted that the choice of sound file or local loop enables the analysis to be performed in the offline mode (for non-interactive applications), because it is not necessary in this situation to synchronize to periodically arriving data packets. This feature allows shortening the analysis time considerably, providing a valuable feature in case of large data series processing.

The next group of blocks makes the most important part of the system, namely the digital signal processing path. It is important to emphasize that the various elements of this group are in fact the algorithms which are subject to evaluation within the system, so that a special attention was paid to designing interfaces in such a way that blocks performing the same function using different algorithms are easily replaceable. The signal processing path is asymmetrical, with the flow oriented "towards the network" and "from the network" being different. The flow "towards the network" begins with the acoustic echo cancelation (AEC) algorithm block, which in this section shall record the signal coming from the near-end-speaker and is supposed to remove the estimated echo signal. At this stage, the system allows detailed analysis of the results of operations and performance of the following AEC algorithms:

- algorithm based on Geigel double-talk detector (DTD) and NLMS adaptive filter [4], [5];
- algorithm available in the open source speex voice codec library [6];
- algorithm based on semi-fragile watermarking DTD and NLMS adaptive filter, developed by the authors [1], [2];
- algorithm based on normalized cross-correlation DTD (NCC) and NLMS adaptive filter [7].

Another element in the processing path is the voice activity detector (VAD), which is used in case the employed speech codec lacks this feature. Its task is to determine whether the currently processed block of data has the characteristics

of voice activity, therefore, whether it is desirable to switch the system to the comfort noise encoding mode (CN). This technique is commonly employed in VoIP systems for bandwidth savings, especially in connection with so-called discontinuous transmission (DTX) mode, involving suppression of the transmission of packets representing the noise with characteristics similar to the memorized state.

The packet classified as active voice is passed to the speech encoder. The choice of audio coding algorithm is determined by the parameters of the application; in the case of interactive applications the codec can be changed during the session. At the present the system supports the following voice coding algorithms:

- PCM with a resolution of 8 and 16 bits/sample,
- ITU-T G.711 A-law and μ-Law [8],
- ITU-T G.722 [9],
- ITU-T G.723.1 [10],
- ITU-T G.726 (in versions 16, 32, 40 and 24 kbit/s) [11],
- ITU-T G.729 (with annexes B, D and E) [12];
- IMA ADPCM (DVI4) [13],
- ETSI GSM 06.10 [14],
- Speex [6],
- Internet low bitrate codec (iLBC, RFC3951) [15].

Figure 2 shows an example screen from the interactive application containing a list of codecs available in the current audio path configuration.

The last block of "towards the network" flow is aforementioned discontinuous transmission controller; whose function is to suppress transmission of the encoded packet in response to a signal from the VAD algorithm, or the sole codec, provided it supports that feature (e.g., G.729B, Speex).

The signal processing path in the direction "from the network" begins with the anti-jitter buffer algorithm. At this stage adaptive-length jitter-buffer implemented in the SpeexDsp library may be used interchangeably with the generic algorithm developed at the GUT. It cooperates with the packet loss concealment (PLC) algorithm, whose purpose is to recover (or to interpolate) packets which were lost during the transmission. The system provides a functionality for simulation of packet loss at a preconfigured ratio, which allows for the evaluation of the quality of speech transmission in a lossy environment. The simulator accepts the probability of packet loss as an input parameter, and the reproducibility of the loss pattern is achieved through the use of a dedicated MLS pseudo-random number generator with user-supplied seed. At this stage, the PLC algorithm described by the ITU-T G.711 Appendix I recommendation has been implemented, as well as several algorithms built-into some speech codecs [16].

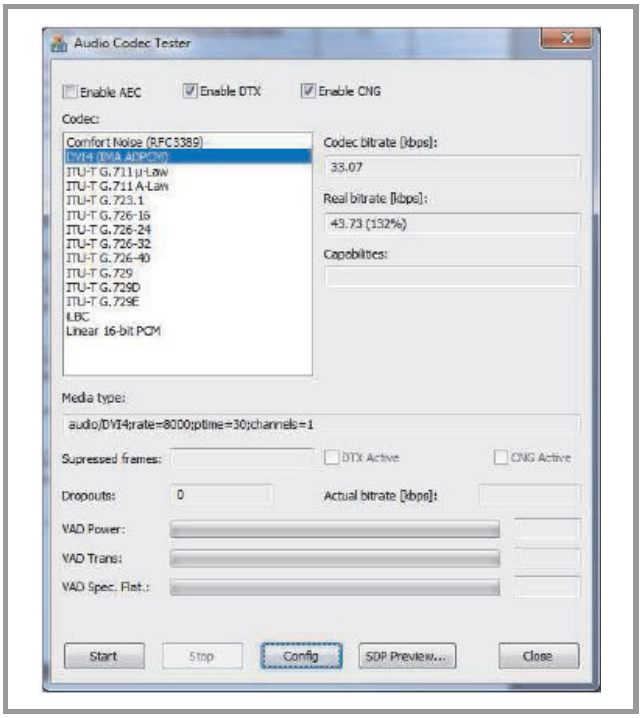


Fig. 2. Application interface allowing interactive selection of audio coding algorithm during the session.

Packets leaving the jitter-buffer are subjected to decoding and the decoded speech signal is delivered to the AEC algorithm, which interprets it as a model of the signal coming from the far-end speaker. Before returning to the output device, the postprocessing is performed, which includes the correction of the dynamics provided by a programmable noise gate, expander, compressor, and limiter blocks.

In the subsequent step the frames of speech signal are fed into the encoder, which produces payload according to the profile corresponding to the applied RTP audio codec. The encoded payload is passed to the RTP stack in order to append the RTP header. This process is called packetization, and the analogous operation "isolating" the payload of the RTP packet received from the transport layer is called depacketization. During depacketization the data obtained from the transport layer is reviewed for accuracy and continuity of the timestamp and sequence number, which allows the detection of loss of the packet, its repetition or change the order. The additional role of RTP stack package is to identify the sender, discarding packets received outside the current session, whose presence may indicate an attack attempt. Also certain statistical measurements are carried out, such as estimation of round-trip delays, delay fluctuation (jitter), packet loss ratio. These data are collected for the control of communication within the RTP session, which is carried out using the RTCP protocol.

RTP transport layer is typically based on sending UDP datagrams over an IP network. In many cases, however, the desired behavior is to work in a "loop", then sent datagrams are transmitted immediately back to the RTP stack. The developed system supports both of these modes. In

UDP/IP mode the system terminal may communicate not only with identical terminal, but also with any RTP client equipped with compatible codecs. This allows the system to use the endpoint for the analysis of data obtained from external applications, such as the popular streaming server VLC. The use of the loop mode allows for a convenient evaluation of algorithms, whose behavior is not dependent on using a distributed configuration. An important complement to the system are the aforementioned packet loss simulator and a "delay line" generating a programmed delay of the packet arrival, useful in a research of acoustic echo cancelation and buffering. The purpose of this delay block is the simulation of round-trip delays introduced by the network, which do not occur in the "loopback". These delays, which can range from single milliseconds up to seconds, are typically responsible for the perceived annoyance of the acoustic echo affecting quality of VoIP calls.

3. Application to Acoustic Echo Cancellation Evaluation

3.1. Robustness Analysis of AEC Algorithms under Time-Variable Echo Conditions

A standard, widely-accepted in the literature method of objective testing, based on the detection theory, was used during the study. This method is based on plotting the receiver operating characteristic (ROC) curves representing the probability of false alarm and miss as a function of relative signal levels of the near- and far-end speakers (NFR, near-to-far ratio). Its detailed description can be found in the literature [17]. During the research a modification of the method was proposed, whose purpose was the simulation of time-varying echo conditions (time-variable echo delay, changing room impulse response).

The evaluation was based on 5 speech excerpts in Polish. 4 recordings of length 1s (2 women, 2 men) represented the near-end speakers, and the recording of the length of 5 seconds (male voice) served as a far-end speaker signal. Fragments were sampled at the frequency of 8 kHz, consistent with common telephony applications. During the evaluation a constant echo delay of 40 ms was applied, with variable component added according to the characteristic plotted in Fig. 3.

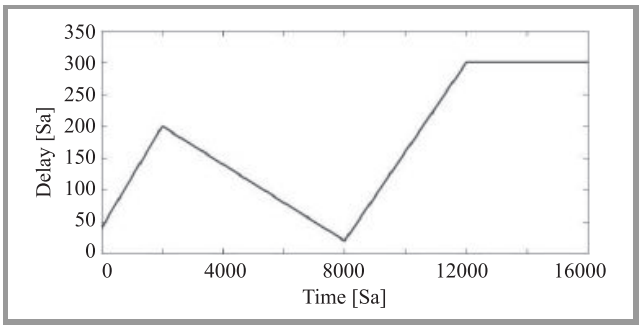


Fig. 3. Characteristics of variation of echo delay time.

The simulation of variable room impulse response involved a weighted sum of 2 impulse responses which were acquired in different locations at the same room. The variation was a simple linear transition from $h_1(n)$ to $h_2(n)$ over the time span of 4 s. The impulse responses are presented in Fig. 4.

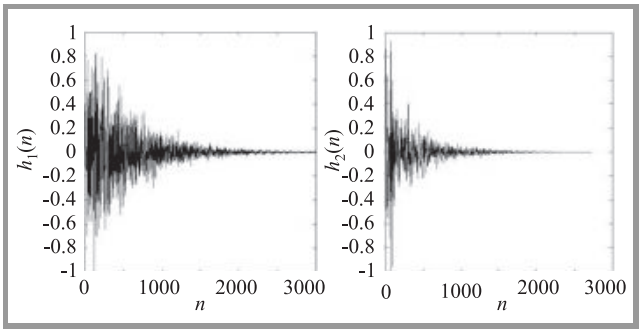


Fig. 4. Acoustic path impulse responses $h_1(n)$ and $h_2(n)$ for the research.

For maintaining consistency with the results presented in the literature the evaluation was conducted for the probability of false alarm $P_f \in \{0.1, 0.3\}$. Robustness of AEC methods based on different DTD algorithms was evaluated through the execution of the same test, first with the constant echo time and the impulse response, and then, with variable ones. Increased probability of DTD algorithm miss (i.e., not detecting the doubletalk when it is present) in these conditions determines the susceptibility of the DTD algorithm to the variability the echo path characteristics. The results obtained are presented in Fig. 5 and in Fig. 6.

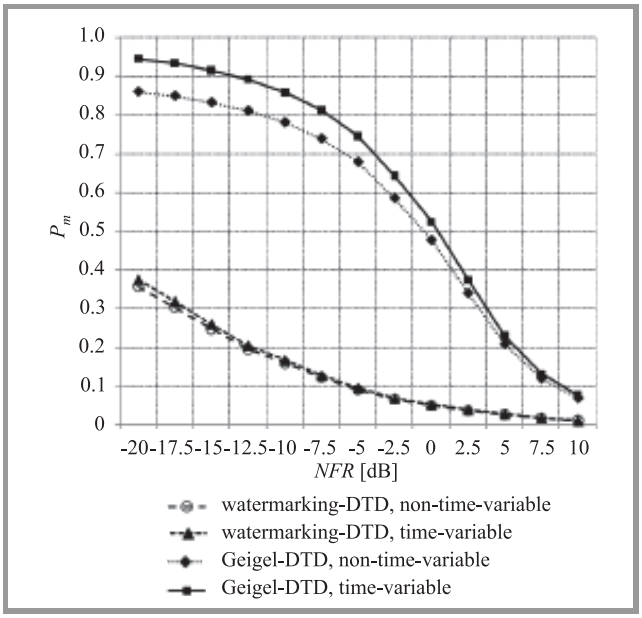


Fig. 5. Probability of DTD algorithm miss for the probability of false alarm $P_f = 0.1$ while running in the variable and "stable"conditions.

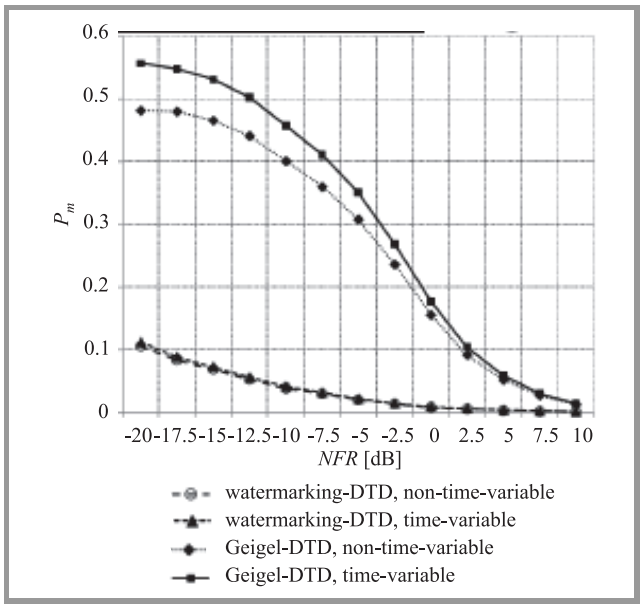


Fig. 6. Probability of DTD algorithm miss for the probability of false alarm $P_f = 0.3$ while running in the variable and "stable"conditions.

Both evaluated algorithms demonstrated a performance deterioration in the "variable" conditions, however, the scale of this phenomenon is different. For comparison, a relative deterioration measure (RDM) was derived which determines how the algorithm performs in "variable"conditions relating to the "stable" ones.

$$\mathrm{RDM} = \frac{P_{m,variable}}{P_{m,stable}} \tag{1}$$

This coefficient values were plotted in Fig. 7.

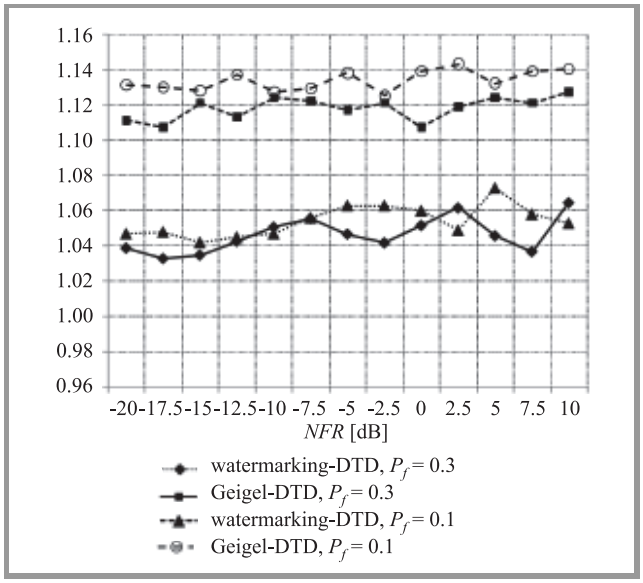


Fig. 7. The relative increase in the DTD algorithm probability of miss while working under time-variable echo delay and changing room impulse response.

3.2. Objective and Subjective Evaluation of Watermarking-Based DTD Algorithm in Relation to NCC Algorithm

The implementation of the normalized cross-correlation DTD algorithm made it possible to conduct a comprehensive evaluation of DTD algorithm developed by the authors against the algorithm representing current state of the art. In the first phase of evaluation, the objective tests were carried out in accordance to the methodology proposed in the literature [17]. The test set used was the same as for the previously presented robustness analysis. Consequently, the listening tests were conducted to investigate as to how DTD misses made by the various algorithms affect the subjective opinion on speech quality.

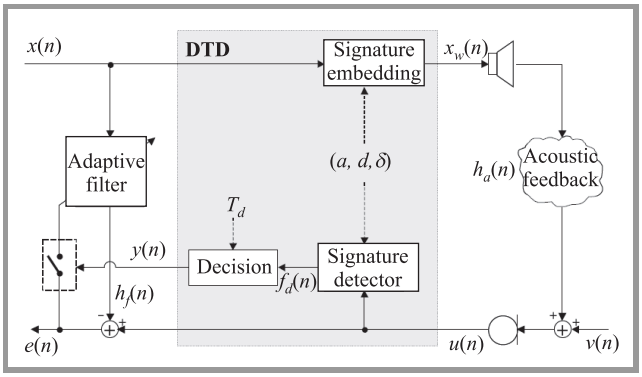


Fig. 8. Acoustic echo cancelation system employing watermarking-based DTD algorithm and adaptive filter.

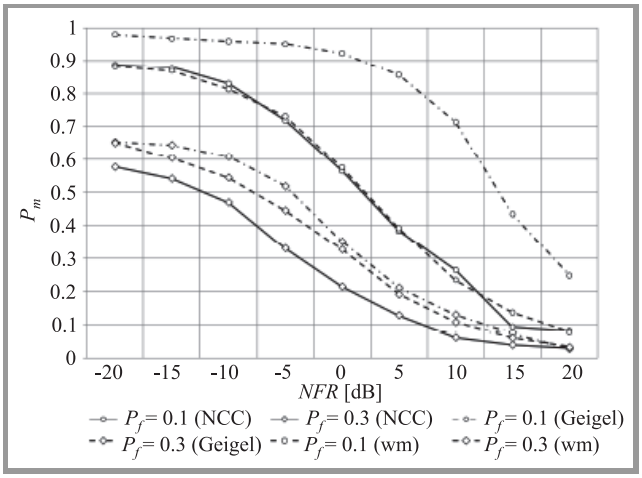


Fig. 9. Probability of DTD algorithm miss in the presence of background noise of –30 dB.

Both DTD algorithms were combined with the NLMS adaptive filter of length $L = 512$ to create a working AEC system during the tests. The example setup of such system with watermarking-based DTD algorithm is depicted in Fig. 8 and for the NCC algorithm only the grayed box labeled DTD is different. The length of the window used by the NCC algorithm to estimate the correlation coefficients between $x(n)$ and $u(n)$ was $W = 500$.

Both of these values were chosen in consistency with the research published in the literature by the authors of the NCC algorithm [7], [18], [19]. The results of objective tests carried out in the first phase are presented in Fig. 9 and Fig. 10.

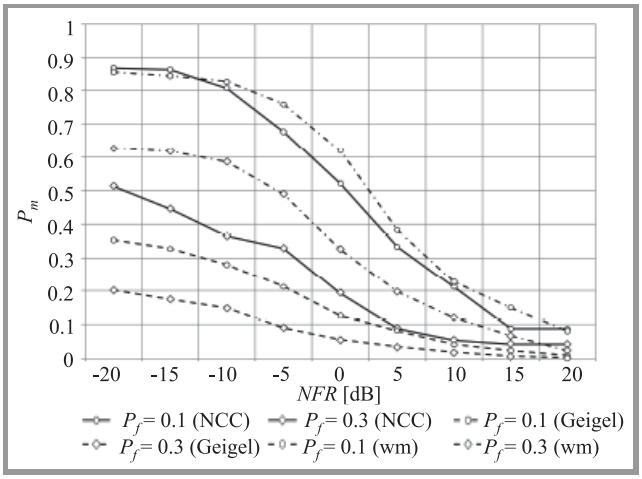


Fig. 10. Probability of DTD algorithm miss in the presence of background noise of –60 dB.

The presented plots were obtained at different levels of background noise in the microphone signal. This allowed assessing the vulnerability of specific algorithms for this type of disturbance. The resulting graphs for the NCC algorithm significantly differ from those published by its authors. This discrepancy may stem from the fact that in the literature description [17] the algorithm has been combined with a real adaptive filter, but the studies were based on the use of the actual impulse response, which was previously used to simulate the echo signal. Therefore, experiments carried out using the developed system are able to model the actual phone call conditions in a more realistic way.

Table 1
MOS values for DTD algorithms; background noise level -30 dB, NFR $= 0$, $P_f = 0.1$

Test signal	MOS
Reference signal (near speaker)	4.83
Reference signal (microphone signal)	1.25
AEC algorithm w/ DTD NCC	3.92
AEC algorithm w/ Geigel DTD	2.58
AEC algorithm w/ watermarking DTD	3.75

Table 2
MOS values for DTD algorithms; background noise level -60 dB, NFR $= 0$, $P_f = 0.1$

Test signal	MOS
Reference signal (near speaker)	4.92
Reference signal (microphone signal)	1.25
AEC algorithm w/ DTD NCC	3.75
AEC algorithm w/ Geigel DTD	3.08
AEC algorithm w/ watermarking DTD	4.33

Table 3
MOS values for DTD algorithms; background noise level -30 dB, NFR $= -15dB$, $P_f = 0.1$

Test signal	MOS
Reference signal (near speaker)	4.75
Reference signal (microphone signal)	1.16
AEC algorithm w/ DTD NCC	2.0
AEC algorithm w/ Geigel DTD	1.25
AEC algorithm w/ watermarking DTD	1.84

Table 4
MOS values for DTD algorithms; background noise level -60 dB, NFR $= -15dB$, $P_f = 0.1$

Test signal	MOS
Reference signal (near speaker)	4.83
Reference signal (microphone signal)	1.16
AEC algorithm w/ DTD NCC	1.84
AEC algorithm w/ Geigel DTD	1.84
AEC algorithm w/ watermarking DTD	3.75

Mean opinion score (MOS) values were obtained in effect of listening tests based on the sound files stored during the objective tests phase, therefore both tests were performed using identical test signals. MOS values presented in Tables 1–4 are the mean of the ratings issued by 12 experts (Ph.D. students and employees of the GUT, Multimedia Systems Department).

4. Summary

The developed system provides a useful tool for comprehensive analysis of various aspects of the voice coding, transmission and quality enhancement algorithms. It has been designed and implemented during the research work, which sought to develop new algorithms for coding and improving the quality of speech transmitted over the Internet. Currently available functionality of the system provides a significant facilitation of the research process, what was practically demonstrated by the results of the evaluation of watermarking-based DTD algorithm proposed and developed by the authors.

Acknowledgement

The research was funded by the Polish Ministry of Science and Higher Education within the grant no. PBZ-MNiSW-02/II/2007.

References

[1] G. Szwoch, A. Czyżewski and A. Ciarkowski, "A double-talk detector using watermarking", *J. Audio Eng. Soc.*, vol. 57, pp. 916–926, 2009.

[2] A. Czyżewski and G. Szwoch, "Method and Apparatus for Acoustic Echo Cancellation in VoIP Terminal". International Patent Application No. PCT/PL2008/000048, 2008.

[3] H. Schulzrinne, et al, "RTP: A Transport Protocol for Real-Time Applications", RFC 3550, IETF, 2003.

[4] D. L. Duttweiler, "A twelve-channel digital echo canceler", *IEEE Trans. Commun.*, vol. 26, pp. 647–653, 1978.

[5] S. M. Kuo, B. H. Lee and W. Tian, "Adaptive echo cancelation", in *Real-Time Digital Signal Processing: Implementations and Applications* (2nd ed), S. M. Kuo *et al.*, Eds. Chichester: Wiley, 2006, pp. 443–473.

[6] J.-M. Valin, "The Speex Codec Manual", May 09, 2011, http://speex.org/docs/manual/speex-manual/

[7] J. Benesty, D. R. Morgan, J. H. Cho, "A new class of doubletalk detectors based on cross-correlation", *IEEE Trans. Speech Audio Process.*, vol. 8, pp. 168–172, 2000.

[8] "Pulse code modulation (PCM) of voice frequencies". ITU-T Recommendation G.711 (11/88), Int. Telecomm. Union, Geneva, Switzerland, 1988.

[9] "7 kHz audio-coding within 64 kbit/s". ITU-T Recommendation G.722 (11/88), Int. Telecomm. Union, Geneva, Switzerland, 1988.

[10] "Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s". ITU-T Recommendation G.723.1 (05/06), Int. Telecc. Union, Geneva, Switzerland, 2006.

[11] "40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)". ITU-T Recommendation G.726 (12/90), Int. Telecomm. Union, Geneva, Switzerland, 1990.

[12] "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)". ITU-T Recommendation G.729 (01/07), Int. Telecomm. Union, Geneva, Switzerland, 2007.

[13] H. Schulzrinne and S. Casner, "RTP Profile for Audio and Video Conferences with Minimal Control", RFC 3551, IETF, 2003.

[14] "GSM Full Rate Speech Transcoding". Europ. Telecomm. Standards Inst. (ETSI), Sophia Antipolis, France, Recommendation GSM 06.10, 1992.

[15] S. Andersen *et al.*, "Internet Low Bit Rate Codec (iLBC)", RFC 3951, IETF, 2004.

[16] "A High Quality Low Complexity Algorithm for Packet Loss Concealment with G.711". ITU-T Recommendation G.711 Appendix I (09/99), Int. Telecomm. Union, Geneva, Switzerland, 1999.

[17] J. H. Cho, D. R. Morgan and J. Benesty, "An objective technique for evaluating doubletalk detectors in acoustic echo cancelers", *IEEE Trans. Speech Audio Process.*, vol. 7, pp. 718–724, 1999.

[18] S. L. Gay and J. Benesty, "An introduction to acoustic echo and noise control" in *Acoustic Signal Processing for Telecommunication*, S. L. Gay and J. Benesty, Eds. Norwell: Kluwer, 2000, pp. 1–18.

[19] J. Benesty *et al.*, *Advances in Network and Acoustic Echo Cancellatation*. Berlin: Springer, 2001.

[20] M. Baughner *et al.*, "The Secure Real-time Transport Protocol (SRTP)". RFC 3711, IETF, 2004.

Andrzej Ciarkowski was born in 1979 in Gdańsk. In 1998–2003 he studied at the Gdańsk University of Technology, where in 2003 he graduated at the Sound Engineering Department. His thesis was related to design of custom, high quality USB audio interface. Since that time he has been a member of the research staff at the Multimedia Systems Department as a Ph.D. student (2003–2008) and is currently working on Ph.D. thesis. Current subjects of his research includes multimedia communications over Internet Protocol, what is reflected by the subject of his Ph.D. thesis, relating to acoustic echo cancelation in VoIP systems.
E-mail: rabban@sound.eti.pg.gda.pl
Multimedia Systems Department
Gdańsk University of Technology
Narutowicza st 11/12
80-233 Gdańsk, Poland

Andrzej Czyżewski is the Head of the Multimedia Systems Department of the Gdańsk University of Technology at the Faculty of Electronics, Telecommunications and Informatics. He received his M.Sc. degree in Sound Engineering from the Gdańsk University of Technology in 1982, his Ph.D. degree in this domain in 1987 and his D.Sc. degree in 1992 from the Cracov Academy of Mining and Metallurgy. In December 1999 Mr. President of Poland granted him the title of Professor. In 2002 the Senate of the Gdańsk University of Technology approved him to the position of Full Professor. He is a member of the Acoustic Committee of the Polish Academy of Sciences, IEEE, International Rough Set Society and Fellow of the Audio Engineering Society. As a researcher, together with his team designed a number of software applications and digital devices; several of which were produced commercially in Poland. The subjects of the mentioned projects concern new methods of speech recognition, audio restoration, beamformers, anti-noise filters, speech communication system for military aircraft pilots, environmental noise monitoring system, advanced surveillance monitoring systems and others.
E-mail: andcz@sound.eti.pg.gda.pl
Multimedia Systems Department
Gdańsk University of Technology
Narutowicza st 11/12
80-233 Gdańsk, Poland

Paper

Video Streaming Framework

Andrzej Buchowicz and Grzegorz Galiński

Institute of Radioelectronics, Warsaw University of Technology, Warsaw, Poland

Abstract—The framework for testing video streaming techniques is presented in this paper. Short review of error resilience and concealments tools available for the H.264/AVC standard is given. The video streaming protocols and the H.264 payload format are also described. The experimental results obtained with the framework are presented in this paper too.

Keywords—error resilience and concealment, H.264/AVC, media delivery over IP networks, video coding.

1. Introduction

The video coding technology has been rapidly developing during the last years. Broadband networks have been growing even faster. Media delivery over the IP networks has been widely accepted and it seems it may replace the traditional media distribution methods in near future. However, the network performance depends on many factors and may not always guarantee the required quality of the transmitted media. It is extremely important in many application to increase the error resilience of the audio or video stream and to effectively conceal any errors that may occur during transmission.

The framework for testing video streaming techniques will be presented in this paper. It has been used as a tool for analysis and development of media adaptation, error resilience and error concealment algorithms. The framework has been limited to the streaming of the H.264/AVC encoded video with the use of RTP/RTCP protocol. However, it can be easily extended for other video codecs and transmission protocols. The experimental results obtained with this framework will also be presented.

1.1. H.264/AVC Bitstream

The international standard MPEG-4 H.264/AVC [1] is currently the most commonly used for video coding. Its first editions was released in 2003. The important enhancements: scalable video coding (SVC) and multiview video coding (MVC) were added in 2008 and 2009 respectively. The H.264/AVC standard is based on the hybrid motion compensation and transform algorithm [2]–[4] implemented in almost all preceding video coding standards, including MPEG-2 Video [5]. Many improvements of the classical algorithm significantly increased the H.264/AVC coding efficiency with respect to its predecessors. However, the coding efficiency was not the only objective for H.264/AVC standard developers. The video stream flexibility and its adaptability for different transmission channels was also an important factor. It has been achieved by separation of the signal processing from the transport-oriented processing – the H.264/AVC codec has been divided into two layers:

– video coding layer (VCL) – contains all compression tools, generates bitstream of the encoded macroblocks organized into slices;

– network abstraction layer (NAL) – encapsulates the bitstream generated by the VCL in units suitable for the transmission.

The H.264/AVC bitstream is a sequence of the NAL units (Fig. 1). Each NAL unit starts with one-byte header containing three fields:

– `F` – error indicator (1 bit), NAL unit with this field set to 1 should not be processed;

– `NRI` – NAL unit priority (2 bits), the value of this fields indicates the importance of the NAL unit for a video sequence reconstruction;

– `TYPE` – type of the NAL units (5 bits), values $0 \div 23$ are restricted to be used only within the H.264/AVC standard, values $24 \div 31$ may be used for other purposes, e.g., in transmission.

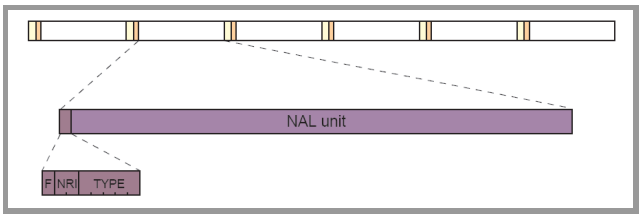


Fig. 1. H.264/AVC bitstream.

NAL units contain only data representing the encoded video sequence. Additional headers must be appended to each NAL unit to separate them. Annex B of the H.264/AVC standard defines such headers (start code – fixed byte sequence) for the transmission in byte-oriented networks (e.g., broadcasting). The headers used in packet-oriented networks will be discussed further in this section.

The H.264/AVC syntax is not as restricted as in its predecessors. There are no layers above the slice layer generated by the VCL. The higher level information are stored in the specific syntax elements: sequence parameter set (SPS) and picture parameter set (PPS). Special NAL unit type are assigned to carry the parameter sets. Several SPSs and PPSs can be defined and used by the encoder. Each macroblock in the H.264/AVC bitstream refers to the SPS and the PPS

which are used to encode it. Usually NAL units with parameter sets precede all other NAL units in the H.264/AVC bitstream, however, it is not required by the standard. They can be transmitted in an additional, more reliable channel, for example. The only requirement is that the parameter sets must be known to the decoder to allow the bitstream processing.
The scalable and multiview extensions of the H.264/AVC standard generally conform to the above concept. The SVC and MVC bitstreams are sequences of the NAL units similarly as the H.264/AVC bitstream. Special NAL unit types (illegal in the AVC bitstream) have been defined to carry additional data (video layers in SVC, views in MVC) introduced by these extensions.

1.2. Error Resilience Tools in H.264/AVC

Error resilience tools are available in many video coding standards. However they are limited to the frame segmentation into slices or group of blocks (GOB) in most cases. The H.264/AVC standard introduces new error resilience tools [3], [6]:

- redundant slices – additional (redundant) data are added to the normal (non-redundant) data representing the entire frame or a part of the frame;
- arbitrary slice order – the frame is divided into slices which are transmitted in non-raster (arbitrary) order;
- slice groups – macroblocks in a frame are allocated to a slice group. Six predefined allocation maps (Fig. 2) can be used, additionally explicit macroblock allocation mode is also available. This technique is also known as flexible macroblock ordering (FMO);
- bit stream partitioning – encoded slice is divided into three partitions containing respectively: slice and macroblock headers, residual data for intra coded macroblocks, and residual data for inter coded macroblocks.

These tools are available only in the Baseline or Extended profile of H.264/AVC standard. It limits their applications since most available codecs conform to the Main or High profile.
The implementations of the H.264/AVC error resilience tools are widely reported in the literature. Dynamic slice group mapping based on a macroblock classification algorithm for prioritized video transmission is presented in [7]. Combined flexible macroblock ordering (FMO) and redundant slices algorithm is presented in [8]. Multiple description coding based on redundant slices is discussed in [9]. The redundant picture coding combined with reference picture selection and reference picture list reordering method is presented in [10]. An interesting approach based on an optimal slicing and unequal error protection is proposed in [11]. Technique based on redundant pictures inserted periodically into encoded sequence is presented in [12].

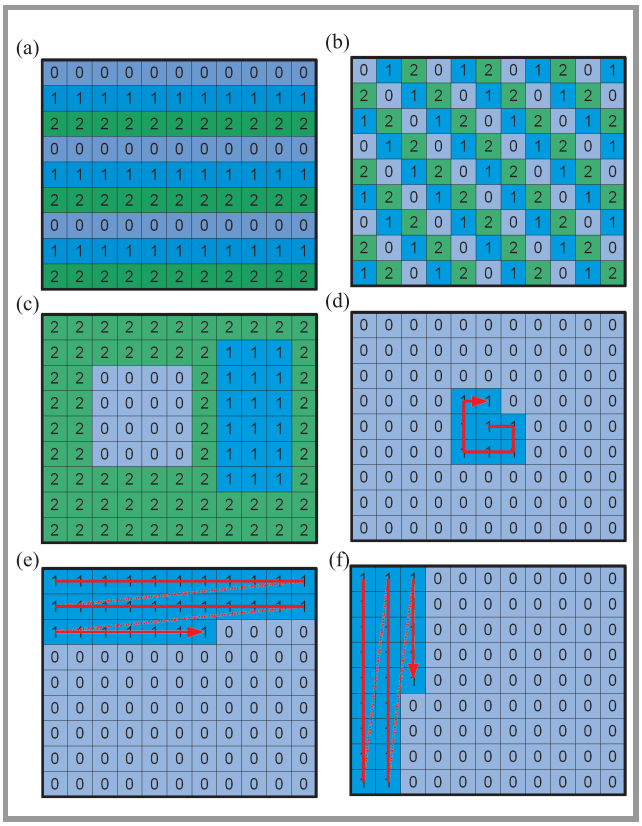


Fig. 2. Standard slice group allocation maps: (a) interleaved, (b) dispersed, (c) foreground and background, (d) box-out, e) raster scan, (f) wipe.

Error resilience tools are usually used jointly with the error concealment techniques which try to reconstruct the parts of the bitstream lost due to transmission errors. Usually perfect reconstruction is not possible. However, even if only approximation of the lost fragments can be found, the overall quality of the reconstructed sequence is improved. Two error concealment algorithms are implemented in the H.264/AVC reference software [13]: frame copy and motion vector copy [14], [15]. The first algorithm simply copies the pixel in the concealed frame from the previous decoded reference frame. The motion compensation with the motion vectors copied from the previous reference frame is used in the second algorithm. The algorithm recovering lost slices in video encoded with the FMO tool, based on the edge-directed error concealment, is presented in [16]. The FMO tool is also used in the algorithm presented in [17] to recover missing motion vectors. Many error concealment techniques utilize spatial and temporal correlation in the video sequences [18]–[20].

1.3. Video Streaming

There are generally two approaches to the media delivery over the IP networks. The first one is based on the transmission protocols utilizing TCP as a transport protocol. The file download with the use of HTTP protocol is the most obvious example. The other option is so called

HTTP progressive download. The file transmitted with the use of HTTP is split into many small fragments in this case. Each fragment is transmitted in a separate HTTP request, allowing media playback after receiving only the small part of the entire file. The most sophisticated HTTP-based solution is the adaptive progressive download [21]–[24]. The several variants of each fragment of the media file are used, each is encoded with different parameter sets, e.g., bit rate. All fragments are transmitted sequentially as in classical progressive download, however, it is possible to switch between variants at fragment boundaries. The variants are selected depending on the actual network throughput. This adaptation scheme provides uninterrupted media delivery with varying quality following the change of the network conditions. The advantage of the HTTP-based solutions is an ability to traverse firewalls so it is widely used in the Internet (e.g., YouTube). However application in real-time systems is limited due to delays introduced by the TCP transmission.

The other approach to the media delivery is based on the user datagram protocol (UDP) as a transport protocol. It is preferred in real-time applications, e.g., videoconferencing or video surveillance. There are usually very strict requirements on the transmission delay in such applications. These requirements can be fulfilled only if the UDP is used. However, since the UDP is an unreliable protocol, some datagrams may be lost, duplicated or may arrive to the destination in the wrong order. The real time protocol (RTP), accompanied by the RTP control protocol (RTCP) [25], [26], were developed to eliminate these drawbacks. The RTP provides data transport mechanism, while the RTCP is a tool for data transmission monitoring. Both protocols are most often used on the top of the UDP, however, it is possible to use them with other transport protocols too. It is worthwhile to mention that the secure enhancement of the RTP has been developed [27]. It defines the media encryption algorithm as well as media integrity and authenticity verification method.

The RTP is very universal and can be used for delivering media of different types. The RTP payload format for the delivery of H.264/AVC bitstream is presented in [6], [28]. It is based on the NAL units concept presented in the Subsection 1.1. Three encapsulation modes are specified:

– single NAL unit in the RTP packet,

– multiple NAL units in the RTP packet (aggregation mode),

– NAL unit split into multiple RTP packets (fragmentation mode).

The first mode is very simple: an entire NAL unit is inserted into the RTP packet as its payload (Fig. 3). The one-byte NAL unit header serves as the RTP payload header. The `NRI` and `TYPE` fields can be used to classify how important the payload is for the sequence reconstruction. The RTP header contains additional data describing the payload:

– `PT` – payload type identifier; certain media types have been assigned fixed identifiers [29]. The identifiers for other media types, including H.264/AVC, must be assigned dynamically, e.g., within the SDP [30] messages;

– `M` – marker bit set to 1 if the payload contains the last NAL unit in the current frame;

– `TS` – timestamp of the NAL unit carried as a payload; the clock frequency for the H.264/AVC video is equal to 90 kHz;

– `SN` – sequence number of the RTP packet; allows detection of the packet loss, duplication or incorrect order;

– `SSRC` – synchronization source identifier; each participant of the RTP session is identified by its unique identifier;

– `CSRC` – contributing source identifier; used only if mixers or translators [25] are used in the RTP session;

– `CC` – CSRC count; number of the CSRC fields in the RTP header; set to zero in most cases;

– `P` – padding flag; if set the last byte of the payload contains number of the padding bytes following the packet payload; used to increase the length of the RTP packet to the fixed value required, e.g., by the encryption algorithm.

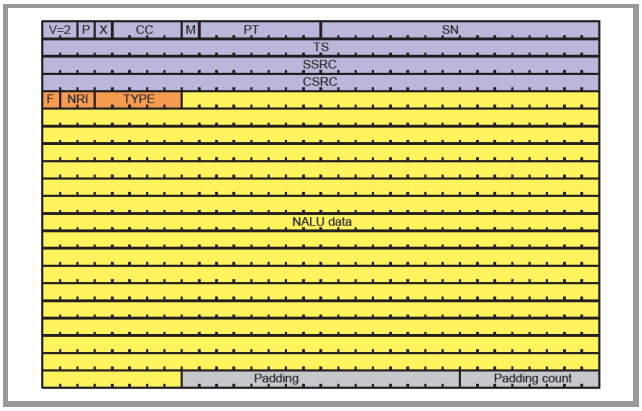


Fig. 3. Single NAL unit in the RTP packet.

Single NAL unit mode is effective if the NAL unit length fits to the network characteristic. The length of the RTP packet must not exceed the maximum length of the UDP datagram equal to 64 kB. It should also not exceed the length of the maximum transfer unit (MTU) for the given network (e.g., 1500 B for Ethernet). If the RTP packet is longer than MTU it will be fragmented by the lower layers in the IP stack. The packet fragmentation increases the probability of the packet loss.

The length of the encoded slice depends on many factors. It may easily exceed the limit of 64 kB, e.g., if the high resolution sequence is encoded with good quality (high bitrate) and no frame segmentation is used (i.e., the entire frame is encoded in one slice). In many cases it exceeds the MTU value too. The fragmentation mode provides the way to handle NAL units containing such long slices. The NAL unit is split into fragments transmitted in consecutive RTP packets. There is also an option to change the order of the NAL unit fragments. Each NAL unit fragment is appended by an additional field containing its order in this option.
The NAL units containing, e.g., parameter sets, SEI messages or encoded slices of fine fragmented frame can be very short. Their transmission in single RTP packet is ineffective due to the header overhead. The aggregation mode allows to join such short NAL units in one longer RTP packet. The aggregated RTP packet can contain either NAL units with identical timestamps or with different timestamps. Similarly, as in the fragmentation mode, NAL units do not have to be inserted into aggregated packet in its decoding order.
The payload format for the scalable extension (SVC) of the H.264/AVC is proposed in the draft specification [31]. Two modes of the SVC bitstream transmission are defined:

– single-session: all layers of the SVC stream are transmitted in a one RTP session. All packetization modes available for the H.264/AVC bitstream may be used in this mode;

– multi-session: layers of the SVC stream are transmitted in different RTP sessions. All sessions are synchronized to the same system clock. Four special packetization modes are defined for this transmission mode.

Multi-session mode is especially suitable for the multicast transmission. Separation of the SVC bitstream layers simplifies the stream adaptation to the network conditions. Specialized network devices, so called media aware network elements (MANE), can simply discard the layers which require higher throughput than is currently available.
The proposed payload format [32] for multiview extension (MVC) of the H.264/AVC is very similar to the specification for the SVC. The views contained in the MVC bitstream can be transmitted in either one RTP session (single-session mode) or in multiple synchronized RTP sessions (multi-session mode).

2. Video Streaming Framework Overview

The following requirements for the framework were specified:

– streaming of the H.264/AVC encoded video with the use of RTP/RTCP;

– monitoring and visualisation of the network parameters during transmission;

– acquisition and real-time encoding of the analogue video signal;

– decoding of the H.264/AVC stream and displaying of the reconstructed video in real time.

The open source software has been extensively used in the framework development. The framework is running under the Linux operating system. It has been written mostly in the C++ programming language, some code fragments directly interfacing with underlying libraries have been written in C. Video4Linux2 (V4L2) [33] application programming interface has been used for video capture. The graphical user interface has been created with the use of Qt library [34]. The classes from the Qwt library [35] have been used to create diagrams for network parameters visualisation. The open source library JRtpLib [36] has been used to sent and receive RTP packets and handling of the RTCP messages. The FFmpeg library [37] has been used for the H.264/AVC stream decoding. The H.264/AVC parsers have been based on the H.264 reference software [13]. The x264 [38] library has been used to encode video.
The most important classes of the framework are:

– `NalUnit` – represents the NAL unit and its timing information;

– `AnnexBReader` – parser for the H.264/AVC bitstream stored in the Annex B [1] format;

– `JmRtpReader` – parser for the H.264/AVC bitstream stored in the RTP format defined in the H.264 reference software;

– `AnnexBWriter` – stores H.264/AVC bitstream in the Annex B format;

– `JmRtpWriter` – stores H.264/AVC bitstream in the JM/RTP format;

– `Rfc3984Packetizer` – encapsulates NAL units in the RTP packets in compliance with the RFC 3984 [28];

– `Rfc3984Depacketizer` – restores NAL units encapsulated in the RTP packets;

– `RtpStreamer` – creates RTP/RTCP session for sending NAL units;

– `RtpReceiver` – creates RTP/RTCP session for receiving NAL units;

– `V4LConfigWidget` – configures the V4L2 video capturing device;

– `V4LStreamerThread` – streams video from the capturing device to the memory buffers;

– `X264ConfigWidget` – configures the x264 encoder;

– `X264EncoderThread` – encodes video stored in the memory buffers;

– `Decoder` – decodes H.264/AVC bitstream and converts decoded YUV frames into RGB images.

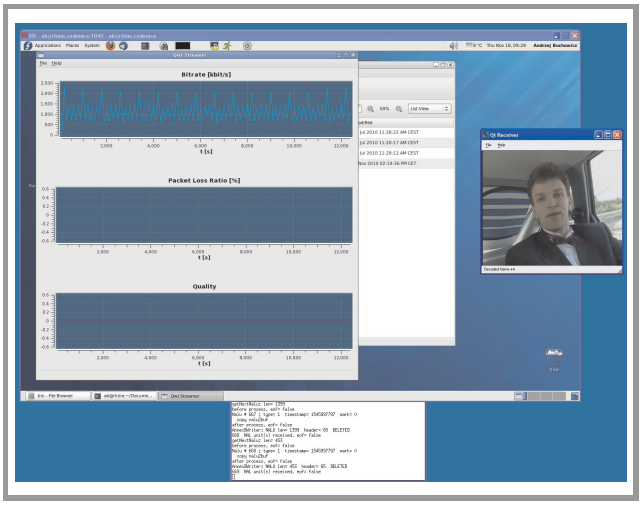


Fig. 4. Screenshot showing two framework applications. The H.264/AVC streamer is shown on the left, diagrams displays bit rate, packet loss ratio and a quality measure for the reconstructed video. The decoded H.264/AVC stream is displayed on the right.

The framework (Fig. 4) has been compiled and tested on the Fedora 10/12/14 and Ubuntu 10.10 distributions. The framework has been developed with the use of standard libraries and development tools so it should be possible to use it on other Linux distributions too. Adaptation for the other operating systems will require the complete rewrite of the classes responsible for video capture.

3. Experimental Results

The framework presented in the previous section has been used for analysis, development and testing of video streaming techniques. The comparison of the H.264/AVC error resilience techniques will be presented as an example of the experimental results obtained with the framework.
The CIF resolution *Carphone* test sequence has been encoded by the H.264/AVC reference software encoder [13] configured for the Baseline profile [1]. Group of pictures composed of 12 I/P frames and a constant value of the quantization parameter QP have been used. Three frame segmentation modes have been used: an entire frame in one slice (denoted as frame), slices containing one row of macroblocks (row) and slices with the length not exceeding the 1400 B which is less then the MTU value. Additionally two slice group modes have been used: interleaved (inter) and dispersed (disp). The rate-distortion (R-D) curves for the selected coding parameters are presented in Fig. 5. The error resilience tools reduce the coding efficiency, especially if the row slices segmentation or the dispersed slice group is used.

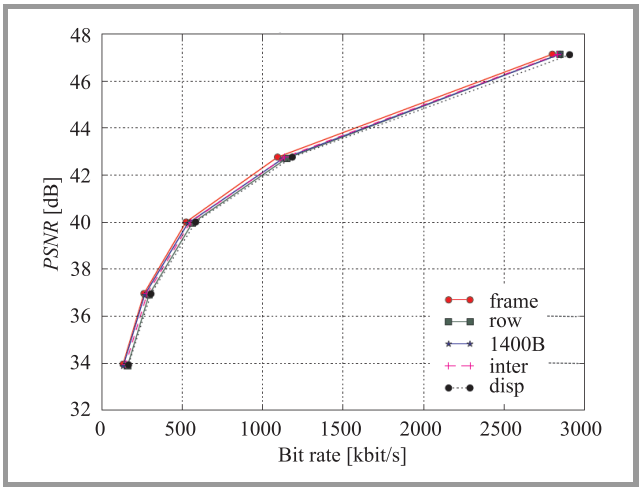


Fig. 5. The R-D curves for coding parameters.

The test sequences encoded with the $QP = 30$ (bit rate $250 \div 300$ kbit/s depending on coding parameters) have been streamed over IP network with controlled throughput. Single NAL unit in the RTP packet has been used in all experiments. Each sequence have been transmitted 5 times for selected network throughputs. The averaged packet loss ratio (PLR) is shown in Fig. 6. The PLR is higher for the row slices segmentation mode than for any other mode.

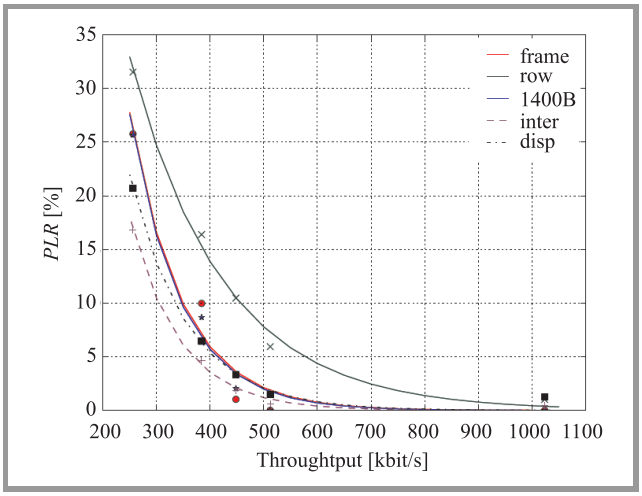


Fig. 6. The PLR for coding parameters.

The received bitstreams have been decoded by the H.264/AVC reference software decoder [13]. The peak signal-to-noise ratio (PSNR) values for each decoded bitstream have been calculated. If the bitstream has not been decoded due to transmission errors it has been assumed that $PSNR = 0$ dB. The averaged PSNR values without any error concealment in the decoder are shown in Fig. 7. The most effective is frame segmentation into slices shorter than MTU, slice group modes are slightly better than coding the entire frame in one slice.
The effectiveness of the slice group increases if the error concealment techniques are used in the decoder.

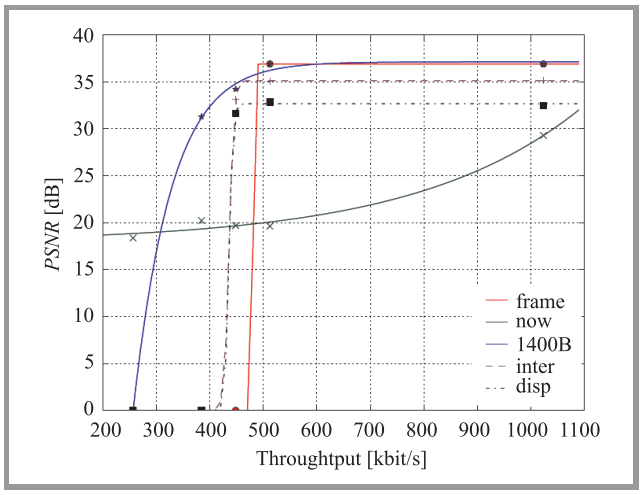


Fig. 7. Averaged PSNR with no error concealment in the decoder.

Figures 8 and 9 present the averaged PSNR for the frame copy and the motion copy error concealment mode respectively.

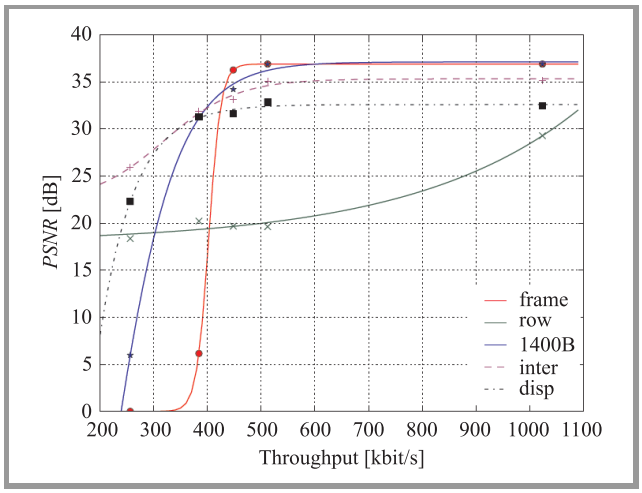


Fig. 8. Averaged PSNR for the frame copy error concealment mode.

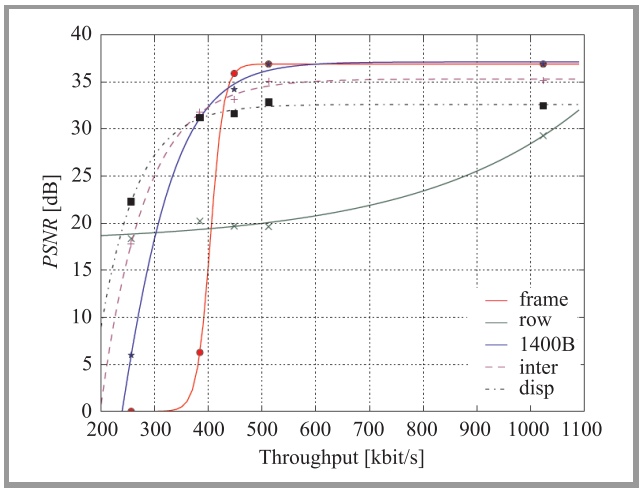


Fig. 9. Averaged PSNR for the motion copy error concealment mode.

The experimental results show that slice groups – new error resilience tool available in the H.264/AVC standard can improve the effectiveness of the transmission if the error concealment techniques are used in the decoder. However, the proper NAL unit encapuluation mode must also be selected. The lengths of NAL units in the test sequences selected for the experiment do not exceed the MTU value in most cases. Therefore, the single NAL unit in the RTP packet has been used. The results for other test sequences, with longer NAL units, would be different. The fragmentation mode would have to be used to achieve comparable effectiveness. It is worthwhile to mention, that frame segmentation into slices of length not exceeding the MTU value provides high effectiveness even if no error concealment algorithms are used in the decoder. This frame segmentation mode is available in all profiles of the H.264/AVC standard and it can always be used with the single NAL unit packetization mode.

4. Conclusions

The framework presented in this paper is a tool for testing video streaming techniques. It is based on the open source software, the H.264 reference software is also used. The framework allows streaming of the H.264/AVC video with the use of RTP/RTCP. The preencoded video stored in the file or real-time encoded video from capturing device can be transmitted. The received video can be decoded and displayed in real-time or stored in the file for further processing. The transmission parameters: bit rate, packet loss ration can be continuously displayed. The framework can be easily extended for other codecs and transmission protocols. It has been developed for the Linux operating system, but most of the libraries are portable, so the adaptation for other operating systems is possible.

References

[1] *Information technology – Coding of Audio-Visual Objects – Part 10: Advanced Video Coding*, ISO/IEC 14496-10, 2008

[2] Y. Q. Shi, H. Sun, *Image and Video Compression for Multimedia Engineering. Fundamentals, Algorithms, and Standards.n*, CRC Press, 2008.

[3] I. E. Richardson, *The H.264 Advanced Video Compression Standard*, Wiley, 2010.

[4] A. Kondoz, *Visual Media Coding and Transmission*, Wiley, 2009.

[5] *Information technology – Generic Coding of Moving Pictures and Associated Audio Information: Video*, ISO/IEC 13818-2, 2000.

[6] S. Wenger, "H.264/AVC over IP", *IEEE Trans. Circ. Sys. Video Technol.*, vol. 13, no. 7, pp. 645–656, 2003.

[7] S. K. Im and A. J. Pearmain, "Error resilient video coding with priority data classication using H.264 exible macroblock ordering", *IET Image Proces.*, no. 1, vol. 2, pp. 197–204, 2007.

[8] B. Katz, S. Greenberg, N. Yarkoni, N. Blaunstien, and R. Giladi, "New error-resilient scheme based on FMO and dynamic redundant slices allocation for wireless video transmission", *IEEE Trans. Broadcast.*, vol. 53, no. 1, pp. 308–319, 2007.

[9] T. Tillo, M. Grangetto, and G. Olmo, "Redundant slice optimal allocation for H.264 multiple description coding", *IEEE Trans. Circ. Sys. Video Technol.*, vol. 18, no. 1, pp. 58–70, 2008.

[10] C. Zhu, Y.-K. Wang, M. M. Hannuksela, and H. Li, "Error resilient video coding using redundant pictures", *IEEE Trans. Circ. Sys. Video Technol.*, vol. 19, no. 1, pp. 9–70, 2009.

[11] O. Harmanci and A. M. Tekalp, "Rate-distortion optimal video transport over IP allowing packets with bit errors", *IEEE Trans. Image Proces.*, vol. 16, no. 5, pp. 1315–1326, 2007.

[12] I. Radulovic, P. Frossard, Y-K. Wang, M. M. Hannuksela, and A. Hallapuro, "Multiple description video coding with H.264/AVC redundant pictures", *IEEE Trans. Circ. Sys. Video Technol.*, vol. 20, no. 1, pp. 144–148, 2010.

[13] K. Suhring, *H.264/AVC JM Reference Software* [Online]. Available: http://iphome.hhi.de/suehring/tml

[14] A. M. Tourapis, A. Leontaris, K. Shring, and G. Sullivan, *H.264/14496-10 AVC Reference Software Manual*, JVT-AE010, 2009.

[15] K.-P. Lim, G. Sullivan, and T. Wiegand, *Text Description of Joint Model Reference Encoding Methods and Decoding Concealment Methods*, JVT-X101, 2007.

[16] M. Ma, O. C. Au, S.-H. G. Chan, and M.-T. Sun, "Edge-directed error concealment", *IEEE Trans. Circ. Sys. Video Technol.*, vol. 20, no. 3, pp. 382–395, 2010.

[17] J. Wu, X. Liu, and K.-Y.Yoo, "A temporal error concealment method for H.264/AVC using motion vector recovery", *IEEE Trans. Consumer Electron.*, vol. 54, no. 4, pp. 1880–1885, 2008.

[18] G.-L. Wu, C.-Y. Chen, T.-H. Wu, and S.-Y. Chien, "Efficient spatial-temporal error concealment algorithm and hardware architecture design for H.264/AVC", *IEEE Trans. Circ. Sys. Video Technol.*, vol. 20, no. 11, pp. 1409–1422, 2010.

[19] K. Seth, V. Kamakoti, and S. Srinivasan, "Efficient motion vector recovery algorithm for H.264 using B-spline approximation", *IEEE Trans. Broadcast.*, vol. 56, no. 4, pp. 467–480, 2010.

[20] X. Chen, Y. Y. Chung, C. Bae, X. He, and W.-C. Yeh, "An efficient error concealment algorithm for H.264/AVC using regression modeling-based prediction", *IEEE Trans. Consumer Electron.*, vol. 56, no. 4, pp. 2694–2701, 2010.

[21] R. Pantos and W. May, *HTTP Live Streaming*, 2010 [Online]. Available: http://tools.ietf.org/html/ draft-pantos-http-live-streaming-04

[22] A. Zambelli, *IIS Smooth Streaming Technical Overview*, Microsoft Corporation, 2009.

[23] *Adobe – HTTP Dynamic Streaming* [Online]. Available: http://www.adobe.com/products/httpdynamicstreaming/

[24] *Information Technology – MPEG Systems Technologies – Part 6: Dynamic Adaptive Streaming over HTTP (DASH), ISO/IEC FCD 23001-6*, ISO/IEC JTC 1/SC 29/WG 11, w11749, 2011.

[25] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, *RTP: A Transport Protocol for Real-Time Application*, RFC 3550, 2003.

[26] C. Perkins, *RTP Audio and Video for the Internet*. Boston: Addison-Wesley, 2006.

[27] M. Baugher, D. McGrew, M. Naslund, E. Carrara, K. Norrman, *The Secure Real-time Transport Protocol (SRTP)*, RFC 3711, 2004.

[28] S. Wenger, M. M. Hannuksela, T. Stockhammer, M. Westerlund, D. Singer, *RTP Payload Format for H.264 Video*, RFC 3984, 2005.

[29] H. Schulzrinne and S. Casner, *RTP Proile for Audio and Video Conferences with Minimal Control*, RFC 3551, 2003.

[30] M. Handley, V. Jacobson, and C. Perkins, *SDP: Session Description Protocol*, RFC 4566, 2006.

[31] S. Wenger, Y.-K. Wang, T. Schierl, A. Eleftheriadis, *RTP Payload Format for SVC Video* [Online]. Available: http://tools.ietf.org/ html/draft-ietf-avt-rtp-svc-21

[32] Y.-K. Wang, T. Schierl, *RTP Payload Format for MVC Video* [Online]. Available: http://tools.ietf.org/html/draft-wang-avt-rtp-mvc-05

[33] *Linux Media Infrastructure API* [Online]. Available: http://www.linuxtv.org/ downloads/v4l-dvb-apis

[34] *Qt – A cross-platform application and UI framework* [Online]. Available: http://qt.nokia.com/products

[35] *Qwt – Qt Widgets for Technical Applications* [Online]. Available: http://qwt.sourceforge.net

[36] *Jori's page – CS/Jrtplib* [Online]. Available: http://research.edm.uhasselt.be/~jori/page/index.php?n=CS.Jrtplib

[37] *FFmpeg* [Online]. Available: http://www.ffmpeg.org

[38] *VideoLAN – x264, the best H.264/AVC Encoder* [Online]. Available: http://www.videolan.org/developers/x264.html

Andrzej Buchowicz received the M.Sc. degree in Electronics and Ph.D. degree in Telecommunication from the Faculty of Electronics, Warsaw University of Technology in 1988 and 1997, respectively. He is currently an Associate Professor in the Institute of Radioelectronics, Warsaw University of Technology. His current research interest include video coding, streaming and adaptation.
E-mail: A.Buchowicz@ire.pw.edu.pl
Institute of Radioelectronics
Warsaw University of Technology
ul. Nowowiejska st 15/19
00-665 Warsaw, Poland

Grzegorz Galiński received his M.Sc. in Electronics in 1997 and Ph.D. in 2003 from Warsaw University of Technology, Poland. Since 2002 he is with Institute of Radioelectronics at Warsaw University of Technology. His main interests include: image and video compression, multimedia indexing and multimedia systems.
E-mail: G.Galinski@ire.pw.edu.pl
Institute of Radioelectronics
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Paper

The Learning System by the Least Squares Support Vector Machine Method and its Application in Medicine

Paweł Szewczyk and Mikołaj Baszun

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

Abstract—In the paper it has been presented the possibility of using the least squares support vector machine to the initial diagnosis of patients. In order to find some optimal parameters making the work of the algorithm more detailed, the following techniques have been used: K-fold Cross Validation, Grid-Search, Particle Swarm Optimization. The result of the classification has been checked by some labels assigned by an expert. The created system has been tested on the artificially made data and the data taken from the real database. The results of the computer simulations have been presented in two forms: numerical and graphic. All the algorithms have been implemented in the C# language

Keywords—classification, Grid-Search, Particle Swarm Optimization, patients diagnosis, Support Vector Machine.

1. Introduction

Recently has been observed efforts to remote patients diagnosis by use of teleinformatic services. It obeys automatic self diagnosis by proper software which is placed on Web servers accessable for patients [1], [2]. Also proper software could be used by the dedicated physicians for automatic preselection of the remote patients who need special attention of the physicians [3]. Such software need be continuously improved to obtain valuable help; software must be based on rules verified by the dedicated physicians.

The aim of the work was to create a system in the form of software that works with a database, enabling efficient classification of data and the initial diagnosis of patients with Least Squares Support Vector Machine (LS-SVM) classifier [4]. In order to find the optimal parameters which define work of the software several techniques have been used. The purpose of the preliminary data processing is the transformation of the input data set, as a result the new set will be obtained, for which the classification algorithm solves the problem with less error or in shorter time [5]. The first step in medical data preprocessing were:

- normalization – the transformation the data after which values of the attributes are in the range [min, max]; in this paper values min = 0, max = 1 have been adopted;
- standardization – the transformation of data, after which values of the attributes have an expected average value of zero and standard deviation equal to unity.

The use of standardization is safer and usually doesn't lead to bad consequences, as it may happen the case of normalization.

The result of the classification has been checked by some labels assigned by an expert. The created system has been tested on the artificially made data and the data taken from the real database. The results of the computer simulations have been presented in two forms: numerical and graphic.

2. Measures of Quality Assessment Classification

The basic problem that appears when we try to assess the ability of generalization of researched models, is the choice of a measure which will be used to estimate this ability [6]. In this research were used two: classification accuracy and confusion matrix.

Classification accuracy determines what part of all cases were correctly classified. It is expressed in percentage. When the accuracy is larger, then the classifier is more effective. In some applications, the distinct between incorrect classifications may have meaning. For example, in medicine pass a sick patient into the health group is much more dangerous than reverse situation. In these situations we may use the confusion matrix. It is the square matrix, where rows correspond to correct decision classes, and the columns refer to the decisions predicted by the classifier. In the case of LS-SVM algorithm, which is binary classifier, the confusion matrix can be written as shown in Table 1.

Table 1
Confusion matrix

Original classes	Predicted classes by the classifier	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

The names used in Table 1 are inspired by the medical terminology, as follows:

- TP (true positive) – number of correctly classified examples from selected class,
- FN (false negative) – number of incorrectly classified examples from this class, negative decision when the example is in fact positive,

– TN (true negative) – number of examples which are not properly allocated to the selected class (correctly rejected),

– FP (false positive) – number of examples which are wrongly assigned to the selected class, when in fact they does not belong to (false alarm).

Using this kind of cases classification, particular attention should be paid to those examples, which are marked as FN. These examples are very important, because they mean not detection the disease, which can have bad consequences.

3. Least Squares Support Vector Machine Simulations

The advantages of the nonlinear SVM classifier are its great ability to solve the classification problems. J. Suykens in [4] proposed the method, which is the modification of the algorithm SVM V. Vapnik's [7], by modifying the cost function. This idea transformed the problem from solving the quadratic programming problem to solving a set of linear equations. This approach will simplify and shorten computation time.

3.1. Optimization of Model Parameters

In the case of the algorithm LS-SVM with radial kernel function [4], [7], optimized parameters are: γ, which is the weight at which the testing errors will be treated in relation to the separation margin and parameter σ, which corresponds to the width of the kernel function. It is not known in advance what combination of these two parameters will achieve the best result of classification. It is impossible to complete the search space of models, therefore the choice of optimal set of parameters is a very complex problem, and the way its solution is a key element of the classification system. In order to find the best values the following techniques were used: Grid-Search [8], K-fold Cross-Validation [5], Particle Swarm Optimization [9]–[15].

3.2. Used Technologies

To create an application development environment Microsoft Visual Studio 2010 has been used. As a programming language was chosen C# 3.0. The following libraries have been used:

– Windows Forms – implementation of the graphical user interface,

– ZedGraph – implementation of graph,

– ILNumerics – mathematical operations.

3.3. Characteristics of the Calculation Results

In order to test the proposed system, several data sets have been selected from UC Irvine Machine Learning Repository [16].

3.3.1. SPECT

A problem of diagnosis perfusion, based on data collected in Medical College of Ohio relies on diagnose of this disease based on 22 attributes [17], [18]. The database consists of 267 cases. All attributes take binary values. A specific case of cardiac perfusion may be classified to two classes: normal and abnormal. The division set to the classes was presented in the Table 2.

Table 2
Partition of the training and the test set – SPECT

Data	Train data		Test data	
Number of instances	80		187	
Perfusion	Normal	Abnormal	Normal	Abnormal
Number of instances	40	40	15	172

It may be noted that the dominant class in the test set are cases of incorrect perfusion – 91.9%. Moreover, in the process of learning classifier was used less examples than during testing. Therefore the subject of analysis it was how the algorithm can handle with a small number of training data, and how the cases will be diagnosed as correct perfusion, because such examples are only 8.1%. For this dataset there was no need for normalization and standardization because all attributes are binary.

Optimization with algorithm Grid-Search. As it has been studied in [13], the best method to find the optimal pair of parameters is changing them by exponential growth method. Using this technique, should be selected a pair of coefficients of γ and σ, for which the classification accuracy is the best.
In order to find the best model, it should be chosen the pair of points for which one side can get the highest accuracy, on the other cases, the best recognition of patients belong to the sick group. A compromise, between perhaps sometimes inconsistent conditions, is necessary.
Making an analysis of the obtained curves the point of coordinates: $\gamma = 2^{52.2}$ and $\sigma = 2^{-1.1}$ has been chosen. In the case someone can observe a slightly lower classification accuracy, but it is maximized the value of parameter true positive. For these data the aim is the best diagnosis of the sick patients. The measure of this recognition is the TP factor, so it should endeavor to a situation when

Table 3
Results of classification on a test set with Grid-Search

Parameter	Value
$\mathrm{Log}_2\gamma$	52.2
$\mathrm{Log}_2\sigma$	–1.1
Classification accuracy [%]	90.9
TP	162
FP	7
TN	8
FN	10

this coefficient has as the greatest value as possible. For this value will be followed a classification with LS-SVM algorithm on a test set.
In Table 3 the results of the classification with LS-SVM algorithm on a test set are presented – data distribution has been shown in Table 2. It is worth noting that only 10 of 172 cases have not been diagnosed as the sick persons. On the other side, the 7 from 15 people have been wrongly diagnosed as sick, although they should be included into the healthy group.

Optimization with algorithm Particle Swarm Optimization. After a large number of simulations with varying parameters of PSO algorithm, the results have been obtained, some of which are presented in the Table 4. It has been studying, how the number of partitions and the number of iterations affect on the obtained results.

Table 4
Results of classification on a test set with PSO

Number of partitions	50	50	100	100
Number of iterations	50	100	50	100
$\text{Log}_2\gamma$	48.8	48.5	24.8	50.4
$\text{Log}_2\sigma$	1.1	–1.9	–1.9	–0.6
Classification accuracy [%]	79.14	89.30	89.83	93.58
TP	138	159	160	168
FP	5	7	7	8
TN	10	8	8	7
FN	34	13	12	4

In the Table 4 the results of the classification with LS-SVM algorithm on a test set have been presented – data distribution has been shown in Table 2. The good level of accuracy has been achieved. Only 4 of 172 cases were not diagnosed as sick. In turn, 8 of 15 people were wrongly diagnosed as sick, although they should be included into the healthy group. The result for the patients from the sick group is 97.6%, whereas for the patients from the healthy group is 46.6%. Making an analysis of the above results, it can conclude, that the algorithm coped very well with the diagnosis of the cases from the sick group. It is evident that the number of particles and iterations allows to get better results. With a small number of particles the algorithm is stagnant and the particles can't escape from local minima. The values of the parameters c_1 and c_2 are not as important for the convergence of the algorithm. The experimental results indicate that it is better to set the value of parameter w to the large one in order to promote a global exploration of the space, and gradually decrease it to obtain more improved solutions. The initial value was set to 0.9 and then was reduced, in each step, to the value 0.4 [14].

3.3.2. Breast Cancer

A problem of diagnosis breast cancer, based on data collected in University Hospital in Madison (Wisconsin) relies on a diagnose of this disease based on 30 attributes [19], [20]. The database consists of 569 cases. It has no missing attribute values. A specific case of cardiac perfusion may be classified to two classes: malignant and benign. The division set of the class has been presented in Table 5.

Table 5
Percentage distribution of classes in the dataset – breast cancer

Class attribute	Number of cases
–1 (malignant cancer)	212 (37.2%)
1 (benign cancer)	357 (62.8%)

In the Table 5 the characteristics of the data set has been presented. The dominant class in these input set are cases of benign cancer – 62.8%. Because the data have not been pre-divided into training and testing datasets, the effect of the percentage partition of the data on the classification accuracy will be investigated. Each time a training set will be selected at random and will contain from 10% to 90% of all data. For each step it had been carried out 10 drawings of the training set, and then the average classification accuracy has been calculated.

Optimization with algorithm particle swarm optimization. In the simulations the following parameters were adopted in the PSO algorithm: $c_1 = c_2 = 0.5$, w-according to the method described above, number of partitions = 100, number of iterations = 100. The obtained results have been shown graphically in Fig. 1. It presents the averaged classification results using several methods of the preliminary data processing.

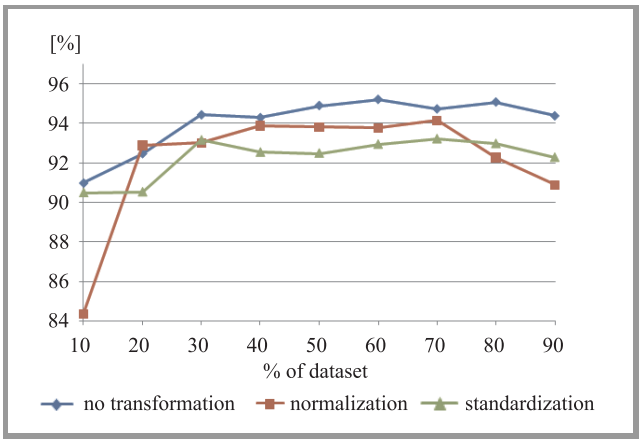


Fig. 1. Comparison of average classification accuracy for three methods of data preprocessing when changes the training set size – breast cancer.

In Fig. 1 it can be seen that the best classification accuracy was obtained when it was not used any preprocessing method. It is evident that when the training set is less than 10%–20%, the obtained results are worse. This is due to the fact the border decision boundary in LS-SVM algorithm, what depends largely on the representativeness of training data. If they are not enough, it can't achieve a satisfactory level of generalization of the model.

3.3.3. Heart Disease

A problem of diagnosis heart disease, based on data collected in Hungarian Institute of Cardiology, University Hospital in Zurich, University Hospital in Basel, V. A. Medical Center in Long Beach, Cleveland Clinic Foundation, relies on a diagnose of this disease based on 13 attributes [21]–[24]. The database consists of 270 cases. It has no missing attribute values. Each specific case may be classified to two classes: healthy and sick. The division set to the class was presented in Table 6.

Table 6
Percentage distribution of classes in the dataset – heart disease

Class attribute	Number of cases
–1 (sick)	120 (44.4%)
1 (healthy)	150 (56.6%)

In Table 6 the characteristics of the data set has been presented. The dominant class in this input set are cases of healthy patients (55.6%). Because the data have not been predivided into training and testing datasets, investigated the effect of the percentage partition of the data on the classification accuracy. Each time a training set will be selected at random and will contain from 20% to 90% of all data. For each step was carried out 10 drawings the training set, and then the average classification accuracy have been calculated.

Optimization with algorithm particle swarm optimization. In the simulations the following parameters were adopted in the PSO algorithm: $c_1 = c_2 = 0.5$, w-according to the method described above, number of partitions = 100, number of iterations = 100. The obtained results have been shown graphically in Fig. 2. It presents averaged classification results using several methods of the preliminary data processing.

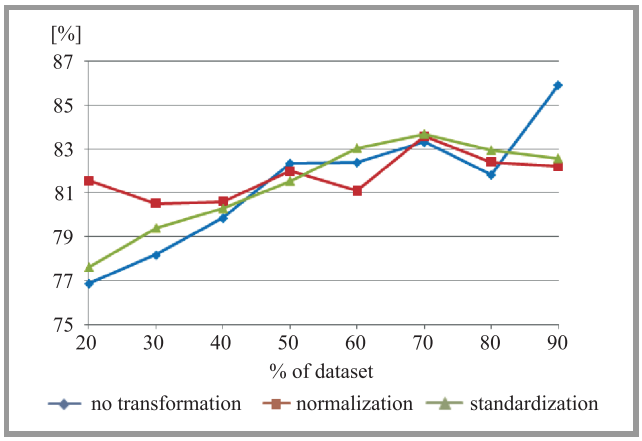


Fig. 2. Comparison of average classification accuracy for three methods of data preprocessing when changing the training set size – heart disease.

In Fig. 2 it can be seen that the best classification accuracy was obtained used the pre-processing method, which are similar. The results accuracy are lower than it was in the case of the breast cancer dataset. This may be due to fewer examples available in the database, and hence smaller number of examples used in the learning process.

4. Conclusions

The aim of the presented work was the adaptation of algorithmic techniques LS-SVM to their most efficient using in the classifying medical data from the patients. The idea here it is primarily to allow the software classification can be reliably put the presumptive diagnosis. The work of this software cannot substitute a real expert – a doctor, but to support his work by patients' selection. The process of classification of patients, based on medical data, was in the work carried out in such a way, that it analyzed data from actual patients, who were already known, what is the correct medical diagnosis. This was possible due to the using of databases available on the Internet, put there by reputable clinics.

The process of adapting the algorithm LS-SVM consisted primarily a very time consuming repetition of the calculations for other sets of parameter values which define the work of computational process, and then the choice of the most favorable version from the point of view of accuracy assessment of quantitative decision making processes results. It can't be predicted analytically for SVM techniques, as well as for other groups of algorithms based on the philosophy of artificial neural networks.

As a result, it seems that the group received the application software suitable for using in a practical analysis of data from a database of medical patients. It was found that the analysis of different groups of medical data classification software by the LS-SVM method has to be differentiated. This was demonstrated by analyzing the sample data on several key areas of treatment.

It was also studied the effect of normalization and standardization of data on the final effect of the allocation of patients. It can't determine what method of data preprocessing is better. It depends on the specific data. And so, in the case of breast cancer database, it was found that by using the normalization and standardization worse results than without preprocessing were achieved. Differences in values are on a level of a few percent. And in the case of heart disease when the training set was smaller, better results were obtained, by using normalization. When the number of examples increased, the results obtained were very close to each other.

Acknowledgement

This work was partly supported by founds on science in 2007–2010 as Ordered Research Project of Polish Ministry of Science and Higher Educations.

References

[1] M. Baszun and B. Czejdo., "An interactive medical knowledge assistant", in *Visioning and Engineering the Knowledge Society*, Berlin: Springer, 2009.

[2] M. Baszun, "Real time medical advising in cyberspace and its security aspects", in *Proc. VI Int. Conf. Cyberspace 2009*, Brno, Czech Republik, 2009.

[3] M. Baszun and B. Czejdo, "Remote patient monitoring system and a medical social network", *Int. J. Social Humanistic Comput.*, vol. 1, no. 3, pp. 273–281, 2010.

[4] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machines*, World Scientific Publishing Company, 2002.

[5] N. Jankowski, "Ontogeniczne sieci neuronowe w zastosowaniu do klasyfikacji danych medycznych", Praca doktorska, Katedra Metod Komputerowych Uniwersytetu Mikołaja Kopernika, Toruń, 1999.

[6] P. Cichosz, *Systemy Uczące się*. Warszawa: WNT, 2000 (in Polish).

[7] C. Cortes and V. Vapnik, "Support-vector network", *Machine Learning*, vol. 20, pp. 273–297, 1995.

[8] C.-W. Hsu, C.-C. Chung, C.-J. Lin, *A Practical Guide to Support Vector Classification*, National Taivan University, March 13, 2010 [Online]. Available: www.csie.ntu.edu.tw/∼cjlin

[9] S. Das, A. Abraham, and A. Konar, *Particle Swarm Optimization and Differential Evolution Algorithms: Technical Analysis, Applications and Hybridization Perspectives*. Berlin: Springer, 2008.

[10] J. Keneddy and R. C. Eberhart, "Particle swarm optimization", in *Proc. IEEE Int. Conf. Neural Netw.*, Piscataway, New York, pp. 1942–1948, 1995.

[11] M.-Z. Lu, C. L. Philip Chen, J.-B. Huo, "Optimization of combined kernel function for SVM by particle swarm optimization", in *Proc. Eighth Int. Conf. Machine Learning Cybernet.*, Baoding, China, 2009, pp. 1160–1166.

[12] M. G. H. Omran, "Particle swarm optimization methods for pattern recognition and image processing", Ph.D. thesis, University of Pretoria, 2004.

[13] C. Sun and D. Gong, "Support Vector Machines with PSO Algorithm for Short-Term Load Forecasting", in *Proc. IEEE Int. Conf. Netw., Sensing Contr. ICNSC '06*, Ft. Lauderdale, USA, 2006, pp. 676–680, 2006.

[14] Q.-Z. Yao, J.e Cai, J.-L. Zhang, "Simultaneous feature selection and LS-SVM parameters optimization algorithm based on PSO", in *Proc. World Congr. Comput. Sci. Informa. Engin. CSIE 2009*, Los Angeles, USA, 2009, pp. 723–727.

[15] Y.g-J. Zhai, H.-L. Li, Q. Zhou, "Research on SVM algorithm with particle swarm optimization", in *Proc. 11th Joint Conf. Inform. Sci. JCIS 2008*, Shenzhen, China, 2008.

[16] *UC Irvine Machine Learning Repository*, Center for Machine Learning and Intelligent Systems, University of California, USA [Online]. Available: http://archive.ics.uci.edu/ml/

[17] L. A. Kurgan, K. J. Cios, R. Tadeusiewicz, M. Ogiela, and L. Goodenday, "Knowledge discovery approach to automated cardiac SPECT diagnosis", *Artificial Intelligence in Medicine*, vol. 23, no. 2, pp. 149–169, 2001.

[18] A. Płachcińska and J. Kuśmierek, *Techniki Obrazowania Serca w Medycynie Nuklearnej*, Zakład Medycyny Nuklearnej Akademii Medycznej w Łodzi, 2001 (in Polish).

[19] K. Polat and S. Güneş, "Breast cancer diagnosis using Least Square Support Vector Machine", *Digit. Sig. Proces.*, vol. 17, pp. 694–701, 2007.

[20] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis", in *Proc. Int. Symp. Electron. Imag.: Sci. Technol.*, San Jose, USA, 1993, vol. 1905, pp. 861–870.

[21] N. Allahverdi and H. Kahramanli, "Extracting rules from neural networks using artificial immune systems", in *Proc. 2nd Int. Conf. Problems of Cybernet. Inform.*, Baku, Azerbaijan, 2008.

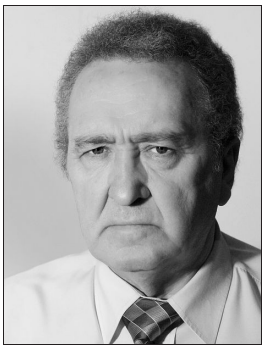
[22] S. Bhatia and P. Prakash, "SVM based decision support system for heart disease classification with integer-coded genetic algorithm to select critical features", in *Proc. World Congr. Engin. Comp. Sci. WCECS'08*, San Francisco, USA, 2008.

[23] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques", *Int. J. Comput. Sci. Network Secur.*, vol. 8, no. 8, 2008.

[24] D. W. Vance, *An All-Attributes Approach to Supervised Learning*, University of Cincinnati, 2006.

Paweł Szewczyk received the M.Sc. degree in Electronics and Computer Engineering from Warsaw University of Technology in 2011. His research interests focus on neural networks, support vector machines, especially in classification and data mining.

E-mail: pawelszewczyk1@gmail.com
Faculty of Electronics and Information Technology
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland



Mikołaj Baszun is an Assistant Professor in the Faculty of Electronics and Information Technology, Warsaw University of Technology. His research focuses on electronic microsystems, computer engineering, artificial intelligence technology, medical informatics, and Web services. He has published more than 50 publications in these areas.

E-mail: mbaszun@elka.pw.edu.pl
Faculty of Electronics and Information Technology
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Paper

Designing Auctions: A Historical Perspective

Michał Karpowicz

Research Academic Computer Network, Warsaw, Poland

Abstract—Auction is a form of organization of competition that leads to the assignment and valuation of resources based on the information obtained from the competing agents. From the perspective of systems science it is a distributed resource allocation algorithm applied in the environment with information asymmetry, i.e., where the interconnected and interacting subsystems have different information about the system as a whole. This paper presents an overview of the historical development of mathematical theory underlying modern approach to auction design. Selected practical applications of the theory are also discussed.

Keywords—auctions, game theory, mechanism design.

1. Introduction

Auctions are used to buy and sell almost anything one can imagine. Number of categories of items being up for the Internet auctions at Allegro, Amazon or eBay web sites is truly astonishing. Auction houses, such as Sothebys, sell art, antiques, books, jewelry, toys, dolls, and other collectible memorabilia. Securities worth billions of dollars are regularly auctioned worldwide by the Departments of Treasury. Directives of the European Parliament recommend application of auctions in awarding of public contracts and coordinating the procurement procedures. Auctions are also widely used to regulate markets of strategic resources such as electric power or radio spectrum. Recently there have also been many attempts to apply auction mechanisms to allocate bandwidth in communication networks, improve industrial supply chain management and efficiency of allocation of landing and take-off time slots in air traffic flow management.

One of the reasons for the popularity of auction is that it provides a convenient way of assigning goods to those who value them the most. The common auction formats used in practice to allocate a single object are the English auction, the Dutch auction, the first-price and the second-price sealed-bid auction. The most popular variant is the English auction in which the auctioneer calls ascending prices until there is only one bidder willing to pay. In the Dutch auction the auctioneer also calls prices, however, he initially starts from the high level and successively lowers the price until there is someone willing to pay. In contrast with the dynamic open bidding formats of the English and Dutch auctions, the sealed-bid auctions are conducted in a single step. The auctioneer determines the outcomes based on the sealed offers submitted by the bidders'. Both in the first-price and the second-price auction the winner is the bidder with the highest bid. The difference is in the amount of money the winner is obliged to pay. In the first-price auction the winner pays his bid. In the second-price auction the winner pays the second highest bid.

Multiple objects can be sold in a sequence of single-object auctions or simultaneously. There are three traditional formats of simultaneous multi-unit auctions: discriminatory (pay-as-bid), uniform and Vickrey auction. In each case bidders submit to the auctioneer a vector of nonincreasing bids (marginal values), which indicate each bidder's willingness to pay for each additional item. In a discriminatory auction a bidder pays the amount of money equal to the sum of his winning bids, i.e., the sum of those bids that belong to the set of K highest bids, where K is the number of goods. In a uniform-price auction all goods are sold at a *market-clearing price*, i.e., a maximal price at which the total amount demanded is greater or equal to the total amount supplied. In a Vickrey auction, each bidder pays an amount equal to the externality exerted on others. If a bidder wins k units of resource, the his payment is equal to the sum of k highest bids of other bidders (defeated by his bids) [1], [2].

The choice of a particular auction format has been a vital problem. On one hand auction may serve as a solution to many problems of decentralized resource allocation. On the other, each format suffers from drawbacks that may negatively influence both efficiency of the outcomes and auctioneer's revenue. In this paper a historical overview of selected aspects of auction design is presented. The key issues that are raised concern contributions of the related game-theoretic analysis.

2. The Systems Science Perspective

Auction is a form of organization of the competition that leads to the assignment and valuation of resources based on the information obtained from the competing agents. From viewpoint of the systems science auction is a distributed resource allocation algorithm applied in the environments with information asymmetry, i.e., where the interconnected and interacting subsystems have different information about the system as a whole. This perspective is taken in the discussion below. First, a survey of results concerning theory of competitive equilibrium is presented. This is justified by the role it plays in the design of resource allocation mechanism, even though its assumptions hardly ever correspond to the reality. Second, we refer to the historical development of the theory of incentives underlying mod-

ern approach to auction design. The theory emerged from the game theoretic analysis of choices made in distributed systems under information asymmetry. As a consequence its models are much more realistic than those derived from competitive equilibrium theory.

2.1. Competitive Equilibrium Theory

The goal of auction design is to take the advantages of competition to solve the problem of resource allocation. In this context the competitive (Walrasian) equilibrium serves as an aspiration point for the auction design. It is defined as a solution of the system of interaction balancing (market-clearing) equations according to which preference maximizing demand equals preference maximizing supply. Static properties of competitive equilibria and conditions that guarantee their existence are described by the fundamental theorems of welfare economics; see Walras [3], Wald [4], [5], Lange [6], Arrow and Debreu [7]. Traditionally, they are viewed as formalization of the Adam Smith's famous conjecture regarding the *invisible hand* of market competition [8]. In essence:

- Pareto-efficiency is consistent with individual self-interest since *price-taking* behavior is reasonable in competitive market, especially if the number of decision makers is large [9].

Stability of competitive equilibrium was first investigated by Samuelson. In [10], [11], [12] he surveyed dynamic models of market-clearing process and examined the relationship between the conditions for stability of competitive equilibrium given by Hicks [13] and general conditions for stability of dynamical systems. Hicks described equilibrium as perfectly stable if an increased demand for a good raises its price even when any subset of other prices is arbitrarily held constant. Samuelson showed that this condition is neither necessary nor sufficient for dynamic stability in Lyapunov sense, except in the case of symmetric matrix of the partial derivatives of *excess demand* – a difference between the value of demand and supply. An extensive exploration of dynamic stability of price adjustment process in perfectly competitive market was later given by Arrow and Hurwicz [14], [15]. The market price adjustment process, described by the system of differential equations defined by continuous and sign-preserving functions of aggregate excess demand, is globally stable if the following assumptions are satisfied:

– agents maximize rational, continuous, monotone and strictly convex preferences,

– agents' preferences are commonly known,

– agents are price-takers (do not anticipate equilibrium prices),

– aggregate demand satisfies the *(weak) axiom of revealed preferences* and has the property of *gross substitution*.

General treatment of the sufficient conditions for the stability of competitive equilibria was also given by Uzawa in [16], [17], [18]. Extensive study of price-based hierarchical control methods was given by Findeisen *et al*. in [19], as well. Saari and Simon [20], [21], on the other hand, investigated local stability of competitive equilibria. They noticed that there is a tradeoff between global stability conditions and information required by the price adjustment procedure to converge to local equilibrium. In particular, they considered Newton algorithm as a price adjustment process and studied the information content it requires for convergence.

Applications in telecommunication. Perhaps the most impressive recent application of competitive equilibrium theory is the design of telecommunication protocols for congestion control. As an illustration of the general approach one can consider the uniform-price auction mechanism proposed by Kelly [22]. Transmission rates of the traffic sources in the computer network are gradually adjusted until their willingness to pay for the introduced congestion equals the corresponding congestion cost. In this model each link in the network acts as an auctioneer, it adjusts its individual congestion price until the demand for the link resources equals the supply. See Srikant [23] and Low [24], [25], [26] for details.
Attractiveness of this approach relates to the common sense of competitiveness of the network environment. Indeed, telecommunication networks consist of a large number of similar traffic sources (characterized by similar preferences) controlled by the same telecommunication protocols. Therefore, the assumptions of competitive equilibrium theory may be regarded as a reasonable description of the traffic exchange process. If all traffic sources calculate transmission rates taking the congestion signals (link prices) as given, then the fixed point of the traffic exchange process can be established in competitive equilibrium maximizing the effectiveness of network utilization. This observation has served as a justification for the design of several recent TCP congestion control algorithms.

2.2. Theory of Incentives

Clearly, assumptions of the competitive equilibrium model are not satisfied in most real-life settings. Namley, economy is rarly a complete system of markets (in which every agent is able to exchange every good with every other agent), externalities are present (prices do not reflect the full costs or benefits), common property resources exist in economy (consumption of such good by one individual does not reduce availability of the good for consumption by others, no one can be effectively excluded from using the good), decision makers anticipate prices, information is imperfect and time delays cannot be ignored, *etc*. From the engineering point of view model inadequacies of this sort can be recognized as a potential source of system distress; see e.g. Stiglitz [27] and Mas-Colell [28].
The shortcomings and failures of competitive equilibrium theory inspired the search for a much more sophisticated

models. General solutions to the resource allocation problems arising in the systems with information asymmetry emerged from the investigations of incentives motivating individuals in decision making. Historically they related to the three streams of thought: theory of market socialism, social choice theory and theory of competitive markets; see e.g. Green, Laffont and Tirole [29], [30]. Currently, the obtained results are included in the theory of incentives (principal-agent models) and mechanism (or game) design theory.

Theory of market socialism, co-founded by Polish economist Oskar Lange in 1930's [31], postulated centralized control in order to reach predefined goals of the economy. The responsibility assigned to the central planner was to determine the values of coordination variables: prices, production inputs and outputs. These were then applied to control performance of local industrial organizations [32]:

> *(...) a market mechanism could be established in a socialist economy which would lead to the solution of the simultaneous equations by means of an empirical procedure of trial and error. Starting with an arbitrary set of prices, the price is raised whenever demand exceeds supply and lowered whenever the opposite is the case. Through such a process of tatonnements, first described by Walras, the final equilibrium prices are gradually reached. These are the prices satisfying the system of simultaneous equations. It was assumed without question that the tatonnement process in fact converges to the system of equilibrium prices. (...) Let us put the simultaneous equations on an electronic computer and we shall obtain the solution in less than a second. The market process with its cumbersome tatonnements appears old-fashioned. Indeed, it may be considered as a computing device of the preelectronic age.*

It seems evident that not only the problem of incentives was ignored but also there was a belief that a government agency could glean and process all the relevant information required to make an economy function well. In practice, on one hand the constraints were imposed on production outputs, but, on the other, the government either provided insufficient inputs or provided more than it was necessary. As a result, with severe conflicts concerning personal freedom and civil rights in the background, the economy strode towards the state of constant struggle to realize production plans. Strategic manipulations to outwit the system, both in order to meet predefined goals of the economy of public goods and to satisfy privately defined interests, arose naturally in effect of recurring coordination failures. Inefficiency of directly coordinated system was largely due to incompatibility of the private interests and goals of the central planner. To assure that they coincide proper incentives were required. However, as it quickly became apparent, without sufficient autonomy, private property or the profit motive, putting democratic procedures aside, incentives were lacking. System's collapse was inevitable. An interesting debate revealing important historical background of the discussed issues can be found in [33]. See also Stiglitz [27], [34].

Social choice theory is concerned with the problem of rational aggregation of preferences within the collective decision rules, including voting systems and competitive markets. Its central result, due to Arrow [35], shows that necessary conditions that preference aggregations should be expected to meet are inconsistent and cannot hold together:

> *If we exclude the possibility of interpersonal comparisons of utility, then the only methods of passing from individual tastes to social preferences which will be satisfactory [i.e. will not reflect individuals' desires negatively and the resultant social tastes will be represented by an ordering having the properties of rationality ascribed to individual orderings] and which will be defined for a wide range of sets of individual orderings are either imposed or dictatorial.*

As it can be noticed, the key concern that motivated the related work grew out of the observation that the concept of preference aggregation, by its very nature, deals with the problem of *interpersonal comparisons* and *measurability* of preferences' intensity. The focus on ordinal preferences, which was largely due to the influential arguments that *no common denominator of feelings is possible* [36], was an attempt to eschew the related controversies. Unfortunately, Arrow's impossibility theorem demonstrated that there are other substantial difficulties that arise as an unavoidable trade off – the impossibility result is the price for the *incomparability* requirement. In an immediate response it was therefore proposed, mostly due to Sen [37], [38], [39], that informational constraints imposed on the collective choice rule should be modified. The line of argumentation was taken that the results of preference aggregation should be *invariant* with respect to the utility signals that provide the same information in terms of the applied notion of measurability and interpersonal comparisons. Consequently, the counterargument gained strong support that the notion of ordinal preferences is inadequate for representing conflicts of gains and losses. These conflicts, however, inevitably occur in many collective choice settings, especially when welfare judgments are involved and the resource constraints are present. When dealing with the considerable number of social choice situations, interpersonal comparisons of intensity of preferences, or weights of interests, provide desirable informational basis for the determination of decision. Conditions imposed on the social choice function by Arrow's theorem may be interpreted as necessary but not sufficient for collective choice. On the other hand, cardinality and full interpersonal comparability of individual welfare units are sufficient but not necessary for rational choice under aggregate welfare maximization. To generate a *complete* and *transitive* aggregation of orderings (preferences) their

partial comparability is sufficient as well; see Arrow [35] and Sen [40] for details. In essence, social choice theory shows how to design a satisfactory procedure for preference aggregation. However, it is not concerned with the question if the aggregated preferences, revealed by the interacting agents, are true or not. This observation inspired investigations of the gaming aspect of collective decision-making, commonly observed in many votings and auctions. Finally, the concept of incentive-based regulation arose as a potential remedy to the wide scope of imperfections of the markets traditionally designed within the framework of fundamental theorems of welfare economics. Spectacular examples intensively discussed in the literature include the global depression of the 1930s, East Asia financial crisis in the late 1990s and California Power Exchange collapse in 2001. The following macroeconomical comment by Stiglitz [27] emphasizes significance of the related issues and places them in somewhat wider perspective of market design for developing economies:

> *even if Smith's theory were relevant for advanced industrialized countries, the required conditions are not satisfied in developing countries. The market system requires clearly established property rights and the courts to enforce them; but often these are absent in developing countries. The market system requires competition and perfect information. But competition is limited and information is far from perfect – a well-functioning competitive markets cannot be established overnight. The theory says that an efficient market economy requires that all of the assumptions be satisfied. In some cases reforms in one area, without accompanying reforms in others, make actually matters worse. (...) economic theory and history show how disastrous it can be to ignore sequencing.*

Inefficiencies arising under asymmetric and imperfect information were first studied by Stiglitz [41]–[43], Akerlof [44] and Spence [45]. For general results see [30].

The above considerations eventually gave rise to the theory of incentives and game design. Its contributions, and especially its rigorous game-theoretic analysis of the incentive compatibility concept introduced by Hurwicz [46], have deepened the knowledge regarding the possibility for achieving Pareto-optimal allocations in decentralized systems and designing efficient auctioning procedures. The following results are often viewed as the most influential [9]:

- When a delegation of tasks occurs within the firm, then because of asymmetric information the firm does not maximize its profit, i.e., allocative inefficiency occurs.
- In markets of private and public goods with a finite number of agents, there are no nonparametric mechanisms (which process only the information received from the agents) that simultaneously yield Pareto-efficient allocations and provide individual agents with incentives to report their true preferences honestly.
- In markets of private and public goods with a finite number of agents, there are nonparametric mechanisms that yield Pareto-efficient allocations when all agents follow their self-interest by playing a Nash-equilibrium strategy.
- In the bilateral trade problem, there is no mechanism that yields efficient allocations, provides individual agents with incentives to report their true preferences honestly, guarantees profitable participation and covers the costs of allocations.

Applications in telecommunication. If the assumption of price-taking behavior is dropped, then in most cases competitive equilibria cannot be reached by means of the decentralized price-based coordination methods, such as uniform-price auctions. This problem was recently investigated in the networking context by Johari [47]–[49]. The major result of his work, focused on the mechanisms of *price-anticipating* bidding, demonstrates that there exist implementations of the uniform-price auctions generating outcomes with bounded loss of efficiency. An interesting conclusion is also due to Roughgarden [50], [51]. Namely, the ratio of efficiency loss, arising in the networks as a consequence of the price-anticipating behavior, is independent of network topology. Following the similar line of argument, Yang and Hajek [52], [53] analyzed the undesirable performance of the algorithm proposed by Kelly [22]. In the settings with strategic bidders competition for network paths is dominated in terms of efficiency by competition for the network links (that form the paths). Finally, sufficient conditions for efficiency of auctions in the environments with price-anticipating agents has been given by Karpowicz in [54].

Anticipation of price effects has been recognized in the literature as an urgent problem of dynamic interconnection management in communication networks. Consider a group of interconnected network service providers (ISPs) exchanging IP traffic between their autonomous systems. The basic observation that one can make about this resource allocation setting suggests that local decisions concerning bandwidth allocations can have a non-negligible influence on the overall network performance. As a result, ISPs may anticipate the effects of their actions on interconnection prices and view these prices as functions of the actions of all interrelated providers. Clearly, in such an environment routing and congestion control protocols applied locally by ISPs can be subject to strategic manipulations. Records of such strategic interactions can be found in the archives of the Polish Office of Electronic Communications (www.uke.gov.pl). For more general treatment of problems related to competition in telecommunications, especially from the viewpoint of interconnection agreements, such as peering and transit, and interconnection pricing, see Laffont and Tirole [55], Laskowski [56], Norton [57], Baake and Wichmann [58].

3. Auction Design and Game Theory

In order to benefit from allocating resources by means of an auction it is necessary that its rules be designed and tailored to the particular allocation setting. To cope with the complexity of this multistage design process it is therefore reasonable to apply convenient modeling tools. Game theory, a branch of applied mathematics, plays an important role in this context. It is a study of mathematical models of interaction (competition or cooperation) of intelligent and rational (in a specified sense) agents making interrelated choices under incomplete (asymmetric) information [30], [59], [60]. On one hand, it aims at providing answers to some of the essential questions regarding properties of different auction formats. On the other, it provides recommendations for the design of resource allocation and pricing rules defining games that are characterized by the desired features.

Properties of outcomes generated by auctions were first identified by means of game-theoretic analysis in the seminal work of Vickrey [61]. Its major conclusions were based on the following observation: information about demand and supply, revealed by the competing agents and used to determine the outcomes, influences the market clearing price, thus encouraging agents to submit price-anticipating bids. As a consequence, investigation of the incentives that agents may have to submit

> *an unbiased report of the marginal-cost (competitive supply) curves (...) and of the marginal-value (competitive demand) curves (...), or at least of the portions of these curves covering a range of prices that will be sure to contain the equilibrium price,*

became the main theme of the auction (game or mechanism) design theory [1], [2], [60]. Major contributions in this field are due to Hurwicz [46], [62]–[66], Myerson [59], [67], [68] and Maskin [69]–[73]. An overview of selected historical attempts to apply the theory in practice is given below.

3.1. Treasury Bill Auctions

An influential investigation of the adverse effects of strategic bidding in auctions of shares was presented in the paper by Wilson [74]. It demonstrated existence of bidding strategies that may lead to the reduction of sale price, which in effect reduces revenue of the resource manager, even with the increasing number of agents placing their bids. The result given by Wilson was next generalized by Back and Zender [75] in the context of the auction of U.S. Treasury bills. Conclusions presented in their paper served as an argument in the debate regarding the merits of different formats of multi-unit auctions that the Treasury could apply for the sale of securities.

Traditionally the discriminatory (pay-as-bid) auction was used, according to which all bidders whose offers exceed the market-clearing price (determined by the auctioneer) are obliged to pay their bids. However, in 1960s suggestion came from Milton Friedman that in order to improve revenues the Treasury should consider switching to the uniform-price format. Back's and Zender's paper supported the resulting debate with the formal arguments against any unconditional and simplified recommendations. In particular, it warned against extrapolation of properties of auctions with single unit demand to the more general situations of multi-unit demand. Interestingly, it was not until recently that the equilibrium properties of the multi-object uniform-price auctions have been thoroughly investigated. The general result was obtained by Ausubel and Cramton [76]. It relates potential inefficiency of the uniform-pricing scheme outcomes to the fact that the scheme creates strong incentives for *demand reduction*: each agent's optimal strategy is to shade bids for units of resource other than the first one; since bids placed on the other units determine the final (clearing) price with positive probability, agents may increase their profits by submitting lowered marginal values. (From the viewpoint of the supply side of the system, the result implies increased marginal production costs revealed to the auctioneer.) The similar result is also given in [77].

Indeed, revenue implications of the potential underpricing has become the subject of intensive studies in the context of Treasury auctions. From 1992 to 1998 the U.S. Treasury, motivated by various academic conjectures and market manipulation scandals (in 1991 a major trader in the U.S. Treasury securities admitted that it had violated auction rules by submitting fraudulent bids [78]), experimented with the sealed-bid uniform-price auctions for selling two-year and five-year notes. Eventually it switched entirely to the uniform price format in the end of 1998. The goal was to verify whether incentives to shade bids would be reduced with uniform pricing rule, which in turn would improve revenues to the Treasury. The experiment did not provide strong support for this conjecture. The impact on revenues of the two pricing formats was demonstrated by Malvey and Archibald [79] to be statistically insignificant. Umlauf [80] and Tenario [81], on the other hand, slightly favor the uniform pricing scheme using data from the Mexican Treasury auctions and Zambian foreign exchange auctions, respectively. One can view this conclusions as consistent with the results of Ausubel and Cramton [76], and Back and Zender [75], which state that the ranking of the two formats is inherently ambiguous. There are cases which show that uniform-price format outperforms in both efficiency and revenue the pay-as-bid format in the particular auction setting, and results which show the reverse. This also seems to correspond to the well known result of the theory of single-object auctions; for models that include both affiliation (log-supermodularity of densities) of bidders' valuations and risk aversion, the first- and second-price auctions of single-objects cannot be generally ranked by their expected prices [1], [82].

Another important result is due to Keloharju, Nyborg and Rydqvist [83] who give an extensive exploration of historical data from the Finnish Treasury auctions. On one hand,

their finding is that individual bidders' demand increases with number of bidders, which is consistent with the argument that bidders exercise market power. On the other hand, however, statistical data show that equilibria with extremely low prices, e.g. predicted by Wilson [74], usually do not occur in practice. The similar conclusion was given by Nyborg and Sundaresan [84] and Goldreich [85]. The practical reason why bidders do not coordinate on the revenue reducing low price equilibria is the strategic behavior of the auctioneer himself. By determining the amount of securities sold in response to the submitted collection of bids, imposing restrictions on the bidding procedures and revealing sufficient amount of information, the Treasury effectively protects itself against revenue reduction. This advocates the important result of auction design theory – games induced by the rules of allocation mechanisms are played not only between the agents but between the agents and the mechanism designer as well.

3.2. Electric Power Auctions

Both discriminatory and uniform-price auctions have also been used in the electricity markets. Interesting examples come from Scandinavia, UK and France. Norway, Sweden, Finland and Denmark buy and sell electricity on the Nordic Power Exchange, Nord Pool, which has been the world's only multinational exchange for trading electric power since 1990s [86], [87]. Since 2001 in the UK electricity generators sell their output on daily basis in the discriminatory auctions, after the switch from the uniform-price format originally adopted in 1990 [88]. The uniform pricing scheme had also been used in the California Power Exchange before its collapse in 2001. Electricité de France (EDF) gave an undertaking to the European Commission in early 2001 to give access to generation capacities in France in the form of contracts conveying the right to purchase energy. Currently contracts with durations between 3 and 48 months are being sold at pay-as-bid auctions conducted approximately every 3 months; see www.edf.com.

Bidding behavior in the electric power auctions has been a growing concern, as it may be related to prices being increased above competitive levels [89]. An intensively studied real-life example that servers as a support of this argument relates to the collapse of the electric power market in California where the uniform-price auctions were used to buy electricity on the power exchange. It is believed that the strategic bidding of the suppliers, extracting the highest possible electricity prices, was among the causative factors of the crisis in the summer of 2000 [90], [91]. Indeed, many mathematical models have been developed to explain and prevent events of this sort. Research that are of great interest in this context concerns especially the ways in which suppliers' bidding manipulations aimed at improving profits may influence allocations of energy production. Indeed, knowledge of the related threats has been playing a role in adjusting regulatory policy around the world. For example, Green and Newbery [92], [93] applied the *supply function equilibrium* approach, originally introduced by Klemperer and Meyer [94], to show that markups on marginal costs may be constituted by the Nash equilibrium of the game induced by the British electricity spot market. Von der Fehr and Harbord [95] reached the similar conclusion with the sealed-bid auction model. However, they also showed that if supply signals are step functions, as it usually is in practice, pure-strategy equilibria do not exist for a wide range of demand distributions. Other results along this line include works of Cramton [96], [97], Engelbrecht-Wiggans and Kahn [98], [99], Baldick and Hogan [100], Day and Hobbs [101], just to name a few examples.

3.3. Spectrum Auctions

The design of spectrum auctions for the Federal Communications Commission (FCC) in the United States[1] is often regarded in the literature as one of the most successful applications of game theory. In fact, Milgrom [2] argues that it was the design that started the era of *putting the theory to work*.

The primary goal of the FCC was the maximization of economic efficiency of spectrum allocation – licenses were to be assigned to those who are capable of providing better services at lower costs. Designers confronted with the regulatory goals turned to game theory for methodological support. Its recommendations narrowed the set of admissible solutions by pointing out the threats related to the expected bidding strategies [2], [102], [103]. Theoretical models also guided the development of experimentation scenarios testing the applicability of the key design judgments [104], [105]. The following conclusions determined the final auction format:

Open bidding is better than a single sealed bid.

Open bidding process reveals information about valuations of goods and provides feedback increasing auction revenues [1], [106].

Simultaneous open bidding is better than sequential auctions.

Sequential auctions of goods requires agents to condition their decisions on the future actions of others. This guesswork is in practice very likely to reduce efficiency of the auction. With simultaneous bidding much of the guesswork is eliminated [1].

Package bids (combinatorial auctions) are too complex.

Once bidding for a combination of goods is admitted, inefficiencies are likely to arise due to threshold problem, a variant of the free-rider problem. The transparency of auction is weakened as well [102], [103], [107].

As a result the simultaneous multiple-round ascending-bid auction was proposed, a multi-object version of the English auction. According to its rules, a number of licenses is auctioned simultaneously in discrete, successive rounds. In every round, a bidder can bid (offering a buy price)

[1]This application of mechanism design theory was indicated by Prof. Eric Maskin in the telephone interview following the announcement of the 2007 Nobel Prize in Economics.

on any license subject to constraints given by the activity rules and bidder's eligibility defined by the upfront payment. Open bidding format gives each bidder information about the highest bids, identities of bidders, their upfront payments and handicaps. The auction stops if a single round passes in which no new bids are placed on any license. Licenses are sold for the price equal to their highest standing bid [2], [102], [103].

Auction rules implemented by the FCC proved its efficiency in series of spectrum auctions and became a worldwide standard. It should be noticed, though, that they did not eliminate the incentives for strategic bidding, potentially decreasing efficiency of allocations. Cramton and Schwartz [108], [109] described several cases of bid signaling that occurred in FCC auctions and identify it as an example of profitable *collusive* behavior – incentives for tacit collusion were especially strong among incumbents and large bidders capable of exerting their market power. Consequently, the experiences gained in practice have guided evolution of the auction. In response to the observed problems many design recommendations have been given to reduce the effectiveness of signaling and collusion, e.g., by concealing bidders identities, offering preferences (handicaps) for small businesses and new entrants, increasing reserve prices, bounding supply by offering licenses that are harder to split up, allowing package bidding. Again, game-theoretic considerations have been often applied in the examination of the refinements.

4. Final Remarks

Game theory has been helpful in explaining bidding behavior in different auction settings. In some cases its qualitative predictions have turned out to be influential enough to affect the regulation policies. The examples presented above may serve as an evidence of its contributions. On the other hand, however, the very same studies unveil its weak points. Clearly, relevance of its recommendations depends on the particular decision setting.

In reality efficiency of auction outcomes depends on many factors that often dominate any influence that a particular allocation or pricing rule may have. Issues that are faced by the auction designer in practice, often playing more important role than the rules of an auction, are listed below.

Auction items. One of the key design problems is related to the choice of an object to be put up on auction. Whether it is divisible or indivisible, homogeneous or heterogeneous may have a decisive influence on the allocation process. This stems from the fact that a particular definition of an allocation determines preference profile of the competing decision makers. Empirical and theoretical evidence show that rules of auction may be irrelevant under particular allocation definitions.

Auction participants. It is essential to define who is eligible to participate in an auction and what approvals are required. As pointed out by Milgrom [2], marketing a sale is often the biggest factor in its success. Announcement of an auction or definition of a resource allocation procedure must provide information targeted to potential participants enabling them to study the opportunity.

Flexible goals. Auctions are conducted to achieve specific economic goals – typically, maximization of efficiency of allocations or maximization of auctioneer's revenue. However, because of the complexity of the auctioning process adjustments are often required. In fact, in many cases it may be reasonable not to conduct an auction, e.g., because of the overall performance of the economy or insufficient legislative support.

Interactions. What to allocate to agents depends on their demand, which depends on who agents are, which in turn may depend on the way the auction is conducted. Decisions made by auction designer are not independent [2]. Interactions occur between agents as well. There are many occasions for them to cooperate before, during and after the auction. Collusion and mergers clearly have a significant influence on the outcomes, as well as possibility of reallocation after the auction.

Information. What information is required to determine allocations and final payments, and what information is revealed to agents may play a decisive role. One of the key motivations for pricing resources by means of an auction is gaining information about agents' privately known valuations and preferences. Under auction bidding process it is not only the resource manager but also agents themselves that are responsible for the final price and allocation of the resource. Related guesswork is therefore distributed between auction designer and auction participants. On the other hand, the responsibility for resource allocation outcomes inevitably creates incentives for agents to manipulate the process. Information about reserve prices, bidding increments, agents' eligibility revealed before the auction, as well as information about submitted bids revealed during the auction may significantly influence competition, bidding behavior and efficiency of outcomes, especially if agents' valuations are interdependent.

To solve at least some of the problems of auction design one may settle the judgments on game-theoretic models *approximating* the auction outcomes. However, any reasoning should be extremely careful and substantiated by experimental verifications of the dominating factors, since arbitrary estimations and behavioral assumptions are inevitable in this context.

Acknowledgement

This work was supported by the Ministry of Science and Higher Education through grant PBZ-MNiSW-02/II/2007.

References

[1] V. Krishna, *Auction Theory*. Academic Press, 2002.

[2] P. Milgrom, *Putting Auction Theory to Work*. Cambridge University Press, 2004.

[3] J. R. Hicks, "Leon Walras", *Econometrica*, vol. 2, no. 4, pp. 338–348, 1934.

[4] A. Wald, "On some systems of equations of mathemetical economics", *Econometrica*, vol. 19, no. 4, pp. 368–404, 1951.

[5] A. Wald, "On a relation between changes in demand and price changes", *Econometrica*, vol. 20, no. 2, pp. 304–306, 1952.

[6] O. Lange, "The foundations of welfare economics", *Econometrica*, vol. 10, pp. 215–228, 1942.

[7] K. J. Arrow and G. Debreu, "Existence of an equilibrium for a competitive economy", *Econometrica*, vol. 22, no. 3, pp. 265–290, 1954.

[8] A. Smith, *An Inquiry into the Nature and Causes of the Wealth of Nations*. Bantam Books, 2003.

[9] *Information, Incentives, and economic mechanisms: essays in honor of Leonid Hurwicz*, T. Groves, R. Radner, and S. Reiter, Eds. University of Minnesota Press, 1987.

[10] P. A. Samuelson, "The stability of equilibrium: comparative statics and dynamics", *Econometrica*, vol. 9, no. 2, pp. 97–120, 1941.

[11] P. A. Samuelson, "The stability of equilibrium: linear and nonlinear systems", *Econometrica*, vol. 10, no. 1, pp. 1–25, 1942.

[12] P. A. Samuelson, "The relationship between Hicksian stability and true dynamic stability" *Econometrica*, vol. 12, pp. 256–257, 1944.

[13] J. R. Hicks, *Value and Capital: An Inquiry into Some Fundamental Principles of Economic Theory*. Oxford: Clarendon Press, 1946.

[14] K. J. Arrow and L. Hurwicz, "On the stability of the competitive equilibrium, I", *Econometrica*, vol. 26, no. 4, pp. 522–552, 1958.

[15] K. J. Arrow and L. Hurwicz, "On the stability of the competitive equilibrium, II", *Econometrica*, vol. 27, no. 1, pp. 82–109, 1959.

[16] H. Uzawa, "The stability of dynamic process", *Econometrica*, vol. 29, no. 4, pp. 617–631, 1961.

[17] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*. SIAM, 1999.

[18] H. Uzawa, "Market mechanisms and mathematical programming", *Econometrica*, vol. 28, no. 4, pp. 872–881, 1960.

[19] W. Findeisen, F. N. B., M. Brdyś, K. Malinowski, P. Tatjewski, and A. Woźniak, *Control and Coordination in Hierarchical Systems*. Wiley, 1980.

[20] D. G. Saari and C. P. Simon, "Effective price mechanisms", *Econometrica*, vol. 46, no. 5, pp. 1097–1125, 1978.

[21] D. G. Saari, "Iterative price mechanisms", *Econometrica*, vol. 53, no. 5, pp. 1117–1131, 1985.

[22] F. P. Kelly, A. K. Maulloo, and D. K. Tan, "Rate control for communication networks: shadow prices, proportional fairness, and stability", *J. Oper. Res. Society*, vol. 49, pp. 237–252, 1998.

[23] R. Srikant, *The Mathematics of Internet Congestion Control*. Birkhäuser Boston, December 2003.

[24] S. H. Low, "A duality model of TCP and queue management algorithms", *IEEE/ACM Trans. Netwo.*, vol. 11, no. 4, pp. 525–536, 2003.

[25] S. H. Low and D. E. Lapsley, "Optimization flow control, I: Basic algorithm and convergence. *IEEE/ACM Trans. Netwo.*, vol. 7, no. 6, pp. 861–874, 1999.

[26] S. H. Low, F. Paganini, and J. C. Doyle, "Internet congestion control", *IEEE Control Sys. Mag.*, vol. 22, no. 1, pp. 28–43, 2002.

[27] J. Stiglitz, *Globalization and its discontents*. Penguin Books, 2002.

[28] A. Mas-Colell, M. D. Whinston, and J. R. Green, *Microeconomic Theory*. Oxford University Press, 1995.

[29] J. R. Green and J.-J. Laffont, *Incentives in Public Decision-Making*. North-Holland Publishing Company, 1979.

[30] J.-J. Laffont and D. Martimort, *The Theory of Incentives*. Princeton University Press, 2002.

[31] O. Lange, *On the Economic Theory of Socialism*. 1938.

[32] C. H. Feinstein, *Socialism, Capitalism and Economic Growth*. Cambridge University Press, 1967.

[33] O. Lange, The practice of economic planning and the optimum allocation of resources. *Econometrica*, 1949.

[34] J. Stiglitz, *Making globalization work*. Penguin Books, 2006.

[35] K. J. Arrow, *Social Choice and Individual Values*. Wiley, 1963.

[36] L. Robbins, "Interpersonal comparisons and utility: A comment", *Economic Journal*, vol. 192, no. 48, pp. 635–641, 1938.

[37] A. Sen, "Interpersonal aggregation and partial comparability", *Econometrica*, 1970.

[38] A. Sen, "Social choice theory: Re-examination", *Econometrica*, 1977.

[39] A. Sen, "On weights and measures: Informational contraints in social welfare", Econometrica, vol. 45, no. 7, pp. 1539–1572, 1977.

[40] A. Sen, *Collective Choice and Social Welfare*. Holden-Day, 1970.

[41] S. J. Grossman and J. Stiglitz, "On the impossibility of informationally efficient markets", *American Econom. Rev.*, vol. 70, no. 3, pp. 393–408, 1980.

[42] M. Rotschild and J. Stiglitz, "Equilibrium in competitive insurance markets: an essay on the economics of imperfect information", *Quarterly J. Econom.*, vol. 90, no. 4, pp. 630–649, 1976.

[43] J. Stiglitz, "The theory of screening, education, and the distribution of income", *American Econom. Rev.*, vol. 65, no. 3, pp. 283–300, 1975.

[44] G. A. Akerlof, "The market for "Lemons": quality uncertainty and the market mechanism", *Quarterly J. Econom.*, vol. 84, no. 3, pp. 488–500, 1970.

[45] M. Spence, "Job market signaling", *Quarterly J. Econom.*, vol. 87, no. 3, pp. 355–374, 1973.

[46] L. Hurwicz, "On informationally decentralized systems", in *Studies in Resource Allocation Processes*, K. Arrow and L. Hurwicz, Eds. Cambridge University Press, 1977, ch. 4, pp. 425–459.

[47] R. Johari, "Efficiency loss in market mechanisms for resource allocation", Ph.D. thesis, MIT, 2004.

[48] R. Johari and J. N. Tsitsiklis, "Efficiency loss in a network resource allocation game", *Mathemat. Oper. Res.*, vol. 29, pp. 407–435, 2004.

[49] R. Johari and J. N. Tsitsiklis, "Efficiency of scalar-parameterized mechanisms", *Oper. Res.*, vol. 57, no. 4, pp. 823–839, 2009.

[50] T. Roughgarden and É. Tardos, "How bad is selfish routing?", in *Proc. 41st Ann. Symp. Foundat. Comp. Sci. FOCS 2000*, Redondo Beach, USA, 2000, pp. 93–102.

[51] T. Roughgarden, "The price of anarchy is independent of the network topology", in *Proc. 34th ACM Symp. Theory. Comput. STOC 2002*, Montréal, Canada, 2002, pp. 428–437.

[52] S. Yang and B. Hajek, "VCG-Kelly mechanisms for allocation of divisible goods: adapting vcg mechanisms to one-dimensional signals", *IEEE J. Selec. Areas Commun.*, vol. 25, no. 6, pp. 1237–1243, 2007.

[53] B. Hajek and S. Yang, "Strategic buyers in a sum bid game for flat networks", in *Proc. IMA Worksh.*, March 2004.

[54] M. Karpowicz, "Coordination in hierarchical systems with rational agents", Ph.D. thesis, Politechnika Warszawska, 2009.

[55] J.-J. Laffont and J. Tirole, *Competition in Telecommunications*. The MIT Press, 2000.

[56] S. Laskowski, "Wspomaganie procesu ustalania cen na rynku usług telekomunikacyjnych", Ph.D. thesis, Politechnika Warszawska, 2007 (in Polish).

[57] W. B. Norton, "Internet service providers and peering", in *Proc. NANOG 19*, Albuquerque, New Mexico, 2000.

[58] P. Baake and T. Wichmann, "On the economics of Internet peering", *Netnomics*, vol. 1, no. 1, pp. 89–105, 1999.

[59] Roger B. Myerson, *Game Theory: Analysis of Conflict*. Harvard University Press, 1991.

[60] D. Fudenberg and J. Tirole, *Game Theory*. The MIT Press, 1991.

[61] W. Vickrey, "Counterspeculation, auctions and competitive sealed tenders", *J. Finance*, vol. 16, no. 1, pp. 8–37, 1961.

[62] L. Hurwicz and M. Walker, "On the generic nonoptimality of dominant-strategy allocation mechanisms: a general theorem that includes pure exchange economies", *Econometrica*, vol. 58, no. 3, pp. 683–704, 1990.

[63] L. Hurwicz and D. Schmeidler, "Construction of outcome functions guaranteeing existence and pareto-optimality of Nash-equilibria", *Econometrica*, vol. 46, no. 6, 1978.

[64] L. Hurwicz and S. Reiter, *Designing Economic Mechanisms*. Cambridge University Press, 2008.

[65] L. Hurwicz, "The design of mechanisms for resource allocation", *American Econom. Rev.*, vol. 63, no. 2, pp. 1–30, 1973.

[66] L. Hurwicz, "On allocations attainable through Nash equilibria", *J. Economic Theory*, vol. 21, no. 1, pp. 140–165, 1979.

[67] R. B. Myerson and M. A. Satterthwaite, "Efficient mechanisms for bilateral trading", *J. Econom. Theory*, vol. 29, no. 2, 1983.

[68] R. B. Myerson, "Optimal auction design", *Mathem. Operation Res.*, vol. 6, no. 1, 1981.

[69] E. Maskin, "Nash equilibrium and welfare optimality", *Rev. Econom. Studies*, vol. 66, no. 1, pp. 23–38, 1999.

[70] J.-J. Laffont and E. Maskin, "A differential approach to dominant strategy mechanisms", *Econometrica*, vol. 48, no. 6, pp. 1507–1520, 1980.

[71] L. Hurwicz, E. Maskin, and A. Postlewaite, "Feasible implementation of social choice correspondences by Nash equilibria", in *Essays in Honor of Stanley Reiter*, J. Ledyard, Ed. Kluwer, 1995.

[72] P. Dasgupta and E. Maskin, "Efficient austions", *Quarterly J. Econom.*, vol. 115, no. 2, pp. 341–388, 2000.

[73] P. Dasgupta, P. Hammond, and E. Maskin, "The implementation of social choice rules: some general results on incentive compatibility", *Rev. Econom. Studies*, 1978.

[74] R. Wilson, Auctions of shares. *Quarterly Journal of Economics*, vol. 93, no. 4, pp. 675–689, 1979.

[75] K. Back and J. F. Zender, "Auctions of divisible goods: on the rationale for the treasury experiment", *The Rev. Financ. Studies*, vol. 6, no. 4, pp. 733–764, 1993.

[76] L. Ausubel and P. Cramton, "Demand reduction and inefficiency in multi-unit auctions", Working paper. University of Maryland, 1996.

[77] M. Karpowicz and K. Malinowski, "Efficiency loss and uniform-price mechanism", in *Proc. 47th IEEE Conf. Decision and Control*, Cancun, Mexico, 2008.

[78] G. J. Miller, "Debt management networks", *Public Adm. Rev.*, vol. 53, no. 1, pp. 50–58, 1993.

[79] P. F. Malvey and C. M. Archibald, "Uniform-price auctions: update of the treasury experience", Tech. rep., Office of Market Finance U.S. Treasury, 1998.

[80] S. R. Umlauf, "An empirical study of the mexican treasury bill auction", *J. Financ. Econom.*, vol. 33, pp. 313–340, 1993.

[81] R. Tenorio, "Revenue equivalence and bidding behavior in a multi-unit auction market: an empirical analysis", *Rev. Econom. Statistics*, vol. 75, pp. 302–314, 1993.

[82] P. Milgrom, "Auction theory" in *Advances in Economic Theory: Fifth World Congr.*, T. Bewley, Ed. Cambridge: Cambridge University Press, 1987.

[83] M. Keloharju, K. G. Nyborg, and K. Rydqvist, "Strategic behavior and underpricing in uniform price auctions", Working Papers 2003.25, Fondazione Eni Enrico Mattei, March 2003.

[84] K. G. Nyborg and S. M. Sundaresan, "Discriminatory versus uniform treasury auctions: evidence from when-issued transactions", *J. Financ. Econom.*, vol. 42, pp. 63–104, 1996.

[85] D. Goldreich, "Underpricing in discriminatory and uniform-price treasury auctions", *J. Financ. Quantitative Analysis* (forthcoming).

[86] H. Eggertsson, "The Scandinavian electricity power market and market power", Master's thesis, Technical University of Denmark, 2003.

[87] A. Botterud, A. K. Bhattacharyya, and M. Ilic, "Futures and spot prices – an analysis of the Scandinavian electricity market", in *Proc. 34th Ann. North American Power Symp. NAPS 2002*, Tempe, USA, 2002.

[88] N. Fabra, N.-H. M. von der Fehr, and D. Harbord, "Designing electricity auctions", *Rand J. Econom.*, vol. 37, no. 1, pp. 23–46, 2006.

[89] H. Gruenspecht and T. Terry, *Horizontal market power in restructured electricity markets*, Office of Policy, US Department of Energy, Washington, USA.

[90] M. Kahn and L. Lynch, "California's electricity options and challenges", Report to the Governor, 2000.

[91] S. Borenstein, 'The trouble with electricity markets: understanding California's restructuring disaster", *The J. Econom. Perspect.*, vol. 16, no. 1, pp. 191–211, 2002.

[92] R. J. Green, "Increasing competition in the British electricity spot market. *Journal of Industrial Economics*, vol. 44, no. 2, pp. 205–216, 1996.

[93] J. R. Green and D. M. Newbery, "Competition in the British electricity spot market", *J. Polit. Economy*, vol. 100, no. 5, pp. 929–953, 1992.

[94] P. D. Klemperer and M. A. Meyer, "Supply function equilibria in oligopoly under uncertainty", *Econometrica*, vol. 57, no. 6, pp. 1243–1277, 1989.

[95] N.-H. M. von der Fehr and D. Harbord, "Spot market competition in the U.K. electricity industry", *Economic Journal*, vol. 103, iss. 418, pp. 531–546, 1993.

[96] P. Cramton, "Competitive bidding behavior in uniform-price auction markets", in *Proc. 37th Hawaii Int. Conf. Sys. Sci. HICSS 2004*, Big Island, Hawaii, USA, 2004.

[97] P. Cramton, "Electricity market design: the good, the bad, and the ugly", in *Proc. 36th Hawaii Int. Conf. Sys. Sci. HICSS 2003*, Big Island, Hawaii, USA, 2004.

[98] R. Engelbrecht-Wiggans and C. M. Kahn, "Multi-unit auctions with uniform prices", *Economic Theory*, vol. 12, no. 2, pp. 227–258, 1998.

[99] R. Engelbrecht-Wiggans and C. M. Kahn, "Multi-unit pay-your-bid auctions with variable award", *Games Econom. Behavior*, vol. 23, pp. 25–42, 1998.

[100] R. Baldick and W. Hogan, *Capacity constrained supply function equilibrium models of electricity markets: Stability, nondecreasing constraints, and function space iterations*. University of California Energy Institute, 2001.

[101] C. J. Day and B. F. Hobbs, "Oligopolistic competition in power networks: a conjectured supply function approach", *IEEE Trans. Power Sys.*, 2002.

[102] P. Cramton, "The FCC spectrum auctions: An early assessment", *J. Econom. Managem. Strategy*, vol. 6, no. 3, pp. 431–495, 1997.

[103] P. Cramton, "Spectrum auctions", in *Handbook of Telecommunications Economics*, M. Cave, S. K. Majumdar, and I. Vogelsang, Eds. Elsevier, 2002, pp. 605–639.

[104] J. Ledyard, D. Porter, and A. Rangel, "Experiments testing multiobject allocation mechanisms", *J. Econom. Manage. Strategy*, vol. 6, no. 3, pp. 639–675, 1997.

[105] C. R. Plott, "Laboratory experimental testbeds: application to the PCS auction", *J. Econom. Manage. Strategy*, vol. 6, no. 3, pp. 605–638, 1997.

[106] P. Milgrom and R. J. Weber, "A theory of auctions and competitive bidding", *Econometrica*, vol. 50, no. 5, pp. 1089–1122 1982.

[107] P. Cramton, *Combinatorial Auctions*. MIT Press, 2006.

[108] P. Cramton and J. A. Schwartz, "Collusive bidding in the FCC spectrum auctions" *The B.E. J. Economic Analysis and Policy*, no. 1, 2002.

[109] R. J. Weber, "Making more from less: Strategic demand reduction in the FCC spectrum auctions", *J. Econom. Manage. Strategy*, vol. 6, no. 3, pp. 529–548, 1997.

Michał Karpowicz received his Ph.D. in Computer Science from the Warsaw University of Technology (WUT), Poland, in 2010. Currently he is assistant professor at Research and Academic Computer Network (NASK). His research interests focus on game theory, network control and optimization.

E-mail: michal.karpowicz@nask.pl
Research Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland

Paper

Personalized Knowledge Mining in Large Text Sets

Cezary Chudzian, Janusz Granat, Edward Klimasara, Jarosław Sobieszek, and Andrzej P. Wierzbicki

National Institute of Telecommunications, Warsaw, Poland

Abstract—The paper starts with a discussion of the concept of knowledge engineering, in particular ontological engineering. Consequently, the paper presents assumptions accepted as a basis for a group research on a radically personalized system of ontological knowledge mining, relying on the perspective of human centered computing and combining ontological concepts of the user with an ontology resulting from an automatic classification of a given set of textual data. The paper presents a pilot system PrOnto that supports research work in two aspects: searching for information interesting for a user according to her/his personalized ontological profile, and supporting research cooperation in a group of users (Virtual Research Community) according, e.g., to a comparison of such personalized ontological profiles. The paper concludes with suggestions concerning diverse applications of ontological engineering tools and future work.

Keywords—human centered computing, knowledge engineering, ontological engineering, personalized ontology.

1. Introduction

During last decade, a special importance in telecommunications and Internet services achieved *data mining* or *knowledge mining* in large data sets describing such services; related terms are called *knowledge management*, *knowledge engineering* or even *knowledge science*. However, *knowledge science* touches philosophy, and *knowledge management*, even if of computer science origin, is today treated as a part of management science; therefore, we shall rather use the term *knowledge engineering* in its broad sense, extending it beyond its classical academic sense of artificial intelligence and learning algorithms.

A research group in National Institute of Telecommunications concentrates on knowledge engineering for over ten years, together with basic research on such disciplines as mathematical logic, multiple criteria decision theory, diverse optimization and statistical methods, also ontological engineering; all these theoretical aspects serve, however, as the basis of development of tools of knowledge engineering, in particular knowledge mining in large sets of data.

Applications of these tools relate to diverse problems. They might consist in diverse data and knowledge mining services for telecom operators, or using advanced statistical methods to analyze diverse indicators of the development of informational society in Poland or in Mazovia region. However, this paper concentrates on applications of ontological engineering to support of knowledge mining, research and knowledge management.

We must add still one explanation. Classical methods of ontological engineering concentrate, similarly as typical work on artificial intelligence, on an automation of knowledge mining from large textual data sets, while the preferences of the user might be taken into account, but typically in a limited extent. The character of the work presented in this paper is different and results from our practical experience in applying data and knowledge mining. We assume a *sovereign position of the user* – explained more specifically in further text – and concentrate on a *radical personalization* of ontological user profile that might use, but should not be dominated by the results of automatic analysis of large sets of textual data[1].

2. Knowledge Engineering and Tacit Knowledge

Experience in applying knowledge engineering tools shows that knowledge mining is aimed not only at finding *logical relations between data*, but as well at discovering *tacit knowledge hidden in large sets of data* and correlated with *tacit knowledge of the user*. We apply here the concept of *tacit knowledge* on purpose, although it denotes usually[2] *preverbal* (difficult to express in words) knowledge hidden in human mind, see [6]–[10].

However, preverbal knowledge is contained also in large data sets, even in textual data sets, and the goal of knowledge engineering is to discover such knowledge – not only

[1]The paper describes results of work in a project called in Polish *Projekt Badawczy Zamawiany "Usługi i sieci teleinformatyczne następnej generacji – aspekty techniczne, aplikacyjne i rynkowe", grupa tematyczna i: Systemy wspomagania decyzji regulacyjnych: Wykrywanie wiedzy w dużych zbiorach danych telekomunikacyjnych (Requested Research Project "Next Generation Services and Networks – technical, application and market aspects", Theme Group i: Systems Supporting Regulatory Decisions: Knowledge Mining in Large Telecommunication Data Sets)* and is a modified version of longer Polish texts [1], [2].

[2]Usually but not exclusively, since there is also tacit knowledge in the *intuitive intellectual heritage of humanity* including *synthetic a priori judgments* [3] and *hermeneutical horizons* (see, e.g., [4]) expressing essential intuitive beliefs propagated by educational systems, as well as *emotional heritage of humanity* including between others *collective unconsciousness* [5]) together with it's parts – myths, archetypes, etc., but also all artworks, say, the emotional load of all films. Hence tacit knowledge can be contained not only in the mind of a single human being, see [6].

in an algorithmic and automatic way, but also with the cooperation or even under the guidance of a human user.
In a broad understanding of knowledge engineering we can distinguish several parts of it:

I. Narrowly understood artificial intelligence and automatic learning engineering.

II. Discovering knowledge (including tacit knowledge) in large data sets, data and knowledge mining.

III. Text processing engineering, including ontological engineering, but also textual knowledge mining.

Part I is described by many books, see, e.g., [11]. Part II relies partly on the tools developed in Part I, but uses also much broader diversity of tools: statistical, decision analytical, etc., and includes to a larger extent the requirements and participation of human users. Part III aims usually at finding or selection of textual explicit knowledge and uses tools of *ontological engineering* and *semantic Web* as well as network *search engines*; in applications, however, decisive is an interpretation of the selected textual knowledge by a human user, hence according to the user's tacit knowledge or hermeneutical horizon [4], [12], [13].
Ontological engineering is also related to *knowledge management*, see, e.g., [14]. The term *ontology* was borrowed from philosophy, where it means *theory of being* (see, e.g., [15]); computer science interprets differently this term as a classification of entities and words representing them. In information technology, we treat today the term *ontology* as an enriched taxonomy, vocabulary with a hierarchy and other (logical, semantic) relations of terms. A significant development of ontological engineering occurred during last two decades, related to the concept of *Semantic Web* and based on the assumption that contemporary WWW network contains (or will soon contain) knowledge corresponding to all intellectual heritage of humanity, thus advanced information technology tools should be able to extract essential part of this knowledge in form of an universal ontology[3].
Ontologies play today, when treated as tools of representation and shared understanding of knowledge about diverse domains, important roles in many applications, such as development of information systems, organizing the content of Internet pages, categorizing commercial products, standardizing vocabularies in given fields, see, e.g., [16]–[19]. However, there are diverse controversies also in ontological engineering, related to several opposite approaches to the construction, application and interpretation of ontologies. There are many methods of constructing ontologies, see, e.g., [20]; we can speak about constructing *lightweight ontologies* with a simple hierarchical structure, or *heavyweight ontologies* including more detailed logical and semantic relations between terms. We can also speak about constructing *local ontologies* characterizing terms used by a local group of researchers or even by a research discipline or a cultural sphere (the same term, such as *ontology*, might have different meaning for different disciplines), as opposed to *universal ontologies* trying to represent all knowledge contained, say, in WWW network. We can also construct an ontology *from scratch*, through *reuse*, or *automated* (using automatic methods of ontological engineering), see, e.g., [19]; the last distinction is not quite precise, since good ontological engineering tools are always *semi-automated*, assume some interaction with the human user that constructs ontology with their help, while an essential problem is the extent and character of this interaction, discussed in detail below.

As the most advanced in ontological engineering, the works of Standard Upper Ontology Working Group (SUO WG) are often cited, aimed at "*forming an upper ontology whose domain is all of human consensus reality*" together with related CYC ontology (see, e.g., [21]. This is an interesting attempt to build a universal vocabulary, but many doubts can be voiced, e.g., to the use of the term "upper" (who is upper – human or network and computer?), or to local applications of such vocabulary (local meaning might not correspond to the popular meaning in the Internet).

Another subdivision of the methods of ontology construction relates precisely to the role of a human constructor of the ontology. If we assume that it is human constructor who should be sovereign and "on top", then we should speak about *top down* way of ontology construction as starting with experience and intuition (as well as emotions) of a human expert or a group of them, while *bottom up* way of ontology construction should denote an automatic construction based on broad textual content. Thus, the "upper" ontology of SUO WG is actually a universal bottom up ontology that might be difficult to apply locally, because it does not take into account the tacit knowledge of a local group of experts.

This distinction is related also to a technical and evolutionary theory of intuition [22], [23] that uses the contemporary knowledge of telecommunications and computer science to show that the use of language (and logic) by humans simplified at least ten thousand times[4] perception and reasoning that was originally immanent (using all senses). This resulted in a tremendous *surplus of brain* that is used in diverse ways, in tacit knowing and tacit knowledge, in intuitive reasoning, existential and transcendental thinking. If only less than 0.1% of neurons in our brains is needed for logical thinking and verbal argumentation, than human intuition can be much stronger (even if still fallible) than

[3]This assumption is debatable, see footnote 2 above and [6] on the role of tacit knowledge in intellectual heritage of humanity, as well as further discussion of the reasons of radical personalization of individual ontological profiles.

[4]The broadband needed for transmission of vision is at least 100 times larger than the broadband needed for transmission of voice, and the computational complexity of processing such large data sets is nonlinear; assuming quadratic increase of complexity gives results close to a lower bound. Therefore, *a picture is worth at least ten thousand words*. Thus, when we developed speech in the evolutionary development of humans, we made a tremendous evolutionary shortcut and obtained *a surplus of brain* (some philosophers call it *surplus of mind*): only less than 0.1% of our brain cells is needed for verbal communication and rational reasoning.

logical argumentation. This, however, implies the need of a *radical personalization* of ontological profiles of users of ontological tools, relying on an increase of the role of personal intuition when defining such profiles; such radical personalization is consistent also with the trend to *human centered computing*.
Such is the perspective that motivated us to search for new approaches to ontology construction (from scratch or by reuse, with lightweight structure, combining top-down and bottom-up, semi-automated approaches) for a local group of researchers. Originally, in the *Theme Group i: Systems Supporting Regulatory Decisions: Knowledge Mining in Large Telecommunication Data Sets* of the Requested Research Project we planned a broader application of such ontological approach to support regulatory decisions on telecommunication markets, but a cut of funding forced us to limit the application to a local research group in telecommunications, affiliated at the National Institute of Telecommunications.

3. Results of the Work on Knowledge Mining

3.1. Preliminary Investigations

Initial investigation involved cooperation with IIASA (International Institute for Applied Systems Analysis) and JAIST (Japan Advanced Institute for Science and Technology, School of Knowledge Science), see, e.g., [24], [25]. A broad survey of literature has shown that there are papers suggesting a combination of bottom-up and top-down methods of ontology construction [26] but not specifying how to combine them. In [24], [25] we proposed the use of *hermeneutic reflection* (expert reflection on the structure of local ontology), of *organizational reflection* (expert reflection on the organizational structure of a research institution); we also considered the use of *mind mapping* to stimulate the intuitive top-down construction of upper layers of an ontology by the user; the lower layers might result from a bottom-up approach and ontology matching tools might be used to combine them.
We also compared diverse available tools of ontological engineering and developed a Polish language modification of the system OntoGen. OntoGen (http://ontogen.ijs.si/) is an open source tool for semi-automated bottom-up text mining and ontology construction. We tested this system on publications of our National Institute of Telecommunication with satisfactory results, see [1], [2], [27], [28]. However, the main result of this preliminary work was an idea how to construct a radically personalized user's ontological profile, leading to the concept of PrOnto system.

3.2. Radically Personalized User's Ontological Profile

We started with an analysis of an important dichotomy in search of textual information in the network. There are two opposite classes of such search problems (and some mixed problems in between):

- searching for an answer to a well defined question of the user (*information retrieval*);
- searching for information interesting for the user, but rather loosely defined (*information filtering*).

Traditional search engines combined these functions to some extent, today we observe a trend to separate them. More important for supporting research is the second class that requires, however, a specification of user's preferences. Such specification can be implicit, resulting from an analysis of the history of behavior of the user (which is a popular tool of supporting internet commerce, with a long own history – see, e.g., [29], or explicit, in the form of a set of keywords, key phrases, or even a simple ontology (which again can be constructed from scratch by the user, or be influenced by the history of user's behavior). Both implicit and explicit specification of user's preferences can be modified for supporting research (see, e.g., serwis CiteULike), but explicit specification makes it possible to preserve the sovereignty of the user while constructing a radically personalized user's ontological profile.
Such a profile (which might be called also a *perspective*, or a *horizon* of the user) is assumed to consist of three layers.

- An upper layer of *concepts* $c \in \mathbf{C}$, defined by the user and treated as her/his intuitive entities (they might be later interpreted logically, but with utmost caution, because, e.g., the concept *Markov chains* can actually mean *these aspects that are now interesting for me in the theory of Markov chains*).
- A lowest layer of *keywords* or *key phrases* $k \in \mathbf{K}$, either defined by the user or by a bottom-up ontological tools (they will be later the main connection of the radically personalized profile with classical ontological tools).
- A middle layer of *relations between* **K** and **C**, or *importance coefficients* $f \in \mathbf{F}$ of a keyword for a concept, defined by the user (later they might be also modified by the history of user's behavior, but the user should be sovereign in their specification), interpreted either as weighting coefficients, or subjective probabilities, or fuzzy logic membership values, or aspiration levels for multiple criteria ranking of documents with respect to the ontological profile, see below.

The radical personalization consists in assuming that only the lower layer **K** is responsible for collaboration with bottom-up ontological engineering tools. The middle layer **F** and the upper layer **C** might form together with the lower layer a kind of personalized ontology (used, e.g., to support cooperation in a research group), but the user is sovereign in using her/his intuition when modifying these two higher layers.

4. Prototype System PrOnto

4.1. A general Structure of PrOnto

Generally, PrOnto system supports research work of a group of users (Virtual Research Community, VRC) using a radically personalized user interface based on profile described above. This radical personalization relies on the assumption that research preferences of a user cannot be fully logically or even probabilistically formalized (at most 0.01% of neurons in our brain work on logical, rational reasoning). Therefore, the interface should preserve and stress an intuitive character of the user choices, while nevertheless supporting her/his collaboration with the tools of ontological engineering. The model of PrOnto system assumes services and support to a research group of users (VCR) with functionalities serving an individual user or group collaboration. The model contains:

- *A radically personalized ontological model of the user*, composed of three layers as described above;
- *Document repository* **D**, containing documents interesting for the user or entire group of them (VRC) in the form of full text or a network link to such text;
- *A method of search and ranking of documents in the repository* for an individual user based on her/his radically personalized ontological profile (many methods are possible and the model of a user does not uniquely define such a method);
- *An agent of network search* (so called hermeneutic agent) that performs search in all accessible network – usually with help of accessible search engines – for new documents in order to enrich the repository, including a ranking method and(or) a decision rule;
- *Functionalities supporting an effective exchange of knowledge between users* that can enrich PrOnto system either for an individual or for group user. Such functionalities might include:
  - cataloguing documents for a group of users (VRC),
  - supporting research collaboration in the group (information about new documents judged as interesting by some users, etc.),
  - search for similarities in user interests, etc.

4.2. Searching for Information in Documents While Using Keywords

Documents in the repository must be indexed with respect to the keywords or keyword phrases. This is a standard problem known as *multiple pattern string matching*, searching for a pattern string (a keyword phrase) in a longer document. Because of large dimensions of documents and large number of pattern strings, it is important to select an algorithm with simplest, linear computational complexity; however, this complexity can be linear either with respect to the number of patterns strings (which can be very large), or, more advantageously, linear with respect to size of documents searched. An algorithm Aho-Corasick [30] was selected, implemented and tested, with the results shown in Fig. 1.

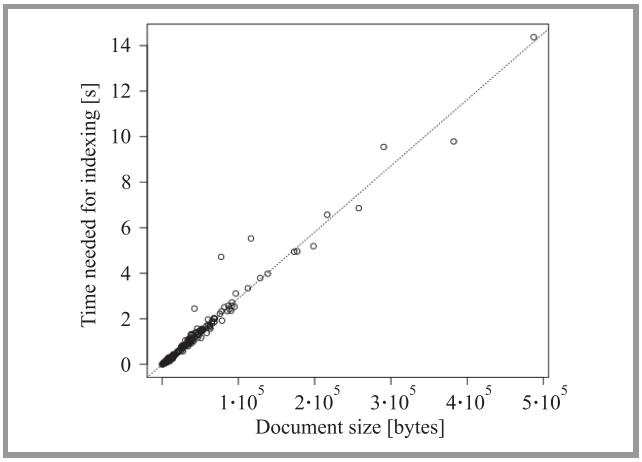


Fig. 1. Time needed for indexing as dependent on document size in bytes.

Another problem is a measure of importance of a document $d \in D$ with respect to a given key phrase $k \in K$. Initially, we selected the classical measure TF-IDF (*Term Frequency – Inverse Document Frequency*). The value of $\mathrm{TF-IDF}(k \in K,\ d \in D)$ grows proportionally to the frequency of occurrence of the phrase k in the document d and decreases inversely to the total number of documents containing k. We plan to investigate also other measures of importance, denoted here generally $g(d,k)$.

4.3. Importance of a Document with Respect to a Concept or a Set of Concepts

Another essential problem is a measure of importance of a document $d \in D$ with respect to a given concept $c \in \mathbf{C}$. If we have:

- set of documents $d \in \mathbf{D}$,
- set of concepts $c \in \mathbf{C}$,
- set of key phrases $k \in \mathbf{K}$,
- set of importance coefficients $f \in \mathbf{F}$ defining the relations between c and k, a function $f : \mathbf{C}\mathrm{x}\mathbf{K} \rightarrow \mathbf{R}$,
- function $g : \mathbf{D}\mathrm{x}\mathbf{K} \rightarrow \mathbf{R}$ defining the results of indexing documents (importance of a document for a given key phrase),

then it is possible to define a measure of importance of a document $d \in D$ with respect to a given concept $c \in C$ as a function $h(d,c)$, e.g. as follows:

$$h(d,c) = \sum_{k \in K} f(c,k) g(d,k)$$

Other formulae as the above weighted sum can be also used, if we interpret differently the importance coefficients $f \in \mathbf{F}$

(as fuzzy logic membership values, or aspiration levels for multiple criteria ranking). We display this measure in the user interface.

However, a more important issue is the use of such measures in overall ranking of a set of documents with respect to entire personalized ontological profile, i.e., the entire set of concepts **C**. A general way of defining a measure of importance of a document $d \in D$ to the entire profile (perspective, horizon) of the user is to treat each concept $c \in \mathbf{C}$ or, rather, each related measure $h(d,c)$ as a separate criterion of importance and then use methods of ranking related to multiple criteria decision making or to fuzzy logic; this will be the subject of further studies. A simple way is just to assume equal importance of each concept and just to sum measures $h(d,c)$ over $c \in \mathbf{C}$, or take a minimum of $h(d,c)$ over $c \in \mathbf{C}$ if each concept is considered essentially important.

4.4. Enriching Document Repository

One of basic functions of PrOnto is to support user's including documents to enrich document repository. A special addition to the Firefox search engine was developed to support this functionality – see Fig. 2.

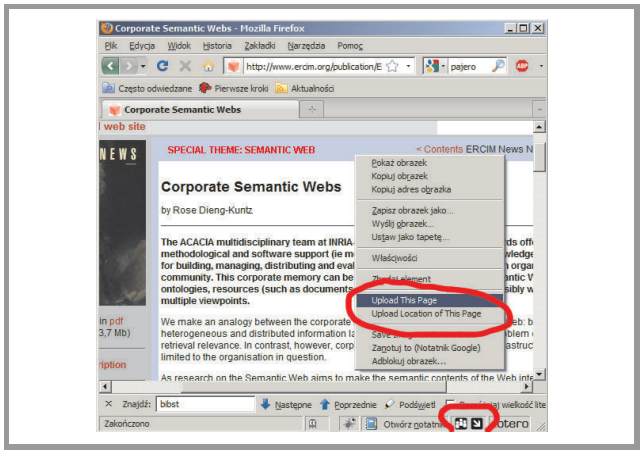


Fig. 2. Suggesting a WWW page for document repository, with marked elements of PrOnto Firefox Extension.

4.5. Multidimensional Search for Documents

PrOnto system is equipped in an advanced search engine (concerning personal names, concepts, documents, key phrases), see Fig. 3, that presents the results of search in a multidimensional structure. The results of search for documents, based on a personalized ontological profile of one of the authors of this paper, are shown on the right side of Fig. 3. The concepts, shown on the left side, come from ontological profiles of many users, but the author of this profile selected those marked by ►. When selecting a concept for more specific definition of importance coefficients f, this icon changes to ▼ (as at the concept *library*) and a set of keywords is displayed, with a simple interface to define subjective values of f. The keywords might come from the profiles of all users, or a set of key phrases originally defined by the specific user.

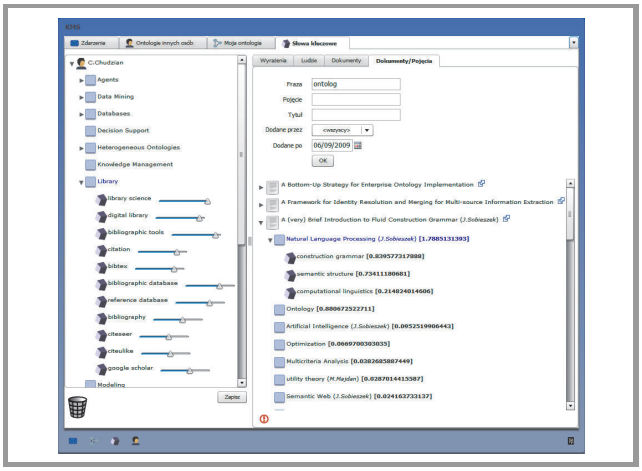


Fig. 3. Documents, concepts and key phrases.

4.6. Sharing Knowledge Using Ontological Models

Problems of accumulating, organizing and sharing sources of knowledge are addressed in computer science for a long time. Recently, however, the interest in these problems is growing because of the importance of internet or intranet as a source of information and knowledge.

This trend has many forms: *social networks*, *communities of practice*, *peer to peer networks*, *virtual research communities*, etc. In these forms, ontological engineering tools are also used. For example, system OntoShare ([31] aims at supporting knowledge exchange in a community of practice, using a common ontology constructed for this community. Users are characterized by profiles selected from this common ontology (this is a difference from our approach: we *start* from individual profiles because we assume the sovereignty of the user). System checks similarity of profiles and suggests document sharing.

Another example is project SWAP (*Semantic Web and Peer-to-Peer*) [32], [33]. The main issue in this project is *Ontology Matching*, see [34]. Another product of this project

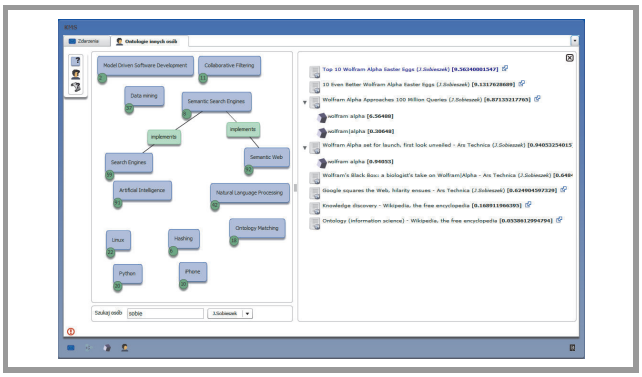


Fig. 4. Documents seen from a perspective of a given ontological profile.

is system Bibster [35] aiming at bibliographic information exchange in a distributed environment.
In the PrOnto system we assume that the users participating in a group (Virtual Research Community, VRC) approve sharing their personalized ontological profiles. Thus, one of functionalities of the system is to analyze importance of a document or a ranking of them *from another perspective* resulting from ontological profile of a different VCR member. This is shown in Fig. 4: on the left size a map of concepts is presented, on the left side a ranking list of documents, together with key phrases and corresponding values of $f(c,k)g(d,k)$.

4.7. Ontology Matching, Off-Line Analysis and Event Information

Another possibility offered by PrOnto is ontology matching. A user can see the concepts used in other ontologies than her/his own or even differences in relations between them. This is illustrated by Figs. 5 and 6.

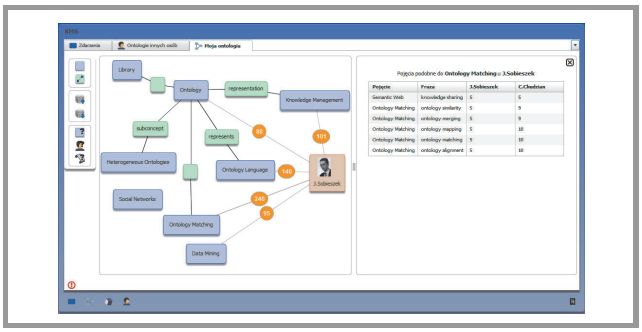


Fig. 5. Similarity of user's profiles.

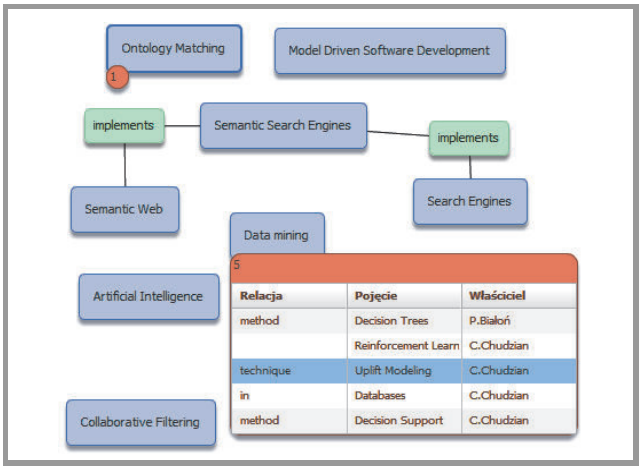


Fig. 6. Checking differences in concept relations.

Beside interactive on-line work, PrOnto system performs also off-line analyses without user's participation. The results of such analyses are presented to users in the form of a list of events, such as occurrence of similar concepts in the profiles of other users, or enriching the document repository by new documents that might be interesting for a user.

4.8. Implementation Issues

PrOnto system was programmed using exclusively open source software. Some of such open source technologies used are already broadly applied, even included into commercial systems. We used a relational data base *PostgreSQL*, *Web Application Framework Django*, script language *Python* and the environment *Adobe Flex* for creating applications *Flash*. Moreover, PrOnto uses original codes written by authors in C language.

5. Conclusions

A prototype system PrOnto was developed in the *Requested Research Project "Teleinformatic Services and Networks of Next Generation – Technical, Applied and Market Aspects", Theme Group i: Systems Supporting Regulatory Decisions: Knowledge Mining in Large Telecommunication Data Sets*. This system realizes the perspective of *human centered computing* and is based on radically personalized ontological profiles of users that, on one hand, express intuition and tacit knowledge of a single user, but on the other hand enable an interaction with ontological engineering tools and with other users in a VRC.
There are many directions of future research on this system, see, e.g., [2]. Recently, these works were included into a new project SYNAT an we started to investigate diverse ways of ranking documents with respect to a personalized ontological profile with interpretations coming from fuzzy logic and multiple criteria decision theory.

References

[1] C. Chudzian, E. Klimasara, J. Sobieszek, and A. P. Wierzbicki, "Wykrywanie wiedzy w dużych zbiorach danych: analiza tekstu i inżynieria ontologiczna. Sprawozdanie PBZ, *Usługi i sieci teleinformatyczne następnej generacji – aspekty techniczne, aplikacyjne i rynkowe, grupa tematyczna i: Systemy wspomagania decyzji regulacyjnych*: Warszawa: Instytut Łączności, 2009 (in Polish).

[2] C. Chudzian, J. Granat, E. Klimasara, J. Sobieszek, and A. P. Wierzbicki, "Wykrywanie wiedzy w dużych zbiorach danych: przykład personalizacji inżynierii ontologicznej", *Telekomunikacja i Techniki Informacyjne*, no. 1–2, 2011 (in Polish).

[3] I. Kant, *Kritik der reinen Vernunft*. 1781 (in Polish: I. Kant, *Krytyka czystego rozumu*. Warszawa: PWN, 1957).

[4] Z. Król, "The emergence of new concepts in science", in *Creative Environments*, A. P. Wierzbicki and Y. Nakamori, Eds. Springer, 2007.

[5] C. G. Jung, *Typy Psychologiczne*. Warszawa: Wydawnictwo KR, 2009 (in Polish).

[6] A. P. Wierzbicki and Y. Nakamori, *Creative Space: Models of Creative Processes for the Knowledge Civilization Age*. *Studies in Computational Intelligence*, vol. 10. Berlin: Springer, 2006.

[7] M. Polanyi, *The Tacit Dimension*. London: Routledge and Kegan, 1966.

[8] I. Nonaka, "The knowledge creating company", *Harvard Busin. Rev.*, vol. 69, pp. 96–104, 1991.

[9] I. Nonaka and H. Takeuchi, *The Knowledge-Creating Company. How Japanese Companies Create the Dynamics of Innovation*. New York: Oxford University Press, 1995.

[10] A. P. Wierzbicki and Y. Nakamori, Eds., *Creative Environments: Issues of Creativity Support for the Knowledge Civilization Age. Studies in Computational Intelligence*, vol. 59. Berlin: Springer, 2007.

[11] C. M. Bishop, *Pattern Recognition and Machine Learning*. Singapore: Springer, 2006.

[12] H.-G. Gadamer, *Warheit und Methode. Grundzüge einer philosophishen Hermeneutik*. Tübingen: J.B.C. Mohr (Siebeck), 1960.

[13] H. Ren, J. Tian, Y. Nakamori, and A. P. Wierzbicki, "Electronic support for knowledge creation in a research institute", *J. Sys. Sci. Sys. Engin.*, vol. 16, no. 2, 2007.

[14] H. Akkermans and J. Gordijn, "Ontology engineering, scientific method and the research agenda", in *Managing Knowledge in a World of Networks*, E. Motta, D. Sleeman, F. van Harmelen, V. Uren and A. Mille, Eds. Berlin-Heidelberg: Springer, 2006, pp. 112–125.

[15] M. Heidegger, *Sein und Zeit*. Halle: Niemayer, 1927.

[16] R. Mizoguchi, K. Kozaki, T. Sano, and Y. Kitamura, "Construction and deployment of a plant ontology", in *Knowledge Engineering and Knowledge Management*, R. Dieng and O. Corby, Eds. of Proc. 12th Int. Conf. EKAW 2000, Juan-les-Pin, France, 2000, pp. 113–128.

[17] O. Corcho, M. Fernández-López, and A. Gómez-Pérez, "Methodologies, tools and languages for building ontologies: where is their meeting point?", *Data & Knowl. Engin.*, vol. 46, pp. 41–64, 2003.

[18] H. S. Pinto and J. P. Martins, "Ontologies: How can They be Built?", *Knowl. Inform. Sys.*, vol. 6, pp. 441–464, 2004.

[19] E. P. Bontas and C. Tempich, "Ontology engineering: a reality check", in *The 5th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE 2006)*, R. Meersman *et al.*, Eds., vol. 4275, *LNCS*. Montpellier: Springer, 2006, pp. 836–854.

[20] A. Gómez-Pérez, M. Fernández-López, and O. Corcho, *Ontological Engineering*. Springer, 2003.

[21] J. Curtis, D. Baxter, and J. Cabral, "On the application of the cyc ontology to word sense disambiguation", in *Proc. Nineteenth Int. FLAIRS Conf.*, Melbourne Beach, USA, 2006, pp. 652–657.

[22] A. P. Wierzbicki, "On the role of intuition in decision making and some ways of multicriteria aid of intuition", *Mult. Crit. Dec. Mak.*, vol. 6, pp. 65–78, 1997.

[23] A. P. Wierzbicki, "Intuicja z perspektywy technicznej: znaczenie zasady multimedialnej i zasady emergencji", *Wiedza a intuicja*, A. Motycka Ed. Warszawa: Wydawnictwo IFiS PAN, 2008, pp. 231–264 (in Polish).

[24] H. Ren, J. Tian, A. P. Wierzbicki, Y. Nakamori, and E. Klimasara, "Ontology construction and its applications in local research communities", in *Modelling and Decision Support for Network-based Services*, J. Granat and D. Dolk, Eds., 2008.

[25] J. Tian, A. P. Wierzbicki, H. Ren, and Y. Nakamori, "Constructing an ontology for a research program", *Int. J. Knowl. Sys. Sci.*, vol. 4, no. 2, pp. 59–67, 2007 .

[26] A. Schmidt and S. Braun, "Context-aware workplace learning support: concept, experiences and remaining challenges", in *Proc. First Eur. Conf. Technol.-Enh. Learning ECTEL 06*, Springer, 2006.

[27] C. Chudzian, "Ontology creation process in knowledge management support system for a research institute", *J. Telecommun Inform. Technol.*, no. 4, pp. 47–53, 2008.

[28] J. Sobieszek, "Towards a unified architecture of knowledge management system for a research institute", *J. Telecom. Inform. Technol.*, no. 4, pp. 54–59, 2008.

[29] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry", *Communications of the ACM*, vol. 35, pp. 61–70, 1992.

[30] A. V. Aho and M. J. Corasick, "Efficient string matching: an aid to bibliographic search", *Communications of the ACM*, vol. 18, pp. 333–340, 1975.

[31] J. Davies, A. Duke, and Y. Sure, "Ontoshare – an ontology-based knowledge sharing system for virtual communities of practice", *J. Universal Comput. Sci.*, vol. 10, no. 3, pp. 262–283, 2004.

[32] M. Ehrig, P. Haase, B. Schnizler, S. Staab, C. Tempich, R. Siebes, and H. Stuckenschmidt, *SWAP: Semantic Web and Peer-to-Peer Project Deliverable 3.6 Refined Methods*, 2003 [Online]. Available: http://swap.semanticweb.org/public/Publications/swap-d3.6.pdf

[33] M. Ehrig, C. Tempich, and Z. Aleksovski, *SWAP: Semantic Web and Peer-to-Peer Project Deliverable 4.7 Final Tools*, 2004 [Online]. Available: http://swap.semanticweb.org/public/public/Publications/swap-d4.7.pdf

[34] M. Ehrig, *Ontology Alignment: Bridging the Semantic Gap (Semantic Web and Beyond)*. New York: Springer, 2006.

[35] J. Broekstra, M. Ehrig, P. Haase, F. Van Harmelen, M. Menken, P. Mika, B. Schnizler, and R. Siebes, "Bibster – a semantics-based bibliographic peer-to-peer system", in *Proc. Third Int. Semantic Web Con.*, Hiroshima, Japan, 2004, pp. 122–136.

Cezary Chudzian received his M.Sc. in computer science from the Warsaw University of Technology in 2002. He is a researcher at the National Institute of Telecommunications. Currently he works on his Ph.D. in the area of knowledge management. His main scientific interests include: practical applications of knowledge discovery techniques, machine learning theory, knowledge management, global optimization, and advanced software engineering.
E-mail: C.Chudzian@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Janusz Granat received his M.Sc. in control engineering (1996) and his Ph.D. (1997) in computer science from the Warsaw University of Technology, Poland. He holds a position as an Assistant Professor at the Warsaw University of Technology, and is the leader of a research group on applications of decision support systems at the National Institute of Telecommunications in Warsaw. He lectured decision support systems and various subjects in computer science. His scientific interests include data mining, modeling and decision support systems, information systems for IT management. Since 1988 he has been cooperating with IIASA. He contributed to the development of decision support systems of DIDAS family and the ISAAP module for specifying user preferences. He has been involved in various projects related to data warehousing and data mining for telecommunica-

tion operators. He was also involved in EU MiningMart project.
E-mail: J.Granat@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Institute of Control and Computation Engineering
Warsaw University Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Edward Klimasara obtained his M.Sc. in Mathematics from Warsaw University, Poland in 1977. He is a senior specialist at National Institute of Telecommunications in Warsaw. Currently he takes part in projects: SIPS, EDFAS, SYNAT. His main scientific interest includes: knowledge management, IT systems for telecommunications and healthcare, network security, quality management, simulation techniques.
E-mail: E.Klimasara@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Jarosław Sobieszek received his M.Sc. degree in computer science from Warsaw University of Technology, Poland, in 2002. Currently he is a researcher at National Institute of Telecommunications, where he prepares his Ph.D. thesis in the area of knowledge management. His research interests include machine learning, artificial intelligence, knowledge management and model-based approaches to software development.
E-mail: J.Sobieszek@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Andrzej Piotr Wierzbicki got his M.Sc. in 1960 in telecommunications and control engineering, Ph.D. in 1964 in nonlinear dynamics in control, and D.Sc. in 1968 in optimization and decision science. He worked as the Dean of the Faculty of Electronics, Warsaw University of Technology (WUT), Poland (1975–1978); the Chairman of Systems and Decision Sciences Program of International Institute for Applied Systems Analysis in Laxenburg n. Vienna, Austria (1979–1984). He was elected a member of the State Committee for Scientific Research of Republic of Poland and the Chairman of its Commission of Applied Research (1991–1994). He was the Director General of the National Institute of Telecommunications in Warsaw (1996–2004). He worked as a research Professor at Japan Advanced Institute of Science and Technology (JAIST), Nomi, Ishikawa, Japan (2004–2007). Beside teaching and lecturing for over 45 years and promoting over 100 master's theses and 20 doctoral dissertations at WUT, he also lectured at doctoral studies at many Polish and international universities. Professor Wierzbicki is an author of over 200 publications, including 14 books, over 80 articles in scientific journals, over 100 papers at conferences; the author of 3 patents granted and industrially applied. His current interests include vector optimization, multiple criteria and game theory approaches, negotiation and decision support, issues of information society and knowledge civilization, rational evolutionary theory of intuition, theories of knowledge creation and management.
E-mail: A.Wierzbicki@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Paper

New SEAMCAT Propagation Models: Irregular Terrain Model and ITU-R P. 1546-4

Dariusz Więcek and Dariusz Wypiór

National Institute of Telecommunications, Wrocław, Poland

Abstract—Implementation of the ITU-R P.1546-4 and ITM propagation models for SEAMCAT prepared and developed in the National Institute of Telecommunications Poland is presented. Results of our research encompasses methodology, implementation and verification of plug-ins into the SEAMCAT software are shown.

Keywords—electromagnetic compatibility, interferences, propagation, SEAMCAT, spectrum engineering.

1. Introduction

Electromagnetic compatibility analyses of wireless systems have important role in process of radio network planning, optimization and frequency availability checking. Greater than ever number of wireless transmission devices work on the same or adjacent radio channels, what leads to necessity of using dedicated computer software carrying out computer calculations and analysis of EM compatibility aspects in very similar conditions as it is in real world. Spectrum Engineering Advanced Monte Carlo Analysis Tool (SEAMCAT) is one of the most advanced and most popular tools in Europe. The SEAMCAT software is generally developed jointly within countries belong to the European Conference of Postal and Telecommunications Administrations (CEPT) under the framework of European Communication Office (ECO). The application is developed by SEAMCAT Technical Group (STG). It is freely distributed software developed on the base of open-source license and provides opportunities for creating scenarios in which radio systems are tested for quantitative and statistic probabilities of interferences. The significant impact on obtained results have different propagation methods dedicated for different propagation conditions. In SEAMCAT besides using a few in-built methods [1] there are also methods developed by user as propagation plug-ins. In this work, the ITU-R Recommendation P.1546-4 and Irregular Terrain Model (ITM) propagation models implementation are presented. The latter one is also called as Longley-Rice method. Those methods were official attached to SEAMCAT software after their successful verification made by ECO's experts and are now available on the SEAMCAT web pages [2].

2. Electromagnetic Compatibility of Radio Systems

Accordingly to the rapid increasing of radio systems working on the same or adjacent frequency channels, in the same or neighboring geographical areas risk of mutual harmful interferences can occur. It is consequence of the propagation of radio waves in all directions, crossing the borderlines and overlapping the same radio spectrum by different radio systems. As a result, receiver receives wanted as well as many interfering signals having various characteristic of electric field strength levels. It often leads to the situations where despite of sufficient wanted signal levels (exceeding the required minimum signal level threshold) in same location or in some time conditions the correct reception is not possible due to interferences.

In order to minimize this effect computer-aided EMC analyses are required before introducing new wireless system to guarantee minimization of the interferences probabilities. In this way, we can obtain an assessment of mutual interferences probabilities depending on varying parameters of individual transmitters (localization, transmitter antenna characteristic, frequency, signal character). Those analyses give guarantee of sufficient EM compatibility and prevent from the occurrence of interferences even at the stage of the preparation or evaluation possibility of allocations frequency bands by administration, international regulators bodies or telecommunications operators. It can be done by defining specific conditions, e.g., territorial separations or specific requirements of radio emissions (e.g., emission masks, receiver selectivity, maximum transmitter highs).

In order to get accurate information what emission parameters should be used or where transmitters ought to be localized, we need to develop new computer-assisted tools corresponding to increasing amount of new radio system and networks and their technical parameters. Also country administration can use such kind of software during multilateral international coordination agreements concerning frequency allotments and assignments. For the purpose of analysis the basic knowledge of input parameters, propagation models, technical aspects of receivers and transmitter stations are required. Also proper methodology of computing in dedicated software is important. In case of congestion of radio spectrum, when we need to use electro-

magnetic spectrum as efficient as never before, continuous researches lead to implementation of improved propagation methods are important. In such situation undertaken work aimed at SEAMCAT development is to be very helpful.

3. SEAMCAT Methodology

Creating an appropriate scenario is the base of conducting desired calculation (simulation) in the SEAMCAT application. Each from those scenarios has to contain one victim link and at least one interfering link. In Fig. 1 relation among various types of radio links are presented. Wanted transmitter (Wt) and victim receiver (Vr) are included in victim link. Besides, there could be any amount of interfering links which contains interfering transmitter (It) and wanted receiver (Wr). Arrows present the directions of radio paths, for which appropriate attenuation of signal levels are computed.

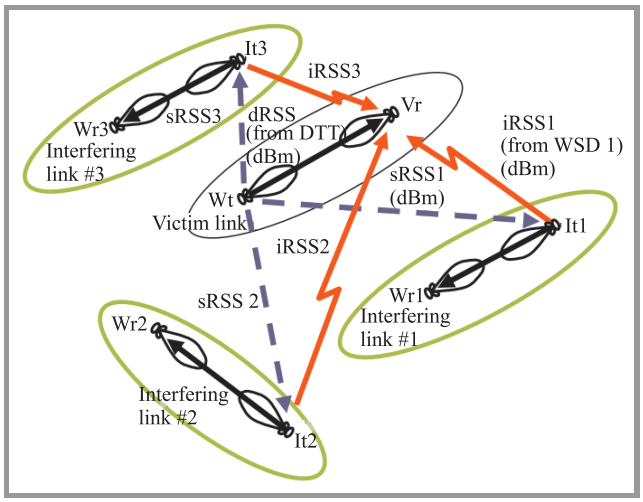


Fig. 1. Signal relations among various types of links [3].

Each of them is described by series of values presenting system parameters, spatial location information and propagation models. SEAMCAT works in Monte Carlo methodology, what means that many events are generated and results return with same statistics. Users have right to choose the number of simulation event. In every scenario, there is a possibility of changing input parameters (particularly station localization, frequency, emission power etc.) according to a pattern set by users. Moreover, in each of the event levels of signals are computed. We can get following results:

- desired received signal strength (dRSS),
- interference received signal strength (iRSS),
- sensing received signal strength (sRSS); it is used only by cognitive radio systems features.

They form the basis for further interference probabilities calculations. The probability of interferences occurring p_I Eq. (1) accordingly to the signal-to-interference ratio (C/I) criterion for single event and form single interfering link is the conditional probability, where $sens_{RX}$ is the receiver sensitivity.

$$p_I = 1 - P\left(\frac{dRSS}{iRSS} > \frac{C}{I}\middle| dRSS > sens_{RX}\right) \tag{1}$$

It is important that SEAMCAT does not give opportunity to use digital elevation/terrain map (DEM/DTM) and all calculation are carried out over a flat terrain, what comes from the fact that the application is to be especially used by government administrations in the work of establishing general rules of coordination, legislative and policy governing the use of radio spectrum on national and international levels. Such simulations are enough detailed for the purposes because it is usually established as worst-case situations from interference point of view.

Besides standard settings of radio networks, there is also possibility of conducting simulation for some predefined radio systems such as CDMA or LTE. What is more, at present STG is working on the introduction of appropriate solutions, which will take part in cognitive radio EMC analyses. Now users have possibility to do only some simple calculation with cognitive radio features, which are still being discussed and developed within ECO forum.

4. Propagation Models

4.1. General

The proper received signal strength prediction is based on using correct propagation models, which are properly chosen, i.e., depends on the type of system or feature of terrain where wireless systems works. During implementation we have analyzed a few methods, focusing on those, which are especially used to calculation over land areas and frequencies in the range 100 MHz till 3 GHz basically. Generally, propagation methods are divided into:

- point-to-area methods,
- point-to-point methods.

It is worth mentioning that first group of methods is used generally for larger areas and general estimations. The result is estimated based on general propagation rules, particular earlier measurements statistics and more or less accurate statistical terrain math descriptions. The latter group of methods contains greater number of input parameters and use sophisticated math dependencies for various physics phenomena and weather conditions (diffraction, scatter, rain cell, vapour attenuation etc.) but usually required detailed DEM/DTM maps and is basically used at detailed network planning stage.

4.2. ITM Propagation Model

Irregular terrain model (ITM) was developed in the 60's of the XX century in the USA for designing and planning

analogue broadcast terrestrial television systems [4]. Later, it has been extended to work in wider frequency range for other systems types. Our implementation into SEAMCAT works as point-to-area method. The average height of profile between receiver and transmitter is described statistically by irregularity terrain parameter. Calculations are conducting for 3 propagation mechanism: line-of-sight, diffraction and scatter. That fragmentation has been made due to the dominant physical phenomena. Furthermore, ITM method contains some empirical dependence. It is very interesting model because of wide frequency range (VHF, UHF, SHF) and wide series of parameters allowing detailed parameterizations listed below.

- System parameters:
  - frequency,
  - height of antenna masts,
  - distance,
  - polarization.
- Terrain parameters:
  - irregular terrain parameter,
  - conductivity of ground,
  - relative permittivity,
  - surface refractivity
  - radio climate.
- Deployment parameters:
  - siting criteria.
- Statistical parameters:
  - percentage of localization,
  - percentage of time,
  - confidence level,
  - mode of variability,
  - signal standard deviation.

ITM model has following limitation:

- frequency: 20 MHz – 20 GHz,
- distance: 1 – 2000 km,
- height of antenna: 0.5 – 3000 m.

Input parameters presented above are treated as quantitative and qualitative parameters. Accordingly, i.e., the first are height of antenna, frequency etc., and the latter there are radio climate, siting criteria etc. In implementation seven radio climate zones are included: equatorial, continental subtropical, maritime subtropical, desert, continental temperate, maritime temperate over land and maritime temperate over the sea. Irregularity parameter dh is given in meters and it is assumed that for a flat terrain is equal 0 m, for plains $dh = 30$ m, for hills $dh = 90$ m, for mountains $dh = 200$ m and rugged mountains $dh = 500$ m. Conductivity of ground, relative permittivity have also such assumption. It is suggested to use following values of conductivity of ground σ, and relative permittivity εr:

- σ = 0.005 S/m, εr = 15 for average ground,
- σ = 0.001 S/m, εr = 4 for poor ground,
- σ = 0.02 S/m, εr = 25 for good ground,
- σ = 0.01 S/m, εr = 81 for fresh water,
- σ = 5 S/m, εr = 81 for sea water.

Surface refractivity depends on a climate zone. Siting criteria is a qualitative description of the care with terminals sites is chosen to improved communications. Percentage of time and localization determine the fraction of time or localization where the attenuation will not exceed the certain value. Confidence level is qualitative parameters describing other variables. For instance, if it is 50% of time, 70% of localization, and confidence levels equal to 90% it means that: in 90% of cases (situations) there will be at least 70% of the locations where the attenuation will not exceed certain value for at least 50% of the time [5]. There are also four mode of variability (single, individual, mobile, broadcast).

4.3. ITU-R P.1546-4 Propagation Model

ITU-R P.1546-4 Recommendation [6] offers field-strength predictions in point-to-area mode for terrestrial services over land and sea in the frequency range 30 MHz to 3 GHz. The method based on series statistically prepared measured field strength values (e.g., values for 50% time and 50% locations) presented as curves included into the Recommendation and widely used for broadcast, land mobile services as well as other wireless systems in this frequency range. Results of statistically prepared measurements data (median, 10% of time, 1% of time exceeding wanted value) in mean temperate climatic conditions obtained for land, cold and warm sea are tabulated and included into the Recommendation. Curves are divided into land and see (cold and warm) respectively and represent the following data:

- frequency (f),
- height of transmitting/base antenna (h),
- time percentage (t),
- distance (d).

Due to lack of ability of using DEM/DTM in SEAMCAT, as mentioned earlier, the discussed implementation of the model omitted sea path calculation and detailed terrain corrections. In this way in our implementation the percentage of sea path length over a profile is not used.
The calculation of field-strength values from curves is based on interpolation/extrapolation algorithm. The more input parameters, which have not fit into exact tabulated field

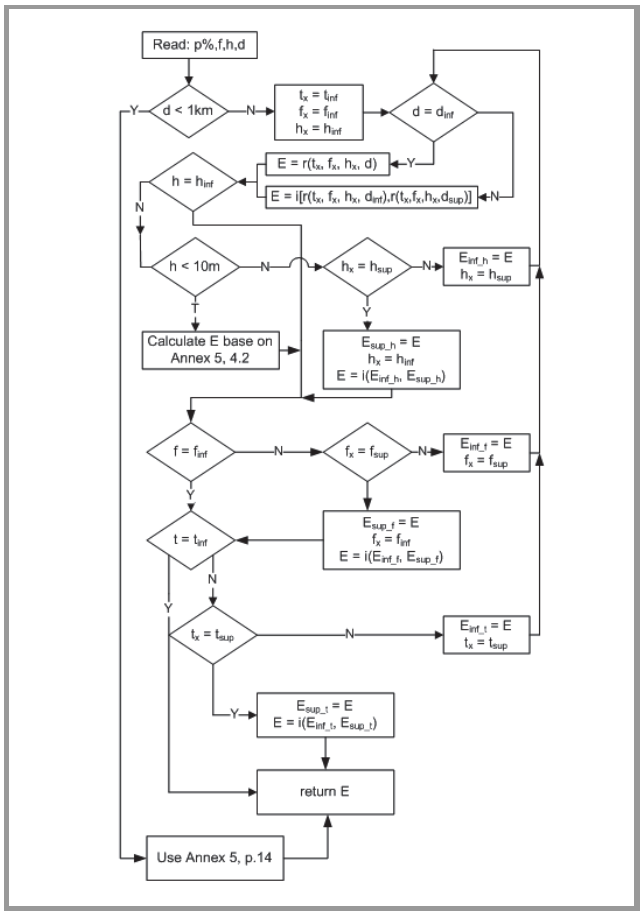


Fig. 2. The interpolation of field-strength value algorithm.

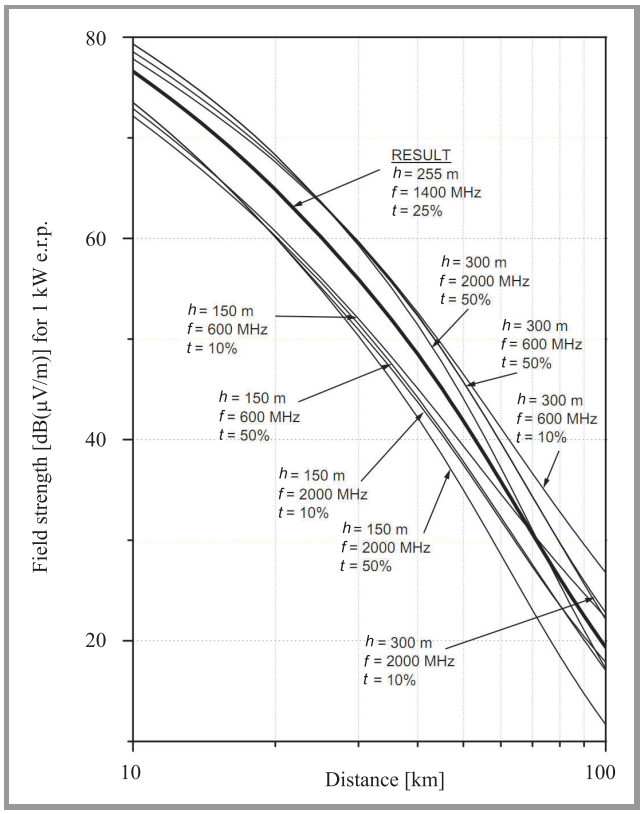


Fig. 3. The sample of quadruple interpolation.

strength values, the more interpolation operations have to be done. In the worst case, if none of those parameters are in direct tabulated form algorithm should perform 15-interpolation. The algorithm for solving the described problem was prepared (Fig. 2). The purpose of it is computing expected field-strength value exceeded at 50% of location within area of 500×500 m. The additional exceptions, as path with distance less the 1 km or transmitting antenna height less then 10 m, are also included. For those cases, the Recommendation introduced additional prediction methods.

Inf and *sup* indexes in algorithm's block boxes denote respectively the nearest tabulation value less then inputted and the nearest tabulation value greater then inputted. Function *i* is discussed interpolation function and function *r* read field-strength value from curves. A sample obtained result of the foregoing procedure is presented in Fig. 3, where following parameters were used: f = 1400 MHz, h = 225 m, t = 25%, distance 10 – 100 km (step 1 km),
In this case, four parameters have not exact corresponding value in the appropriate tabulated data and should be interpolated. The algorithm was implemented and all necessity actions and the results for the example were returned. Figure 3 shows the result and also all data, read directly from tables, which are used for interpolation calculations.

4.4. Practical Implementation

SEAMCAT offers possibility to use plug-in propagation models. Each of them has to be written in Java programming language, compiled and put into SEAMCAT home directory. The template for plug-ins development with some basic code containing sample class is able to download from the official online user manual [2]. Users are able to make optional changes in that code adequacy to their needs. Propagation plug-in activity can be illustrated as a "black box" (Fig. 4). On its input, the basic system parameters are

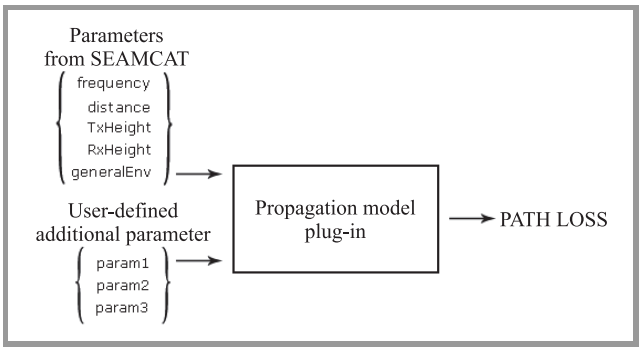


Fig. 4. Propagation plug-in treated as "black box".

introduced – reading directly from the workspace. What is more, every user has ability to introducing up to 3 additional user-defined optional parameters. The radio path loss is designated after each event and returned to workspace as preliminary value of dRSS, iRSS or sRSS depending on which signal attenuation was computed.

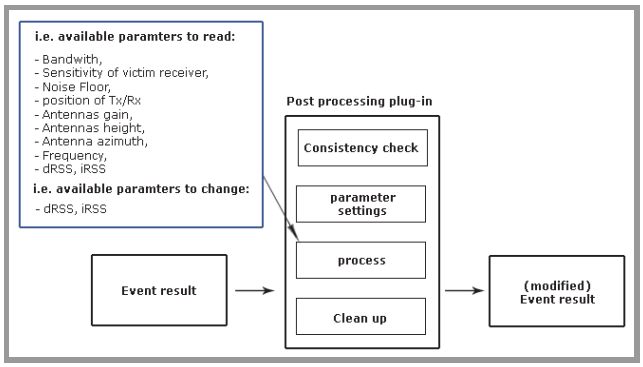


Fig. 5. Postprocessing plug-in treated as "black box".

In case of implementation the ITM propagation plug-in some problems were occurred. They were connected mainly to maximum 3 user-define parameters, which developer may use. That issue was solved by using second type of plug-in, which SEAMCAT offers namely postprocessing plug-in. It allows performing operations for various parameters, both input and output as instance dRSS (Fig. 5). Through this extension postprocessing plug-in is treated in our solution as a configuration panel for all additional input parameters. Presented solution is not the best because by changing at least one input parameter it can be introduced an incorrect result in first event – all parameters shall be set after first event. However, it can be omitted if simulation event number is very large or in case, when first one event simulation will be made, which will be treated as a pre-configuration simulation. After that manipulation, the event in second simulation will be correct.

5. Tests

5.1. ITM Model

During testing phase many simulations were conducted for various input parameters and type of terrain as well as radio climates. They aimed to verifying obtained results and source code correctness. Results, for certain input parameters (Table 1), derived from our implementation was compared with results from ITM – Irregular Terrain Model 1.5.5 software (Fig. 6). A very good correlation of both applications was received. Details can be found in the PBZ Report [7]. Other sample results are presented in Figs. 7 and 8. Test workspace was configured in all simulation in the same way (Table 1). All radio system works on the frequency 900 MHz, the heights of transmitter and receiver antennas have respectively 100 m and 10 m. In the particular calculation, input parameters have been changed to those presented in Figs. 7 and 8. Furthermore, for radio climate changing surface refractivity was also varied and set on 280, 301, 320 respectively for desert, continental temperate, maritime temperate over land radio climate.

Table 1
Input values for ITM model test

Parameter	Value	Units
Frequency	900	MHz
Transmitting antenna height	100	m
Receiving antenna height	10	m
Surface refractivity	301	N-units
Terrain irregularity parameter *dh*	90	m
Conductivity of ground	0.005	S/m
Relative permittivity	15	
Antenna polarization	Horizontal	
Siting criteria	Random	
Radio climate zone	Continental temperate	
Percentage of time	50	%
Percentage of localization	50	%
Confidence level	90	%
Mode of variability	Broadcast	
Standard deviation	0	dB

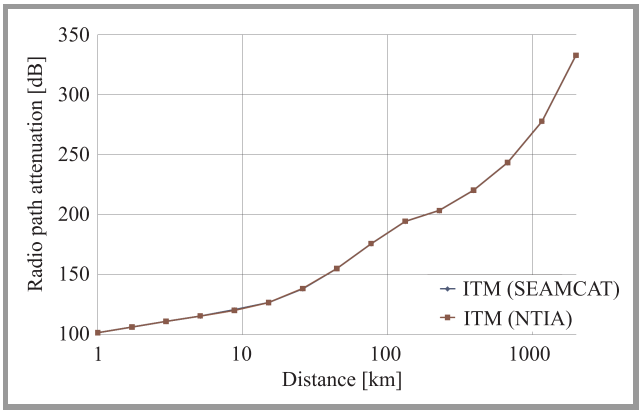


Fig. 6. Radio path attenuation against distance for comparison of ITM model applications.

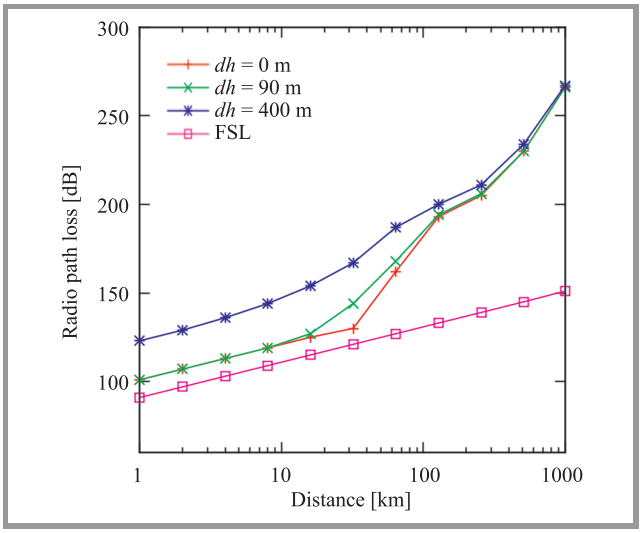


Fig. 7. Results of radio path loss for various irregular terrain parameters *dh*.

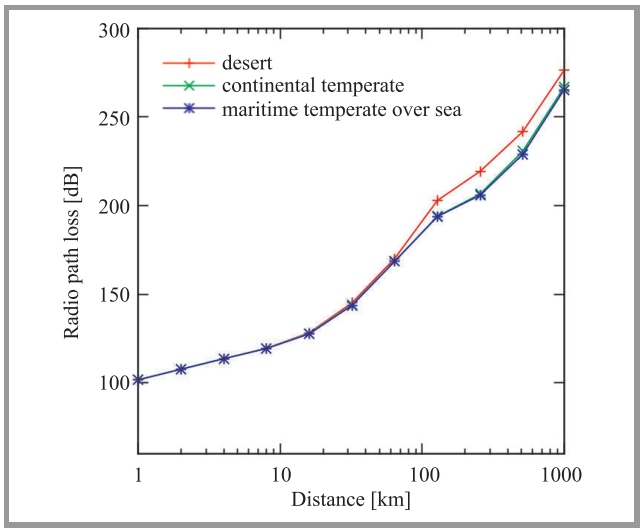


Fig. 8. Results of radio path loss for various radio climates.

5.2. ITU-R P.1546-4

The test procedure was oriented on correct radio path loss returning in dependencies on various input parameters. Obtained results were compared firstly to the version P.1546-1 from previous version of SEAMCAT. Some sample results are shown in Figs. 9–12.

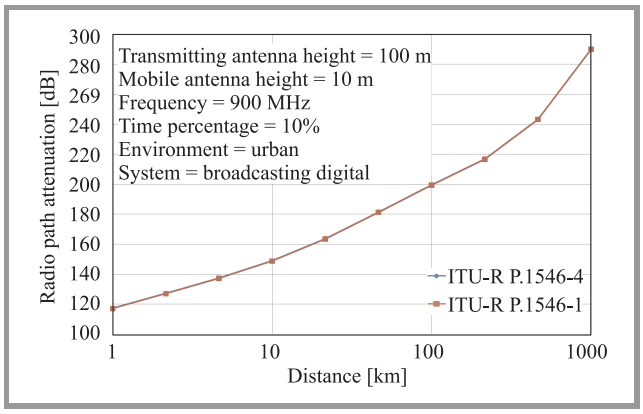


Fig. 9. Radio path attenuation against distance for P.1546-1 and P.1546-4.

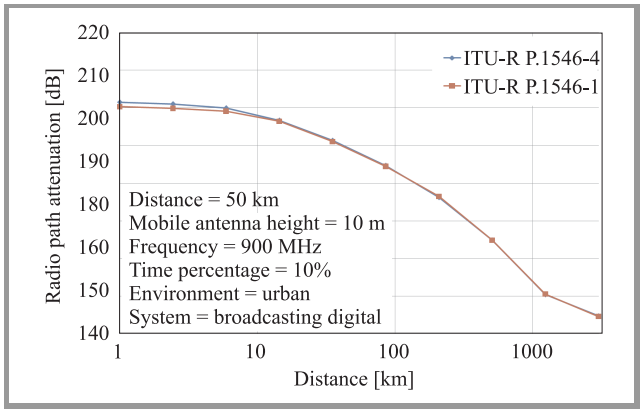


Fig. 10. Radio path attenuation against transmitting antenna height for P.1546-1 and P.1546-4.

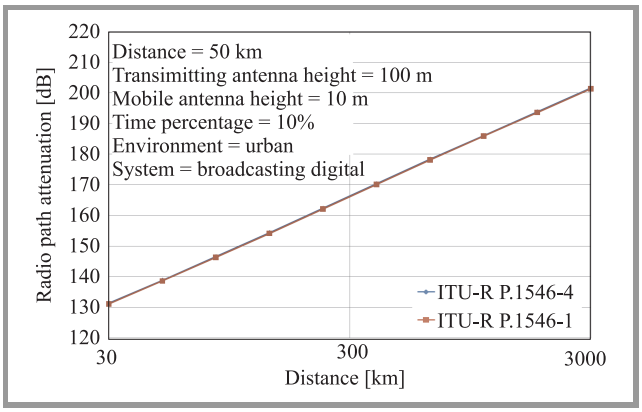


Fig. 11. Radio path attenuation against frequency for P.1546-1 and P.1546-4.

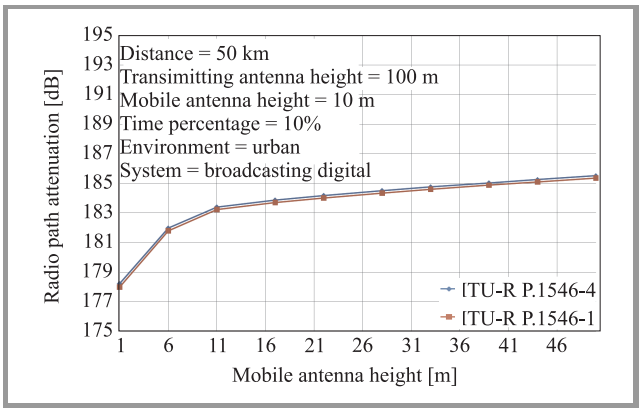


Fig. 12. Radio path attenuation against percent of time for P.1546-1 and P.1546-4.

The result obtained from comparison of both models confirmed proper implementation of the P.1546-4. It is worth to note that in P.1546-4 interpolation between tabulated points taken from curves is used while the implementation of recommendation P.1546-1 in SEAMCAT computes the curves from some equations for certain coefficients, but both methods should generate the same value for single case calculations of basic variables (e.g., distance, frequency, transmitting antenna high) as it is shown above in Figs. 9–12.

The second tests phase led to calculations of results for some group of input parameters (like time percentage, environment in vicinity of receiver etc). The workspace, during tests, was configured default as follow:

- frequency: 900 MHz
- transmitting antenna height: 100 m
- receiving antenna height: 10 m
- environment: urban
- time percentage: 50%
- area: 500 × 500 m
- number of event: 20 000.

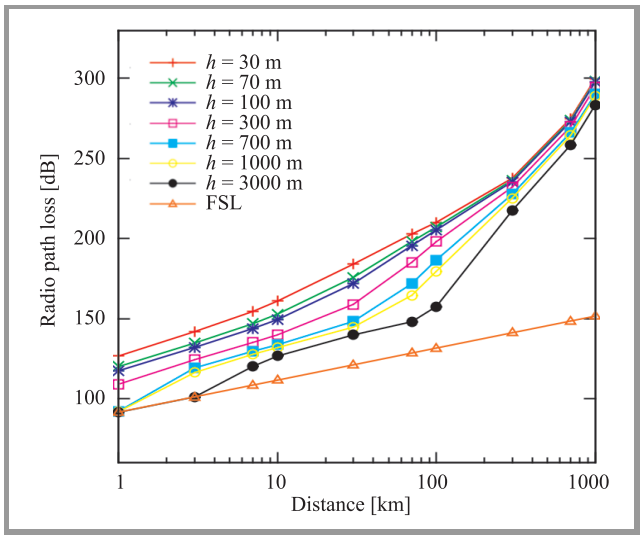


Fig. 13. Radio path loss for various transmitting/base station antenna heights *h*.

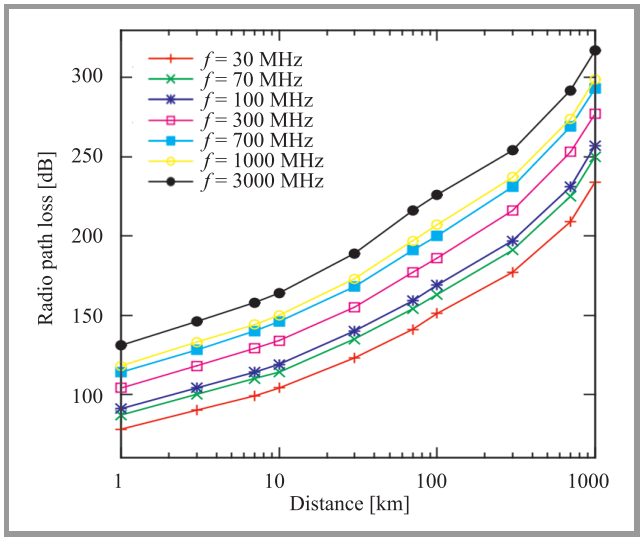


Fig. 14. Radio path loss for various frequencies *f*.

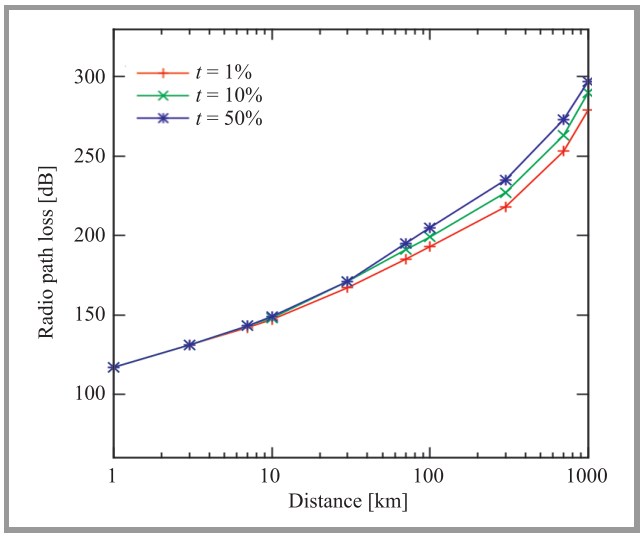


Fig. 15. Radio path loss for various time percentages *t*.

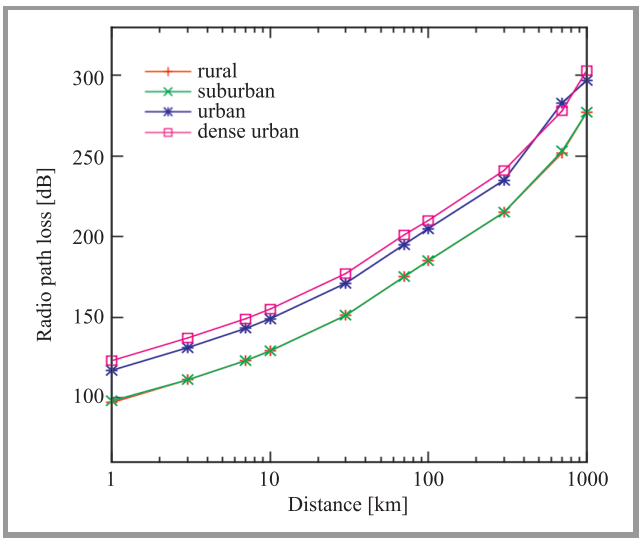


Fig. 16. Radio path loss for various clutter types.

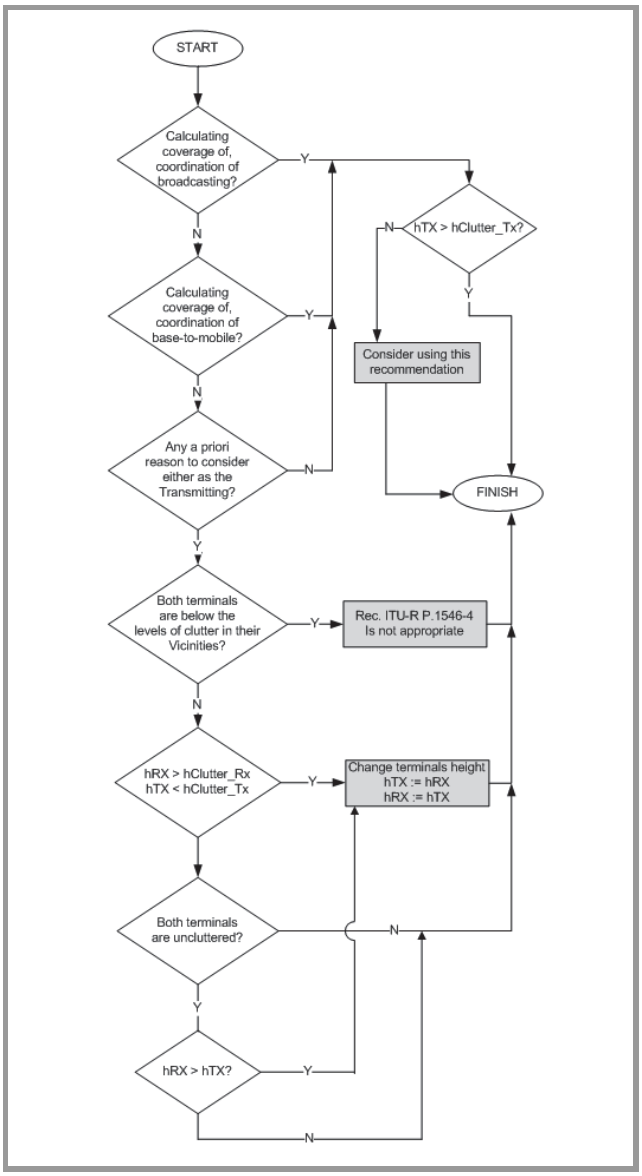


Fig. 17. The terminal designation algorithm.

Results which are presented (Figs. 13–16) were obtained by changing parameters accordingly to those written in figures.
Additionally ITU-R P.1546-4 Recommendation has introduced *the terminal designations* in Annex 5 Paragraph 1.1. In any cases where there are not *a priori* reason to consider terminal as transmitting/base, and user is aware this paragraph can be omitted. This way, all results shown earlier omitted that outlines. Algorithm (Fig. 17) presented step-by-step procedure how the application of the procedure ought to be use. It is important to use them, because of lack of reversible with respect to designations of transmitting and receiver station and in view of wording from Recommendation "It is primarily intended for use with broadcasting and mobile services where the transmitter/base antenna is above the level of local clutter" [6].

6. Application Example

Some analyses using developed implementation of new SEAMCAT propagation methods were conducted during the validating test stage. Below some examples of the results are presented.

6.1. The Influence of BS PMR into DVB-T

Typical example of analysis which can be carrying on into SEAMCAT software is impact of one wireless system into another, working in the same or adjacent frequency bands. Below modified scenarios taken from the ECC Report 104 [8] were used. Influence of BS PMR on DVB-T receivers were analysed there. DVB-T receivers work on channel 21. The values of the system setting are present in Table 2 and Table 3.

Table 2
DVB-T parameters

Height of TV transmitting antenna	100 m
Height of TV receiving antenna	10 m
EIRP	43 dBW
Bandwidth	8 MHz
TV channel	21 (470–478 MHz)
RX sensitivity	–79 dBm

Table 3
PMR parameters

Height of BS antenna	30 m
Height of MS antenna	1.5 m
EIRP	22 dBm
Bandwidth	12,5 kHz
Centre frequency	469.99375 MHz

The digital television receivers were at distance 30 km from transmitting antenna (DVB-T). Mobile base station of PMR was located in random distances from television receiver, but no further than 10 km. In the presented results there are probabilities of the interferences. The unwanted and blocking signals calculation models used in SEAMCAT were taken into account.

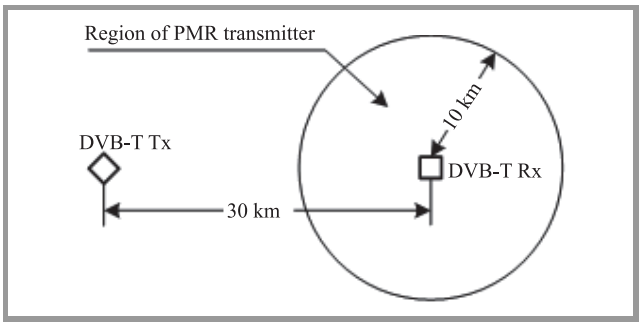


Fig. 18. Location of transmitters and receivers in analysis scenario.

The radio path loss between DVB-T transmitter and receiver was computed by ITM method. Input parameters are presented in Table 4. The signal attenuations between PMR base stations and DVB-T receivers were calculated by using Extended-Hata propagation model. Probabilities of interferences were computed for modification of individual parameters of ITM method. In this simulation 100 000 events were taken into account.

Table 4
Probability of interferences for different terrain irregularity parameters

Terrain irregularity parameter	Interference probability
0 m	0.03%
90 m	0.1%
200 m	0.2%
450 m	1.38%

As we can observe, the higher terrain irregularity parameter value the less signal level at receiver input is, what resulting in decreasing of the signal to noise plus interference ratio C/(N+I) and corresponding probability of interference. In Table 5 results for different radio climate zones are also shown. The terrain irregularity parameter was set to 250 m. In this simulation 500 000 events were used.

Table 5
Probability of interferences for different terrain radio climate zones

Radio climate zone	Interference probability
Equatorial	0.26%
Continental subtropical	0.29%
Maritime subtropical	0.24%
Desert	0.32%
Continental temperate	0.30%
Maritime temperate over land	0.28%

It can be noticed that in such propagation zones where additional propagation effects occurs (and better propagation conditions are), higher useful signal levels at the same distances exist and then interference probability decrease (e.g., maritime versus land).

7. Summary

SEAMCAT as an open-source licensing software delivers many additional manners to make its closer to user needs either by using plug-ins or editing the source code.

The ITU-R P. 1546-4 and ITM propagation models were implemented into official SEAMCAT application. Those are important methods both in Europe and in the USA. The enrichment of the software allows to extending SEAMCAT ability. It seems to be interesting to use ITM model because it allows introducing some quantitative values which describing details about terrain in case where no DEM/DTM maps in SEAMCAT are used. Such solution may lead to more detailed EMC calculations increasing spectrum efficiency in the areas where some special terrain obstruction exists and where no such detailed terrain descriptions were possible.

In the National Institute of Telecommunication Poland, the P.1546-4 method with digital terrain map reading ability for SEAMCAT have been also developed as an additional function however such method is not used in official SEAMCAT version. In such case the postprocessing plug-ins were used. It allows introducing coordinates information and attaches maps files and gives an opportunity to make calculation with very good precision between transmitters and receivers, as we can compute the exact great circle distances. It may be interesting for others to use such solution and it may be worth to extend SEAMCAT with such functionality.

In future, it could be expected that SEAMCAT developers offer ability of using more then three user-defined parameters for propagation plug-ins in order to offer usage of advanced many-parameters propagation models as it was in the case of ITM model.

These days where more and more different wireless systems are introduced in the same or adjacent frequency bands such open, powerful and flexible software for electromagnetic compatibility analysis is very useful on research, scientific or radio spectrum policy levels as well as for preparation the decisions about introducing new radio systems which should be well prepared after such detailed EMC analysis performed and proper results-evaluation.

Acknowledgment

The paper is prepared under Research Project entitled: "Next Generation Services and Networks – technical, application and market aspects" contracted by the Polish Ministry of Science and Higher Education (no. PBZ-MNiSW-02/II/2007), authors: J. Sobolewski, D. Więcek, J. Wroński, B. Gołębiowski, D. Niewiadomski, R. Strużak, and D. Wypiór.

References

[1] D. Więcek, B. Gołębiowski, J. Sobolewski, and D. Wypiór, "Opracowanie i oprogramowanie modułów aplikacji SEAMCAT do wykonywania symulacyjnych badań kompatybilności międzysystemowej", *Conference PBZ-MNiSW-02/II/2007*, Warszawa, Polska, 2010.

[2] www.seamcat.org

[3] SEAMCAT Handbook, January 2010, www.ero.dk/seamcat

[4] A. G. Longley, "Radio propagation in urban areas", OT Report 78-144, 1978.

[5] G. A. Huffor, A. G. Longley, W. A. Kissick, "A guide to the use of the ITS irregular terrain model in the area prediction mode", NTIA Report 82-100, 1982.

[6] "Method for point-to-area predictions for terrestrial services in the frequency range 30 MHz to 3000 MHz", ITU-R 1546-4, Geneva, 2009.

[7] J. Sobolewski, D. Więcek, J. Wroński, B. Gołębiowski, D. Niewiadomski, R. Strużak, and D. Wypiór, "Sprawozdanie merytoryczne nr 4 i nr 5 z realizacji części Projektu Badawczego Zamawianego Usługi i sieci teleinformatyczne następnej generacji – aspekty techniczne, aplikacyjne i rynkowe", IŁ PIB, O/Wrocław, 2011.

[8] "Compatibility between mobile radio systems operating in the range 450-470 MHz and digital video broadcasting – terrestrial (DVB-T). System operating in UHF TV channel 21 (470–478 MHz)", ECC Report 104, Amstelveen, June 2007.

Dariusz Więcek received the M.Sc. and Ph.D. degrees in Telecommunications Engineering from Wrocław University of Technology, in 1992 and 2006, respectively. He joined National Institute of Telecommunications Poland in 1993 where currently he is Head of Spectrum Engineering and Management Section. His research areas include radio systems networks planning and optimization, broadcasting systems, compatibility analysis of different radio systems, cognitive radio, opportunistic and white space radio, dynamic spectrum access and spectrum engineering and management on national and international (CEPT, ITU, WRC, IEEE) levels. He was involved in preparation of the digital broadcasting switchover strategy in Croatia and digital plans in Poland. Dr. Więcek is member of Management Committee COST IC0905 (TERRA – Techno-Economy and Regulatory Aspects of Cognitive Radio Systems). He manages many projects for companies from telecommunications industry sector. He

was delegate of the Republic of Poland to the conferences: ITU WRC2007, ITU RRC06, ITU RRC04, ITU WRC2000, CEPT Chester97, CEPT Wiesbaden95. He is member of Section of Electromagnetic Compatibility of Electronic and Telecommunications Committee of the Polish Academy of Science (PAN), senior member of IEEE and v-Chair of Polish Chapter of the IEEE EMC Society. He takes part in reviewing and evaluation process of EU research projects. Dr. Więcek chaired organizing committees of the Wrocław International Symposium and Exhibition on EMC in years 2002 and 2004. He was published more than 40 scientific papers in journals and conference proceedings.

E-mail: D.Wiecek@il.wroc.pl
National Institute of Telecommunications
Swojczycka st 38
51-501 Wrocław, Poland

Dariusz Wypiór received the B.Sc. in Electronics and Telecommunications from Wrocław University of Technology, in 2010. Currently he is with the National Institute of Telecommunications Poland where he works as an research engineer in the Electromagnetic Compatibility Department, Spectrum Management and Engineering Section. His interests cover propagation modeling, radio networks planning, spectrum engineering and software development.

Email: D.Wypior@il.wroc.pl
National Institute of Telecommunications
Swojczycka st 38
51-501 Wrocław, Poland

Paper

Technical Aspects Outline for the Strategy of Launching Digital Broadcasting in Poland on Wave Bands Below 30 MHz

Andrzej Dusiński and Jacek Jarkowski[a,b]

[a] Institute of Radioelectronics, Warsaw University of Technology, Warsaw, Poland
[b] National Institute of Telecommunications, Warsaw Poland

Abstract—The article discusses the state of art knowledge concerning the introduction of DRM in the world and prospects for its further development. It presents the possibility of introducing this system in Poland.

Keywords—digital radio broadcasting, Digital Radio Mondiale, DRM features, technical aspects, DRM in Poland.

1. Introduction

To present the strategy for transition to digital broadcasting in Poland we need to determine the possibility of undertaking such a task. There is no doubt that the digital broadcasting system is well defined and documented both in technical and legal terms. The Digital Radio Mondiale (DRM) system is approved by radio broadcast regulators such as ETSI (European Telecommunications Standards Institute) and ITU (International Telecommunication Union). It is supported by international organizations such as the EBU (European Broadcasting Union), ABU (Asia Pacific Broadcasting Union) and IEC (International Electrotechnical Commission) too. There are adequate facilities for the transmission of radio signals such as RF high-power amplifiers and transmitters, modulators, antennas and informatics apparatus like PCs or servers, as well as a wide range of receivers both independent and associated with the receivers of other radio systems, digital and analog ones.
For the smooth introduction of the DRM system into operation, the DRM Consortium was formed in 1998 [1]. The Consortium is an international non-profit making association of broadcasters, network operators, manufacturers of transmitters and receivers, broadcasters, universities, research institutes and other organizations. Its purpose is to promote and distribute a digital system suitable for use in the 148.5 kHz – 174 MHz frequency range. Currently the consortium brings together 93 members and 90 "fans" from 39 countries.

2. The Definition of DRM, Documentation, Regulations

DRM digital system is designed to improve reception quality, reliability and ease of use at long, medium and short wavelengths, and enables further use of these ranges following analog broadcasting. DRM digital broadcasting technology is documented in detail through a series of technical specifications approved and published by ETSI. The basic description, ES 201 980 [2] contains all details of the DRM system: system architecture, coding systems and modes of transmission allowing operation in different propagation conditions at the maximum width of the channel. The V3.1.1 version of August 2009, contains a detailed description of both DRM30 and DRM+ systems.
ITU recommends implementation of digital DRM system in the frequency bands below 30 MHz [3]. ITU has established conditions for the use of digital DRM system in the electronic environment through a series of reports and recommendations. The most important are: ITU-R BS.2144 Report [4] and ITU-R BS.1615 Recommendation [5].
Both documents provide guidance for planning digital broadcasting in the bands below 30 MHz. Parameter sets included there provide useful planning field strengths such as the minimum usable field, strengths and RF protection ratios. There are a number of additional support standards related to distribution and communication protocols.

3. Features of the DRM System

The concept of DRM digital radio for frequency ranges of below 30 MHz with its implementation and possibilities of further use have been described above. These opportunities are also provided by the system that offers:

- competitive sound quality,
- additional data transmitted to the radio as "now and then"and other broadcast,
- EPG screen with a list of all digital radio available services,
- the ability to stop receiving in real-time and the ability to scroll backwards,
- the ability to successfully launch additional channels to reach new DRM listeners,
- use of other software to enrich information service.

The main characteristics of the DRM system are [6]:

- Access to four services on one frequency and a convenient choice of all currently received broadcasts:
  - audio broadcast in each service,
  - text information and/or use of multimedia applications [Journaline, Diveemo (see below) and others],
  - the ID (worldwide unique) easily scanning the specific services to make automatic frequency switch possible.
- List of stations (in Unicode – the system designed to handle the worldwide exchange, processing and displaying of texts written in different languages).
- Especially important for DRM:
  - possibility to be received in any country,
  - regular frequency changes.
- Alternative frequency may be limited territorially or temporarily.
- Checking the availability of services without interruption.
- Service announcements:
  - types of ads: traffic information, news, weather, warnings, alarms, an additional maximum of 6 types of ads reserved for the system,
  - active ads may be transferred to other services: to another DRM service within the same multiplex, to another site under another DRM multiplex, to other broadcast system (DAB/DMB, FM-RDS, AM, ...) services.
- Transmission of practical information:
  - current date and time (local/UTC),
  - the language used (ISO code),
  - information on a country of origin (IS code).
- Information can be selected, scanned and displayed.

4. Radio Equipment

4.1. Transmitters

There are two possibilities to receive/get a transmitter to broadcast DRM digital signals:

- adaptation of an existing analog transmitter system to work in the DRM system,
- purchasing a digital signal transmitter.

Thus, the senders who provide digital radio broadcasting are facing a very serious question concerning the transmitter: to buy or modify the old one. The transmitter can be modified for DRM broadcasting in a simple and cost effective way. At the moment there is no difficulty in meeting each of these solutions. There are several manufacturers of such devices. And the same may be modified in a short period of time, not exceeding a few hours. In general, manufacturers are producing equipment for broadcasting signal in door transmitters including DRM exciters, modulators, servers and even the antennas.

4.2. Receivers

DRM technology is very demanding and places great emphasis on the quality of the receiver, e.g., a very stringent requirement imposed on phase and noise parameters. Therefore, the receivers are expensive. Although there is a choice of receivers to enable reception of DRM digital radio, but because of their price these receivers do not enjoy too much attention [7].
The easiest way to receive a digital signals is to apply the DRM IF signal from any analog received and shifted it to 12 kHz and apply it to PC sound card input. Computer software will demodulate and decode the digital signal [8].

"Di-Wave 100" DRM receiver [9] is the first one with a color screen. It has been in mass production since 2009. The receiver has all the multimedia features offered by DRM technology: provides the name of the station, information about the Journaline program, MOT slides and time shift listening. The radio can receive DRM and analog stations in the SW, MW and LW bands as well as FM stations. User can store 768 stations. The receiver also has a USB port, SD-reader and mp3/mp4 players. 3.5-inch TFT color screen can display text in multiple languages.

The DR111 is a new receiver designed for receiving DRM, FM, AM signals developed with minimal production costs. Both in DRM, and AM systems the receiver works in the MF and SW. It meets all the minimal requirements which was specified by DRM consortium. DR111 receiver is one of the best solutions for the existing analog AM radio, which evolves toward the digital radio. The receiver has a 16-character LCD screen with two lines. Additionally it plays the recordings from an SD card and USB "pendrive" memory [1].

4.3. Communication Capacities of DRM System

Digital radio has a wide range of extra functions gained through new technologies in the field of semiconductors and the software. The receiver can be implemented with a range of functions provided by electronic program guide applications (EPG), such as:

- schedule view, with different levels of detail for the programs in the area of services,
- view of schedules, programs and events, as expected by various groups of listeners,
- navigation and selection of services and programs,

– search for current programs and the ones planned in the near future,

– choosing individual programs or groups of applications to record and to program selected specific or similar topics,

– careful program selection and recording via PNUM (Programe Number) signaling.

In addition, additional software, such as Journaline and Diveemo, has been created.
Journaline is a relatively new service of data transmission, which has been internationally standardized by the World DMB Forum for use in DAB and DRM [10]. Journaline is an application of data for DAB and DRM digital radio, with a hierarchical structure which provides text information. It is "teletext for digital radio" and is immediately available for interactive use. The listener can easily and quickly access the topics that are currently interested in.
Journaline provides:

– flexible menu structure,

– details of the text (headline, content), the list of messages (automatic update: sports scores, stock market, etc.),

– changing (ticker) messages (classifying information),

– bookmarks (favorite features) [10], [11].

Journaline supports two ways of organizing the transfer of objects, and both options can be easily mixed within a Journaline single site: carousel and transmission in real time [10].
Diveemo is an application of a new video on a small scale based on a DRM standard [12], [13]. It is intended for distribution to large areas of educational and informational programs. Diveemo information can be transmitted by one transmitter operating in one of the ranges: long, medium and short wave. It is an ideal platform for customers scattered over a wide geographical area. The transmission of DRM on shortwave provides virtually unlimited coverage of 100 square kilometers to more than 5 million square kilometers, depending on the transmission system. The system also has all the advantages of DRM, such as a choice of services by a Unicode-compatible labels, alternative signaling and switching frequency, features of announcements and warnings, etc.
Diveemo offers convenient mobile Internet services, a small-scale video, allowing users to quickly switch between channels and listen to the full audio and video, even in poor reception conditions. The video stream may be accompanied by one or more audio streams, allowing for synchronous, multiple languages, features of announcements and warnings. Diveemo provides cost-effective distribution of video programs, education and information by DRM.

Diveemo application was developed by Fraunhofer IIS, and its performance was presented by the Digital Radio Mondiale at IBC 2010 [13].

5. The Situation in the World

Digital Radio Mondiale Consortium has achieved a great technical success in developing the DRM system and its effective implementation. This system, despite the obvious limitations imposed by the need to adapt the occupied transmission bandwidth (9/10 kHz) to the arrangements for the allocation of spectrum [14], is promoted as a complementation to digital radio, and not as a competitor to DAB. Narrowed to about 6 kHz bandwidth of the original signal, while the digital sound quality is sufficient for musical and verbal broadcast. DRM+ system implemented in the frequency bands greater than 30 MHz is a European alternative to American HD-Radio, and can be used to replace FM broadcasts. It is recognized that digital radio has to be unified as one solution. An example of such unification are radios designed to receive radio signals broadcast in a variety of DRM, DAB digital systems, and AM and FM analogue systems.

A great opportunity for the DRM development are local information systems for cities, municipalities, tourist centers, social, cultural and commercial organizations as well as in public buildings and during big events (stadiums, rallies, etc.). Predictions are that digital radio systems may be used for safety and civil protection against extraordinary threats as a convenient means of information for the population at risk. At the beginning of October 2010 there were 41 multilingual programs broadcast in the world of, including one in Polish by Vatican Radio (7320 kHz from Santa Maria) and the Polish Radio program broadcast in German on 6135 kHz frequency from Skelton in Britain and in English on 7265 kHz frequency from Kvitsoy in Norway [1]. Three other stations fit experimental programs.

Thus, DRM digital radio has already started in most European countries (unfortunately, not in Poland) in special applications, but did not reach a significant level of universality. A different course of large-scale development is visible in countries where the possibility of getting with the program to large areas is essential. This applies especially to countries such as India, Russia and China. DRM signals are in total broadcast regularly in the world 75 stations, including two long-term and 14 medium-term. Most transmitters use more than 20 kW power, shown in Fig. 1. The most often used power is 90 kW. It is used by 21 transmitters.

A particularly strong interest in DRM is recorded in the Asian and Pacific region. According to ABU assessment, medium-wave digital radio has great potential in Asia-Pacific regions to ensure effective coverage of large areas. Significant progress in the implementation of universal DRM system in such countries as China, India, Pakistan, Indonesia and Iran is being watched with great interest. Hence, special attention given by ABU to the use of medium and short wave [14].

According to ABU, in 2009 India had 42 transmitters operating on medium wave in the AM system, which they intend to convert to DRM digital broadcasting. Further-

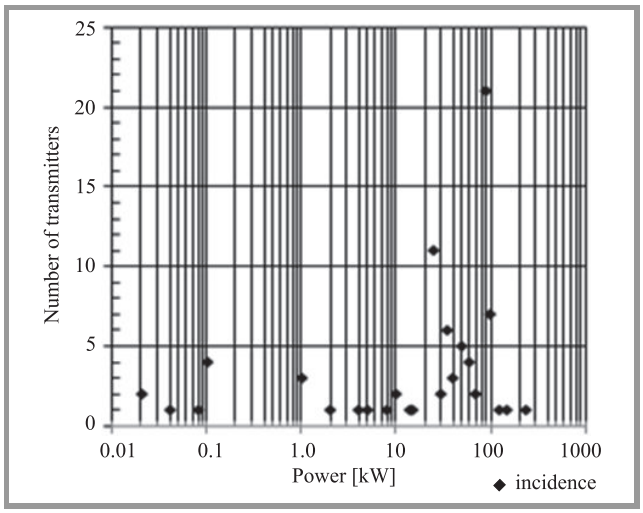


Fig. 1. Relationship between the number of existing transmitters versus their power.

more, 32 high-power transmitters are to be started/run. The main task of broadcasting is to cover the country. The Conversion from analogue to digital signal will take place smoothly with simulcasting.
Currently, the aim of DRM is to achieve better audibility and to enrich the radio reception by:

- optimizing the reception quality in accordance with the requirements of the recipient,
- enhancing technical flexibility to meet all the specific needs of broadcasting,
- introduction of additional features such as dual language programming and related to multimedia access and web content,
- introduction of a wider/additional offer of transmission through better use of available radio spectrum [12].

6. The Situation in Poland

Based on the GE75 plan, Poland has the right to use 18 medium-term frequencies and 123 station locations, which means that some low-power stations could operate on the same frequency [14]. Today a significant part of the spectrum is not used, and most of the stations do not longer exist. Some of the frequencies and locations allocated to Poland in the GE75 plan [14], mainly with the permission to broadcast with the power up to 1 kW in AM system, are used by a company called Polskie Fale Średnie S.A. [7], which uses 8 frequencies in 31 locations [16], [17].
On the LW there are two frequencies: 198 and 225 kHz available in Poland. We promote the concept of reconfiguration LW Polish stations assuming that the Polish Radio SA in Solec Kujawski (225 kHz) transmitter will working in a dual channel mode, transmitting AM signals in basic channel and digital signals in a neighboring channel or with time division between two systems. However, Raszyn transmitter (198 kHz) can start transmitting in a single digital channel [18]. Virtually, all awarded to Poland and desirable locations can be used to broadcast in the new DRM technology. We can change the location indicated in the GE75 plan, which would greatly facilitate the creation of a single frequency network SFN stations.

7. An Outline of Technical Aspects of the Strategy to Launch Digital Broadcasting in the Waves Below 30 MHz in Poland

7.1. The Technical Capacity to Start Broadcasting in DRM System

Currently, there are several manufacturers of transmission equipment, receivers and measuring equipment who accepted the challenge of digitization of the spectrum already using analog modulation system, such as Transradio AM, Harris, Nutel, Fraunhofer IIS, Thomson Grass Valley, NewStar, Digidia. Additional applications such as Journaline, Diweemo intended for data transmission and presentation greatly increase the usability of DRM. There are known results of propagation for different joint configuration transmissions in the AM and DRM systems.
On the technical side, there are different solutions to the broadcasting of the track. On the receiving side there are a number of receivers, both professional and designed for ordinary listeners. Their only one fault is the price at the moment.

7.2. Take the Necessary Action on Medium Wave

The introduction of digital radio services on medium wave in Poland will require such works as:

- Complementation of the current methods of forecasting the propagation of radio waves:
  - the value of laboratory-set protective factors for digital radio may change,
  - forecasting techniques for network ranges are needed.
- Identification of prospects for the current analogue broadcasting scenes and digital radio:
  - reception of digital signals is exposed to the noise interference of analog, and digital DRM signals,
  - gradual introduction of stations with DRM digital system, changes the interference situation,
  - comparing these forecasts with the forecasts for the expected scene with digital broadcasting onl.
- Proposing new solutions in the plan of the location of radio stations with digital services.

- Developing a plan for a single frequency station network (SFN) considering the following proposals [19]:
  - to cover the country with a network of low power transmitters which allows you to create regional or local networks,
  - to use several high power transmitters carrying the idea to develop a national synchronous network,
  - with plans for full coverage of the country the concept of dual frequency synchronized network may be necessary to use.
- Ceasing to issue NBC permits to broadcast on medium wave in analogue AM technology – in the current situation in the field of radio waves in Poland there is no need for simulcasting as a transition phase.

7.3. Opportunities for the Introduction of DRM in Poland

Purchasing an adequate set of DRM digital transmission path or its components ranging from studio equipment to the antenna should not cause any other problems but cash. There are several manufacturers that offer relevant equipment. Installation and running a digital transmission system in the long waves at a frequency of 198 kHz is practically possible at any time. However, at a frequency of 225 kHz utilized by long wave central transmitter, it requires a transitional period in a form of simulcasting or broadcast time distribution. The easiest way is to run the radio transmissions in the DRM system on medium wave, on the frequencies allocated to Poland under the GE75 plan.
Due to great interest of the neighboring countries in the DRM system there is a risk that they may hold additional spectrum for research purposes. In the future this could hinder obtaining approval for additional channels for the purpose of DRM digital system broadcasting in Poland. Losing any frequency will be an irretrievable loss for Poland. The already owned frequencies may prove useful in the future for various currently unknown reasons. Then, gaining access to the desired frequency may be impossible.

7.4. The Benefits of Broadcasting in the DRM System in Poland

The benefits of digital radio have been explained in the above mentioned characteristics of DRM. The benefits of broadcasting and reception in a digital system can be divided into benefits for broadcasters, both consumers/listeners and governments.
For broadcasters, it means creating new profit-making opportunities, through the use of new forms of transmission of information content- broadcasting of varied, not only audio but also video ads.

For analogue broadcasting listeners it means:

- improving the quality of reception,
- receiving a variety of information parallel with the tuned broadcast,
- enriching experience with radio,
- practical considerations.

For the government and various departments – audio and visual information concerning:

- risks of floods, hurricanes or fires,
- accidents and difficulties on the road,
- important events in the country area.

Polish Radio One now boasts that its anchors are viewed on TV HD for half an hour in the morning. "TVP HD" may be watched on following channels: "n" platform on item 5, "Cyfra+" platform on item 12 and "Cyfrowy Polsat" on position 101 as well as in UPC cable networks, "TP", "Stream Communications", "Sun Film", "Promax", "Petrus", "Telefonia Dialog" [20]. This means that it can only be watched by stationary recipients, and only those who watch the TVP HD channel. DRM system offers to its customers transmission of low resolution pictures 24 hours a day. Of much poorer quality, true, but the video in real time. The question is: which is more important, quality or information? Permanent watching presenters talking can be boring, therefore they can give other information in the intervals between sessions of the video from the radio studio.

8. Summary

Replacement of analogue with digital transmission is just a matter of time. DRM has been launched in most European countries (unfortunately, not in Poland) to be used in special applications, but has not yet reached a significant level of universality. A different direction of development is observed in big countries, where the possibility of getting with the program to large areas is essential. This applies particularly to countries such as India, Russia and China.
In Poland, there is potential for rapid mobilization of digital signal transmission with DRM. Frequencies, locations, forecasting programs, access to educated specialists and sets for transmission testing are available.
A vast opportunity for the development of DRM system are applications related to the local scope of use, i.e., local information for cities, municipalities, tourism centers, social, cultural and commercial organizations, as well as inside public buildings and during big events gathering high numbers of participants (stadiums, rallies, etc.).
The digital radio systems can also be used as a convenient means of information for safety and civil protection against extraordinary threats.

KRRiT can play a positive role in implementing the digital system in Poland by giving licenses to broadcast on medium wave, but only to DRM broadcast since now.

References

[1] www.drm.org

[2] "Digital Radio Mondiale (DRM); System Specification". Europ. Telecom. Standards Inst. Sophia antipolis, France, ES 201 980 V3.1.1., 2009.

[3] "Signal-on-the-air". Int. Telecomm. Union, Geneva, Switzerland, ITU-Recommendation BS.1661, 2003.

[4] Planning Parameters and Coverage for Digital Radio Mondiale (DRM) Broadcasting at Frequencies Below 30 MHz". Int. Telecomm. Union, Geneva, Switzerland, ITU-Recommendation BS.2144, 2009.

[5] "Planning Parameters for Digital Sound Broadcasting at Frequencies Below 30 MHz". Int. Telecomm. Union, Geneva, Switzerland, ITU-Recommendation BS.1615, 2003.

[6] P. Charron, DRM-what's going on? Thomson Grass Valley

[7] E. Wielowieyska, A. Dusiński, J. Jarkowski, T. Keller, and K. Kurek, "Radiofoniczne sieci cyfrowe, narzędzia i metody ich projektowania oraz emisje doświadczalne", raport z *Metody i Narzędzie Projektowania Pokrycia Radiowego Radiofonii Cyfrowej na Falach Długich i Średnich* – etap 1. Warszawa: Politechnika Warszawska, 2008.

[8] H. Chaciński and W. Kazubski, "Metody odbioru sygnału DRM". KKRRiT, Warszawa, 2009 (in Polish).

[9] www.uniwave.fr

[10] www.worlddab.org

[11] A. Zink, "DAB Surround & Journaline – Enhancing the Digital Experience", Fraunhofer Institut Integrierte Schaltungen, 2009, www.radioacademy.org/

[12] R. Obreja, "Welcome note", presentation DRM Consortium. Int. Broadcasting Conf. IBC, Amsterdam, The Netherlands, September 2010.

[13] M. Stoll, Diveemo – small-scale video service over DRM, Presentation DRM Consortium. Int. Broadcasting Conf. IBC, Amsterdam, The Netherlands, September 2010.

[14] "GE75: Plan for mf broadcasting and lf broadcasting", in *Frequency Assignment Plans on CD-ROM*. Int. Telecomm. Union (ITU), Geneva, Switzerland, 1997.

[15] S. Sadhu, *DRM Digital Radio in MW: Take-Off in Asia-Pacific*. Asia Broadcasting Union, Technical Department. Erlangen, 26-27 March 2009.

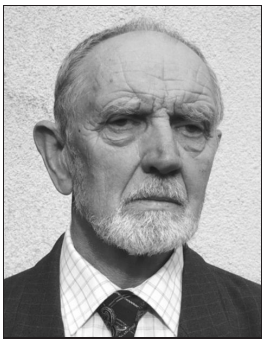
[16] "Radiofonia – programy nadawców koncesjonowanych", http://www.krrit.gov.pl/nadawcy/programy/rk.htm

[17] "Polskie rozgłośnie AM", www.polskaam.radiopolska.pl/polskaam.htm

[18] J. Jarkowski, A. Dusiński, and E. Wielowiejska, "Koncepcja uruchomienia w Polsce emisji cyfrowej DRM". Instytut Łączności, Warszawa, 2003.

[19] J. Jarkowski, A. Dusiński, K. Kurek, T. Keller, and K. Bryłka, "Radiofoniczne sieci cyfrowe, narzędzia i metody ich projektowania oraz emisje doświadczalne", raport z *Metody i Narzędzie Projektowania Pokrycia Radiowego Radiofonii Cyfrowej na Falach Długich i Średnich* – etap 2. Warszawa: Politechnika Warszawska, June 2008.

[20] "Gdzie i dlaczego TVP HD", www.tvp.pl/hd/



Andrzej Dusiński received the title of engineer in electronics Evening Engineering School in Warsaw in 1965 and in 1972 obtained the vote of acceptance of part-time studies in mathematics on Department of Mathematics and Mechanics UW. Since 2004 he is the retired engineer and Senior R&D Specialist of the National Institute of Telecommunications (NIT). Recently he dealt with tools and methods of network planning for digital broadcasting service DRM within the research project granted by the Ministry of Science and Higher Education of Poland. He has spent the majority of his career in NIT, working in Radio Waves Propagation Department and last in Radio Communications Department. His research interests include several aspects of radio propagation within the frequency range from 150 kHz to 60 GHz. In this role, he focuses on propagation measurements and prediction tools for terrestrial services among others as prediction software for digital sound broadcasting at frequencies below 30 MHz.
E-mail: adrezer1@x.wp.pl
Plutonowych st 21
04-404 Warsaw, Poland

Jacek Jarkowski was born in Warsaw, Poland. He received M.Sc. from Warsaw University of Technology (WUT) in 1963 and Ph.D. degree in Radiocommunication Science in 1975. Since 1962 was employed at the Faculty of Electronics WUT, and since 2003 he is with the National Institute of Telecommunications, Warsaw. His primary research interest are antennas, propagation and radiocommunication systems and currently wireless cognitive sensor networks.
E-mail: J.Jarkowski@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

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