JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

4/2015

(*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 the *Journal of Telecommunications and Information Technology* contains thirteen papers that deal with diverse problems of network security, or various issues related to wire and wireless communication networks including wireless sensor networks.

The first two papers are devoted to intrusion detection in computer networks. Damian Jankowski and Marek Amanowicz in the paper *Intrusion Detection in Software Defined Networks with Self-organized Maps* consider the new opportunities enabled by the Software Defined Network (SDN) architecture to implement security mechanisms in terms of unauthorized activities detection. They describe a novel approach to threat detection based on the machine learning integrated with the SDN controller. The movement-assisted threat detection system using mobility to enhance a global threat assessment in networks created by resource-limited things is presented in the paper *Intrusion Detection in Heterogeneous Networks of Resource-Limited Things*. Adam Kozakiewicz *et al*., describe the architecture of the threat monitoring system for a wireless sensor network that provides a separate physical secure channel to deliver collected information.

Mariusz Dzwonkowski and Roman Rykaczewski in the paper *Quaternion Feistel Cipher with an Infinite Key Space Based on Quaternion Julia Sets* concentrate on the quaternion encryption. The authors claim that the application of modular quaternion to perform rotations of data sequences in 3D space to Feistel Cipher can bring an efficient encryption scheme. The proposed encryption algorithm is described and compared with the AES method.

The next three papers deal with secure IP communication provision for the power system management. The architecture and functionality of the cyber security system for a power grid control are presented by Jacek Jarmakiewicz *et al*., in the paper *Evaluation of the Cyber Security Provision System for Critical Infrastructure*. The results of the system verification and validation in a testbed network are presented and discussed. Przemysław Malec *et al*., in their paper *Detecting Security Violations Based on Multilayered Event Log Processing* start with premise that correlating data from multiple event log sources increase the accuracy of threat detection. The authors propose the multilayered event log analysis approach for managing and handling security incidents. Kamil Kołtyś and Robert Gajewski in the paper *SHaPe: A Honeypot for Electric Power Substation* describe the concept, architecture,

and implementation of the SCADA (Supervisory Control and Data Acquisition) honeypot supporting the IEC 61850 standard. The presented system is open source software publicly available under GNU GPL.

Kazumasa Oida in the paper *Uniqueness and Reproducibility of Traffic Signatures* considers the feasibility of applying a traffic signature to the user machine identification. The author describes the novel approach with the signature calculated from HTTP-based video stream transmitted by the authentication server, and evaluates his proposal through numerous experiments.

The problem of limitations of a mobile cloud network and possible fault events at clouds are discussed by Jordi Mongay Batalla *et al*., in the paper *On Providing Cloud-awareness to Client's DASH Application by Using DASH over HTTP/2*. The concept and implementation of the end-to-end framework for cloud congestion identification for DASH-capable video application are presented. The performance of the system for cloud-aware and its applicability to real clouds are discussed.

The applicability of the burst ratio parameter to multi-channel scenarios in transmission networks is discussed in the paper *Analysis of Burst Ratio in Concatenated Channels*. To confirm and demonstrate the validity of the burst ratio analysis Jakub Rachwalski and Zdzisław Papir present the results of numerous simulation experiments performed with NS2 network simulator.

The issue of coexistence of DVB-T and LTE communication systems operating in contiguous UHF frequency bands is raised in the paper *Measured Interference of LTE Uplink Signals on DVB-T Channels*. Massimo Celidonio *et al.*, survey and investigate the potential LTE influence on a DVB-T reception. The performance parameters that should be considered for assessing the corresponding interfering effects, i.e. protection ratio and protection distance parameters are under consideration. The numerous measurements carried out at the laboratory are analyzed.

The following two papers deal with various issues related to sensing systems. The paper *The Integration, Analysis and Visualization of Sensor Data from Dispersed Wireless Sensor Network Systems using the SWE Framework* is concentrated on integration and interoperability of measurements gathered by various sensing devices. Yong Jin Lee *et al.*, describe the extensions to the SWE (Sensor Web Enablement) framework to integrate disparate and disperse wireless sensor networks and to support standardized access to sensor data. Furthermore, the proposed software system introduces web-based data visualization and provides statistical analysis services. The next issue considered is the detection of location of the radiation in space to localize given data source. Youssef Khmou *et al.*, in the paper *Lorentzian Operator for Angular Source Localization with Large Array* propose a novel high resolution algorithm utilizing the Lorentzian function for Direction of Arrival (DoA) of narrowband and far field punctual estimation. The profitability of application of this algorithm is demonstrated through simulation study.

The last paper is concerned with maintenance of lead-acid batteries to achieve a high reliability of telecommunication services. Ryszard Kobus *et al.*, in the paper *Maintenance of Lead-acid Batteries Used in Telecommunications Systems* survey the primary types of batteries used in telecommunication systems and describe the techniques for monitoring and measuring their current capacity. The main attention is focused on an universal module for charging/discharging batteries (TBA-A) that was developed by the authors. They claim that the TBA-A module integrated with a control unit TBA-W creates a novel ATE device that can be successfully used in power systems for remote telecommunication facilities.

We wish our Readers an interesting reading time.

Ewa Niewiadomska-Szynkiewicz
Guest Editor

Paper

Intrusion Detection in Software Defined Networks with Self-organized Maps

Damian Jankowski and Marek Amanowicz

Institute of Telecommunication, Faculty of Electronics, Military University of Technology, Warsaw, Poland

Abstract—The Software Defined Network (SDN) architecture provides new opportunities to implement security mechanisms in terms of unauthorized activities detection. At the same time, there are certain risks associated with this technology. The presented approach covers a conception of the measurement method, virtual testbed and classification mechanism for SDNs. The paper presents a measurement method which allows collecting network traffic flow parameters, generated by a virtual SDN environment. The collected dataset can be used in machine learning methods to detect unauthorized activities.

Keywords—IDS dataset, machine learning, metasploit, network security, network simulation, open flow, virtualization.

1. Introduction

The Software Defined Networks (SDNs) allow to implement and control network functionality through software. Such an approach allows to deploy new services and applications in a virtual environment and to share resources with the performance and isolation from the rest of processes [1]. However, with the flexibility and scalability provided by SDNs, it is important to maintain an adequate security level. The intrusion detection technologies, integrated with the SDNs environment, can provide an additional security element, besides the classical Intrusion Detection System (IDS) and Intruder Prevention System (IPS). SDN architecture creates new opportunities to increase the security level, especially in the context of the unauthorized activities detection. Nevertheless, there are certain risks associated with this technology.

2. The Software Defined Network Technology

The basic idea of the SDNs is the separation of data plane from the control plane. In contrary to the classical network solutions, the network devices are here supervised by SDN controllers in a centralized manner. Such a solution enables configuration and programming from a host to match the service requirements in a distributed network. Moreover, the centralized logic and management allows for comprehensive monitoring network [1].

SDNs are associated with Network Function Virtualization (NFV). The current rapid development of hardware server platforms gives sufficient performance of services operating on virtual machines. Servers became more efficient and have better functionality than previously, and are suitable for use in a virtual environment. NFV is a network architecture that implements virtualization of network nodes. It enables for the functionality implementation based on the available servers, switches, storage devices, without using dedicated hardware devices. To sum up, the functionality of network hardware devices can be implemented in software technologies [2].

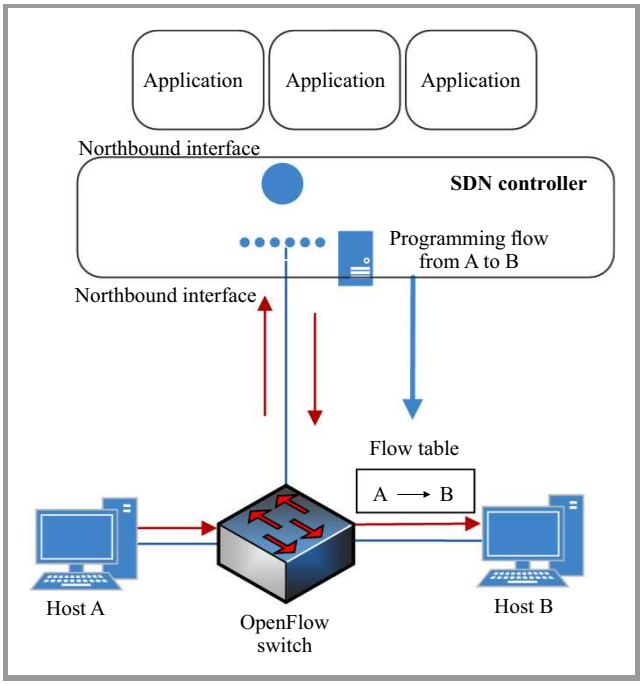


Fig. 1. Packets forwarding scheme in SDNs.

The SDN controller communicates with network devices using the OpenFlow protocol and controls network traffic according to the programmed rules. The forwarding packets methods are defined in a flow table, which is stored in SDN controllers and in switches memory supporting the OpenFlow protocol. The operation order, which describes how SDN controllers set the traffic flow, is presented in Fig. 1. In the case, that packet is forwarded to the switch

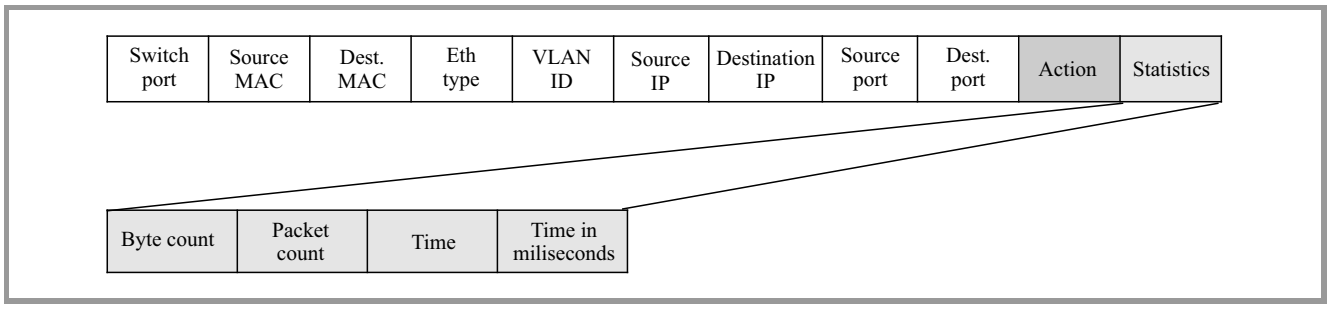


Fig. 2. Flow entry structure in SDNs.

and there is no entry in the flow table, the packet is transmitted to the SDN controller. Applications and modules, which run on controller, determine manner of packet processing. After that, the controller transmits the new entry in the flow table to the switch. That network traffic flow is then defined and established using algorithms developed on application modules in SDN controller. The structure of the traffic flow is shown in Fig. 2.

The main flow element specifies the parameters that are taken into consideration in the process of matching packets to the flow. An action field defines a way of forwarding that can be performed on packets from the particular flow. In addition, flows are linked with specific network traffic statistics, which latter can be used for traffic features extraction [2].

Thanks to the open architecture, the network logic is established with application modules running on the SDN controller. Hence, it is possible to develop algorithms fulfilling specific user functionality. Programmers can use classical programming languages, frameworks, frameworks, APIs and libraries for the process of developing application in the SDN environment.

3. Vectors of Attacks in SDNs

The SDN architecture has an important impact on the class of attack that can be performed, as presented in Fig. 3. The most dangerous situation takes place when the SDN controller is compromised. This can be done by the exploitation of vulnerabilities of processes and services running on the controller. Consequently the entire SDN domain is compromised, and the attacker has the ability to take control of all network devices. The degree of vulnerability of such attacks mainly depends on the hardware implementation of the SDN controller, programming languages and libraries used. The threats prevention can utilize IDS and IPS techniques, as well as methods of replication and recovery status of SDN servers from time before attack. Due to the nature of SDN technologies, classical IDS may be insufficient [3].

Potential security vulnerability may also exist on the administrative station, which is used to manage network operating system (SDN controller). Such terminals are used for developing applications for the control logic of network devices and ensure the monitoring of activities in the network

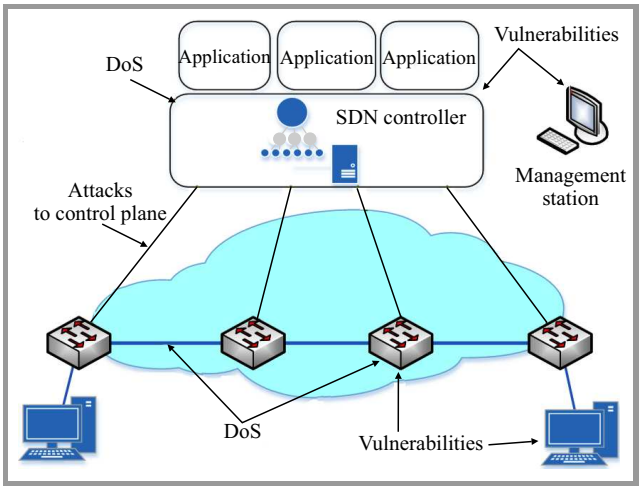


Fig. 3. Potential vector of attacks in SDNs.

environment. Attackers could potentially exploit vulnerability at the supervisor station or its connection with the controller. Reducing the risk can be achieved by making use of mutual SDN server authentication and the terminal or IDS and/or IPS techniques [4].

Other threats are attacks on the communication stream between data and control plane. For instance, the security protocol between controllers and network devices can be TLS. A potential vector of attack would exploit vulnerabilities in its implementation.

The classical network threats are still present in the SDN technologies. It is possible to generate malicious activities in IT systems, for instance deny of services or exploit vulnerabilities on servers, host or network devices [5].

4. Attacks Detection in SDNs

Despite the security vulnerabilities, SDN creates new opportunities for the implementation of more effective intrusion detection methods. Moreover, it allows for the integration of threat detection methods with the SDN environment. Due to the openness of platforms supporting SDN technologies, it is possible to use existing mechanisms and protocols. An important factor associated with intrusion detection is the possibility of aggregating statistics logs from network devices memory and forwarding them to the controller. Collected parameters can be used as source data

Table 1
Selected intrusion detection methods for SDNs

Approach	Principle of operation	Extraction of attack symptoms or features	Machine learning method or logical reasoning	Detected attacks for dataset	Network traffic	Accuracy [%]	False positive [%]
Method 1	Assessment of the first packet transmitted to the SDN controller	Maximum entropy detector, TRW-CB, Rate-limiting, NETAD	None	DoS, probe	Benign – real SDN network, attacks – artificial traffic	80–90	0–70
Method 2	Evaluation of the threats level	TRW-CB, Rate-limiting	Fuzzy logic	DoS	Artificial traffic	95	1.2
Method 3	Creating profiles using sFlow and OpenFlow	TRW-CB, Entropy level	None	DDoS, worm propagation probe	Benign – real SDN network, attack – artificial traffic	100	23–39.3
Method 4	Flow statistics collection	Flow based statistics and features	Self-organized maps of Artificial Neural Network	DDoS from botnet	Artificial traffic	98.57–99.11	0.46–0.62

for intrusion detection algorithms. In recent studies, there are a few proposals to use SDN's capabilities for intrusion detection mechanism. The four sample solutions are shown in Table 1:

- Method 1 – revisiting traffic anomaly detection using software defined networking [6];
- Method 2 – a fuzzy logic-based information security management for SDNs [7];
- Method 3 – combining OpenFlow and sFlow for an effective and scalable anomaly detection and mitigation mechanism on SDN environments [8];
- Method 4 – lightweight DDoS flooding attack detection using NOX/OpenFlow [9].

The methods listed above detect common types of malicious activities, i.e., denial of service, port scan, and attempts to propagate malicious software. However, there is no papers describing SDN solutions, which would enable the detection of more sophisticated groups of attacks. These attacks can rely on the use of vulnerabilities in sophisticated services, and one of the phases of attack is to inject a malicious code. The presented methods have a very good detection accuracy rate, but the false positive rate is poor in approaches given by method 3. In method 1 the false positive rate may vary from 0 to 70 due to the detection technique used. Hence, it is problematic to compare the effectiveness of selected solutions because their performance tests were carried out in different environments, according to various methodologies and using different data sets.

5. The Architecture of the Proposed Approach of Intrusion Detection

The presented idea is based on the assumption that it is possible to classify whether network traffic flows represent normal operation or attack (Fig. 4). The flows classification is based on features obtained through the functionality

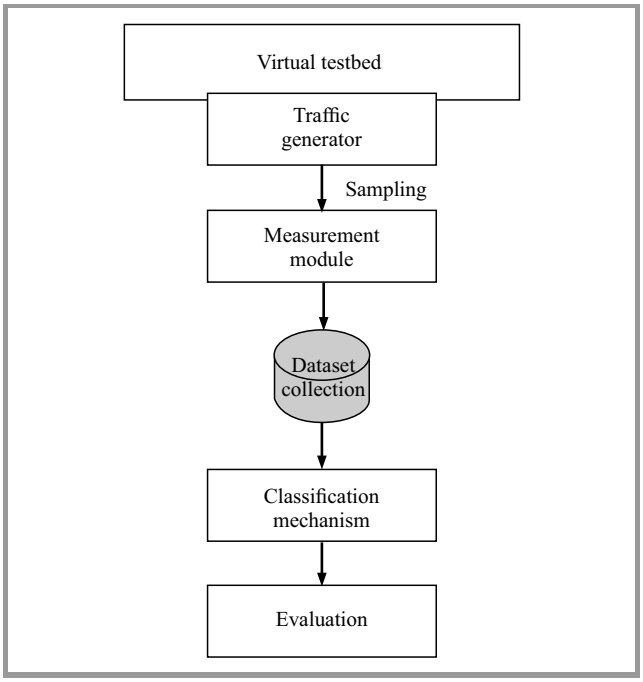


Fig. 4. Architecture of presented mechanism.

available in the SDN technology. The virtual testbed is used for generating certain classes of traffic, benign and malicious. Generated network traffic is sampled by the measurement module implemented as OSGi bundle and resides on the Opendaylight SDN controller memory. The module also programs flows. At the same time, the REST client communicates with the SDN controller and collect statistics related with flows. The present mechanism operates in the control plane. The collected data can be stored in a database or an external file. At this stage, the features for further classification can be calculated and extracted. The collected data are then converted to a dataset for testing machine-learning methods. At the current stage of research, the classification process is performed by self-organized maps of Artificial Neural Network (SOM ANN), but in future, studies would be extended to other classification methods.

6. Principle of Measurements

The measurement software module works on the SDN Opendaylight controller as an OSGi bundle [10] and implements the switch functionality (Fig. 5). The traffic flows are matched by the following criteria:

- destination IP address,
- source IP address,
- destination port of transport layer,
- source port of TCP/UDP layer,
- protocols – ARP, IP, TCP, UDP or unknown.

For each flow, an idle timeout parameter is set that defines the period after the entries are deleted from the flow table, and an identification number. Such matching network traffic distinguishes traffic in the context of different port numbers. As a result, it is possible to measure the parameters and define relationships between connections at the transport layer, which are refreshed within a specified period.
Another component used in the measurements is a REST client. This module communicates with the Opendaylight server. The following parameters permit to change the resolution of measurements:

- time between queries (in the REST client),
- time between refresh of array status and statistical parameters defined configuration flow controller SDN.

For each collected flow, a set of parameters is determined. These data values constitutes the input vector for the machine learning method:

- the measurement results contain value of n vectors features $X_i(x_1, x_2, \ldots, x_n)$ at i time of sampling. The vector features x_i are parameters and statistics retrieved from network flows;
- the input vector $X(\max\ x_1,\ \max\ x_2, \ldots, \max\ x_n)$ includes maximum values of features from all X_i vectors;
- labels defining classes are assigned to vectors X, due to IP host address.

In a review of the existing solutions in threat detection technology, SDN indicates that the described method shall detect the attack time from flow table. The presented approach is based on the flows classification by the transport layer level discrimination. It allows identifying a specific connection representing the unauthorized action. The Opendaylight software environment enables the measurement of selected parameters, which can be potentially used as features for threat detection methods.
The primary parameters obtained from the OpenFlow protocol are IP addresses, port numbers, duration, number of packets and bytes in the flow. More, the collected statistics

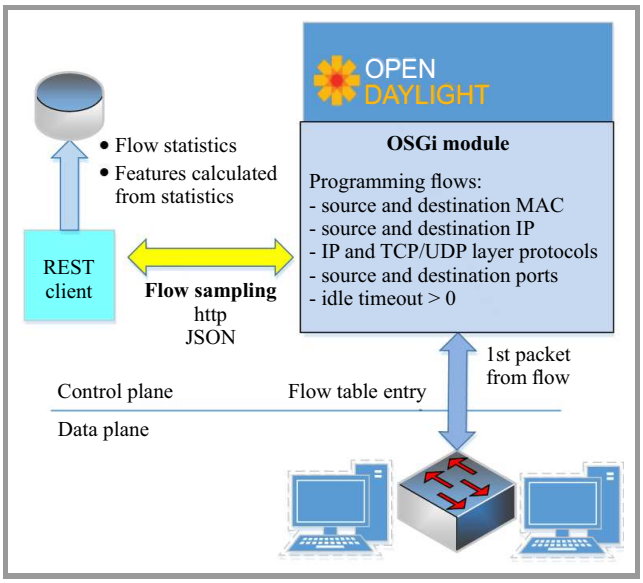


Fig. 5. Measurement method based on SDNs.

can be used to extract other information. Such values are calculated in the context of the entire array, for the sample time stamp. It reflects the dependence of connections between hosts. Example additional features are:

- single flow coefficient,
- flow rate to the host,
- multiple flows with the same host coefficient,
- flow rate of the same service to a host from multiple hosts.

The primary and additional parameters represent the x features of the input vector X. Features can be determined based on the first packet transmitted to the controller. However, at this stage of research, this mechanism is not implemented. The approach of analyzing the header and the first package contents may have a positive influence on the detection performance of attacks on specific groups. However, it is necessary to improve the performance without packets inspection, because packets can be obfuscated. Therefore, with the development of the presented method, it will be evaluated which feature is more important.

7. Virtual SDN Testbed

The most of the machine learning methods for intrusion detection is based on the KDD99Cup dataset. It is used in the process of learning and testing, allowing comparing the performance of different methods. Unfortunately, there are no such datasets that could be used to evaluate the detection methods. The proposed concept is based on the mechanism of SDN flow classification. Therefore, an important component of the presented approach is the test environment for the generation of SDN network traffic, allowing verifying the effectiveness of the presented method. In this

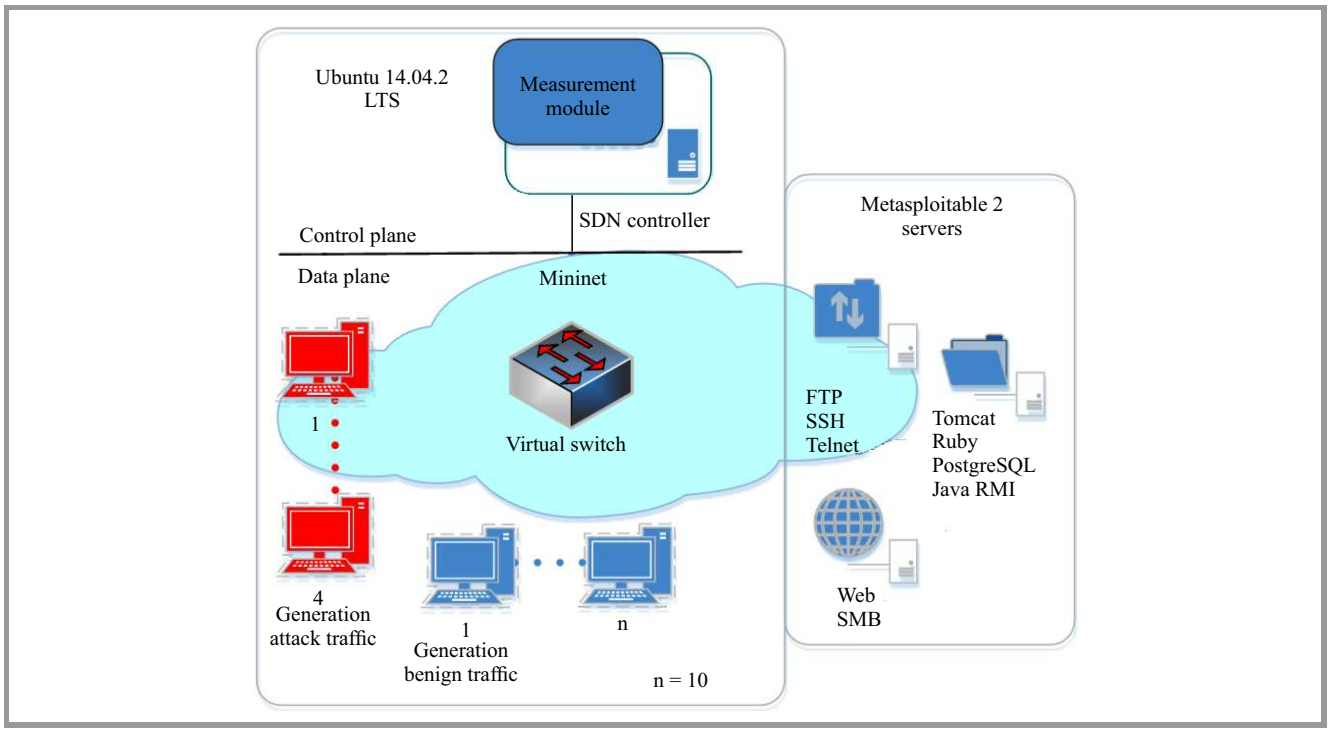


Fig. 6. Virtual SDN testbed architecture.

Table 2
Traffic classes which are performed in virtual testbed

Activities	Examples of activities	Tools for traffic generation
DoS	Denial of service	metasploit, hping3, nping
Probe	Port probe, vulnerability scan, version scan	metasploit, nmap
User2Root	Exploit vulnerabilities, shell control, backdoor	metasploit
Remote2Local	Password crack	metasploit, hydra
Normal	Communication between clients and servers	FTP, SSH, SMB, Apache, Web, Tomcat, RMI Ruby, Java RMI, Postgres, Telnet

research a SDN network emulator mininet, developed in Python and C is used. In the connection with the Opendaylight controller functionality, it is possible to model the SDN environment with different network topologies. At this stage of the implementation, the SDN has a star topology with a single switch. The architecture of the testbed is shown in the Fig. 6. The testbed covers normal traffic generation and selected groups of attacks using tools presented in Table 2.
The following classes of activities are generated in this virtual environment:

- normal – benign traffic generated between hosts and servers;
- DoS – denial of service attacks, performed against the transfer, network or computation resources of IT system;
- Probe – port, version or vulnerability scanning. Such activities give information to intruders about the potential targets of the attacks;
- U2R – attacks work by exploitation of vulnerability;
- R2L – this class covers credentials guessing and unauthorized access to IT resources.

The server side is emulated by metasploitable virtual machines with the Ubuntu Linux operating system. Vulnerabilities of services and the OS are intentionally left on the server environment. Simultaneously, the clients generate requests to the server. At the same time, the malicious host performs unauthorized activities directed to servers by using attack tools. The client activities are automated by Python scripts [11]. Generated traffic is probing by the measurement module. The servers reside on separate virtual machines and clients are virtualized on the mininet OS level.

8. Self-organizing Maps as Machine Learning Attack Detector

Machine learning methods are commonly researched concerning intrusion detection mechanisms [12]. The presented approach would use self-organized maps (SOM) to perform the unauthorized activities detection. It is a method of unsupervised machine learning, based on artificial neural

networks, useful for a graphical representation of datasets. However, when input vectors are labeled, this method can be used as a classification mechanism. The Kohonen algorithm is used as SOM learning method. As a result of the applied input signals, network indicates the activation of neurons in varying degrees, as a result of adaptation to changes in the synaptic weights, during the process of learning. Some neurons, or groups of neurons, are activated in response to stimulation, adapting to the form of specific patterns. Because of this, the test vectors activate neurons of a trained network, which are the most similar. After the initialization process, networks are reliant based on the parameters, each of the neurons is assigned to a specific position of the multidimensional space. Distribution of neurons can be created in a random way. They are associated with neighbors in a hexagonal manner. In a further stage, the network is learned by a training set. The most stimulated neuron and neighboring neurons update the weights, in response to learning vectors using the scheme (see Fig. 7):

$$W_i(k+1) = W_i(k) = \eta_i G(r)\left[X - W_i(k)\right], \qquad (1)$$

where: X – input vector of features, W_i – i weight vector of the neuron at k time, η_i – learning rate, $G(r)$ – neighborhood function given by Eqs. (3) and (4).

The distances of input vector $X(x_1, x_2, \ldots, x_j)$ to winner neurons $W(w_1, w_2, \ldots, w_j)$ are calculated on base of the Euclidean distance:

$$d(X, W_i) = \left\|X - W_i\right\| = \sqrt{\sum_{j=1}^{N}(x_j - w_{ij})^2}, \qquad (2)$$

where: X – input vector of features, x_j – j feature in X input vector, W_i – i weight vector of the neuron, w_{ij} – j value of weight in i weight vector of the neuron.

The weight adaptation degree $G(i)$ of winner and neighbors neurons is calculated by Gaussian formulas:

$$r = d(i, W) = \left\|i - W\right\| = \sqrt{\sum_{j=1}^{N}(i_j - w_j)^2}, \qquad (3)$$

$$G(r) = \mathrm{e}^{\frac{-r^2}{2\lambda^2}}, \qquad (4)$$

where: r – Euclidean distance of i neuron from winner neuron, W – winner neuron, λ – neighborhood radius.

The SOM network allows creating a type of structure, which can be represent input vectors in the best way [13], [14], [15]. It can be said, that the single neuron represents many vectors from the dataset. The class is assigned to the neuron with the consideration of which class is the most numerous, from stimulating vectors. The classification step is preformed after learning. This involves determining which labeled neuron is activated under the input vector. The dataset is normalized in the range $0 \ldots 1$. The SOM input vector records are collected data in the measuring module with assigned labels specifying the

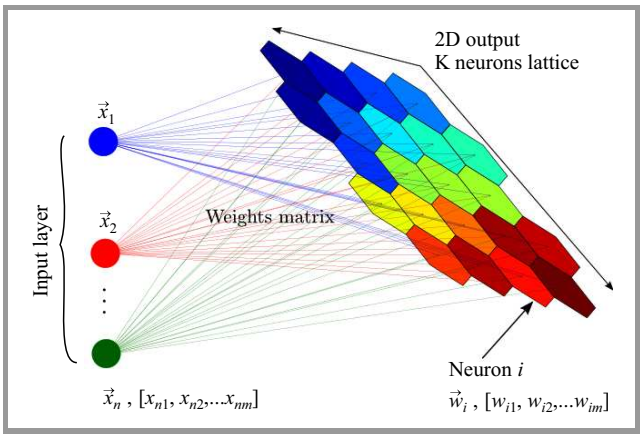


Fig. 7. Self-organizing map structure.

type of traffic. For instance, the 11 dimensional input vector $X(x_1, x_2, \ldots, x_{11})$ elements can be represented by following features:

- x_1 – destination IP address,
- x_2 – source IP address,
- x_3 – destination TCP/UDP port,
- x_4 – source TCP/UDP port,
- x_5 – duration,
- x_6 – number of packets in flow,
- x_7 – number of bytes in the flow,
- x_8 – single flow rate,
- x_9 – flow rate to the host,
- x_{10} – multiple flows rate with the same source host,
- x_{11} – rate of connections on the same port to a host from other computers.

In addition, the feature vector is linked to a label that defines the class of activities. Due to research scenario, the composition of features in input vector can vary. To evaluate performance, the cross validation method with 10 folds will be used. After the process of learning, neural network can be presented in low dimension space by using Sammon mapping. The results of the classification will be evaluated by the confusion matrix, ROC curves and typical coefficients used in machine learning methods evaluation process.

9. Summary

The presented research describes the intrusion detection method integrated with the SDN controller. This conception classifies unauthorized activities performed in SDN environment. Further studies cover the implementation of all modules and performance tests of the detection mechanism. Realized implementation and evaluation of the effectiveness will be described in further publications. In the case that the proposed mechanism will not be effective, it is necessary to research and implement additional features, especially based on the parameters and data of

the first packet in the flows. Another important stage is the research on the most significant feature selection and features extraction.

The aim of the work is also the investigation of the ability to detect a wider range of network attacks, especially those that are not identified in other technologies. It is important to study new classes of attack specified for the SDN environment. This aspect especially includes attacks on SDN controllers and the control plane. The proposed method may have potential performance disadvantages because of the high-grained traffic matching in a flow table. The use of the measuring module developed for all controllers in the network can be problematic in terms of performance. Therefore, an exemplary architecture can assume that only selected SDN controllers in the network will implement functions of intrusion detection. Another point is the comparison of other machine learning methods.

References

[1] J. Kleban and M. Puciński, "Sieci sterowane programowo SDN w centrach danych SDDC" (SDN network software controlled data centers SDDC), in *XVII Poznań Commun. Worksh. PWT 2013*, Poznań, Poland, 2013 (in Polish).

[2] D. Kreutz *et al.*, "Software Defined Networking: A Comprehensive Survey", *Proc. of the IEEE*, vol. 103, no. 1, pp. 14–76, 2015.

[3] S. Shin and G. Gu, "Attacking software-defined networks: A first feasibility study", in *Proc. 2nd ACM SIGCOMM Worksh. Hot Topics in Softw. Def. Netw. HotSDN'13*, Hong Kong, China, 2013, pp. 165–166.

[4] D. Kreutz, F. M. V. Ramos, and P. Verissimo, "Towards secure and dependable software-defined networks", in *Proc. 2nd ACM SIGCOMM Worksh. Hot Topics in Softw. Def. Netw. HotSDN'13*, Hong Kong, China, 2013, pp. 55–60.

[5] V. Tiwari, R. Parekh, and V. Patel, "A survey on vulnerabilities of openflow network and its impact on SDN/Openflow controller", *World Academic J. Eng. Sci.*, vol. 1, no. 01:1005, 2014.

[6] S. Akbar Mehdi, J. Khalid, S. A. Khayam, "Revisiting traffic anomaly detection using software defined networking", in *Recent Advances in Intrusion Detection*, R. Sommer, D. Balzarotti, and G. Maier, Eds. *LNCS*, vol. 6961, pp. 161–180. Berlin Heidelberg: Springer, 2011.

[7] S. Dotcenko, A. Vladyko, and I. Letenko, "A fuzzy logic-based information security management for software-defined networks", in *Proc. 16th Int. Cof. Adv. Commun. Technol. ICACT 2014*, Pyeongchang, South Korea, 2014, pp. 167–171.

[8] K. Giotis, C. Argyropoulos, G. Androulidakis, D. Kalogeras, and V. Maglaris, "Combining OpenFlow and sFlow for an effective and scalable anomaly detection and mitigation mechanism on SDN environments", *J. Comp. Netw.*, vol. 62, pp. 122–136, 2014.

[9] R. Braga, E. Mota, and A. Passito, "Lightweight DDoS flooding attack detection using NOX/OpenFlow", in *Proc. 35th Ann. IEEE Conf. Local Comp. Netw. LCN 2010*, Denver, Colorado, USA, 2010, pp. 408–415.

[10] OpenDaylight Platform [Online]. Available: https://www.opendaylight.org/

[11] Mininet – An Instant Virtual Network on your Laptop (or other PC) [Online]. Available: http://minimet.org

[12] A. S. Subaira and P. Anitha, "Efficient classification mechanism for network intrusion detection system based on data mining techniques: a survey", in *8th IEEE Int. Conf. Intell. Syhst. & Control ISCO 2014*, Coimbatore, India, 2014, pp. 274–280.

[13] K. Choksi, B. Shah, and O. Kale, "Intrusion detection system using self organizing map: a surevey", *Int. J. Engin. Res. Appl.*, vol. 4, no. 12, pp. 11–16, 2014.

[14] S. Osowski, *Sieci neuronowe do przetwarzania informacji* (*Neural Networks for Information Processing*). Warsaw: Publishing House of Warsaw University of Technology, 2013 (in Polish).

[15] "SOMz: Self Organizing Maps and random atlas" [Online]. Available: http://lcdm.astro.illinois.edu/static/code/mlz/MLZ-1.2/doc/html/somz.html

Damian Jankowski received B.Sc. and M.Sc. degrees from the Military University of Technology, Warsaw, Poland in 2010 and 2011, in Telecommunication Engineering. His research interests include programming, system virtualization, system administration, IT security, machine learning, and data mining.

E-mail: damian.jankowski@wat.edu.pl
Military University of Technology
S. Kaliskiego st 2
00-908 Warsaw, Poland

Marek Amanowicz received M.Sc., Ph.D. and D.Sc. degrees from the Military University of Technology, Warsaw, Poland in 1970, 1978 and 1990, respectively, all in Telecommunication Engineering. In 2001, he was promoted to the professor's title. He was engaged in many research projects, especially in the fields of communications and information systems engineering, mobile communications, satellite communications, antennas & propagation, communications & information systems modeling and simulation, communications and information systems interoperability, network management and electronics warfare.

E-mail: marek.amanowicz@wat.edu.pl
Military University of Technology
S. Kaliskiego st 2
00-908 Warsaw, Poland

Paper

Intrusion Detection in Heterogeneous Networks of Resource-Limited Things

Adam Kozakiewicz, Krzysztof Lasota, and Michał Marks

Research and Academic Computer Network (NASK), Warsaw, Poland

Abstract—The paper discusses the threats to networks of resource-limited things such as wireless sensors and the different mechanisms used to deal with them. A novel approach to threat detection is proposed. MOTHON is a movement-assisted threat detection system using mobility to enhance a global threat assessment and provide a separate physical secure channel to deliver collected information.

Keywords—client honeypot, Internet of Things, intrusion detection, wireless sensor network.

1. Introduction

The terms like computer or network are becoming less clear as the technology advances. No more than ten years ago, most of the nodes in the Internet were stationary computers with a wired connection. On the server side of the network this is still an accurate depiction at least of the physical setup, although admittedly more and more inaccurate on the logical side, as virtualization advances and the additional cloud layer isolates the servers from their hardware.

On the client side, the network changed completely. Most of new devices are wireless. Also, the name "device" is quite appropriate, as more and more of them do not look like traditional computers (even if that's what they essentially are). The trend is not only directed at mobility of computing, as in case of laptops, smartphones, tablets, etc., but also towards expanding the computational abilities of other things, leading to ideas such as smart home or smart city.

The side effect of this approach is that the network is becoming full of devices with at least one of the following limitations: battery power, meaning that energy conservation becomes crucial factor, or limited computing power, due to lowering costs, lowering energy consumption or preventing heating. These limitations, the fact that many of the devices (especially smart things) are designed by companies with little experience in computing and the thing status, meaning that users are unlikely to participate in installation of updates (so either the things will never be updated or will have a fully automated update mechanism, creating a tempting target for attacks) lead to a rather difficult situation from the security standpoint. While most of the things are seen as not worth attacking, the situation becomes worse when the entire heterogeneous network is seen as a single system. Unsafe devices are points of entry to the network, threatening other resources. They can also be used in orchestrated attacks, e.g. providing multiple consistent but false data streams leading to wrong decisions. With diminishing isolation, security of things becomes crucial.

The paper focuses on the client side of such a heterogeneous network – the (logically) local network of things, using multi-hop ad-hoc connections if transmission range is too short. The energy and power limitations, specialized hardware, wireless communication and minimal manual configuration characterizing most of smart things are also typical in wireless sensor networks (WSNs), making them a proto-example of a network of things. Many of the results of research in WSNs security are therefore almost directly applicable to other devices. The main difference is that a single WSN is usually rather homogeneous, while in case of the Internet of Things (IoT) the devices may be completely different in both hardware and software.

Detection of compromised nodes is most complicated in this resource- and energy-limited part of the heterogeneous network, where the extra load introduced by the detection mechanisms becomes too large. The tradeoff between having an insecure network or expending energy and resources on protective measures could be eliminated by introducing more powerful nodes dealing with this task. Unfortunately, providing sufficient network coverage would require many such nodes, effectively multiplying the system cost beyond sensible limits.

In this paper a workaround limiting the cost of detection nodes is proposed by allowing each detector to monitor multiple locations through mobility.

The paper starts with a discussion of major threats classes to such networks of small devices in Section 2. Section 3 provides an overview of the approaches toward securing such networks. Section 4 presents authors idea of a mobile intrusion detection system (IDS). A short conclusion is given in Section 5.

2. Threats to the Network

There are many possible modes of attack against a sensor-like network of things [1]–[3]. In general, they can be grouped depending on several factors, such as the activity of the attacker (active or passive), computing power (sensor class or laptop class), location (logically inside or outside

the network), target layer and attack goals (communication obstruction, data capture, modification or data fabrication).

Attacks in physical layer are hard to prevent, as neither the electromagnetic medium nor the sensors themselves are (usually) physically protected. Active jamming attacks are therefore effective, although rather easy to detect. Passive sniffing is effective unless encryption is used. Physical attacks on nodes are possible even without any tools (destruction or theft of nodes). More advanced physical attacks are dangerous to the network as a whole, because of the virtually unlimited possibility of tampering with the hardware and programming (e.g. using JTAG interface).

Attacks on the data layer are more limited, usually focusing on flooding the medium with messages, or using standard violations such as long frames to cause collisions.

Attacks in the network layer are potentially very effective, but made more difficult by the fact that routing in such networks is not part of the standard and may be done using a variety of algorithms. Attacks in this layer usually focus on affecting the routing decisions in order to either obstruct communication as such, or to maximize the effectiveness of the limited number of malicious nodes in the network. In the first case, providing false information in the path building phase or sending many unnecessary path queries are simple and quite effective sensor-class attacks, causing additional unnecessary communication and computation by network nodes, draining batteries. The second group of attacks uses advertising great connection quality (or other methods) in order to direct as much of the network's traffic as possible through a malicious node. Then, after routing is established, the node can be used to monitor the traffic (sniffing) or obstruct it, either by blackholing the communication or by selective forwarding increasing loss frequency.

Finally, attacks in the higher layers (transport – application) are also possible and potentially most useful in case of targeted attacks. The range of possibilities is too wide to describe here. As simple examples consider attacks on time synchronization algorithms, node location or key distribution. An attack in this layer, conducted with a deep understanding of the goals and implementation of the network, can turn the network into an extremely dangerous misinformation tool.

3. Security Measures

Due to the limited computing power of nodes and their need to conserve energy, any security measure that requires a lot of computing activity on part of the network nodes is a mixed blessing. Another problem is the broadcast-based, self-organizing dynamic nature of such networks – even if not mobile, they must reorganize to allow for node malfunctions, etc. There are no natural policy enforcement points apart from the base station – any node in the network may be routed around. These problems result in a reduced choice of security solutions for networks of things.

3.1. Intrusion Prevention

The first layer of defense is provided by protection measures aimed at preventing successful penetration. In case of wireless networks of resource-limited devices this layer is unfortunately not as strong as in wired computer networks. Since the medium is freely accessible, the prevention must be applied at every point in the network. However, application of advanced mechanisms is made difficult by the limited computing resources and the need to preserve energy. Still, some steps have been made towards provision of important information security protections.

Proper application of cryptographic techniques can provide privacy, authentication and data integrity. Unfortunately, software implementations require a lot of operations, lowering battery life. Hardware support reduces this impact and is available in radio modules implementing the IEEE 802.15.4 standard [4]. An unfortunate limitation of this solution is the use of a single symmetric key. A lot of work towards introducing cryptographic protections to higher layers and enabling efficient and secure key distribution has been performed in recent years, including e.g. TinySec [5], MiniSec [6], ContikiSec [7], ZigBee, LEAP/LEAP+ [8].

Another protective measure, most effective not in prevention of attacks, but in network protection against already malicious nodes, is trust management [9]. Due to limited memory in network nodes the most practical approach is reputation based. An example of its application to sensor networks can be found in [10].

3.2. Intrusion Detection

Once a successful attack has been performed, the network might be operating with one or more malicious nodes. This constant threat to information security is often more dangerous than the initial attack. Detection of misbehaving nodes allows proper mitigation techniques to be applied, including blacklisting the node and routing around it or even physically removing it from the network (if possible). Many methods have been proposed to detect malicious nodes. Most of them have a common problem – secure delivery of detection information to the base station or secure propagation of information between non-malicious nodes. If a single malicious node is detected, it might not be the only one. If another one is in path of the warning message, it can easily render the detection system powerless. Therefore IDS alerts require either a separate secondary channel for propagation, or effective protection if the primary channel is used (separate path, encryption, etc).

The simplest form of detection of malicious behavior is the watchdog mechanism [11], using the shared medium aspect of wireless networks. The node sending a message can monitor the medium to verify whether the receiving node forwarded it correctly. There are many variations to this mechanism in the literature. A different, more sophisticated approach is using more advanced, rule-based intrusion detection systems [12], capable of detecting many different kinds of attacks. Both approaches require modification of

all or selected nodes. A standard watchdog mechanism only detects malfunctioning neighbors, so it must be active in most of the nodes to cover the entire network. An IDS may be somewhat effective even if only implemented at the base station, but delivers less information.
One more approach, often used in classical networks, is a server honeypot – a service existing only as a target for attacks. This approach might be applicable to WSNs by emulating a node on a more powerful device, waiting for messages modifying its state in illegal ways. This detection method would be effective against previously unknown types of attacks as long as the identity of the honeypot node remains secret.
All of the previously described approaches are generally passive – they either base detection on received traffic only or (in case of the server honeypot) respond to messages, but never initiate communication as part of the detection activity. An active alternative is a client honeypot – a node, which sends messages in order to verify whether they are properly forwarded to the base station or other target. The actual detection is performed at the receiving node, where any changes to the message or variation from the normal loss rate can be easily identified. The mechanism requires that both the sender and the receiver agree on the nature of the testing message or sequence of messages. This can be done through predefinition or by using a secondary channel for transmission of this information. Note that predefined messages are easier to learn and avoid at malicious nodes.

4. Movement-assisted Threat Monitoring in WSN

Taking into account limited resources of sensors, collection and analysis of data concerned with network security are usually performed in a periodic manner and carried out by selected devices implementing threat monitoring capabilities. However, in general, it is possible to extend the functionality of all nodes and to implement permanent monitoring. Regardless of the selected monitoring scheme, a common objective of all security systems is establishing a safe and reliable communication channel for exchanging security information i.e. reporting, alerting, control traffic, etc. between nodes. Therefore the communication can be organized in several ways:

- by utilizing the transmission channels already set up to propagate data in the system,
- by creating channels using disjoint logical connections within existing networks,
- by adding extra nodes to create a separate sensor network which shares the same transmission medium,
- by equipping sensor nodes with additional hardware modules (Wi-Fi, GSM, etc.) that can be used to establish an additional communication channel.

Finally, threat monitoring can be successfully supported by controlled geographic migration of sensors that have locomotion. A mobility of sensors is leveraged recently for many WSN applications. Using mobile platforms to assist sensor placement in a working space can significantly enhance the capability to monitor the data and detect attacks.

4.1. MOTHON System Overview

The authors have proposed a novel approach to threat monitoring in WSN. In presented threat detection system one or several mobile sensors implementing threat detection functionality are forced to move to desired directions. As it is presented in Fig. 1, due to the ability to change the location of a sensor node, information about security events can be passed directly to the network sink (IIIa – after completion of all tasks, IIIb – after threat detection) or indirectly via other nodes from another area of the monitored network.

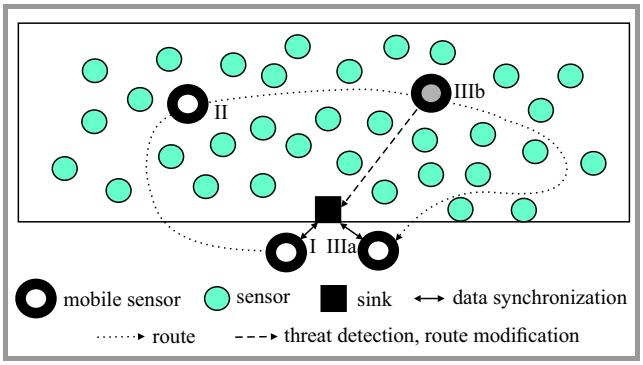


Fig. 1. The concept of movement-assisted threat monitoring: I – task order phase, II – performing actions phase, III – reporting phase (a – after completion of all tasks, b – after threat detection).

The MOTHON (MObile THreat mONitoring for WSN) system implements the concept depicted in Fig. 1. It is composed of three components: a mobile platform (MP), threat monitoring sensor (TMS) and management station (MS) responsible for controlling of MP and TMSs (see Fig. 2). It is assumed that the mobile platform can carry one or several TMS sensors. All these sensors can be placed in any locations in a workspace.
The MOTHON operates in three stages. The management station initiates the first stage. The aim of the this stage is to synchronize and gather data about a given network (topology, characteristics of nodes, etc.) and define monitoring mode and plan. Passive and active modes of threat detection are considered. Next, all data related to decisions done by MS are transferred to a MP and TMS sensor (or sensors) dedicated to threat monitoring, and the system switches to the second stage.
Threat monitoring sensor is carried to the desired destination by mobile platform (MP). After placing at the target location TMS starts to perform assigned tasks concerning the threats detection. The third stage relates to re-synchronization of information between TMS and MS which can occur in two cases – after completion of all

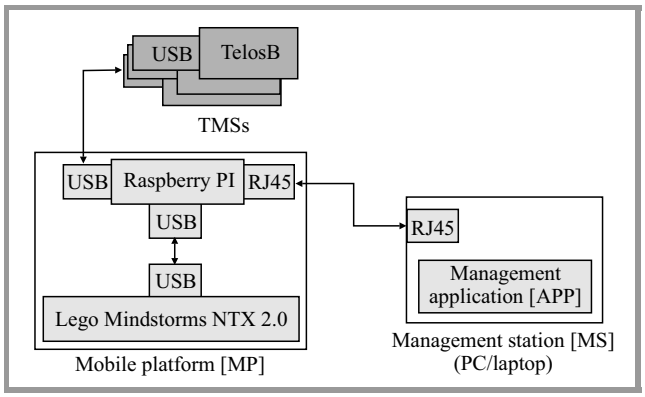


Fig. 2. MOTHON prototype architecture.

tasks or after threat detection. TMS transfers collected information, e.g. detected threats, additional statistics about traversed route, etc. to MS. It can be implemented in two ways:

- all data gathered by TMS are stored in MP, which transfers this data to MS directly – just after threat detection or after completion of all tasks,
- using existing communication channel via other nodes from another area of the monitored network.

In both cases MP can leave the TMS (with default network software) to avoid the occurrence of "temporary" nodes in the system. Moreover one mobile platform can carry multiple TMS sensors and place them in different locations in workspace.

4.2. Detection Methods

MOTHON can employ either active or passive methods for threat detection. Data analyses can be carried out either on-line by the TMS and mobile platform, or post factum by the mobile platform and management station.

Passive methods, which are based on analysis of information received from neighbors (IDS) or data sniffed from a shared medium (watchdog IDS), can keep a copy of the observed traffic for further analysis. This is not effective approach in case of threat monitoring with static nodes, but can be very valuable in case of mobile platform returning to sink from time to time. Moreover, simultaneous monitoring of the communication channel from several locations in the workspace can ease analysis by allowing detection of hidden and exposed nodes problems.

In contrast to passive methods, active solutions are not limited only to verification of individual sensor actions (correct operation of protocols, transmitted information, etc.). Active methods provide tools for verification of the network operation as a whole, e.g. verification the service packet forwarding over the network to the sink by sending a specific content, at the specific time and from specific place.

4.3. MOTHON Prototype Architecture

The prototype system composed of three elements is presented in Fig. 2. The management application provided by MS is a console tool is written in C++.

Mobile platform consists of two hardware components: Lego Mindstorms NTX 2.0 and Raspberry Pi single board microcomputer. Moreover it is expected, that MP will be equipped with GPS module or other localization system [13], [14].

Fulfilling all the tasks assigned by management application requires from this platform significantly greater capabilities in terms of processing power and energy resources. The key software components of MP are:

- **Control module** – the main module, responsible for communication with all other modules and creating the logic of solution based on information obtained from the management station.
- **Mobility module** – responsible for motion trajectory planning movement and speed calculation.
- **Threat analysis module** – used for data gathered from TMSs modules analysis and threat detection.

TMS is implemented over TelosB platform using Contiki OS. Eventually, to complete the system, an automation process enabling wireless communication between MP an MS must be added.

5. Conclusion

Starting with a review of threats and security measures applicable to wireless networks of resource-limited things, a new approach, introducing mobility as a way of overcoming the limitations of existing methods has been presented.

Mobility of a threat detection sensor should improve overall security state of monitored networks without any need to perform their reconfiguration or upgrade. The approach can be used in existing networks without any modifications to installed devices. The implementation details, such as means of mobility, depend on the target network – obviously a different solution is appropriate inside a building than in case of a network of oceanic drones.

Various methods of threat detection in MOTHON are currently under development. In future work authors plan to conduct experiments in testbed network to show the effectiveness of detection against different kinds of attacks.

References

[1] H. K. Kalita and A. Kar, "Wireless sensor networks security analysis", *Int. J. of Next Gener. Netw.*, vol. 1, no. 1, pp. 1–9, 2009.

[2] Z. S. Bojkovic, B. M. Bakmaz, and M. R. Bakmaz, "Security issues in wireless sensor networks", *NAUN Int. J. Commun.*, vol. 2, no. 1, pp. 106–115, 2008.

[3] I. Butun, S. D. Morgera, and R. Sankar, "A survey of intrusion detection systems in wireless sensor networks", *IEEE Commun. Surveys & Tutorials*, vol. 16, no. 1, pp. 266–282, 2014.

[4] IEEE standard for part 15.4: IEEE Std. 802.15.4, IEEE, New York, Oct. 2011.

[5] C. Karlof, N. Sastry, and D. Wagner, "TinySec: A link layer security architecture for wireless sensor networks", in *Proc. 2nd ACM Conf. on Embedded Netw. Sensor Syst. SenSys 2004*, Baltimore, Maryland, USA, 2004, pp. 162–175.

[6] M. Luk, G. Mezzour, A. Perrig, and V. Gligor, "MiniSec: a secure sensor network communication architecture", in *Proc. 6th Int. Conf. on Inform. Process. in Sensor Netw. IPSN'07*, Cambridge, MA, USA, 2007.

[7] L. Casado and P. Tsigas, "ContikiSec: A secure network layer for wireless sensor networks under the Contiki operating system", in *Proc. 14th Nordic Conf. on Secure IT Syst.: Identity and Privacy in the Internet Age NordSec 2009*, Oslo, Norway, 2009, pp. 133–147.

[8] S. Zhu, S. Setia, and S. Jajodia, "LEAP+: Efficient security mechanisms for large-scale distributed sensor networks", *ACM Trans. on Sensor Netw.*, vol. 2, no. 4, pp. 500–528, 2006.

[9] A. Felkner, "How the Role-based trust management can be applied to wireless sensor networks", *J. Telecommun. Inform. Technol.*, no. 4, pp. 70–77, 2012.

[10] G. Saurabh, L. K. Balzano, and M. B. Srivastava, "Reputation-based framework for high integrity sensor networks", *ACM Trans. on Sensor Netw.*, vol 4, no. 3, 2008.

[11] F. Hu, J. Ziobro, J. Tillett, and N. K. Sharma, "Secure wireless sensor networks: problems and solutions", *J. Syst., Cybernet. and Inform.*, vol. 1, no. 4, pp. 90–100, 2003.

[12] G. Huo, X. Wang, "DIDS: A dynamic model of intrusion detection system in wireless sensor networks", in *Proc. IEEE Int. Conf. on Inform. Autom. ICIA 2008*, 2008, ZhangJiaJie, China, pp. 374–378.

[13] M. Marks, E. Niewiadomska-Szynkiewicz, and J. Kolodziej, "An integrated software framework for localization in wireless sensor network", *Comput. and Inform.*, vol. 33, no. 2, pp. 369–386, 2014.

[14] M. Marks, E. Niewiadomska-Szynkiewicz, and J. Kolodziej, "High performance wireless sensor network localization system", *Int. J. Ad Hoc and Ubiquitous Comput.*, vol. 17, no. 32, pp. 122–133, 2014.

Adam Kozakiewicz got his M.Sc. in Information Technology and Ph.D. in Telecommunications at the Faculty of Electronics and Information Technology of Warsaw University of Technology (WUT), Poland. Currently he works at NASK as Assistant Professor and Manager of the Network and Information Security Methods Team, also as part-time Assistant Professor at the Institute of Control and Computation Engineering at the WUT. His main scientific interests include security of information systems (especially industrial networks), parallel computation, optimization methods and network traffic modeling and control.

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

Krzysztof Lasota works as Research Associate at Network and Information Security Methods Team in the Research Division of NASK. He received his B.Sc. in Telecommunications (2010) and M.Sc. in Telecommunications (2011) from Warsaw University of Technology, Faculty of Electronics and Information Technology and is currently a Ph.D. student there. He participated in security-related projects at NASK, including FISHA, HoneySpider Network and Secure workstation for special applications. Currently he participates in the project "The system of secure IP communication provision for the power system management". Furthermore, his research aims at developing new methods for threat detection in wireless sensor networks.

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

Michał Marks received M.Sc. (2007) in Computer Science and Ph.D. (2015) in Automation and Robotics from the Warsaw University of Technology. Since 2007 with Research and Academic Computer Network (NASK). The author and co-author of over 30 journal and conference papers. His research area focuses on wireless sensor networks, global optimization, distributed computation in CPU and GPU clusters, decision support and machine learning.

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

Paper

Quaternion Feistel Cipher with an Infinite Key Space Based on Quaternion Julia Sets

Mariusz Dzwonkowski[1,2] and Roman Rykaczewski[1]

[1] *Faculty of Electronics, Telecommunications and Informatics, Department of Teleinformation Networks, Gdańsk University of Technology, Gdańsk, Poland*

[2] *Department of Radiological Informatics and Statistics, Medical University of Gdańsk, Gdańsk, Poland*

Abstract—In this paper Quaternion Feistel Cipher (QFC) with an infinite key space based on quaternion Julia sets is proposed. The basic structure of the algorithm is based on the scheme proposed in 2012 by Sastry and Kumar. The proposed algorithm uses special properties of quaternions to perform rotations of data sequences in 3D space for each of the cipher rounds. It also uses Julia sets to form an infinite key space. The plaintext is divided into two square matrices of equal size and written using Lipschitz quaternions. A modular arithmetic was implemented for operations with quaternions. A computer-based analysis has been carried out and obtained results are shown at the end of this paper.

Keywords—cryptography, lossless scheme, multimedia encryption, security.

1. Introduction

Quaternion encryption as presented in [1]–[3] uses the unique properties of quaternions in order to rotate vectors of data in a three-dimensional space. The concept, however, lacks general interest and is not popular, thus it is difficult to find a paper that would introduce any significant contribution to the field. However, an authors' papers [4]–[8] show different, possible implementations of quaternion encryption and they discuss the security aspect of each proposed algorithm.

The cipher proposed by the authors in [9] is a modification of the Feistel Cipher with the implementation of modular arithmetic. In this paper an another modification of the Feistel Cipher is proposed. In this modification, modular quaternion rotations are used to encrypt subsequent rounds without the need to use matrix multiplication as introduced in [9]. The proposed quaternion model features very fast computation advantages over its matrix-based counterpart in [9], thus this makes it a perfect match for encrypting multimedia. The specific properties of computations in the field of quaternions are covered more extensively in [10], [11]. Additionally, when encrypting a color image in RGB representation, it is possible to increase encryption efficiency even further because in that case a single quaternion can successfully store information about all three RGB channels.

It is important to note that the algorithm proposed here is part of an ongoing project, thus further studies on the method are necessary and are highlighted in the paper. The algorithm proposed here was originally designed to encrypt multimedia data, thus the security aspect is not the focus of this paper.

The paper is organized as follows. In Section 2 a brief introduction to quaternion calculus and quaternion Julia sets is provided. Section 3 describes the quaternion rotation concept as well as the application of quaternion Julia sets. Section 4 concerns the proposed encryption scheme with an illustration of the Quaternion Feistel Cipher. In Section 5 the simulation results are shown, the avalanche effects are discussed and the computation speed of the proposed algorithm with the AES algorithm is compared. Finally, in Section 6, the conclusions are drown.

2. Quaternion Calculus

Quaternions are hyper-complex numbers of rank 4 and have two parts – a scalar part and a vector part, which is an ordinary vector in a three-dimensional space $\mathbb{R}^3$. A quaternion q is defined by formula [10], [12]:

$$q = w + xi + yj + zk, \tag{1}$$

where w, x, y, z are real coefficients of quaternion q, and i, j, k are imaginary units with the following properties [10], [12]: $i^2 = j^2 = k^2 = ijk = -1$, $ij = -ji = k$, $jk = -kj = i$, $ki = -ik = j$. A quaternion could also be considered as a vector represented by a column matrix (all vectors in this paper are represented by column matrices) or as a composition of scalar part w and vector part $\vec{v}$

$$q = [w\ x\ y\ z]^T \quad or \quad q = (w,\ \vec{v}) = \left(w,\ [x\ y\ z]^T\right). \tag{2}$$

The sum of two quaternions q_1, q_2 is defined by adding the corresponding coefficients of those quaternions, i.e., in the same manner as for complex numbers [10], [13]:

$$q_1 + q_2 = (w_1 + w_2) + (x_1 + x_2)i + (y_1 + y_2)j + (z_1 + z_2)k. \tag{3}$$

The product of two quaternions is more complex due to the anti-commutativity of the imaginary units of quaternions. The product of the two quaternions q_1, q_2 consists of scalar

and vector products ("$\circ$" denotes the scalar product and "$\times$" denotes the vector product) [10], [13]:

$$q_1 \cdot q_2 = (w_1 w_2 - \vec{v}_1 \circ \vec{v}_2,\ w_1 \vec{v}_2 + w_2 \vec{v}_1 + \vec{v}_1 \times \vec{v}_2). \quad (4)$$

In this paper "$\cdot$" denotes the quaternion multiplication. Furthermore, it is important to define the other properties of quaternions: a conjugate q^*, a norm $\|q\|$ and an inverse q^{-1} of a quaternion q:

$$q^* = w - xi - yj - zk, \quad \|q\| = \sqrt{w^2 + x^2 + y^2 + z^2}, \quad (5)$$

$$q^{-1} = \frac{q^*}{\|q\|^2} = \frac{w - xi - yj - zk}{w^2 + x^2 + y^2 + z^2}. \quad (6)$$

It is important to notice that in the case of a unit quaternion, for which the norm is equal to 1, there is the following relation: $q^{-1} = q^*$.

2.1. Quaternion Julia Sets

Julia sets are produced by a procedure of repeated iterations [14], [15]. The polynomial used in the process of iteration is quadratic, cubic, quartic or any higher order degree [15]. A Julia set consists of all points $p \in \mathbb{C}$ for which a recursive sequence:

$$z_0 = p, \quad (7)$$

$$z_{n+1} = z_n^2 + c, \quad (8)$$

does not approach infinity. The parameter c in Eq. (8) is a complex number, i.e., a parameter determining the shape of the produced set. For a quaternion Julia set, the starting point p is determined in a three-dimensional space (1, i, j) for a constant dimension k. Parameter c is then considered a quaternion. An exemplary quaternion Julia set is shown in Fig. 1.

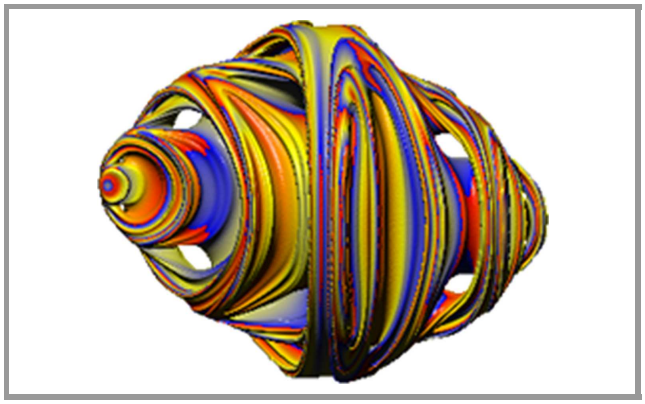


Fig. 1. Exemplary quaternion Julia set, number of iterations = 12, $c = 0.0882 + 0.1251i - 0.7555j + 0.1552k$, control number = 16, without an intersection plane.

In order to generate a quaternion Julia set, first must be set: all of the necessary initialization parameters, such as the number of iterations, to perform rule given by Eq. (8), the coefficients of quaternion c, a control number determining convergence of the starting points and the intersection plane.

3. Quaternion Rotation

The quaternion rotation can be performed by possessing a quaternion around which we will be rotating another quaternion. If the rotated quaternion as a data vector in three-dimensional space is considered, then the idea of quaternion encryption could be implemented.
Let us consider two quaternions $q = [w\ x\ y\ z]^T$ and $P = [0\ a\ b\ c]^T$, where a vector $[a\ b\ c]^T$, which represents a vector part of the quaternion P with a zero scalar part, will store information about a piece of data to be rotated around quaternion q. The obtained quaternion P_{rot} will be a spatial mapping of the rotated data vector $[a\ b\ c]^T$. The quaternion rotation is written as

$$P_{rot} = q \cdot P \cdot q^{-1}. \quad (9)$$

If we possess a tool, which can handle quaternion calculations, it is possible to implement encryption according to the Eq. (9).

3.1. Encryption Concept

The encryption method that was implemented in presented algorithm is entirely based on the quaternion rotation (9). It is possible to optimize the rotation process by extending the vector part of quaternion P in order to obtain a new quaternion B, as is shown in Formula (10):

$$P = \left(0, \begin{bmatrix} a \\ b \\ c \end{bmatrix}\right) \rightarrow B = \left(0, \begin{bmatrix} [a_1 & a_2 & a_3] \\ [b_1 & b_2 & b_3] \\ [c_1 & c_2 & c_3] \end{bmatrix}\right). \quad (10)$$

The encryption and decryption process for the quaternion method with the new extended quaternion B (meant to store data information) is shown in Eqs. (11) and (12), respectively.

$$B_{rot} = q \cdot B \cdot q^{-1}, \quad (11)$$

$$B = q^{-1} \cdot B_{rot} \cdot q, \quad (12)$$

where B_{rot} is the rotated (encrypted) quaternion B and q is the quaternion-key (encryption key).

3.2. Infinite Key Space

In order to generate an infinite key space it is necessary to first calculate a rotation matrix [1]–[3]. By using the Formula (9) and applying the Formulas (4)–(6) it is possible to introduce a rotation matrix [3], [12], [13], [16] and write:

$$\mathbf{P}_{\text{rot}} = \mathbf{\Gamma}(q)\mathbf{P}, \quad (13)$$

$$\mathbf{P} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}, \quad \mathbf{\Gamma}(q) =$$

$$\begin{bmatrix} w^2 + x^2 - y^2 - z^2 & 2xy - 2wz & 2xz + 2wy \\ 2wz + 2xy & w^2 - x^2 + y^2 - z^2 & 2yz - 2wx \\ 2xz - 2wy & 2yz + 2wx & w^2 - x^2 - y^2 + z^2 \end{bmatrix},$$

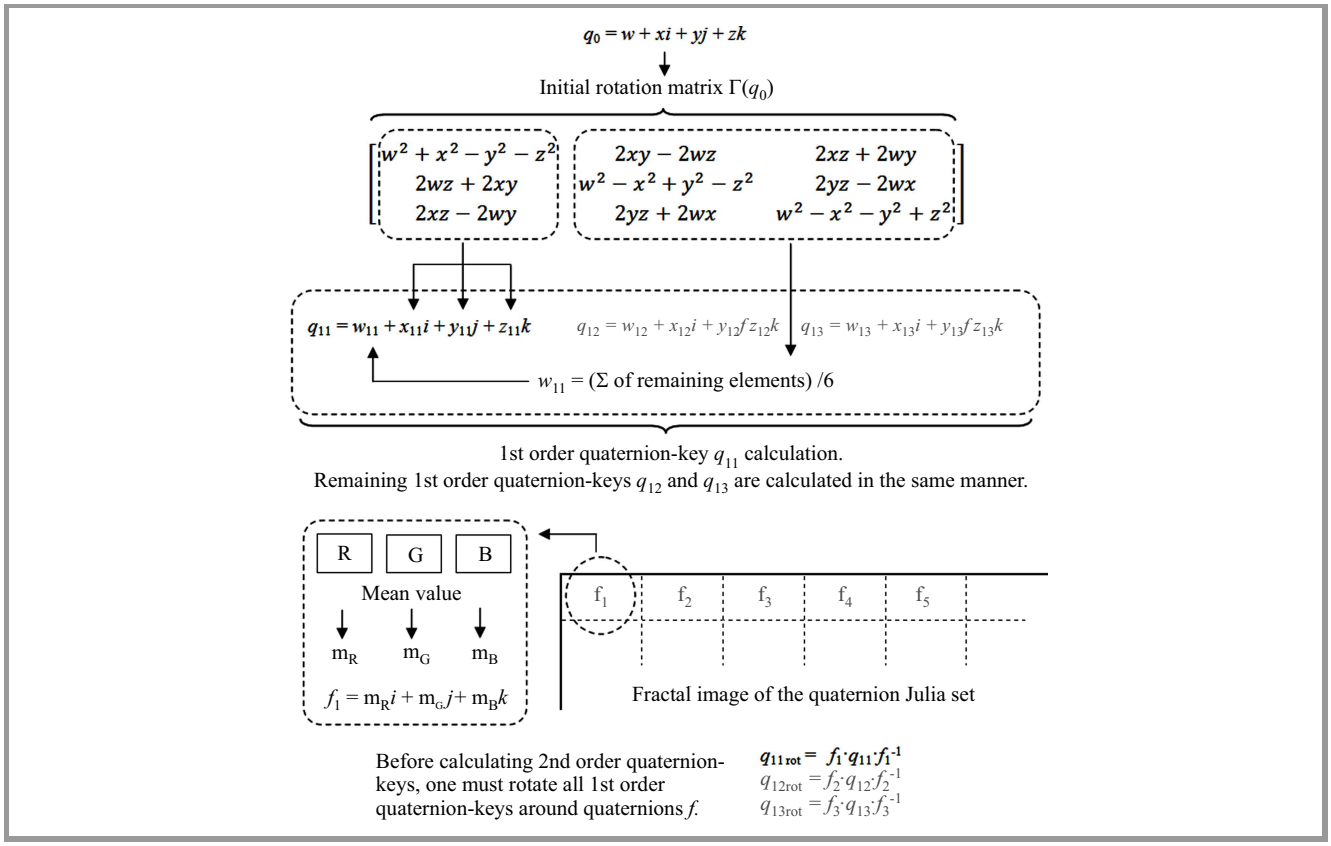


Fig. 2. Process of calculating 1st order quaternion-keys from an initial rotation matrix.

where $\mathbf{\Gamma}(q)$ is the rotation matrix calculated from the vector part of quaternion P_{rot}, which is defined by Formula (9). The rotation matrix is directly linked to quaternion q from which it was calculated.

The authors propose a key-generation algorithm based on concept [1], but additionally a quaternion Julia sets is introduced. The idea of the process is to treat elements of each column of the rotation matrix as coefficients (x, y, z) of subsequent quaternions from which other rotation matrices can be generated. In order to obtain coefficient w of subsequent quaternions, the mean value from six elements of the rotation matrix must be calculated, which were not used to determine coefficients (x, y, z), see Fig. 2.

Let us assume that the first quaternion q from which a rotation matrix was generated (13) is called an initial quaternion q_0. After grouping the elements of initial rotation matrix $\mathbf{\Gamma}(q_0)$ three quaternion-keys of 1st order (q_{11}, q_{12}, q_{13}) can be obrained. From these quaternions three rotation matrices can be generated, from which there is possibility to generate nine quaternion-keys of 2nd order. If we assume n as a rotation order, then the number of obtained quaternion-keys for the appropriate order n is equal to 3^n. The process is iterative, which means that in order to obtain higher order quaternion-keys we first need to define the rotation matrices of the lower orders.

However, before generating rotation matrices from quaternion-keys, we must first rotate every quaternion-key around quaternion f_i. Quaternions f_i are calculated from a random quaternion Julia set (Fig. 2).

The authors used the Quat generator [17] for visualization of quaternion Julia sets in 3D space as color images. The fractal image is divided into smaller fragments. The number of fragments is based on the size of the key space, e.g., for order = 2 nine quaternion-keys of 2nd order would be produced, thus we will need to divide the fractal image into 12 fragments (9 for quaternion-keys of 2nd order and 3 for quaternion-keys of 1st order). From each fragment a different quaternion f is calculated. Its coefficient w is always equal 0 and coefficients (x, y, z) represent a mean value calculated from the pixels values in the R, G and B channels of the fragments (Fig. 2).

The process of calculating 1st order quaternion-keys is shown in Fig. 2. The key generation process can be summarized by the following steps:

1) define q_0,
2) calculate $\mathbf{\Gamma}(q_0)$,
3) calculate $q_{11}\ q_{12}\ q_{13}$,
4) calculate $f_1 \cdot q_{11} \cdot f_1^{-1}\ \ f_2 \cdot q_{12} \cdot f_2^{-1}\ \ f_3 \cdot q_{13} \cdot f_3^{-1}$,
5) calculate $\mathbf{\Gamma}(q_{11\,rot})\ \mathbf{\Gamma}(q_{12\,rot})\ \mathbf{\Gamma}(q_{13\,rot})$,
6) calculate $q_{21}\ q_{22}\ q_{23}\ q_{24}\ q_{25}\ q_{26}\ q_{27}\ q_{28}\ q_{29}$,
7) ...

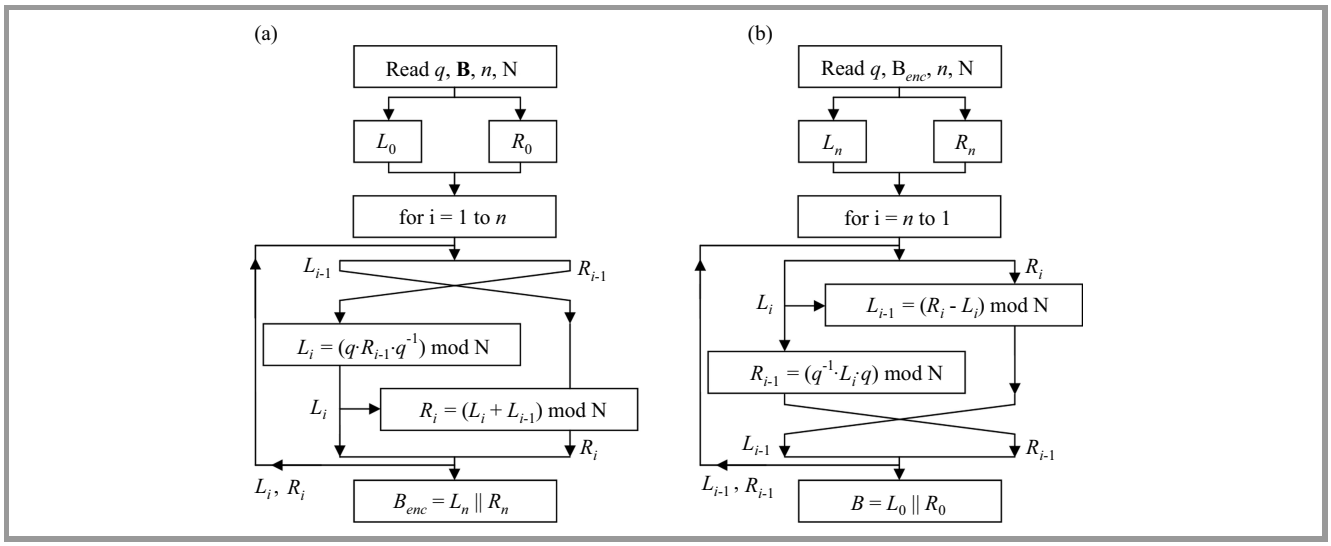


Fig. 3. Process of encryption (a) and decryption (b) for the proposed Quaternion Feistel Cipher.

4. Proposed Scheme

The proposed algorithm is designed to encrypt images (both color and gray-tone) but it can also be used to encrypt textual data. For the purpose of this paper an implementation for RGB color images is presented.

Let us now consider a plaintext B, which will be treated as RGB color image data. The plaintext can be written as three matrices **B**: $\mathbf{B_R}$, $\mathbf{B_G}$, $\mathbf{B_B}$. Each matrix **B** is of equal size with the image. Each element of all three matrices **B** is a value in the range of 0–255. For the purpose of the algorithm, all three matrices **B** should be rewritten as matrices with m rows and $2m$ columns, where m is calculated according to the rule:

$$m = \left\lceil \sqrt{\frac{width_B \cdot height_B}{2}} \right\rceil . \tag{14}$$

If the number of elements in such matrices exceeds the original amount of the images' pixels, then the additional elements are filled with random numbers in the range of 0–255. The three obtained matrices **B** ($m \times 2m$) are split into three square matrices: $\mathbf{L_{0R}}$, $\mathbf{L_{0G}}$, $\mathbf{L_{0B}}$ and three square matrices: $\mathbf{R_{0R}}$, $\mathbf{R_{0G}}$, $\mathbf{R_{0B}}$, each of size $m \times m$. All six square matrices are then written as components of two quaternions, L_0 and R_0 using the following rule:

$$L_0 = w_0 + x_0 i + y_0 j + z_0 k, \quad \text{where} \tag{15}$$
$$w_0 = 0, \quad x_0 = [\mathbf{L_{0R}}], \quad y_0 = [\mathbf{L_{0G}}], \quad z_0 = [\mathbf{L_{0B}}].$$

$$R_0 = w_0 + x_0 i + y_0 j + z_0 k, \quad \text{where} \tag{16}$$
$$w_0 = 0, \quad x_0 = [\mathbf{R_{0R}}], \quad y_0 = [\mathbf{R_{0G}}], \quad z_0 = [\mathbf{R_{0B}}].$$

The basic equations governing the encryption (17) and decryption (18) in proposed scheme are very similar in concept to the one presented in [9]. In this work, matrix multiplication is substituted by quaternion multiplication:

$$L_i = (q \cdot R_{i-1} \cdot q^{-1}) \mod N, \tag{17}$$
$$R_i = (L_{i-1} + L_i) \mod N \text{ for } i = 1 \text{ to } n,$$

$$R_{i-1} = (q^{-1} \cdot L_i \cdot q) \mod N, \tag{18}$$
$$L_{i-1} = (R_i - L_i) \mod N \text{ for } i = n \text{ to } 1.$$

The flowcharts depicting both the encryption and decryption processes of the proposed cipher are presented in Fig. 3.

According to Fig. 3, it should be noted that the symbol $\|$ is used for placing the vector part of a quaternion L adjacent to the vector part of quaternion R. The value n indicates the number of rounds in the cipher. Each round is encrypted with a different quaternion-key q_i, $i = 1, 2, \ldots, n$. The unique round keys are provided by the key generation algorithm (see Section 3.2).

4.1. Modular Arithmetic

Modular arithmetic operations were implemented in order to remain in the same field of values for data and cipher text. For that reason it was necessary to use Lipschitz integers (quaternions with integer components).

In order to calculate a modular inversion of any integer from the range of 0–255, and also a modular inversion of a Lipschitz integer, which is needed according to rotation rule (9), it was necessary to choose a special modulus value (a prime number) that together with all the integers in 0–255 would yield their Greatest Common Divisor (GCD) equal to 1. That is why for the RGB color images we cannot go for the most obvious choice and select a modulus of value 256, as such a number is not prime ($256 = 2^8$) and will not make it possible to calculate a modular inversion for most cases. Instead, a modulus equal to

prime integer N = 257 is used. Therefore values of the range 0–256 for all quaternions L_i and R_i can be obtained. It is important to note that despite the fact that in the encrypted image we obtained modified values from the range (0–256), the algorithm is so constructed that we will still be able to obtain the exact same image without any errors after the decryption process. The values obtained in the decryption process will be set into the appropriate range of 0–255.

The modular arithmetic with modulus 257 was implemented not only for the encryption/decryption process but also for the key generation algorithm (Section 3.2).

5. Simulation Results

The proposed scheme was scrutinized by computer-based simulation. The results of the encryption and decryption processes are shown in Fig. 4b and 4c, respectively.

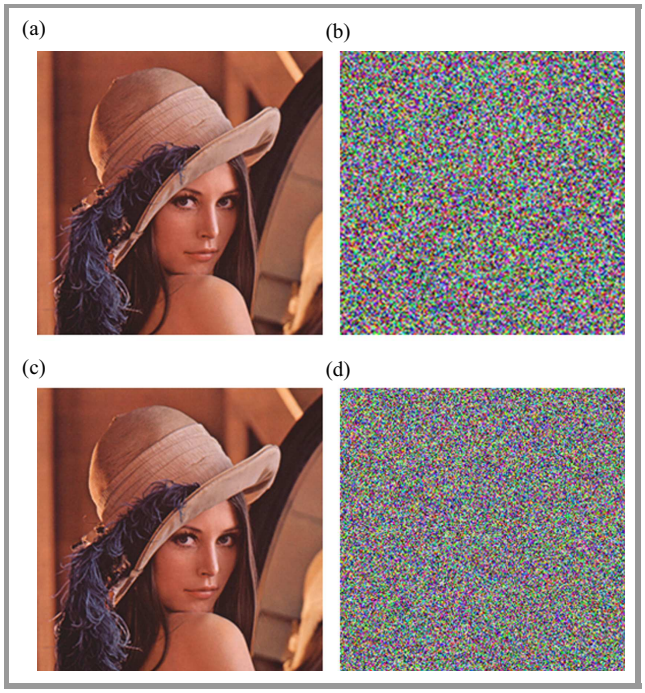


Fig. 4. Encryption and decryption of a color Lena image (https://en.wikipedia.org/wiki/Lenna) performed by the proposed quaternion cipher: (a) original image, (b) encrypted image, (c) decrypted image, (d) decrypted image using one key that is different in its binary form (1 bit) from the original one.

In Fig. 5, two histograms, i.e. of the original and the encrypted color Lena image, are shown. One can notice that the pixel values of the encrypted image are governed by a uniform distribution.

5.1. Randomness Tests

Chi-squared tests of randomness based on the diehard package [18] as well as on the freeware software Cryptool [19] were implemented in order to estimate the security level

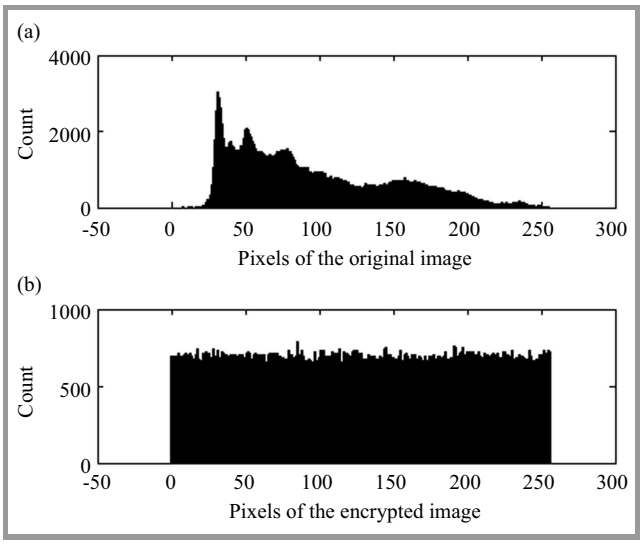


Fig. 5. Histograms for a color Lena image: (a) original image, (b) encrypted image.

of the proposed encryption model. The primary factor, which informs us about the effectiveness of the obtained randomness in encrypted data is a parameter called the p-value. Its values differ from 0 to 1, and the authors always aim for values as close to 0.5 as possible. If the p-value equals 0 or 1, that means that the encrypted data fails a particular randomness test. The obtained results for selected diehard randomness tests are shown in Table 1.

Table 1
p-values for different randomness tests from the diehard package

	Birthday spacings	Binary rank	Parking lot
p-values	0.613	0.585	0.209
	The 3Dsphere	Up-down runs	Craps
p-values	0.293	0.147	0.692
	DNA	Count-the-1's	OQSO
p-values	0.726	0.241	0.363
	OPERM5	Minimum distance	Overlapping sums
p-values	0.588	0.811	0.534

The authors also analyzed the randomness of proposed algorithm by using the Cryptool [19]. It offers 6 statistical tests: the Frequency Test, Poker Test, Runs Test, Serial Test and two additional tests embedded in Cryptool's battery test: Long-Run Test and Mono-Bit Test. Presented algorithm successfully passes all of them.

5.2. Avalanche Effect

In order to study the avalanche effect, let us consider a color Lena image as a plaintext. If 1 bit in one of the calculated

round keys (quaternion-keys of higher order) is changed, a new cipher text (encrypted image) will be obtained. Such a cipher text will yield a nearly 50% difference in its binary form in reference to the original cipher text.

The same situation is achieved when changing 1 bit in the initial encryption key (initialization quaternion). The difference in binary form of the modified cipher text and the original cipher text is also very close to 50%.

Let us now consider an example shown in Fig. 4. The aim is to encrypt a color Lena image. In the encryption process 9 rounds of the proposed algorithm are used. Each round has its own unique key (quaternion-key of higher order). Assume that the attacker knows 8 unique keys and possesses substantial knowledge about the last key (let us assume that the only difference is 1 bit in the binary form in the value y of the last quaternion-key). Decryption of the image performed by the attacker in such a scenario is shown in Fig. 4d. The presented result proves to be a strong avalanche effect.

5.3. Computation Speed

The authors analyzed the computation speed of proposed algorithm in comparison to the Advanced Encryption Standard (AES). For the purpose of this test the fastest AES version is implemented, i.e. AES-ECB, and its capabilities on the same machine as for our quaternion algorithm are measured. The machine used for the test was: Intel Core i5-3570 CPU @ 3.40 GHz, 16 GB RAM, and simulation environment was Matlab. The results of the comparison for a color Lena image are presented in Table 2. The expected values with confidence intervals calculated from 20 simulations are shown. The expected values presented in Table 2 are an estimation of the expected value μ, determined according to equation [20]:

$$P\left(\overline{X} - t_\alpha \frac{S}{\sqrt{N-1}} < \mu < \overline{X} + t_\alpha \frac{S}{\sqrt{N-1}}\right) = 1-\alpha, \quad (19)$$

where $\overline{X}$ is the expected value of the sample, S is the standard deviation of the sample, N is the size of the sample (20 simulations), t_α is a value obtained from the t-Student table for $N-1$ degrees of freedom and $1-\alpha$ is the confidence coefficient equal to 95%.

The initialization time refers to the time needed to calculate all of the necessary initialization parameters/values/matrices in order to perform the encryption/decryption. For AES, the initialization parameters refer to: the substitution box and its inverse, an arbitrary 16-byte cipher key, key expansion, a polynomial transformation matrix and its inverse. For the proposed algorithm, the initialization parameters refer to: the components of the initialization quaternion, initialization values for the quaternion Julia fractal, and the rotation order for which an appropriate key space is calculated (Fig. 2).

According to the results as presented in Table 2, one can see one of the main advantages of proposed algorithm, i.e. its fast computation speed. Moreover, in a practical scenario, encryption with AES could take even more time, especially considering the fact that a more secure implementation will be required, e.g., AES-CBC, AES-CTR, AES-OFB, or AES-OCB.

Table 2
Comparison of the computation speed of AES and the proposed algorithm for a color Lena image of size 243×243 pixels

AES-ECB	$\overline{X}$	$\frac{t_\alpha S}{\sqrt{N-1}}$
Initialization time [s]	0.6226	4.851E-03
Encryption time [s]	68.05	0.1416
Decryption time [s]	98.55	0.2095
Total time [s]	167.2	–
Proposed algorithm	$\overline{X}$	$\frac{t_\alpha S}{\sqrt{N-1}}$
Initialization time [s]	0.05428	9.101E-03
Encryption time [s]	0.2643	2.018E-03
Decryption time [s]	0.2642	1.270E-03
Total time [s]	0.5829	–

6. Conclusions

According to the presented simulation results, it is relatively easy to show that a very good randomness of bit sequences in an encrypted image can be obtained using the proposed encryption scheme. Moreover, one of the main advantages of QFC is its fast computation speed. Because of the structure of the algorithm, the specific properties of quaternions and enormous key space, the algorithm's resistance to cryptanalytic attacks should significantly exceed multimedia encryption requirements. Of course, many possibilities exist according to which the proposed model could further be improved, e.g., main research focus now is to introduce additional operations to the cipher which will further increase its robustness capabilities.

It is important to note that when using the proposed key generation scheme, the number of possible encryption keys for each round is particularly large because of the possibility of setting the rotation order, the values of the 4 parameters of the initialization quaternion and the values for all quaternions f calculated based on a random quaternion Julia fractal image [1]–[3], [12].

References

[1] T. Nagase, M. Komata, and T. Araki, "Secure signals transmission based on quaternion encryption scheme", in *Proc. 18th Int. Conf. Adv. Inform. Netw. Appl. AINA 2004*, Fukuoka, Japan, 2004, vol. 2, pp. 35–38.

[2] T. Nagase, R. Koide, T. Araki, and Y. Hasegawa, "A new quadripartite public-key cryptosystem", in *Proc. Int. Symp. on Commun. and Inform. Technol. ISCIT 2004*, Sapporo, Japan, 2004, pp. 74–79.

[3] T. Nagase, R. Koide, T. Araki, and Y. Hasegawa, "Dispersion of sequences for generating a robust enciphering system", *Trans. Commun. Inform. Technol. (ECTI-CIT 2005)*, vol. 1, no. 2, pp. 9–14, 2005.

[4] M. Dzwonkowski and R. Rykaczewski, "A new quaternion encryption scheme for image transmission", *ICT Young 2012 Conf.*, Gdańsk, 2012, pp. 21–27.

[5] M. Dzwonkowski and R. Rykaczewski, "Quaternion encryption method for image and video transmission", *Przegl. Telekom. + Wiad. Telekom.*, vol. 8–9, pp. 1216–1220, 2013.

[6] B. Czaplewski, M. Dzwonkowski, and R. Rykaczewski, "Digital fingerprinting based on quaternion encryption for image transmission", *Przegl. Telekom. + Wiad. Telekom.*, vol. 8–9, pp. 792–798, 2013.

[7] B. Czaplewski, M. Dzwonkowski, and R. Rykaczewski, "Digital fingerprinting for color images based on the quaternion encryption scheme", *Pattern Recogn. Lett.*, vol. 46, pp. 11–19, 2014.

[8] M. Dzwonkowski and R. Rykaczewski, "A quaternion-based modified feistel cipher for multimedia transmission", *Przegl. Telekom. + Wiad. Telekom.*, vol. 8–9, pp. 1177–1181, 2014.

[9] V. U. K. Sastry and K. A. Kumar, "A modified feistel cipher involving modular arithmetic addition and modular arithmetic inverse of a key matrix", *Int. J. Adv. Comp. Sci. Appl. (IJACSA 2012)*, vol. 3, no. 7, pp. 40–43, 2012.

[10] R. Goldman, "Understanding quaternions", *Graphical Models*, vol. 73, no. 2, pp. 21–49, 2011.

[11] R. Goldman, *An Integrated Introduction to Computer Graphics and Geometric Modeling*. New York: CRC Press, 2009.

[12] F. Zhang, "Quaternion and matrices of quaternions", *Linear Algebra and its Applications*, vol. 251, pp. 21–57, 1997.

[13] D. Eberly, "Quaternion Algebra and Calculus", Geometric Tools, LLC, 2010 [Online]. Available: http://www.geometrictools.com/Documentation/Documentation.html

[14] G. M. Julia, "Memoir on iterations of rational functions", *J. de Mathématiques pures et appliquées*, 4th tome (83th volume of the collection), pp. 47–246, 1918.

[15] A. Douady, "Julia Sets and the Mandelbrot Set", in *The Beauty of Fractals: Images of Complex Dynamical Systems*, H.-O. Peitgen and P. H. Richter, Eds. Berlin: Springer, 1986, p. 161.

[16] C. Corrales-Rodrigáñez, "Rotations and units in quaternion algebras", *J. Number Theory*, vol. 132, no. 5, pp. 888–895, 2012.

[17] Quat – A 3D-Fractal-Generator, Version 1.20 [Online]. Available: http://www.physcip.uni-stuttgart.de/phy11733/quat_e.html

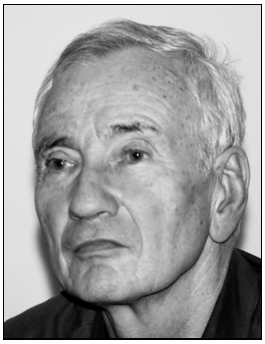
[18] Robert G. Brown's General Tools Page [Online]. Available: http://www.phy.duke.edu/∼rgb/General/dieharder.php

[19] Cryptool Portal [Online]. Available: http://www.cryptool.org/en/

[20] P. Armitage, G. Berry, and J. N. S. Matthews, *Statistical Methods in Medical Research*, 4th ed. (revised). Chichester: Wiley – Blackwell, 2001.

Mariusz Dzwonkowski received his M.Sc. Eng. degree in Telecommunication from Department of Teleinformation Networks, Gdańsk University of Technology. He is now working toward his Ph.D. degree in Telecommunication from Gdańsk University of Technology. He is currently a lecturer at Medical University of Gdańsk, Poland in Department of Radiological Informatics and Statistics. His research interests include steganography, cryptography with emphasis on quaternion encryption, network security and image processing.
E-mail: mar.dzwonkowski@gmail.com
Gdańsk University of Technology
Faculty of Electronics, Telecommunications and Informatics
Department of Teleinformation Networks
Gabriela Narutowicza st 11/12
80-233 Gdańsk, Poland

Medical University of Gdańsk
Department of Radiological Informatics and Statistics
Tuwima st 15
80-210 Gdańsk, Poland



Roman Rykaczewski received his M.Sc. (1968) and Ph.D. (1975) from Gdańsk University of Technology – Faculty of Electronics, Telecommunications and Informatics. From 1968 to the present he is working as an academic teacher at Gdańsk University of Technology, Faculty of Electronics, Telecommunications and Informatics. His current research mainly focuses on cryptography, watermarking and steganography.
E-mail: romryk@eti.pg.gda.pl
Gdańsk University of Technology
Faculty of Electronics, Telecommunications and Informatics
Department of Teleinformation Networks
Gabriela Narutowicza st 11/12
80-233 Gdańsk, Poland

Paper

Evaluation of the Cyber Security Provision System for Critical Infrastructure

Jacek Jarmakiewicz, Krzysztof Maślanka, and Krzysztof Parobczak

Faculty of Electronics, Military University of Technology, Warsaw, Poland

Abstract—The paper presents an assessment of the functional mechanisms that are part of the security system for the power grid control. The security system, its components, and the real time processes for the control of electricity supply were defined. In particular, SCADA protocols used in the control system and mechanisms for transferring them between the control center and actuators were identified. The paper also includes presentation of a test environment that is used for developed security mechanisms evaluation. In the last fragment of the paper, the test scenarios were formulated and the results obtained in the cyber security system were shown, which cover security probes reaction delay, forged malicious IEC 60870-5-104 traffic detection, DarkNet and HoneyPot interception of adversary actions, and dynamic firewall rules creation.

Keywords—critical infrastructure, cyber security system, power system security, SCADA system.

1. Introduction

The Critical Infrastructure (CI) includes supply of energy, raw materials and energy consumption, communication, computer networks, financial services, food and water supply, healthcare system, transport, emergency medical services. A CI ensures the continuity of public administration, production, storage, handling and use of chemicals and radioactive substances, including pipelines of hazardous substances. CI also comprises real and cyber systems (and devices or facilities included in these systems) necessary for minimal operation of the economy and the state.

In many countries energy supply is controlled in real-time from the Load Frequency Control (LFC) system [1]. The electricity is generated on the basis of electrical devices requests and is adjusted to their load. The power supply realization is centralized in the so called "secondary control process", which means that the produced power is controlled via the Central Control System (CCS). Any disturbance in this system can have significant impact on all industries and citizens.

Recently, the Supervisory Control And Data Acquisition (SCADA) systems have been used to control power supply processes. In the past, such systems run over dedicated analog lines and networks with vendor specific protocols, hardware and software. The network for power generation control was, and still should be, isolated from the public networks. Control systems such as SCADA, power transmission management system, centralized LFC system and intelligent field devices, e.g. Remote Terminal Unit located in the Control and Supervisory Substation (CSS) and Intelligent Electronic Devices (IED) create new concerns for the cyber security.

Today open transmission protocols are broadly used and computers with commercial operating systems work as IED. It significantly improves automation efficiency and decreases costs, but certainly it also increases system vulnerabilities and decreases the security level. Nowadays SCADA control commands and responses flow across IP networks and over IP protocol stack. Control processes run in real-time and are managed by power station generators.

Cyber security in information technology is used to protect computers and networks from intentional and unintentional events and malicious attacks. Many research and development programs in SCADA security assessment and analysis have been conducted, including risks analysis, vulnerability and security assessment, penetrating testing and evaluation, system simulation and emulation. Many works and articles [2]–[7] related to this subject have been written, but it seems that they don't investigate the essence of the problem. Certainly the importance of this issue and security restrictions don't allow publications related to it. Due to the differences in equipment and technologies used in the industrial control networks, the security solutions are unique and must be adapted to the specific CI, because of the limitations relate to the system generation environment, tools and software libraries used for its development, as well as predispositions of design and generation teams documented with security certificates.

The paper presents a developing process of security system for the critical infrastructure. As a result of the design works, solutions were developed and implemented that are aimed at detection and reaction to cyber incidents such as attacks from outside the system and authenticated, but unauthorized actions from inside the power system. The elaborated solutions and mechanisms were combined into a cyber security and incident response system. At present, the security system mechanisms developed by authors are being assessed in the test environment [8]. A number of tests and tools are prepared that are used for evaluation of the efficiency of presented solutions. The research is carried out in quasi-real conditions, whereby the threats

and attacks are detected in the ICT traffic that comes from the real control system of the national energy sector. The power control network environment was very accurately mapped. The research is conducted with the use of tools elaborated for performance of attacks on SCADA control systems that authors use in order to adjust the sensitivity of probes and solutions. The developed system can operate in both a multi-domain, dispersed environment and in a multi-domain, centralized environment, depending of the stakeholders' requirements. Soon after mechanisms adjustment, the system will be used in a real control and supervision station. The authors believe that the developed cyber security system will be successfully implemented in the national power system.

2. Related Works

Many research centers develop and adapt the security systems for the critical infrastructure in an environment mapping the real control mechanisms. At least several approaches to develop a SCADA system testbed were identified, varying from high-level modeling and simulation frameworks interconnecting simulation environments, to specialized tools recreating client-server interactions on protocol level. These environments are prepared to test the solutions for attacks detection and IT protection of the critical infrastructure systems.
The need of European SCADA Security Testbed creation is subject of [9]. SCADA LAB project [10] was an example of such an initiative, which lasted for 2 years, and constituted coordinated efforts of many European partners. The benefits of the project include:

- definition of security requirements for industrial control systems and a methodology for security testing,
- development of a security laboratory reflecting real environment,
- creation of tools to facilitate efficient testing channels and remote testbed as a service, and for effective sharing of results and experiences,
- smart online and offline dissemination of results for beneficiaries in public and private sectors in the EU.

Research work described in [11] presents a high abstraction level method of modeling CI, as transformation from a detailed "potential" system model to a "specific" model reflecting a particular instance of the system.
Authors of [12] describe the Critical Infrastructure Protection and Resilience simulation (CIPRsim) modeling and simulation framework, which has the capability for simulation and visualization of effects and interdependencies associated with a hazard or threat event. The elements of a simulation model communicate through High Level Architecture (HLA) bus [13], which provides a common architecture for distributed modeling, component-based simulation and linking to real systems. Thousands of modeled objects took part in the simulation process, while threats were modeled analytically in order to stress HLA bus performance capabilities, rather than to assess security features of the evaluated CI system.
Only part of the existing solutions was created with security characteristics assessment in mind, the rest being focused on studying the interdependencies in critical infrastructures, information management performance and/or reliability, etc.

3. Analysis of the Domestic Power Distribution Control System

Efficient and proper work of complex power generation, transmission and distribution system is required for common access to the benefits of electricity. The system is characterized by simultaneous generation and consumption of energy, where practically there is no energy storage. The main task of the energy services is to constantly maintain appropriate settings in the system in order to generate the right amount of power to fulfill the ever-changing consumers needs, with appropriate quality parameters and in the agreed quantities. The process of energy supply is realized by power plants within less than 30 s since the moment the demand occurs. This process is controlled in real-time from the CCS. Generators are activated in the power plants by the LFC mechanism in the process of frequency and power control [1]. In CCS, the controller operates in real-time in a loopback. Currently, the power grid consists of 114 centrally controlled energy sources [14].
In LFC, the command and response process time equals less than 10 s (Table 1) The control is implemented using SCADA protocols IEC 60870-5-104, IEC 61850, which are encapsulated in TCP/IPv4 packets. In the case of failure, the control is taken over by redundant systems and manual control is possible as well.

Table 1
Processing and transmission estimated time of control systems by LFC

Name	Estimated time [s]
Downloading data from RTU CSS on the control area exchange lines	5
Front-end to LFC transmission	0.8
LFC processing	2
Transmission of control LFC to ICCP generator	2
Total	10

Generator turbines are controlled from the CCS by sending data through an independent wide area network based on the SDH technology (Fig. 1). Telecommunication cables are suspended on poles along with high voltage cables.

LFC CCS computer network is based on standard IPv4 and Ethernet with VLANs. The exchange of commands and responses with the generators and energy consumption readouts is performed in CSS. Control commands from the CSS are directed through switches and routers to WAN and are transferred to target routers of the power plants generators [15].

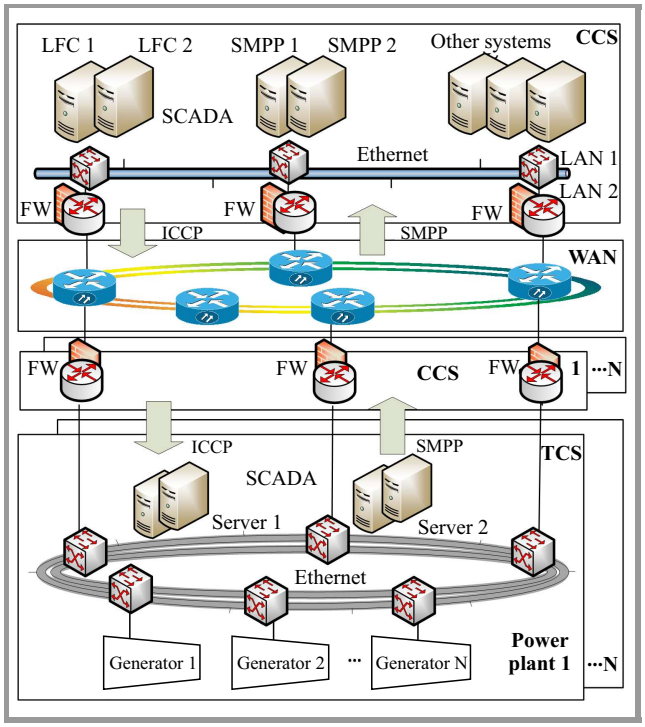


Fig. 1. The power distribution control system architecture.

General example of the station is shown in Fig. 2. Control data are sent to the CSS station, from where they are directed to servers, which send the commands to generators' drivers. The connection equipment and automation systems for the fields are controlled by the CSS as well. The CSS is also used for supervision and monitoring of the stations equipment and systems. At the CSS side of the power plant, the traffic flows by separated VLANs. In the CSS, there are field controllers which are responsible for connection processes of electrical circuits, protection of circuits and cooperation with power plants. Commands from the CCS are delivered to the CSS through routers and modems which are used to perform readouts from IEDs [2]. In addition to remote control from the CCS, it is possible to manually control field automation systems from the CSS, 400 kV protections from Human Machine Interface (HMI) and internal elements of the station from HMI Substation Control. The entire CSS is physically protected using an alarm and supervisory system. Control commands are performed by SCADA systems in the CSS station. SCADA systems are hard real-time systems because the completion of an operation after its deadline is considered useless and potentially can cause cascading effect and severe damage to expensive facilities.

4. Cyber Security System Objectives

The purpose of the developed CI security system is to ensure secure IP communication within the power grid management. The results of the works include a security system prototype providing:

- probing and correlating information with the use of probes and network sensors, aimed at handling,
- automated detection and tracking of threats and appropriate response measures,
- ensuring the security of ICT infrastructure of stations and technological communication through:
  - authentication,
  - advanced access control, e.g. with the use of security policies,
  - monitoring and filtering of management and control IP traffic transferring IEC protocols,
  - encryption of management messages,
  - monitoring of the status of the protected facility and secure storage of information,
  - honeypots and SCADA hardware emulation,
  - secure communication with the central device of Security Information and Event Management (SIEM) and Graphic User Interface (GUI),
  - documentation of management operations and detection of potential unauthorized inside operations.

5. Security System and Testbed Environment Overview

Electricity supply control network is an object of a too high strategic value for direct conducting of tests related to attacks and detection of threats because it could impose a high risk for the power system. Therefore, testing of protective mechanisms of control networks in the power industry requires organization of environment similar to the real one. This is way the environment for testing the cyber security system for the power grid control was developed in the project. Such an environment should reflect power control network elements and imitate processes implemented therein as reliably as possible [3]–[8]. A number of requirements could be formulated in relation to the testbed for cyber security, i.e.:

- similar network resources and protocols should be used that characterize the same vulnerabilities of the real system. The mechanisms should enable reflection of the network structure, elements configuration, routing mechanisms and set of the computer network protocols used;

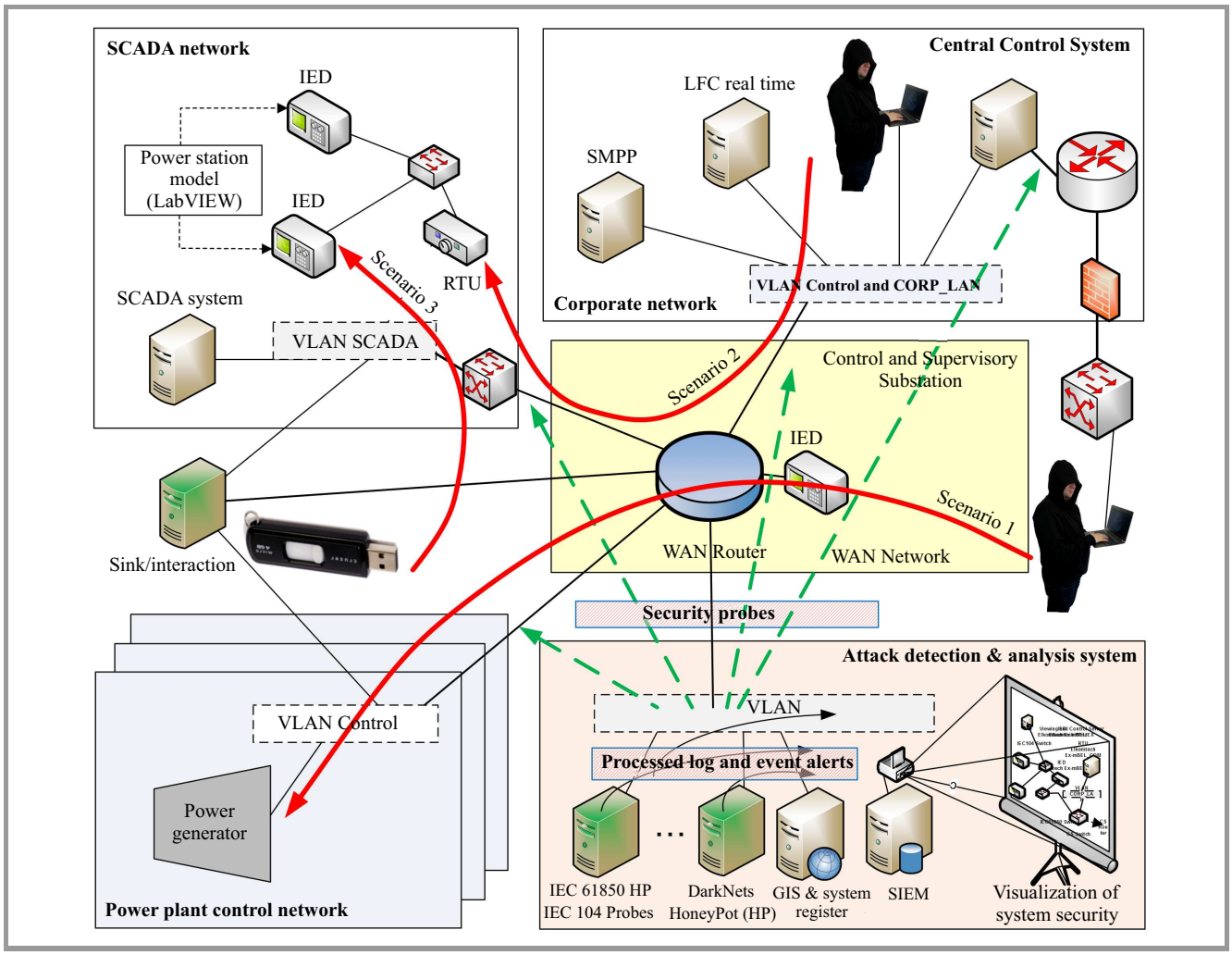


Fig. 2. The CSS station structure and attack scenarios.

- the processes realized in the real system should be reflected as accurately as possible. Controlled SCADA resources should be used as in the real system;
- data traffic exchanged in the testbed should correspond to flows in the real network.

In order to identify, and later reflect the real system properties and solutions for its protection, it is convenient, at the initial stage, to use the ontological model of the critical infrastructure system. It will allow identification of important system resources of the real environment and error prevention consisting in omission of elements influencing the entire system security. Afterwards, the system vulnerability should be determined and, as a result, solutions increasing the system security level should be proposed. Increased security of the control system requires identification of those system properties that will be controlled by the protection system [9].

The authors built a testbed consisting of one control center CCS and several substations CSS, as illustrated in Fig. 2. A communication subsystem is modeled in the form of switch and router is the central node. These elements provide communication within the entire power station, supporting the individual VLANs, and they create virtual network in the entire network using the VRF technology [16]. The individual subsystems of the CCS station operate in the independent VLANs, and they can communicate through a router and CSS firewall. Elements of the CSS system were developed using Cisco devices. The testbed includes real IED subsystem and fields emulators as part of the CSS, communicating using IEC 61850 and 60870-5-104 protocols. The Elkomtech devices were used for its construction: communication hub (RTU) Ex-mBEL_COM [17] and Ex-mBEL field controller [18] (Fig. 3). Extortions for IEDs (single-line-to-ground, symmetrical, short-circuit current) and readouts (measured values) are generated using a custom device and a station model prepared using the LabVIEW environment. Field controllers and the communication hub communicate through an industrial Ethernet switch using IEC-104 protocol. Display and management of the model power system is performed using WindEx software [19] located in the corporate network segment. The control and management processes are performed using IEC 61850 protocol.

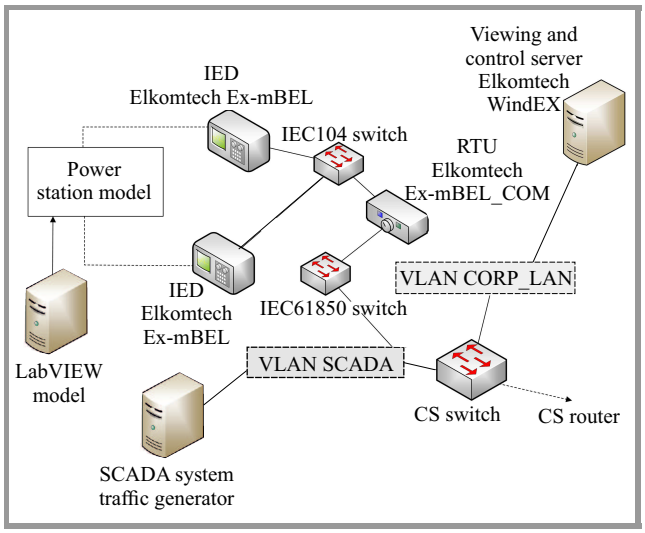


Fig. 3. The power station model and the SCADA system.

SCADA network includes also the SCADA probe (Snort, Bro) which enables monitoring of SCADA traffic. The remaining probes of the security system allow monitoring of i.e.:

- traffic incoming to the power station network, of which copy is fed to one of the probes from CS router,
- both the generation and non-generation traffic from the enterprise network and monitoring and control network – a copy of this traffic is fed on-line from CS switch.

The generated IP network traffic streams are based on traffic samples analysis captured at the boundary of the real industrial power system. The gathered traffic types covered LFC, corporate network, energy trading, measurement and control and network management.
Traffic samples were obtained using FPGA-based hardware probe with nanosecond accuracy and wirespeed record capability. Traffic capture includes corporate network segment traffic and control traffic data (IEC 60870-5-104). Collected samples were regenerated [8] prior to use in testbed generators and targeted to appropriate sinks.
Preliminary analysis of the collected traffic shows the possibility of data packets fields structure and values modification and therefore a chance to slightly overcome protocols inconveniences and limitations (required by industrial equipment manufacturers). These proprietary modifications can be analyzed with customized protocol analyzers, such as Sisco open source Ethernet analyzer based on WireShark for IEC-61850, IEC60870-6 TASE.2 [20].
In such a network (Fig. 2), which reflects the condition of the real power management network (Fig. 1), the functional tests of the developed protection mechanisms were conducted. The researchers intend to verify the efficiency of anomalies and attack detection by tools developed by us, i.e.:

- probes based on Snort and Bro software that are adapted for analysis of SCADA protocols in order to detect anomalies in the power control and management systems,
- commercial IDS/IPS probes that were previously purchased and are currently used in the power control and management network,
- HoneyPots, SCADA HoneyNets and DarkNets for monitoring and logging of all of the threats activities in ICS network,
- mediation device developed to normalize the messages obtained from the other security systems and elements,
- SIEM system gathering, analyzing and aggregating information received from abovementioned elements,
- databases gathering the history of power control and management conditions,
- Cyber security Visualization and Management System processing data developed in SIEM in real-time,
- developed tools and open source tools designed for verification of resilience of the power control and management systems.

6. The Use Cases for Evaluation of the Security System Elements

Functional tests were implemented in a quasi-real environment which is described in Section 5. The experiments were designed to verify the system ability to detect cyber attacks and to protect against them, as well as to adjust the sensitivity of probes and decoys developed in the project.
For the purpose of the experiments, test scenarios were defined, in which the probable directions and sources of attacks were provided. In particular, the scenarios relate to the following directions of attacks were defined:

1. from the Internet and over PSTN with the use of unauthenticated and unauthorized measures by hackers,
2. from the enterprise network, the attacks coming from authorized users of this network who, due to various reasons, attack the power control system,
3. from the control network by persons who know the effects of the attacks and due to personal and/or external reasons conduct attacks on the infrastructure,
4. from the control network by users who are not aware of the threats, authorized to resources, e.g. during a software update, a malware is installed and transferred along with the useful applications.

The attacks may be carried out from outside and inside the control system. They can be performed not only by external attackers, but also by e.g. bribed or intimidated employees, or those unaware of the threat.

(a)

```
[root@pl.bipse.mut.gen1 ~]#nmap -Pn -p 1-100 --min-rate 1000
172.15.102.104
```

Event Time	Phy Element	CSS	Domain	Msg
2015-04-01 13:02:27.797	sonda2-hnp	si2	mut	Many Rejected TCP Packets for IP Addr Pair
2015-04-01 13:02:23.971	ssin1	si2	mut	TCP Packet Rejected

(b)

```
[root@pl.bipse.mut.gen1~]#wget test@172.15.102.100 /  --ask-password
100%[============>] 204         --.-K/s   w  0s
2015-04-01 13:27:42 (10,4 MB/s) - saved `index.html' [204/204]
```

Event Time	Phy Element	CSS	Domain	Msg
2015-04-01 13:27:42.589	sonda2-hnp	si2	mut	TCP Connection Terminated
2015-04-01 13:27:42.589	sonda2-hnp	si2	mut	TCP Connection Terminated
2015-04-01 13:27:42.108	sonda2-hnp	si2	mut	TCP Connection Establish
2015-04-01 13:27:42.108	sonda2-hnp	si2	mut	TCP Connection Establish

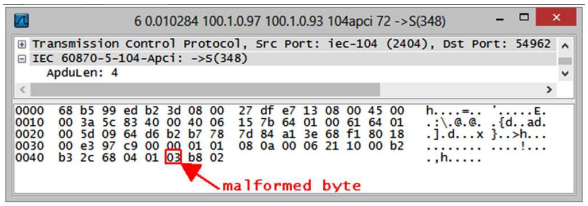
Fig. 4. Functional tests results of DarkNet and HoneyPot.

Let us analyze the scenario of attack from the Internet (no. 1) with the use of PSTN or from the direction of the enterprise network, in which the firewall protecting the control network access was breached (no. 2, Scenario 1 in Fig. 2). In this case, the symptoms of attack are traces left by the control station environment recognition applications. In the scenario with the hacker poorly acquainted with the environment, the attacker has to find out the control station structure, functions performed by the devices, their addresses and protocols used. Two-directional traffic monitoring, from and to the station, will be performed then. Collected features of monitored traffic will be sent toward the attacker. Such actions could be detected on the routers in the form of increased traffic. Scanning of addresses and/or ports may be detected by SCADA decoys – darknets (Fig. 4a) and honeypots (Fig. 4b), emulating operation of station devices.

The authors present the test results that confirm detection of unauthorized operations on the security elements (Fig. 4). The network is scanned directly to the unused DarkNets and HoneyPots addresses, and, as a result of referring to honeypots in GUI of the server with SIEM, the system service is informed on the attempt of unauthorized access to the resources. The next scenario relates to unauthorized operations performed from the control network (no. 3, Scenario 2 in Fig. 2). Unfortunately, this access is possible, e.g. as a result of the CSS personnel carelessness. Suppose that the engineering interface serving for updating the software is not secured and the attacker connected through an external telecom box. It would be also possible to access the CSS control network through a GPRS modem, which, despite prohibitions, was installed by the control devices manufacturer to facilitate the software update process. A dangerous attack would be that performed by an authenticated user authorized to operate within the CSS but, for e.g. religious reasons, wants to cause failure of the unfaithfuls' power supply system. There was once a case of an authorized employee of a large power plant suffering from a heartbreak, who attempted to shutdown the generators. In this scenario an alert information from HoneyPot, HoneyNet or DarkNet systems won't probably be received. The attacker is familiar with the station structure and probably knows that, apart from the regular devices, decoy devices operate there as well. Disconnection of the station devices would immediately result in generation of alerts in the supervision center. Therefore, the attacker wishing to be successful, will seek to incorrectly control the station devices of the power plant generator and possibly forge responses from the controlled device. In this class of attacks in the security system, SCADA probes adjusted to the device and analyzing the status history of the protected systems will be helpful. The SCADA probe is connected in parallel with the protected facility, but it exchanges alerts over a path isolated from the control network to the station security system, so its operation cannot be noticed even by an authorized employee.

Figure 5 presents functional test of SCADA device security. The traffic probe is connected to the IED and it analyzes the traffic and its responses – this is the Probe learning stage. It collects the behavior patterns of the server issuing commands to the IED and feedback sent to the server. After some time, the Probe enters the detection mode. The Probe constantly reports its activity in the facility security system. If it is switched off, the security system will generate an alert. If the Probe detects incorrect data in the control commands, an alert will be generated as well. The second case is presented in Fig. 5. The incorrect control signals are sent to the IED (c), due to which an alert occurs in the security system (d).

(c)



(d)

Event Time	Phy Element	CSS	Domain	Msg
2015-04-01 12:35:16	SCADA Probe	si1	mut	IP_VIOLATION
2015-04-01 12:35:15	SCADA Probe	si1	mut	WRONG_BYTE_2_OF_ACPI_S. Value: 3.

Fig. 5. Functional tests results of SCADA traffic probe.

The use of malware is the last presented scenario (case no. 4, Scenario 3 from Fig. 2), which is in fact not the last possible scenario of attacks. Such software may enter the control station along with the installed hardware or

updated software, or as a result of infection from USB drive or the station personnel laptop containing a virus. The older devices and those used in the station may already contain malware, which is waiting for the right moment to activate. Such cases already happened in the past, an example of which is the Stuxnet worm and its more advanced forms from Flame or Gauss platforms.

(e)
```
[root@pl.bipse.mut.si1.trgen test_mms]# ./kf-mrn-test-mms -s
172.15.102.150
MMS connected.
```

(f)

Event Time	Phy Element	CSS	Domain	Msg
2015-04-01 13:47:52	sonda2-hnp	si1	mut	TCP Connection Terminated
2015-04-01 13:47:51	sonda2-hnp	si1	mut	MMS Association Setup
2015-04-01 13:47:51	sonda2-hnp	si1	mut	TCP Connection Establish

Fig. 6. Functional tests results of DarkNet and HoneyPot.

In this scenario, detection of malware will be possible for each of the developed cyber security measures. If the device recognizes the network environment, it will be noticed on HoneyPot or DarkNet during scanning of addresses or ports in the control network. The alerts will be delivered by channels isolated from the control network to the network protection center. In the case of control sequences sent to SCADA (Fig. 6e) devices by malware, SCADA probe will report anomalies, which will force appropriate action by the station personnel. The result of malware impact will be similar to that obtained in scenario (Fig. 6f). If the malware starts to cooperate with SCADA decoy, an alert will be reported in SCADA DarkNet, similar to that presented in Fig. 6. The changes in the control process will be detected by SCADA probes present in the security system, and the effect will be identical to that provided in the scenario.

7. Summary

The achieved readiness state of the security system allows for its installation in the real power station environment. The paper presents developed the supervision system of the CSS set through the use of domain cooperation mechanisms within one entity and inter-domain cooperation mechanisms of different entities with the use of security policies. Due to the limitations in the volume of the article, they could not be included in the presented content. In addition to the issues covered in this paper, authors have been working on automation of security mechanisms implementation, and they plan to not only passively, but also actively influence certain, selected processes carried out in the control system. In the near future, authors intend to develop the system toward adaption to other critical infrastructure environments, such as the gas industry or smart-grid.
It may be anticipated that the challenges in the smart-grid environment associated with the security level will be even greater.

Acknowledgements

This work is sponsored by the National Centre for Research and Development as a part of a research project for national security and defense of Poland – "System of secure IP communication assurance in power control network", no. ROB 0074 03 001. Project is realized by Military University of Technology, Research and Academic Computer Network, Asseco Poland and Military Communications Institute.

References

[1] "Wymogi wobec JWCD na potrzeby wdrażania systemu LFC (Requirements toward power sources for implementation of the LFC system)", PSE Operator S.A., 2011 (in Polish).

[2] "Vulnerability Analysis of Energy Delivery Control Systems", Idaho National Laboratory, Idaho Falls, Idaho, USA, Sept. 2011.

[3] G. Giannopoulos, R. Filippini, and M. Schimmer, "Risk assessment methodologies for Critical Infrastructure Protection", JRC Technical Notes, European Commission, Joint Research Centre Institute for the Protection and Security of the Citizen, 2012.

[4] "National SCADA test bed", U.S. Department of Energy [Online]. Available: http://energy.gov/oe/technology-development/energy-delivery-systems-cybersecurity/national-scada-test-bed

[5] "About the Cybersecurity for Energy Delivery Systems Program", Office of Electricity Delivery & Energy Reliability [Online]. Available: http://energy.gov/oe/services/technology-development/energy-delivery-systems-cybersecurity

[6] D. Kuipers, "Idaho National Laboratory National SCADA Test Bed", Idaho Falls, IO, USA, Oct. 2010 [Online]. Available: http://www.inl.gov

[7] "Common Cyber Security Vulnerabilities Observed in Control System Assessments by the INL NSTB Program", U.S. Department of Energy Office of Electricity Delivery and Energy Reliability, Idaho National Laboratory, Nov. 2008.

[8] J. Jarmakiewicz, K. Maślanka, K. Parobczak, "Development of cyber security testbed for critical infrastructure", in *Int. Conf. Milit. Commun. and Inform. Syst. ICMCIS 2015*, Cracow, Poland, 2015.

[9] *Critical Infrastructure Protection*, E. Goetz and S. Shenoi, Eds., IFIP Advances in Information and Communication Technology, vol. 253. Springer, 2008.

[10] SCADA LAB Project Homepage, https://www.scadalab.eu, Sept. 2012–2014 by INTECO, The Innovative Business Association for Network Security and Information Systems (AEI), Everis Consultancy Ltd, The National Centre for Critical Infrastructure Protection (CNPIC), EFB, Telvent Energía S.A., C Global Services (CGS), Zanasi and Partners (Z&P) and Nisz.

[11] M. Rybnicek, R. Poisel, M. Ruzicka, and S. Tjoa, "A generic approach to critical infrastructures modeling and simulation", in *Proc. Int. Conf. Cyber Secur. CYBERSECURITY 2012*, Washington, DC, USA, 2012, pp. 144–151.

[12] S. Walsh, S. Cherry, and L. Roybal, "Critical Infrastructure Modeling An Approach to Characterizing Interdependencies of Complex Networks", in *Proc. 2nd Int. Conf. Human Syst. Interact. HSI'09*, Catania, Italy, 2009.

[13] 1516-2010 – IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA) – Framework and Rules, IEEE Standard Association, Aug. 2010.

[14] "Informacje o pracy KSE (Information on KSE production resources)" (as of 01.12.2014), PSE Operator S.A. (in Polish).

[15] "Standard architektury sieci IP na stacjach elektroenergetycznych PSE S.A (IP Network Standard Architecture for Energetic Station in PSE)", standard tech. specif. PSE PSE-SF.LAN_IP_SE/2014v1, Konstancin-Jeziorna, Poland, 2014 (in Polish).

[16] Configuring Virtual Routing and Forwarding [Online]. Available: http://www.cisco.com

[17] Communication controller Ex-mBEL_COM [Online]. Available: http://www.elkomtech.com.pl/produkty/p/ex-mbel-com-koncentrator-danych/ko/1/3.html (retrieved 12.01.2014).

[18] Communication controller Ex-mBEL [Online]. Available: http://www.elkomtech.com.pl/produkty/g/ex-mbel-1/ko/1/2.html

[19] LabVIEW System Design Software [Online]. Available: http://www.ni.com/labview/ (retrieved 12.01.2014).

[20] Wireshark tool for IEC61850 (8-1, 9-2, 90-5, GOOSE), IEC60870-6 TASE.2(ICCP), UEEE C37.118 and MMS [Online]. Available: http://www.sisconet.com/downloads/Wireshark-win32-1.11.3-SkunkWorksIEC61850.exe

Jacek Jarmakiewicz received his M.Sc. and Ph.D. degrees in Telecommunications and Computer Networks from Military University of Technology (MUT), Warsaw, Poland in 1994 and 2004, respectively. He finished post graduated studies in Telecommunication Management Networks in 2000. He works on MUT as an assistant professor since 2004. At the same time he worked in Military Communication Institute (2009–2014), where he was a member of technical team in Cryptology Dept., and afterwards in C4I Systems Dept. He has held position as a senior researcher, project manager, and head of research groups in national and international, especially NATO RTO IST Panel Support. He was involved in projects in the field of IPv6 tactical networks, resources management in military mobile networks, cryptography, multilevel security, security of power control systems, and Internet of Nano-Things in telemedicine applications. He is author and co-author of 4 books, several chapters in monographs, more than 30 technical papers. He was recipient of Best Paper Award at International Academy, Research, and Industry Association.
E-mail: jjarmakiewicz@wat.edu.pl
Faculty of Electronics
Military University of Technology
Gen. Sylwester Kaliski st 2
00-908 Warsaw, Poland

Krzysztof Maślanka received his M.S. degree in Telecommunication Engineering in 1999 from the Faculty of Electronics, Military University of Technology (MUT), Warsaw, Poland. He is currently working as Assistant Lecturer in Telecommunications Institute, Faculty of Electronics, MUT. He engages in problems of communications and information systems (CIS) modeling and simulation, IP networks problems, telecommunication systems engineering, systems design and implementation for power grid, innovative routing algorithms, and Internet of Nano-Things in telemedicine applications.
E-mail: kmaslanka@wat.edu.pl
Faculty of Electronics
Military University of Technology
Gen. Sylwester Kaliski st 2
00-908 Warsaw, Poland

Krzysztof Parobczak graduated from Military University of Technology in 2009 after achieving M.Sc. degree in area of Electronics Engineering, Teleinformatics speciality. Currently works as Assistant Lecturer at MUT's Institute of Telecommunications, Electronics Faculty. He has participated in ICT security related projects. He is a secretary at AFCEA Polish Chapter, co-author of several publications in field of mobile network security, covert communication channels, power control systems security, and Internet of Nano-Things in telemedicine applications.
E-mail: kparobczak@wat.edu.pl
Faculty of Electronics
Military University of Technology
Gen. Sylwester Kaliski st 2
00-908 Warsaw, Poland

Paper

Detecting Security Violations Based on Multilayered Event Log Processing

Przemysław Malec, Anna Piwowar, Adam Kozakiewicz, and Krzysztof Lasota

Research and Academic Computer Network (NASK), Warsaw, Poland

Abstract—The article proposes a log analysis approach to detection of security violations, based on a four layer design. First layer, named the event source layer, describes sources of information that can be used for misuse investigation. Transport layer represents the method of collecting event data, preserving it in the form of logs and passing it to another layer, called the analysis layer. This third layer is responsible for analyzing the logs' content, picking relevant information and generating security alerts. Last layer, called normalization layer, is custom software which normalizes and correlates produced alerts to raise notice on more complex attacks. Logs from remote hosts are collected by using rsyslog software and OSSEC HIDS with custom decoders and rules is used on a central log server for log analysis. A novel method of handling OSSEC HIDS alerts by their normalization and correlation is proposed. The output can be optionally suppressed to protect the system against alarm flood and reduce the count of messages transmitted in the network.

Keywords—HIDS, log analysis, NIDS, syslog.

1. Introduction

Events occurring in the operating system, like software installation, managing system services, as well as successful and failed login attempts, are preserved real-time in the form of logs. Every log stores data regarding its origin, priority and time of appearance, which allows use of event logs as a reliable source of information when building systems for alerting about security violations. Raising alarm after detecting every single malfunction would lead to frequent false positives. That is why receiving and correlating data from multiple event log sources would increase accuracy of detection and allow to reveal violations with more complex indicators.

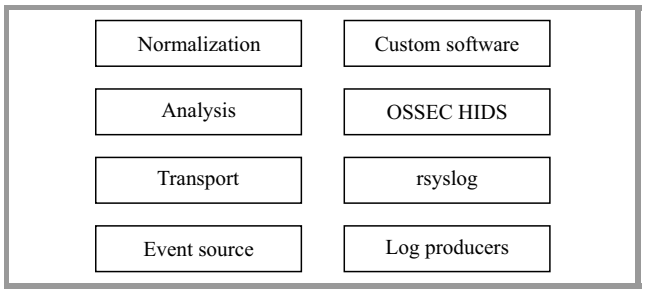


Fig. 1. Multilayered event log processing.

The design of the log-based system for detecting security violations consists of four layers, presented in Fig. 1.

The bottom layer, named event source, specifies the log sources in Linux and network environment relevant in the process of detecting events. The essential information may originate from typical system services and network devices, as well as from security dedicated services like audit and integrity check tools.

Transport layer is responsible for collecting log messages from various sources and passing them to the log-collecting server, where the analysis is done. The transport must guarantee confidentiality and transmitted data integrity, achieved by using *rsyslog* software [1]. This layer secures a copy of all incoming logs (crucial in security log management according to the guidelines [2]), which enables discovery of data tampering attempts. Preserving logs in remote localization enables incident reconstruction even after unrecoverable machine failure [3]. Moreover, preserving three timestamps for every log (generation, server reception, database insert) can indicate server downtime or communication disruption, which can be relevant for further investigation.

The role of the analysis layer is to generate alerts based on incoming log entries by decoding key information and filtering events that are relevant, while detecting malicious behavior in the network. The first analysis is executed by *OSSEC HIDS* [4] engine with a set of custom-written decoders and rules.

The need to decrease the number of repetitive alerts was widely discussed in [5], [6] and some of the existing strategies of alarm suppression were presented in [7]. The normalization layer, implemented by dedicated software, examines the output of the analysis layer and suppresses excess alerts. The need to combine many sources of information expressed in [8] is satisfied by performing nonlinear correlation to detect more complex attacks. As a result, this layer creates alarms that are normalized to a protocol, and can be used by security system consumer.

2. Event Source Layer

2.1. Log Sources

Log messages containing knowledge about events taking place in the operating system or the network can be ob-

tained from various process sources. There are several services and modules that gather information about security events and store it in the form of logs, shown in Fig. 2. This article focuses on events generated by following sources:

- *auditd*,
- Advanced Intrusion Detection Environment (AIDE),
- *sshd*,
- *OSSEC rootcheck*,
- *racoon*,
- *iptables*,
- network devices,
- event logs coming from dedicated processes related with the system's security.

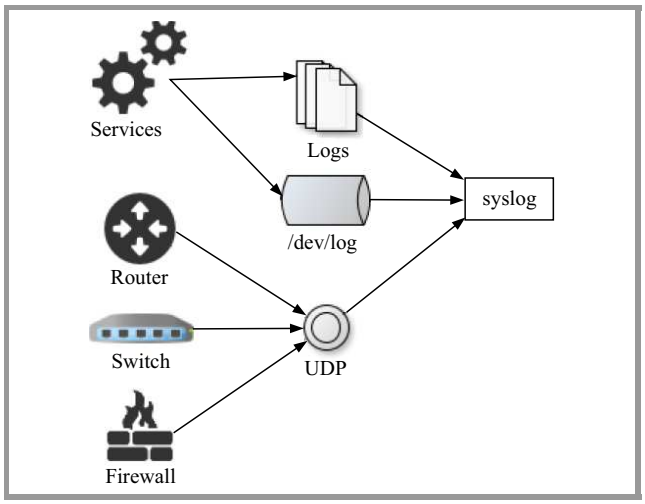


Fig. 2. Event sources.

auditd [9] is based on pre-configured rules and generates log entries recording a large variety of information about the events occurring in the system. According to [10], auditd can discover policy violations by monitoring file activities and collecting system calls.
AIDE [11] is a file and directory integrity checker. It creates a database from the state of the system, register hashes, modification times, etc. This database is later used to verify the integrity of files by comparing it against the real status of the system. All of the usual file attributes can also be checked for inconsistencies. AIDE generates syscall ANOM_RBAC_INTEGRITY_FAIL, when change in the monitored file structure is detected and received by auditd from kernel.
sshd [12] is an OpenSSH Daemon, which provides secure encrypted communications between two untrusted hosts over an insecure network. It can refer to Pluggable Authentication Modules (PAMs) [13], which provide a common authentication framework for applications and services in a Linux system. The errors in communication can be a valuable source for log analysis.
OSSEC's rootcheck is an OSSEC HIDS module for rootkit detection, which runs at regular intervals querying the system for information and comparing the results with a list of known rootkits and trojans. When the rootcheck module finds discrepancies in information about a file, a process, port or network interface, it will raise an alert about a suspected rootkit.
racoon [14] is an Internet Key Exchange (IKE) daemon for automatically keying IPsec connections to establish safe associations between hosts. Reported errors in communication can be used to detect suspicious behavior of nodes.
iptables [15] is used as a firewall to set up, maintain, and inspect the tables of IP packet filter rules in the Linux kernel.
Network devices (i.e., routers, switches, firewalls) send information about their activity in UDP messages.

2.2. Log Formats

Log records are an essential source of information and can be written by a process to a dedicated text file, although the majority of logs are a product of sending data to local /dev/log socket, from where the messages are received by a log collecting service called syslog. All information collected by syslog should by written in a simple syslog format described in RFC 5424 [16] with a such structure:

```
<PRI> TIMESTAMP HOSTNAME APP-NAME: MSG
```

The value of the PRI part is assigned based on two properties, which are severity and facility. Severity is a numerical value of event priority and varies from 0 (emergency) to 7 (debug). Facility value depends on the log supplier, where different kinds of system services have taken their own codes, e.g., kernel messages (kern) have 0 facility value, authentication events (auth) have facility value equal to 4, etc. In addition, every log contains the timestamp of the logged event, hostname of the machine which produced the message, program source of the event (optionally with PID in brackets) and message content of registered log.

An example of a syslog protocol log is shown below:

```
Jul 7 14:37:10 pl.bipse.wil.si.kbk1
sshd[15308]:  Accepted password for root from
192.168.56.1 port 56440 ssh2
```

Audit has an independent log processing system, which writes events to /var/log/audit/audit.log file and also provides a rotation mechanism for its journals. However, with the usage of audisp daemon the audit output can be redirected to the standard /dev/log socket for further processing. The audit log contains information about the type of registered event (system call, login information, etc.), its timestamp and a list of audit event fields suitable for given occurrence.

The example of audit log is as follows:

```
type=SYSCALL msg=audit(1441374914.091:1975):
arch=c000003e syscall=82 success=yes
exit=0 a0=1668900 a1=4a4fa8 a2=1647c00
a3=6d617473656d6974 items=5 ppid=1 pid=1840
auid=4294967295 uid=0 gid=0 euid=0 suid=0
fsuid=0 egid=0 sgid=0 fsgid=0 tty=(none)
ses=4294967295 comm="NetworkManager"
exe="/usr/sbin/NetworkManager" key="LINKING"
```

3. Transport Layer

3.1. Open Source Systems for Log Processing

A mechanism for centralized logging can be set up by configuring existing open source solutions for log processing. The leading software utilities are *syslog-ng* Open Source Edition, developed by BalaBit IT Security Ltd.[17] and *rsyslog* from Rainer Gerhards and Adiscon.

The *syslog* implementation used in a security system should meet the requirements described in [18], such as high availability of service and confidentiality and integrity of transmitted data. Both of the chosen solutions fulfill these assumptions and offer the ability to send and collect remote log messages by encrypted transmission using the Transport Layer Security (TLS) mechanism. Another key feature that both *rsyslog* and syslog-ng provide is database support. In addition, syslog-ng represents a highly customizable solution that can gather information from many different sources beside the standard operating system events, e.g. additional text files and the content of binary files such as /var/log/btmp, which contains records of failed user login attempts). The syslog-ng's database support is flexible and allows storing logs in daily tables without additional administrational procedures. The main disadvantage of syslog-ng is its fix-sized queue mechanism, which does not guarantee avoiding log loss during server unavailability in case of client's queue overflow.

On the other hand, rsyslog is capable of collecting standard system messages and events written to external text files and transferring them via Reliable Event Logging Protocol (RELP) to the central log server, ensuring zero message loss among the network nodes. rsyslog also supports disk buffers for not forwarded log messages, which protects log journals from loss and inconsistency.

For the purposes of the designed system rsyslog was chosen to create the centralized log processing system due to the reliability of its built-in mechanisms for log transfer between the hosts.

3.2. Proposed Architecture of the Transport Layer

The transport layer is based on client-server architecture. On every client machine data is collected from typical sources, such as the /dev/log socket (including the output of audispd redirecting audit events), system services like AIDE or sshd. Moreover, clients can gather information from network devices by receiving UDP messages and passing the relevant data further. Collected information is sent to the log server via the RELP communication, which guarantees no message loss. RELP mechanism uses a Transmission Control Protocol (TCP) connection, which is an advantage over plain syslog communication, which uses User Datagram Protocol (UDP) as a default.

The server machine receives messages from remote clients and processes them next to the logs from the local sources described above. Remote logs are written to separate files in catalogues named after client's hostnames, reflecting the file structure on origin machine. The copy of all messages is sent to the database, from where logs can easily be accessed with SQL queries. The database output stores all information included in the log, server receive timestamp and database insert time as well. The rsyslog, based on properties from syslog header, filters messages and forwards them to the named pipe /dev/ossec, from where they are received by OSSEC HIDS analyzer. The flow of information in the system is displayed in Fig. 3.

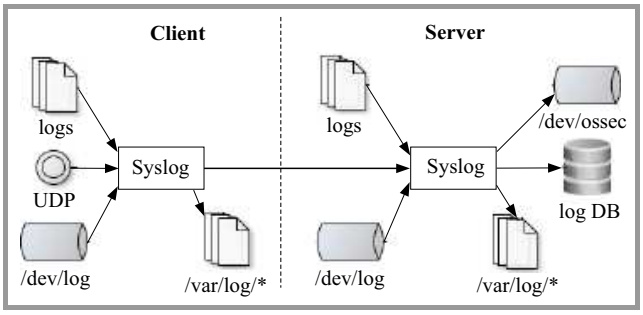


Fig. 3. Event log flow.

4. Analysis Layer

OSSEC HIDS is a platform to monitor the status of network elements. It offers the functionality of Host Intrusion Detection System (HIDS), Security Information and Event Management (SIEM), log monitoring, rootkit detection and checking the integrity of system files. It is frequently used as a part of more complex security systems, due to its flexibility and facility in adapting to own needs [19], [20]. In the proposed architecture OSSEC HIDS with custom rules and decoders is used to build a real-time log analyst monitor. This approach enables correlation of events from every node in the network. The stages of the analyst process are shown in Fig. 4.

Process steps are as follows:

- Pre-decoding – as mentioned, *rsyslog* is used as a transport channel for logs, which adds a syslog header. This step decodes the syslog format and pulls information about hostname and time from log;
- Decoding – custom decoders extract information based on program name, relevant for event type. It detects data like login name, address and ports for source and destination;

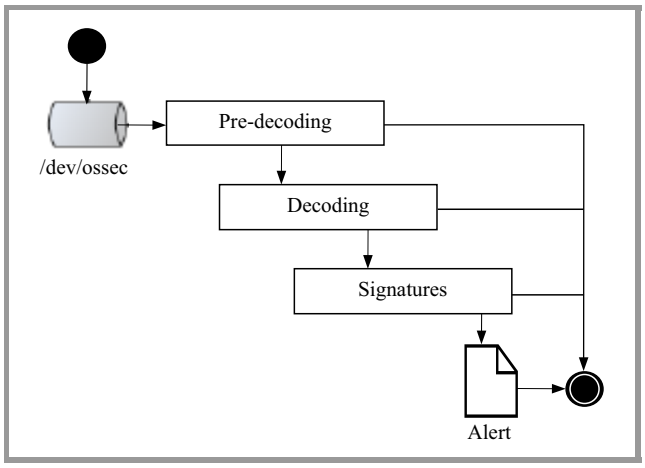


Fig. 4. OSSEC HIDS flow.

- Signatures and alerting – a set of custom rules is matched to the log event. Composite rules correlate hostnames, IPs, users etc. Every rule has an assigned priority level. When a single event matches more than one rule, the generated alarm has the priority of the rule with the highest level or with the level of the first matched rule when levels have the same value. If the incident violates adopted policies, an alarm is raised. Generated alarms precisely define the detected security events.

4.1. Pre-decoding

OSSEC HIDS recognizes different log formats. In the proposed system, the most common syslog format is used. Based on the syslog format, specified fields are decoded, such as:

- hostname – DNS/IP address of component which originated the event,
- time – time of the event from the element,
- message – message which is used in analysis,
- program – information which process generated the event.

Table 1 shows the result of the pre-decoding of a sample log.

Table 1
The result of the pre-decoding of a sample log

Time	Jul 17 08:34:40
Hostname	pl.bipse.nask.element2
Program	sshd[19721]:
Message	Accepted keyboard-interactive/pam for root from 192.168.56.1 port 51499 ssh2

4.2. Decoding

Based on the results of pre-decoding, the field named program determines the process from which the log comes. This approach minimizes the number of decoding attempts from the various decoders to only those that meet the criteria. The matching decoder will later be used to decode relevant information, for example the source IP address of the user. The decoding process consists of three stages:

- decoder selection,
- log content matching using regular expressions,
- decoding declared fields.

Example of decoding an *sshd* log is shown in Fig. 5. Based on process name sshd, the sshd decoder is used in the first stage. This decoder has a set of sub-decoders, that are used in regular expressions matching to determine which of them can decode specific field. It is possible to use several decoders simultaneously. Figure shows that the sshd-success decoder has been selected. The last stage presents that fields root and IP address were chosen and decoded as USER and SRCIP (names of fields used in the inner OSSEC HIDS logic).

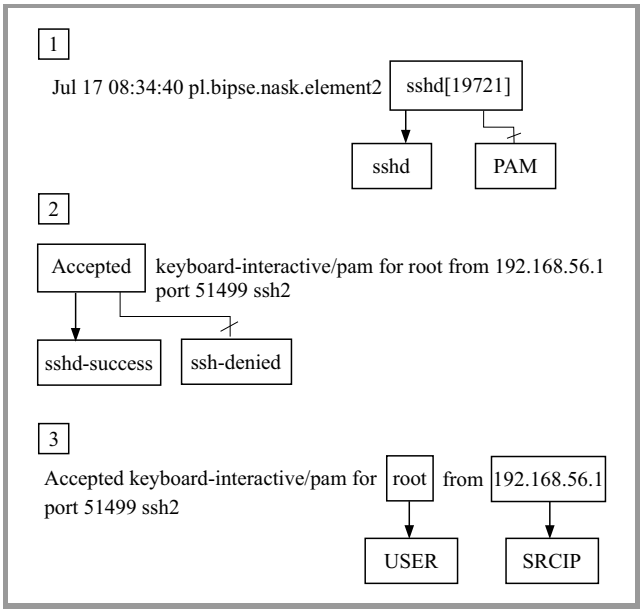


Fig. 5. OSSEC HIDS decoding process.

4.3. Event Correlation and Signatures

OSSEC HIDS distinguishes two categories of rules that generate alarms:

- atomic – based on a single event which occurred in the system,
- composite – correlated over time, based on patterns of other logs.

Figure 6 shows the logic of matching a rule to a decoded event. Log entries are analyzed in a sub-rule only after matching its parent rule. This approach accelerates the process of analysis and minimizes the tested rules amount.

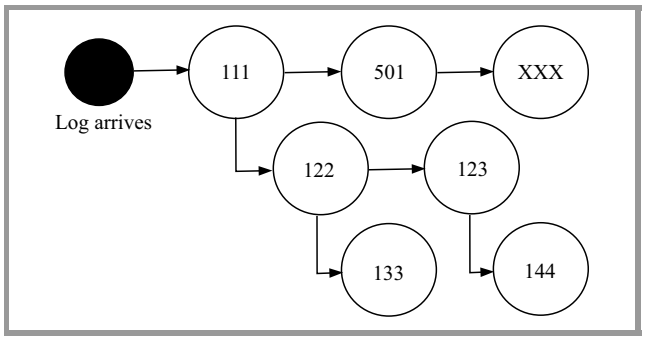


Fig. 6. OSSEC HIDS rules match.

Pessimistic number of rules to process is the sum of all parent type rules.
In the proposed solution the grouping of rules is a base for event correlation. For example, events coming from two different sources are decoded by two separate decoders and getting in two independent branches of rules. However, those rules have the same group id. While appearing individually none of the detected events generate alarms, but their correlation leads to signs of an incident being detected and raises an alarm. The grouping logic based on decoders sshd and login is shown in Fig. 7.

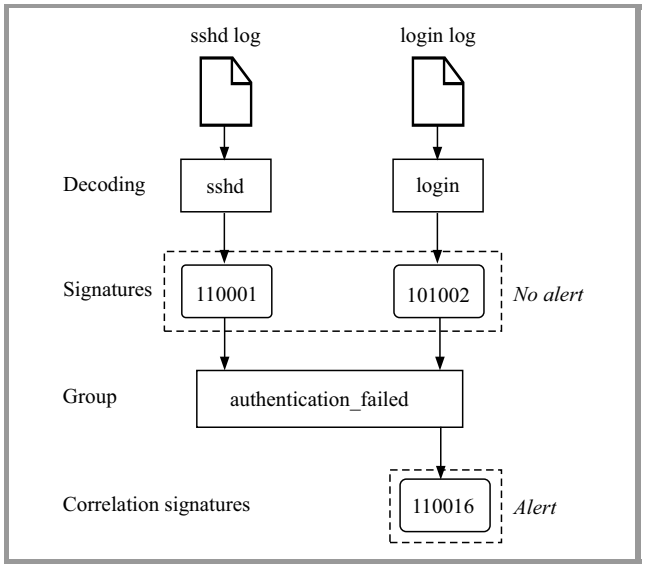


Fig. 7. OSSEC HIDS correlation and grouping logic.

4.4. Proposed Abuse Detection

Custom rules and decoders allow detection of events, which can be categorized into groups of security violation alarms like:

- file modification – files that are watched by auditd or AIDE, mostly configuration files of system or security processes,
- authentication abuse – local or remote, based on login and sshd proccess,
- rejected connection attempts – system firewall reports about rejected connection attempts, this information can be grouped and detected as a host scan,
- successful system user login – for example user apache or mysql,
- malfunctions services – for example, possible error of racoon negotiation between elements or invalidity of certificates,
- file system and hardware errors – can cause unnoticed relevant events in system, like overflow of hard drives,
- sudo abuse – system user and group modification,
- occurrence of unknown errors – events which should draw attention of the operators, those can be symptoms of attacks or reconnaissance,
- network equipment errors and events – errors and changes in the network topology are crucial information about inappropriate network activity for further investigation.

5. Normalization Layer

Previous layers process relevant event logs that can possibly generate alarms. This layer adds extra functionality, which makes events more useful and foolproof. This is performed by custom-written software. Main objectives of this layer are:

- alarm suppression,
- normalization for other security systems,
- guaranteed delivery,
- nonlinear correlation,
- rejection of irrelevant alarms.

5.1. Alarm Suppression

In the proposed solution OSSEC HIDS can aggregate and correlate events. Such aggregation can easily generate too many alerts if events incoming in a short time are a huge number of identical logs. As OSSEC HIDS suppresses identical logs, this layer suppresses OSSEC HIDS identical alarms. Each subsequent alarm which is propagated, contains the number of events summed in the suppression, so that the number of individual events is not lost. Suppression affects only the repeating alarms. The suppression time is configurable, depending on the type of event to which it refers to. This process is shown in Fig. 8. The proposed approach minimizes the amount of identical alarms sent and processed by central alarm analyzer, which results in improvement of performance without losing any information.

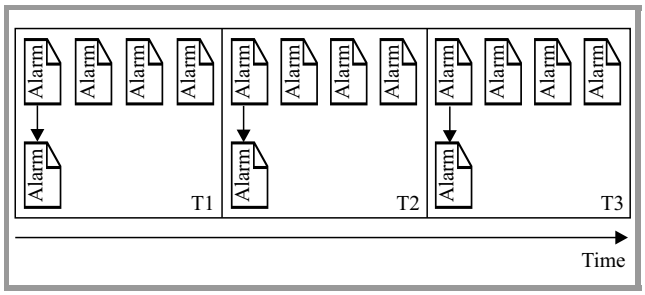


Fig. 8. OSSEC HIDS alarm suppression.

5.2. Normalization for Security System Protocol

Security systems often use their own protocols to communicate between components. Additionally, redundant information generates unnecessary network traffic, which can lead to internal Denial of Service (DoS) of security systems. Triggered alarms must be standardized to meet the needs accepted as a norm for reporting security incidents. Adopted policy may require additional data. At this point, the normalization layer enriches the message with additional data based on the configuration or correlating this event with information from knowledge bases. This is shown in Fig. 9.

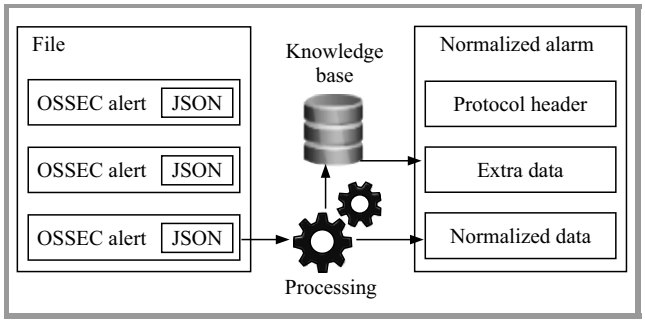


Fig. 9. Normalization for other security systems.

5.3. Guaranteed Delivery

Ability to communicate with another component that is a consumer of alerts requires normalization of forwarded information. The second step is to ensure that the detected event has been received by the central processing system. Each of the outgoing events has an identifier of an incident and its time validity. The mechanism checks whether the consumer has sent a reception acknowledgment or the validity of the message has expired. Based on this feedback information, it is possible to send a message again and guarantee delivery.

5.4. Nonlinear Correlation

The normalization layer adds a second correlation that can combine facts from various areas of events. This layer, by focusing on brokering, is able to expand the analysis based on the rules of event occurrence. This can provide historical events analysis and detection of patterns in alarms based on concept of states. Combining the facts of undesirable incidents occurring in various parts of the network and applying security modeling, could result in producing security rules that may protect other nodes, which are not endangered yet.

5.5. Rejection of Irrelevant Alarms

The normalization layer adds the ability for the complex security system to work in various modes, which may require additional alarm suppression. To reject alarms that are irrelevant, when system's state is taken into consideration, following modes can be distinguished:

- learning mode – detected incidents are a base for rule creation, which will preserve similar events from propagating in normal mode,
- reconfiguration mode – alarms are temporarily suppressed when the system is under configuration and modifications could result in false-positives,
- normal mode – every event classified as an alarm is propagated further, unless there was a corresponding rule created in the learning mode.

6. Summary

The article presented a multilayered approach to managing and handling security incidents based on event logs. Each layer presented key aspects of its functioning with examples of implementation. Use of ready-made solutions allows to focus attention on upper layers associated with event processing logic. The division of responsibilities into layers allows easier modification and implementation as part of security systems.

Acknowledgements

This work was supported by the National Centre for Research and Development (NCBiR) as a part of the research project "The system of secure IP communication provision for the power system management" (no. ROB 0074 03 001).

References

[1] Rsyslog – The rocket fast system for log prosecing [Online]. Available: http://www.rsyslog.com

[2] K. Kent and M. Souppaya, "Guide to Computer Security Log Management", National Institute of Standards and Technology (NIST) Special Publication 800-92, 2006.

[3] K. Julisch and M. Dacier, "Mining intrusion detection alarms for actionable knowledge", in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, Edmonton, Alberta, Canada, 2002, pp. 366–375.

[4] OSSEC documentation [Online]. Available: http://ossec-docs.readthedocs.org

[5] H. W. Njogu and L. J. Wei, "Using alert cluster to reduce IDS alerts", in *Proc. 3rd IEEE Int. Conf. Comp. Sci. Inform. Technol. ICCSIT 2011*, Chengdu, China, 2011, pp. 467–471.

[6] H. T. Elshoush and I. M. Osman, "Alert correlation in collaborative intelligent intrusion detection systems – A survey", *Appl. Soft Comput.*, vol. 11, pp. 4349–4365, 2011.

[7] T. H. Nguyen, J. Luo, and H. W. Njogu, "Improving the management of IDS alerts", *Int. J. Secur. Its Appl.*, vol. 8, no. 3, pp. 393–406, 2014.

[8] A. Oliner, A. Ganapathi, and W. Xu, "Advances and challenges in log analysis", *Commun. of the ACM (CACM)*, vol. 55, no. 2, pp. 55–61, 2012.

[9] auditd – security guide [Online]. Available: https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/6/html/Security_Guide/chap-system_auditing.html

[10] "Guide to the Secure Configuration of Red Hat Enterprise Linux 5", National Security Agency, Revision 4.2, pp. 87–94, 2011.

[11] AIDE – Advanced Intrusion Detection Environment [Online]. Available: http://aide.sourceforge.net/

[12] sshd deployment guide [Online]. Available: https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/6/html/Deployment_Guide/ch-OpenSSH.html

[13] Using Pluggable Authentication Modules (PAM) [Online]. Available: https://access.redhat.com/documentation/en-US/Red_Hat _Enterprise_Linux/6/html/Managing_Smart_Cards/Pluggable _Authentication_Modules.html

[14] racoon – Linux daemon [Online]. Available: http://linux.die.net/man/8/racoon

[15] iptables [Online]. Available: https://wiki.centos.org/HowTos/Network/IPTables

[16] R. Gerhards, The Syslog Protocol, RFC 5424 [Online]. Available: https://tools.ietf.org/html/rfc5424

[17] Syslog-ng – open source log management [Online]. Available: https://syslog-ng.org/

[18] K. E. Nawyn, "A Security Analysis of System Event Logging with Syslog", SANS Institute, no. As part of the Information Security Reading Room, 2003.

[19] L. Ying, Z. Yan, and O. Yang-jia, "The design and implementation of host-based intrusion detection system", in *Proc. 3rd Int. Symp. Intell. Inform. Technol. Secur. Informat.*, Jinggangshan, China, 2010, pp. 595–598.

[20] J. Timofte, "Intrusion Detection using Open Source Tools", *Revista Informatica Economică*, no. 2, vol. 46, pp. 75-79, 2008.

Przemysław Malec got his M.Sc. in Computer Science in 2015 at Polish-Japanese Academy of Information Technology. At present he is a research assistant at NASK. His main scientific interests concern network programming and information security.

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

Anna Piwowar got her M.Sc. in 2013 in Electrical and Computer Engineering from Warsaw University of Technology, Poland. At present she is a research assistant at NASK. Her main scientific interests include information security, especially intrusion detection and prevention systems.

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

Adam Kozakiewicz, **Krzysztof Lasota** – for biographies, see this issue, p. 14.

Paper

SHaPe: A Honeypot for Electric Power Substation

Kamil Kołtyś and Robert Gajewski

Research and Academic Computer Network (NASK), Warsaw, Poland

Abstract—Supervisory Control and Data Acquisition (SCADA) systems play a crucial role in national critical infrastructures, and any failure may result in severe damages. Initially SCADA networks were separated from other networks and used proprietary communications protocols that were well known only to the device manufacturers. At that time such isolation and obscurity ensured an acceptable security level. Nowadays, modern SCADA systems usually have direct or indirect Internet connection, use open protocols and commercial-off-the-shelf hardware and software. This trend is also noticeable in the power industry. Present substation automation systems (SASs) go beyond traditional SCADA and employ many solutions derived from Information and Communications Technology (ICT). As a result electric power substations have become more vulnerable for cybersecurity attacks and they need ICT security mechanisms adaptation. This paper shows the SCADA honeypot that allows detecting unauthorized or illicit traffic in SAS which communication architecture is defined according to the IEC 61850 standard.

Keywords—cybersecurity, IEC 61850, honeypots, SCADA.

1. Introduction

An electric power substation is a part of a critical infrastructure, an electrical grid, which delivers essential services that many industry sectors and millions of individual consumers depend on. Substations are used to distribute electrical energy to consumers, transform voltage to different levels, supervise and protect the distribution network. In the modern substations theses functions are performed with the support of a substation automation system (SAS).

SAS realizes common tasks of a Supervisory Control and Data Acquisition (SCADA) system and also provides additional features enhancing operation and maintenance efficiency, e.g. alarm processing or substation integration [1]. Figure 1 presents a typical architecture of SAS that is divided on three levels: station, bay and process. In the station level there is a so-called Human Machine Interface (HMI) used to control, supervise and monitor the substation, a workplace for engineering and configuration purposes and interfaces for the remote communication, e.g. with a control center. The bay level consists of control, protection and monitoring units of each bay and the process level provides devices that directly interface the primary substation equipment, i.e. smart sensors, actuators.

The devices in the bay and process levels are mostly intelligent electronic devices (IEDs). IED is a device that implements particular function in a substation and has a microprocessor and communication ports to be able to transmit data and execute control commands. The examples of IEDs are circuit breakers, voltage regulators, protection relays and Programmable Logic Controllers (PLC). Big substations may have more than 100 bay level IEDs and a similar amount of process level IEDs. Those IEDs are usually made by different vendors. To provide interoperability between them the IEC 61850 standard defining common communication architecture has been proposed.

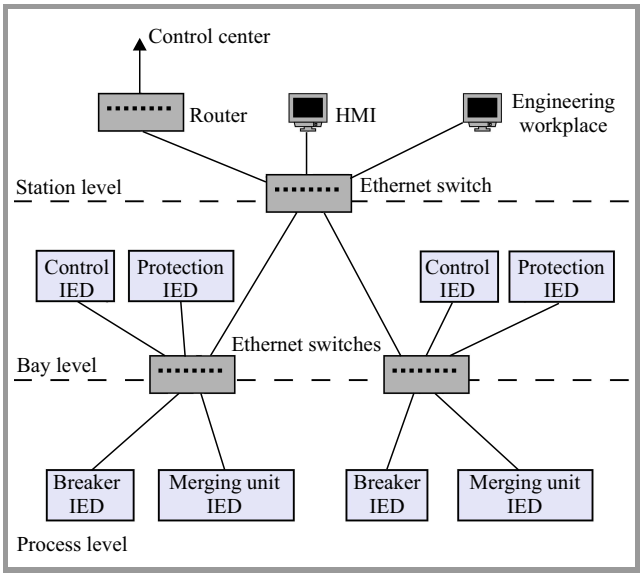


Fig. 1. A typical architecture of SAS.

In most countries the electrical grid contains typically very few substations and a failure of one of them may have severe consequences. Thus any single substation has to be carefully protected. In particular, a special attention should be devoted to the cybersecurity of SAS. The application of Ethernet and other Information and Communications Technology (ICT) solutions, as indicated by IEC 61850, makes SAS more exposed to cyberattacks.

A disclosure of cybersecurity incidents in SCADA systems confirms that they are not free of security issues. The prominent example of an attack on the SCADA system was an attempt to sabotage Iran's nuclear project by means of a computer worm known as Stuxnet. Stuxnet was released in 2009. Chen *et al.* in [2] show an overview of the Stuxnet's architecture. They point on the considerable effort needed to develop such a malware as well as on the fact that the attack would not be possible to succeed without insider knowledge and the support from a large team of

experts. The Stuxnet targeted attacks are able to penetrate into the isolated part of the SCADA system that were traditionally separated from the parts connected to the Internet. Stuxnet contains modules that attack PLCs in the target system and may cause physical damage to the equipment.
Fortunately, the awareness of cybersecurity threats is growing. According to the recent ICS-CERT report [3] in 2014 there were reported 232 incidents and 167 vulnerabilities concerning SCADA systems. It is widely recognized that the protection based on network isolation and an obscurity of proprietary communication protocols is no longer suitable for today's such systems. From the analysis of SCADA security standards presented in [4] it results that the most important cyber countermeasures are authentication and cryptography. On the other hand the most frequently mentioned threat in those standards is a malicious code. The leading SAS vendors try to address these security issues offering their products with additional security features adapted from ICT systems such as firewalls, antivirus software, advanced account management systems or intrusion detection and prevention systems supporting SCADA signatures.
A useful security mechanism that is becoming more popular in ICT systems are honeypots. As defined by Spitzner [5] a honeypot is an information system resource whose value lies in unauthorized or illicit use of that resource. The honeypot is a trap for the attackers. One of the honeypot's aim is to maintain the attacker's interest and thus observe the attack methods. This way previously unknown attack methods can be revealed and analyzed to improve the system security. Honeypots surely can help to better protect the SCADA systems. Their application is considered in the one of big research project concerning the cybersecurity of critical infrastructures [6].
This paper presents the honeypot named SHaPe that is able to emulate any IED conforming to the IEC 61850 standard. SHaPe can be used to detect unauthorized or illicit traffic in SAS, which communication architecture is defined according to IEC 61850.
The remainder of the paper is organized as follows. In Section 2 the related work concerning SCADA honeypots is discussed. In Section 3 the main principles of the IEC 61850 standard are presented. Section 4 describes the SHaPe honeypot. Finally, Section 5 summarizes the paper.

2. SCADA Honeypots

The literature mentions only few honeypots designed especially for SCADA systems. Each of them belongs to one of the two traditional honeypot classes [7]: low-interaction or high-interaction. A high-interaction honeypot usually uses a real resource and let an attacker to interact with it, e.g. log into the operating system. A low-interaction honeypot operates by emulating a resource or some part of it making an attacker convinced that he interacts with the real resource. On the one hand the high-interaction honeypot is able to induce and thus detect any type of attack against the particular resource while the efficiency of the low-interaction honeypot is limited by the accuracy of the emulation. On the other hand the low-interaction honeypot is usually easier to deploy and maintain and involves a lower risk of the honeypot to become compromised.

One of the first initiatives concerning SCADA honeypots is the SCADA Honeynet Project [8] that was started in 2004. It aims to create a SCADA honeypot based on the low-interaction honeypot Honeyd [9]. Honeyd simulates a number of network protocols such as HTTP, SMTP and FTP but it can be extended to simulate other network protocols using simple scripts. The developers of the SCADA Honeynet Project create a number of scripts emulating a PLC device with HTTP, FTP, Telnet and Modbus services. They also implement a Java applet that shows the status of a SCADA device. The project being at the proof of concept stage has not been developed since 2005.

Based on the SCADA Honeynet Project, Digital Bond [10] develops a low-interaction SCADA honeypot that emulates a popular PLC device with SNMP and all services provided by the SCADA Honeynet Project honeypot. Moreover, Digital Bond proposes a security mechanism called SCADA Honeywall. It uses IDS with special SCADA signatures to detect known attacks and is able to stop the outbound traffic from the compromised honeypots. The SCADA Honeywall can be placed in front of either a low-interaction honeypot like the one provided by Digital Bond or a high-interaction honeypot using e.g. a real PLC.

Two different honeypot systems that have been used to collect statistical data about the SCADA cyberattacks are described in [11]. One system is a high-interaction honeypot that utilizes an actual PLC device and a physical server. The PLC mimics a temperature controller in a factory and has temperature, fan speed and light settings that can be modified. The physical server that is connected with the PLC operates as a HMI and hypothetically modifies the PLC settings. The second system is a low-interaction honeypot realized on the Amazon EC2 cloud Web service. One Amazon EC2 instance is configured as a Web page emulating the interface of a water pressure station. The another Amazon EC2 instance connected with the first one simulates PLC with DNP3 and Modbus services.

Another low-interaction SCADA honeypot emulating PLC is presented in [6]. It implements three communication protocols: Modbus, FTP and SNMP. Moreover it has a special module for detecting probing activity at the remaining TCP ports. The honeypot also provides additional features such as filtering and aggregating the security events.

One of the latest SCADA honeypots is Conpot [12] on which work began in 2013. Conpot is a low-interaction honeypot that at the default configuration emulates Siemens SIMATIC S7-200 PLC. It provides an implementation of Modbus and SNMP. The response times of emulated services can be artificially delayed to mimic the behavior of a system under constant load. Conpot can be deployed with a custom HMI. It is an open source software that can

be easily extended to emulate more complex SCADA systems. The project is actively developed under the auspice of the Honeynet Project.
At the end, it should be noted that beside the aforementioned typical SCADA honeypots there are other more general honeypot solutions that may be employed to protect SCADA systems. For example, GhostUSB [13] is a low-interaction honeypot that emulates a USB storage device. Although it does not focus on the SCADA network protocols it can be used in the SCADA system to detect malware that propagates through USB devices, e.g. Stuxnet.
Concluding, the SCADA honeypots known in the literature allow monitoring traffic involved with a HMI and typical PLC devices. They focus on the traditional SCADA communication protocols such as Modbus, SNMP, FTP and HTTP. Taking into account that these protocols are not included in the IEC 61850 standard none of the existing SCADA honeypots is suitable for modern SASs compliant with this standard.

3. Overview of IEC 61850

IEC 61850 is an international standard that defines layered communication architecture for a SAS to provide an interoperability between IEDs from different vendors. The communication architecture is based on abstract information and service models.

3.1. Information Model

The IEC 61850 information model is presented in [14]. It is an object oriented model that specifies a set of basic data types and data objects with strict naming conventions. The data objects have a fixed hierarchical organization illustrated in Fig. 2.

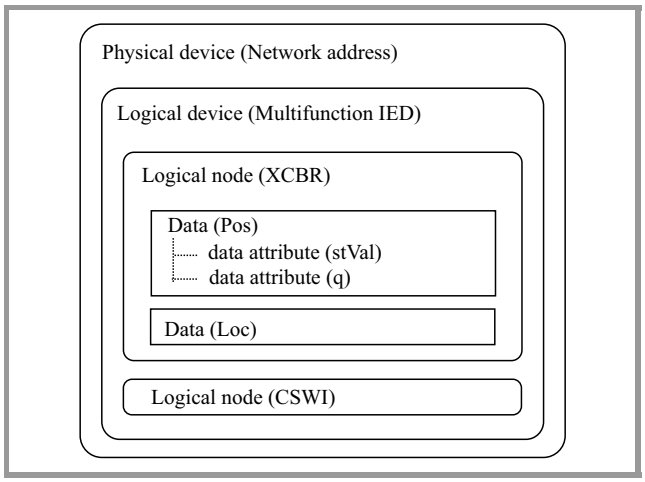


Fig. 2. The IEC 61850 information model.

At the top level of the hierarchy there is a physical device that represents an IED connected to a SAS network. The physical device is identified by its network address. It may contain one or more logical devices. The logical devices are used to form a group of some power system functions which are defined as logical nodes. Typically, the physical device has one logical device. However the possibility of having multiple logical devices allows a single physical device to act as a proxy or gateways for several IEDs.
The logical nodes contained in the logical device are the key objects in IEC 61850 representing the smallest entities of a SAS functionality used to exchange information between IEDs. A logical node is a named grouping of data objects that are logically related to the specific function. IEC 61850 defines more than 100 kinds of logical nodes covering the most common applications of SAS equipment. They are classified into 19 groups. The names of all logical nodes from the same group begin with the same character. For example the logical node XCBR that is used to model switches with short circuit breaking capability belongs to the Switchgear group which all logical nodes have names beginning with the letter X.
The semantic of the logical node is defined by its data and data attributes. Each data in the logical node has a unique name determining its purpose. IEC 61850 specifies about 500 data types with different semantic definitions. A data object may have multiple data attributes each one having name and attribute type. Data attribute names are standardized and carry specific semantic. For example the logical node XCBR has several data objects, e.g. Pos that describe a position of the circuit breaker, Loc indicating a switchover between local and remote operations or OptCnt representing an operations counter. In turn, the data object Pos has many data attributes, e.g. stVal representing a position of the real breaker that is an enumerated type taking one of the following values: intermediate-state, off, on or bad-state.

3.2. Service Model

The IEC 61850 service model is described in [15]. Like the information model it is also object oriented. The service model defines several classes with related services. The class GenServer represents the external behavior of a device. Each GenServer object contains one or more instances of GenLogicalDeviceClass class. GenLogicalDeviceClass together with three other classes GenLogicalNodeClass, GenDataObjectClass and GenDataAttributeClass represent the generic logical device, logical node, data and data attribute, appropriately. GenDataAttributeClass has an important property named functional constraint that indicates what services can be performed on the particular data attribute. For example value ST of the functional constraint means that the data attribute represents status information whose value may be read, but cannot be written.
In the service model there are also defined functional constraint data and functional constraint data attribute. The functional constraint data is an ordered collection of data attributes of the data object having the same functional constraint. The functional constraint data attribute is a data attribute having the specific functional constraint. An ordered set of elements being either a functional con-

straint data or a functional constraint data attribute is called a data set and is represented by the DATA-SET class. Data sets allow for more efficient information exchange between IEDs.
The services related with the aforementioned classes are the following:

- GenServerClass:
  - GetServerDirectory: retrieves a list of all logical devices;
- GenLogicalDeviceClass:
  - GetLogicalDeviceDirectory: retrieves a list of all logical nodes;
- GenLogicalNodeClass:
  - GetLogicalNodeDirectory: retrieves a list of all instances of a given object class,
  - GetAllDataValues: retrieves a list of all data attribute values (optionally having a given functional constraint) of all data objects;
- GenDataObjectClass:
  - GetDataValues: retrieves a list of all data attributes values,
  - SetDataValues: sets a value of a given functional constraint data or functional constraint data attribute,
  - GetDataDirectory: retrieves a list of all data attribute names,
  - GetDataDefinition: retrieve a list of all data attribute definitions (names, types and functional constraints);
- DATA-SET:
  - GetDataSetValues: retrieves a list of the values of all data attributes of the data set,
  - SetDataSetValues: sets values of all data attributes of the data set,
  - CreateDataSet: creates a data set with a given list of members being either a functional constraint data or a functional constraint data attribute,
  - DeleteDataSet: deletes a given data set,
  - GetDataSetDirectory: retrieves a list of all data set members.

IEC 61850 defines also other classes with different services that are described in detail in [15]. All services are divided on several categories. For example one category contains services supporting the device self-description. Another category is related with fast peer-to-peer exchange of status information between IEDs and yet another involves the control of an IED.

3.3. Mapping to MMS

The objects defined in the information and service models are independent of any protocol stack. However to enable real communication between IEDs these abstract objects need to be implemented in a form that can practically operate in a SAS network. IEC 61850 has established that the Manufacturing Message Specification (MMS) protocol over TCP/IP should be used for this purpose. Nonetheless another protocol can be chosen in the future to follow the evolution in ICT.
MMS is a public ISO 9506 standard that specifies the ways in which real time process data and supervisory control information is transferred between networked devices and computers. The key element of MMS is a Virtual Manufacturing Device (VMD) that models an MMS device (server) from the viewpoint of an MMS client. The VMD defines the objects (e.g. variables, domains) that are contained in the MMS server, the services (e.g. read or write a variable) a client can use to access or manipulate the objects and the behavior of the server upon receipt of those service requests. Tables 1 and 2 present the mapping of the IEC 61850 objects and services to MMS as defined in [16].

Table 1
The mapping of IEC 61850 objects to MMS objects

IEC 61850 object	MMS object
GenServerClass	VMD
GenLogicalDeviceClass	Domain
GenLogicalNodeClass	Named variable
GenDataObjectClass	Named variable
DATA-SET	Named variable list

Table 2
The mapping of IEC 61850 services to MMS services

IEC 61850 service	MMS service
GetServerDirectory	GetNameList
GetLogicalDevice-Directory	GetNameList
GetLogicalNodeDirectory	GetNameList
GetAllDataValues	Read
GetDataValues	Read
SetDataValues	Write
GetDataDirectory	GetVariableAccess-Attributes
GetDataDefinition	GetVariableAccess-Attributes
GetDataSetValues	Read
SetDataSetValues	Write
CreateDataSet	DefineNamedVariableList
DeleteDataSet	DeleteNamedVariableList
GetDataSetDirectory	GetNamedVariableList-Attributes

According to ISO 9506 a VMD and thus every instance of *GenServerClass* must also implement the following services:

- Initiate – establishes an application association, i.e. an agreement between the MMS client and the MMS server governing their communication,
- Conclude – terminates an existing application association in a graceful manner,
- Abort – terminates an existing application association in an ungraceful manner that may result in the loss of data,
- Reject – notifies about reception of an unsupported service request,
- Cancel – cancel an outstanding MMS service request,
- Identify – obtains information and status about the MMS server.

MMS services are grouped into two categories: the services requiring confirmation (so-called confirmed services) and the services that do not require such a confirmation (so-called unconfirmed services). The confirmed services contain an invocation identifier that identifies the service instance.

3.4. Configuration Description Language

Proper substation operation requires appropriate configuration of all its IEDs. IEC 61850 provides a Substation Configuration Language (SCL) that allows to describe the substation topology, the communication system, e.g. how IEDs are connected to networks and subnetworks, how data objects are grouped into data sets. SCL also allows to describe the particular IED capabilities in terms of logical nodes and the relation between substation structure and the SAS functions represented by the logical nodes. The SCL is based on XML. Its specification is given in [17].
A configuration process of substation involves many SCL files to be created. The structure and functions of the substation is defined in a System Specification Description (SSD) file that for example may contain the required types of logical nodes and data. The configuration of all IEDs with the communication section and the substation description is included in a System Configuration Description (SCD) file. Each IED to be compatible with IEC 61850 must have an IED Capability Description (ICD) file describing its functional and engineering capabilities or an Instantiated IED Description (IID) file including its project specific configuration. Moreover, it must be able to use the SCD file to set its communication configuration that is saved as a Configured IED Description (CID) file.

4. SHaPe

The IEC 61850 standard was published in 2004 and since then it has gained popularity among electric utilities and power system authorities in many countries. It can be taken for granted that more and more substations will be upgraded to conform the IEC 61850 standard. Thus a future SCADA honeypot to be useful in the electric power industry must support the communication protocols specified in IEC 61850.

4.1. General Concept

SHaPe is a low-interaction honeypot that is able to emulate any IED compliant with IEC 61850.
The low-interaction approach allows to achieve several important goals. Firstly, SHaPe can be easily configured to emulate different devices. What is needed is to provide an ICD or IID file with the requested IED configuration. The SCL file can be prepared using some IEC 61850 configuration tool or simply obtained from the existing IED if the similar one has to be emulated by SHaPe.
Secondly, SHaPe does not require any specialized equipment or much computing resources. A typical personal computer may run several instances of SHaPe. Each instance can listen on multiple IP addresses. Taking into account that a traffic coming into honeypots is rather low one machine should be enough to deploy a farm of SHaPe honeypots emulating many IEDs of different types.
Finally, SHaPe as a low-interaction honeypot involves lower risk of being compromised by an attacker than a high-interaction solution [18].

4.2. Detection Scope

SHaPe emulates IED behaviour that is involved with the MMS communication over TCP/IP connection. The generic substation events based on GOOSE or transmission of sampled values that according to IEC 61850 are mapped to other protocol stacks are not handled by the SHaPe honeypot. Nonetheless all IEC 61850 services mapped to MMS are supported and executed by SHaPe. If a service creates, modifies or deletes some object the state of the emulated IED will be accordingly updated.
SHaPe allows for detecting many important events that may appear during the communication with the emulated IED. These events are the following:

- an establishment of a TCP connection,
- a termination of a TCP connection,
- an establishment of an MMS application association (Initiate service),
- a graceful termination of an MMS application association (Conclude service),
- a receipt of a Reject service request,
- a receipt of an Identify service request,

- a receipt of a GetNameList service request,
- a receipt of a Read service request,
- a receipt of a Write service request,
- a receipt of a GetVariableAccessAttributes service request,
- a receipt of a DefineNamedVariableList service request,
- a receipt of a DeleteNamedVariableList service request,
- a receipt of a GetNamedVariableListAttributes service request,
- a receipt of an unknown MMS request before establishing an MMS application association,
- a receipt of an unrecognized MMS request after establishing an MMS application association.

Note that all IEC 61850 services are captured by SHaPe in terms of appropriate MMS services. Moreover, SHaPe detects the establishment and termination of every TCP connection and the receipt of MMS services related to VMD except Abort and Cancel services.

4.3. Monitoring Multiple Network Addresses

One instance of SHaPe can emulate an IED of a particular type. However SHaPe is able to run many copies of the emulated IED each one having assigned different IP address. In this way SHaPe allows for easily increasing the number of monitored IP addresses and thus the attack surface involved with the particular type of IED.
For each running copy of IED SHaPe maintains a separate state, i.e. the values of all data attributes in all logical nodes. The IED state can be modified by some IEC 61850 services, e.g. *SetDataValues* or *SetDataSetValues*. All MMS clients connected with the same IED copy see the modifications made by any of them. SHaPe keeps the modified state until there is no MMS client connection over a predefined period of time. After this idle time the IED state is restored to the initial one.

4.4. Implementation

SHaPe has been implemented as a module of Dionaea [19], which is a general purpose low-interaction honeypot running on the Linux platform. Dionaea has several modules that emulate different services prevalent in typical computer networks, e.g. HTTP, FTP, SMB. None of these modules can emulate a typical SCADA device. SHaPe is the first Dionaea module that is designed for SCADA networks.
Dionaea provides two useful mechanisms for their modules: a communication mechanism handling TCP connections and a logging mechanism that registers security events. Thanks to the communication mechanism SHaPe does not have to operate directly on the TCP sockets as it can handle appropriate events related to the transport communication layer, e.g. establishing a new TCP connection, terminating a TCP connection or receiving data. The events concerning the establishment or termination of a TCP connection are automatically forwarded to the Dionaea logging mechanism. For each establishment of a TCP connection the following information is registered: the TCP connection identifier, the timestamp, the source IP address, the source port, the destination IP address and the destination port. An event of the termination of a TCP connection contains the identifier of the TCP connection and the timestamp.
The Dionaea logging mechanism saves event information in a log file and optionally sends it to a specific XMPP server. SHaPe uses this logging mechanism to register events related to MMS protocol layer in the same way as the TCP layer events are registered. For each MMS protocol layer event the following information is provided: the identifier of TCP connection within the event has occurred, the event timestamp, the type and body of the MMS request in which the event has been detected. The type of MMS request indicates one of the unconfirmed services (Initiate, Conclude or Reject) or that a requested service is confirmed. In the latter case SHaPe registers additional information – the subtype corresponding to the particular confirmed MMS service and the invocation identifier of the service instance.
To handle MMS requests SHaPe utilizes library libiec61850 [20] that for integration purposes has been slightly modified.
Both Dionaea and libiec61850 are an open source software. Also SHaPe is publicly available under GNU GPL at the SHaPe project website [21].

5. Summary

In this paper the honeypot named SHaPe is proposed. SHaPe opposed to other SCADA honeypots supports the IEC 61850 standard. Thus it can be used to protect a modern SAS conforming to this standard.
SHaPe can be easily configured to emulate any IED by providing an appropriate SCL file. One SHaPe instance may listen on multiple IP addresses maintaining many copies of the particular IED. Several SHaPe honeypots allow to create a network of different IED decoys significantly increasing the chance of detecting an unauthorized or illicit traffic in SAS.
SHaPe has been implemented as a module of Dionaea which is a general purpose low-interaction honeypot. To handle MMS requests according to IEC 61850 SHaPe uses library libiec61850. SHaPe along with Dionaea and libiec61850 is an open source software publicly available under GNU GPL.

Acknowledgements

This work was supported in part by the National Centre for Research and Development (NCBiR) under the research project “The system of secure IP communication provision

for the power system management" (no. ROB 0074 03 001). The authors would like to thank Tomasz Pałka for his contribution in developing the SHaPe software.

References

[1] W. Rebizant, J. Szafran, and A. Wiszniewski, *Digital Signal Processing in Power System Protection and Control*. Springer, 2013.

[2] T. M. Chen and S. Abu-Nimeh, "Lessons from stuxnet", *IEEE Comp.*, vol. 44, no. 4, pp. 91–93, 2011.

[3] "ICS-CERT Year in Review 2014", Industrial Control Systems Cyber Emergency Response Team, 2014 [Online]. Available: https://ics-cert.us-cert.gov/Year-Review-2014

[4] T. Sommestad, G. N. Ericsson, and J. Nordlander, "SCADA System cyber security – A comparison of standards", in *Proc. IEEE Power Energy Soc. General Meet.*, Minneapolis, MN, USA, 2010.

[5] L. Spitzner, "Honeypots: catching the insider threat", in *Proc. 19th Ann. Comp. Secur. Appl. Conf. ACSAC 2003*, Washington, DC, USA, 2003, pp. 170–179.

[6] P. Simões, T. Cruz, J. Gomes, and E. Monteiro, "On the use of Honeypots for detecting cyber attacks on industrial control networks", in *Proc. 12th Eur. Conf. Inform. Warfare Secur. ECIW 2013*, Jyväskylä, Finland, 2013.

[7] L. Spitzner, *Honeypots: Tracking Hackers*. Boston, MA, USA: Addison-Wesley, 2002.

[8] V. Pothamsetty and M. Franz, "SCADA HoneyNet Project: Building Honeypots for Industrial Networks", 2005 [Online]. Available: http://scadahoneynet.sourceforge.net/

[9] The Honeyd website [Online]. Available: http://www.honeyd.org

[10] The SCADA Honeynet website [Online]. Available: http://http://www.digitalbond.com/tools/scada-honeynet

[11] K. Wilhoit, "Who's Really Attacking ICS Equipment?", Trend Micro Research, Cupertino, CA, USA, 2013.

[12] The Conpot website [Online]. Available: http://www.conpot.org

[13] The Ghost USB honeypot website [Online]. Available: http://code.google.com/p/ghost-usb-honeypot

[14] "Communication networks and systems for power utility automation – Part 7-1: Basic communication structure – Principles and models", IEC 61850-7-1, 2011.

[15] "Communication networks and systems for power utility automation – Part 7-2: Basic information and communication structure – Abstract communication service interface (ACSI)", IEC 61850-7-2, 2010.

[16] "Communication networks and systems for power utility automation – Part 8-1: Specific communication service mapping – Mappings to MMS (ISO 9506-1 and ISO 9506-2) and to ISO/IEC 8802-3", IEC 61850-8-1, 2011.

[17] "Communication networks and systems for power utility automation – Part 6: Configuration description language for communication in electrical substations related to IEDs", IEC 61850-6, 2010.

[18] K. Gorzelak, T. Grudziecki, P. Jacewicz, P. Jaroszewski, Ł. Juszczyk, and P. Kijewski, "Proactive Detection of Security Incidents", Tech. Rep., ENISA, 2012.

[19] The Dionaea website [Online]. Available: http://dionaea.carnivore.it

[20] The libiec61850 website [Online]. Available: http://libiec61850.com

[21] The ShaPe project website [Online]. Available: https://www.assembla.com/spaces/scada-honeypot

Kamil Kołtyś received the M.Sc. and Ph.D. degrees in Computer Science from Warsaw University of Technology (WUT) in 2007 and 2012, respectively. He was a research assistant at WUT from 2011 to 2012. From 2012 to 2015 he worked as a lecturer in Institute of Control and Computation Engineering of WUT. Since 2012 he has been an associate professor at Research and Academic Computer Network (NASK). His research interests include honeypots, data analysis, graph theory and operations research.

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

Robert Gajewski received the M.Sc. degree in Control Engineering Science from Warsaw University of Technology in 2002. He was designing and building IT tools to optimize operations in NUKAT union catalogue. Since 2012 he has been an research assistant at Research and Academic Computer Network (NASK). His present area of interests includes spam data analysis, data mining, and security mechanisms.

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

Paper

Uniqueness and Reproducibility of Traffic Signatures

Kazumasa Oida

Department of Computer Science and Engineering, Fukuoka Institute of Technology, Fukuoka, Japan

Abstract—Usable user authentication is an important research topic. The traffic signature-based approach is a new authentication technology that identifies the devices used by online users based on traffic signatures, where the traffic signature is a statistic of the video stream delivered by the authentication server to the user device. This approach has two advantages. First, users need not do any operations regarding the device identification. Second, users need not be sensitive to the privacy loss and computer theft. In this paper, an author evaluates the uniqueness and reproducibility of the signature by introducing a function that quantifies the distance between two signatures. Through number of experiments is demonstrated that the process interference approach has the advantage of generating new signatures that are sufficiently distinguishable from one another.

Keywords—user authentication, traffic signature, HTTP streaming, packet capture, variance plot.

1. Introduction

User authentication is mostly based on passwords. A password-hacking exercise, however, demonstrated that a large number of passwords can easily be cracked [1]. Accordingly, users who place high value on their accounts should adopt a more robust authentication strategy. Since the United States Federal Financial Institutions Examination Council officially recommended the use of multi-factor authentication in 2005 [2], various authentication technologies have been proposed, where multi-factor authentication requires a prover to provide more than one distinct factors to a verifier and there are three distinct authenticating factors: what you have (e.g. house keys), what you know (e.g. passwords), and what you are (e.g. fingerprints) [3], [4].

Current what you have authentication schemes add an additional hardware device to a desktop/laptop PC. Such a device is, for example, a security token, smartphone, or trusted platform module (TPM). Unfortunately, they are not widely used today because they are complex, lead to a loss of privacy, reduce control of the computer, or need to be protected against device theft [3]. Number of authors argue that users' capabilities and understanding should be factored into the design of security technologies [5], [6].

In [7] is proposed another what you have authentication scheme, which identifies the machines users are operating to access their accounts. Presented in [7] approach is based on video traffic analysis. The authentication server delivers a video stream and the user device records packet arrival times to calculate a traffic signature. The server then verifies whether the obtained signature agrees with the one obtained before or the one registered previously. Recently, some online banks request users to show their countersigns, when users try to sign-in using devices that are different from those they used to use. The difference can be detected using the client environment carried by the HTTP protocol, which includes an IP address, a browser type, etc. This scheme roughly distinguishes user computer platforms, whereas author's approach, shown in [7], precisely distinguishes them based on their unique signatures. As long as user machines are correctly identified, users do not need to be aware of anything about the machine identification since signatures are calculated and verified without intervention from users.

Contrarily, the current three major what-you-have authentication technologies, which deliver codes via the security token, email (or SMS), or an app running on a portable device (e.g. smartphone) [8], direct users to do some operations, such as starting the device/app and typing in the code. In addition, as opposed to previous authentication schemes, in which a single device/app generates codes based on some algorithm, the traffic signature is formed through the interactions among numbers of elements, which include not only hardware and software components of the user platform but also the server, video coding techniques, communication protocols that affect statistics of video traffic [9]. Since the interactions are not simple, it is difficult to infer the signature even if detail specifications of the user and server machines are given.

Meanwhile, a TPM chip into which unique RSA keys have been burnt can strictly identify the user machine [10]. This PKI-based approach strongly connects a user device to its owner, so that owners must pay careful attention to a privacy loss and unit theft. Some mechanisms that minimize the risk when stolen are a priori integrated (e.g. platform integrity). Unfortunately, these extra attention and mechanisms may cause the usability problems. Contrarily, in author's approach, users do not have to be sensitive to these risks since the signatures appear only in the authentication process and there is no personal identity-related information on the user device.

The PKI-based approach is also costly. As discussed in [5], authentication solutions must be accessible to all online users not just in terms of knowledge and effort but also in terms of cost. The authors in [5] quote that older users, who have much to gain from online participation, might be unable or unwilling to own a smartphone, which is the second

factor of choice in many authentication solutions currently deployed or planned. Presented approach demands video delivery. The HTTP-based streaming technologies are used because they are inexpensive and widely used today [11]. They avoid NAT and firewall traversal issues and provide cost effectiveness since there is not need dedicated streaming servers for video delivery.

The previous author's work [7] introduced the traffic signature and discussed its sensitivity to the user machine. This paper evaluates the signature through numbers of experiments and clarifies its uniqueness and reproducibility. Section 2 introduces the traffic signature and briefly outlines the result in [7]. Section 3 defines the distance between two signatures, based on which uniqueness and reproducibility are discussed in Section 4. Section 5 investigates which components of the user machine or interactions among them dominantly participate in forming the signature. The findings in this section are effective not only in enhancing reproducibility, but also in allowing users to have different signatures even if their machines consist of the same hardware components. Finally, Section 6 presents the conclusions.

2. Traffic Signature

First decay rate is defined, which is derived from the variance plot [12]. Decay rate $\beta(m)$ indicates how fast traffic variability declines at time scale m. It depends on various factors (e.g., computer hardware and software implementation, protocols, propagation delays, and bandwidth) and the dominant factors vary with m. The following describes how to calculate $\beta(m)$. Let X_k denote the number of arriving packets during the k-th time interval of length δ, where $\delta = 10^{-5}$ s. The m aggregated series $\{X_k^{(m)}\}$ are obtained by dividing time series $\{X_k\}$ into blocks of length m and averaging the series over each block as

$$X_\ell^{(m)} = \frac{1}{m} \sum_{i=\ell m-m+1}^{\ell m} X_i, \ \ell = 1,2,\ldots,\lfloor N/m \rfloor, \qquad (1)$$

where m is a positive integer, N is the size of series $\{X_k\}$, and $\lfloor x \rfloor$ is the largest integer that does not exceed x. The sample variance of $\{X_k^{(m)}\}$ is given by

$$V^{(m)} = \frac{1}{\lfloor N/m \rfloor - 1} \sum_{k=1}^{\lfloor N/m \rfloor} (X_k^{(m)} - \bar{X})^2, \qquad (2)$$

where $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$. Hereafter it is assumed that aggregation levels m_i, $i = 0,1,\ldots$, take real numbers. The decay rate at level m_i is defined as

$$\beta(m_i) = \log\left(V^{(\lfloor m_{i+1} \rfloor)} / V^{(\lfloor m_i \rfloor)}\right), \qquad (3)$$

where $m_0 = 1$ and $m_{i+1} > m_i$. Throughout the paper, $\log \frac{m_{i+1}}{m_i} = \log \frac{N}{50} \cdot \frac{1}{21}$ for all i, and $N = 6 \cdot 10^6$. The traffic signature is twenty decay rates $\beta_1, \beta_2, \ldots, \beta_{20}$, where β_i is used to indicate $\beta(m_i)$ for simplicity.

2.1. Experimental System

Unless otherwise mentioned, all results in this paper are obtained using the experimental system in Fig. 1. In the figure, the client PC (C-PC) accesses France 24 live (IP = 213.205.104.131), a news channel based in France, using the Internet Explorer Flash Player add-on. This online news is delivered at a constant rate (448 Kb/s) using the TCP protocol. In Fig. 1, all packets destined for C-PC are copied to the attacker's PC (A-PC) by the port-mirroring hub, so that not only C-PC but also A-PC collects video packets from the server with WinDump [13] to calculate signatures. Hereinafter, signatures calculated on C-PC (resp. A-PC) are referred to as user (resp. attacker) signatures. The news channel France 24 live was used because there are 30 routers between the news server and C-PC and the round-trip time is approximately 283 ms. Such a long-distance communication generates highly variable video traffic. Practical signatures should be stable in this case.

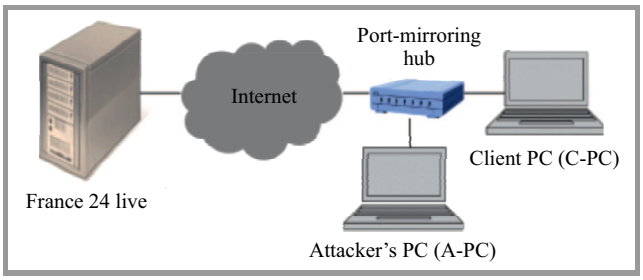


Fig. 1. The port-mirroring hub copies all packets destined for C-PC to A-PC.

2.2. Previous Results

In this subsection the results presented in [7] are briefly introduced. Four machines in Table 1 are used as C-PC in Fig. 1. Although they are all Windows machines, their software and hardware components are somewhat different. Figures 2 and 3 show their attacker and user signatures, respectively. Throughout the paper, ten samples are obtained for each signature $\{\beta_i\}_{1\le i\le 20}$ to see the stability of each decay rate β_i. From the figures, decay rates $\{\beta_i\}_{1\le i\le 16}$ are mostly stable, while $\{\beta_i\}_{17\le i\le 20}$ are not. This is mainly because the number of samples $X_k^{(\lfloor m_i \rfloor)}$ decreases as an increase in i. It can be seen that attacker signatures (Fig. 2) are all similar. In contrast, distinct differences exist between any two user signatures in Fig. 3.

Table 1
Four Windows machines *a-d* used in the experiment

C-PC	Model	Purchase	OS
a	XPS420	Jun. 2008	Vista, 32 bits
b	XPS435T	Jul. 2010	Windows 7, 64 bits
c	XPS9100	Jul. 2011	Windows 7, 64 bits
d	Inspiron	Jan. 2012	Windows 7, 64 bits

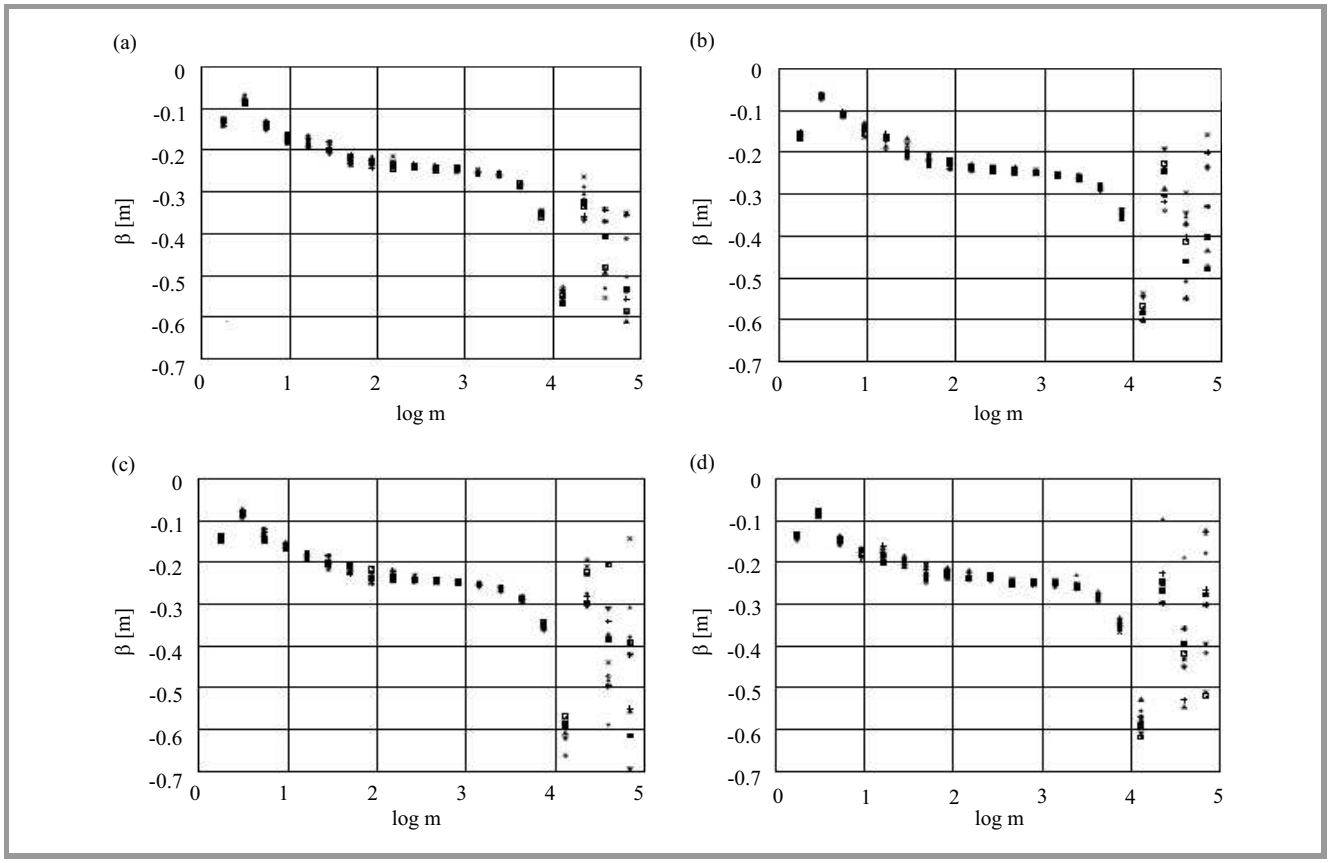


Fig. 2. Attacker signatures (measured on A-PC): (a), (b), (c), and (d) correspond to machines *a*, *b*, *c*, and *d* from Table 1, respectively.

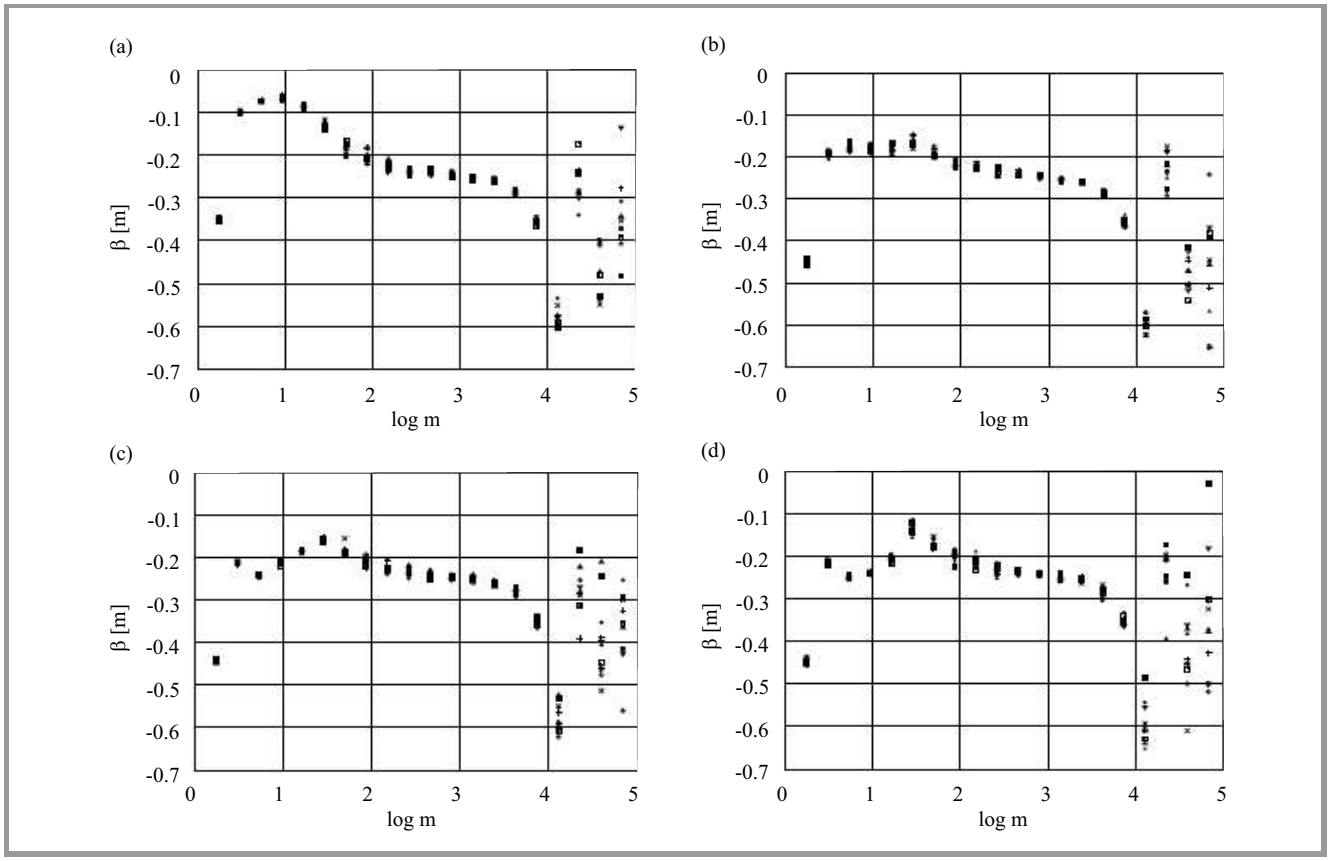


Fig. 3. User signatures (measured on C-PC): (a), (b), (c), and (d) correspond to machines *a*, *b*, *c*, and *d* from Table 1, respectively.

Next the difference between user and attacker signatures is shown. From Figs. 2 and 3, user signatures are different from attacker signatures typically at levels m satisfying $\log(m) < 2$. Note that $m = 10^2$ corresponds to the time scale of one millisecond since $10^2\delta = 10^{-3}$ s. The port-mirroring hub never affects variances at this large time scale. The following is author's explanation for this phenomenon. The difference occurs because A-PC performs only packet collection, while C-PC performs both packet collection and packet processing. On C-PC, the two jobs are executed in parallel on every packet arrival. The two jobs interfere with each other and this interference makes the execution time to obtain the current time fluctuate. In brief, the difference is due to inaccurate packet arrival timestamps caused by resource (memory, CPU, etc.) competition between two jobs. Hence, signatures in Fig. 3 differ only at small time scales. Furthermore, the TCP protocol intensifies the competition because packets tend to arrive in batches when the protocol is used. Since the interference is influenced by various factors (e.g. I/O controllers, device drivers, and job scheduling), it is conjectured that different machine models in Table 1 generated different signatures.

3. Signature Distance

Previous work does not quantify the difference between two signatures [7]. This section first defines the distance between them. Let $\{\beta_i^a\}$ and $\{\beta_i^b\}$ denote user signatures of machines a and b, respectively. They are distinguishable if there exists at least one integer i at which $|\beta_i^a - \beta_i^b|$ is sufficiently large. Therefore, the research is focused on the distance between the i-th decay rates β_i^a and β_i^b.

Figure 4 shows the histogram of 100 β_1 samples and the normal Q-Q plot, which compares the 100 β_1 samples with the theoretical normal distribution. From the figure, β_1 has a distribution close to the normal distribution. The normality tests were performed using 100 β_i samples. If $1 \le i \le 14$, both the Shapiro-Wilk and Anderson-Darling tests do not reject the null hypothesis stating that the β_i samples are normally distributed when the significance level α is 0.01. Therefore, this paper uses only $\{\beta_i\}_{1\le i\le 14}$ for authentication and assumes that β_i^a and β_i^b for $1 \le i \le 14$ are independent and each has a normal distribution.

Let μ_i^a and σ_i^a be the sample mean and standard deviation obtained from ten β_i^a samples. If $\mu_i^a > \mu_i^b$, the distance between β_i^a and β_i^b must be a decrease function of $\Pr(\beta_i^a < \beta_i^b)$, the probability that a sample of β_i^a is smaller than that of β_i^b. Let $F(x;\mu,\sigma^2)$ be the cumulative distribution function (CDF) of normal distribution $N(\mu,\sigma^2)$. If $\mu_i^a > \mu_i^b$, the probability $\Pr(\beta_i^a < \beta_i^b)$ is given by

$$\Pr(\beta_i^a < \beta_i^b) = F(0; |\mu_i^a - \mu_i^b|, (\sigma_i^a)^2 + (\sigma_i^b)^2). \quad (4)$$

Meanwhile, if $\mu_i^a < \mu_i^b$, $\Pr(\beta_i^a > \beta_i^b)$ is equal to the right hand side of Eq. (4). Therefore, independent of whether

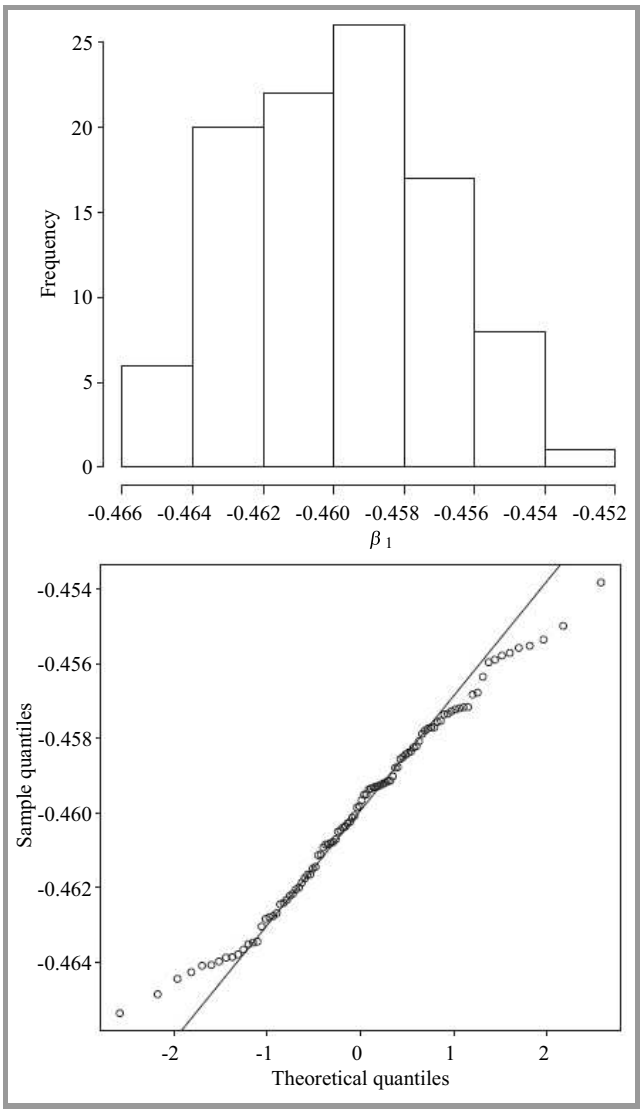


Fig. 4. The histogram and normal Q-Q plot obtained from 100 β_1 samples.

$\mu_i^a < \mu_i^b$ or not, $d_i(a,b)$, the distance between decay rates β_i^a and β_i^b, is defined as

$$d_i(a,b) = -\log F(0; |\mu_i^a - \mu_i^b|, (\sigma_i^a)^2 + (\sigma_i^b)^2). \quad (5)$$

Note that $d_i(a,b)$ is a decrease function of $\Pr(\beta_i^a < \beta_i^b)$ if $\mu_i^a > \mu_i^b$.

Also the $D(a,b)$, the distance between two signatures $\{\beta_i^a\}$ and $\{\beta_i^b\}$, is defined as

$$D(a,b) = |\{i \in \{1,2,\ldots,14\} | d_i(a,b) \ge L_i\}|. \quad (6)$$

Briefly, $D(a,b)$ is the number of integers i that satisfy $d_i(a,b) \ge L_i$, where threshold L_i, which is obtained later, determines whether the i-th decay rates of two signatures are the same or not. Two distance functions d and D have the following features: $d_i(a,b) = d_i(b,a)$, $D(a,b) = D(b,a)$, $d_i(a,b) \ge d_i(a,a) = -\log 0.5 (\approx 0.3)$, $D(a,b) \ge D(a,a) = 0$, and they do not satisfy the triangle inequality.

3.1. Signature Verification

Algorithm 1 shows the signature verification procedure A_s. In the algorithm, $\{\tilde{\beta}_i^b\}$ indicates the most recently obtained sample signature of machine b. Let G be the set of all user machines that request signature verification. Given $a(\in G)$ and $\{\beta_i\}$, procedure A_s returns "accept" if $\{\beta_i\}$ is considered as a signature of machine a; otherwise, it returns "reject". If accepted, $\{\tilde{\beta}_i^a\} = \{\beta_i\}$.

Algorithm 1: Signature verification procedure A_s.

```
    Require: For any b ∈ G_a, D(a,b) ≥ 1.
 1: procedure A_s(a, {β_i})
 2:     while G_a is not empty do
 3:         Select b ∈ G_a
 4:         G_a ← G_a \ {b}
 5:         for i = 1 to 14 do
 6:             if d_i(a,b) ≥ L_i then
 7:                 if (μ_i^a − μ_i^b)(β_i − β̃_i^b) ≤ 0 then
 8:                     return reject
 9:                 end if
10:             end if
11:         end for
12:     end while
13:     return accept
14: end procedure
```

The verification is performed by comparing the signature $\{\beta_i\}$ with signatures of other machines. Let G_a be the set of machines that are used for verifying a signature of machine a. The procedure works under the condition that for any $b \in G_a$, $\{\beta_i^a\}$ and $\{\beta_i^b\}$ are distinguishable, i.e.

$$D(a,b) \geq 1, \text{ for any } b \in G_a. \tag{7}$$

In the 7th line of Algorithm 1, an inequality

$$(\mu_i^a - \mu_i^b)(\beta_i - \tilde{\beta}_i^b) \leq 0 \tag{8}$$

implies that the magnitude relation between μ_i^a and μ_i^b is different from that between β_i and $\tilde{\beta}_i^b$. Note that P_{error}, the probability that inequality (8) holds, is

$$P_{error} = 10^{-d_i(a,b)}. \tag{9}$$

This seldom occurs if $d_i(a,b) \geq L_i$ (in the 6th line of Algorithm 1) holds for a large L_i. Thus, the procedure considers that $\{\beta_i\}$ is not a signature of machine a and returns reject. Procedure A_s returns accept after it makes t comparisons, where $t = \sum_{b \in G_a} D(a,b)$. Therefore, P_{forge}, the probability that an attacker successfully forges a signature that is accepted by the procedure, is

$$P_{forge} = 2^{-t} \tag{10}$$

if the attacker has no information about the signature of machine a. The forgery probability exponentially decreases with t.

Algorithm 1 needs statistics μ_i^a, μ_i^b, σ_i^a, and σ_i^b for calculating $D(a,b)$ and $d_i(a,b)$. It is recommended that μ_i^a, σ_i^a, and G_a should be updated by using the latest samples because these statistics may be affected by various software updates (e.g. Windows update).

3.2. Requirements

The traffic signature-based authentication has the same targets of challenge as biometric-based authentication, where biometric information (e.g., fingerprint, iris, etc.) is required to hold three requirements [14]:

- R1 – it is sufficiently different between any two users,
- R2 – it is reproducibly captured repeatedly,
- R3 – it is hard to be faked.

This paper focuses on R1 and R2. Before verifying whether traffic signatures satisfy R1 and R2, there is need to determine the values of $\{L_i\}$, which are criteria for determining whether the i-th decay rates of two signatures are different or not. This paper decomposes L_i into two parts as

$$L_i = L_s + \Delta L_i, \tag{11}$$

where $L_s(>0)$ is the distance required by security strength and $\Delta L_i(>0)$ denotes the fluctuation range of distance d_i caused by changes in CPU and network loads, etc. Requirement R1 demands that an integer i that satisfies $d_i(a,b) \geq L_i$ for any a and b should exist, and R2 insists that ΔL_i should be sufficiently small. In the next section, ΔL_i values are experimentally derived.

First the value of L_s such that the authentication system satisfies the following capability is determined. A user accesses the system every one minute and A_s returns reject once in a year on average. Assume that $\Delta L_i = 0$ and that for all $b \in G_a$ and for all i satisfying $d_i(a,b) \geq L_s$, $d_i(a,b) = L_s$. Then, the mean of the binomial distribution indicates that

$$60 \cdot 24 \cdot 365 \cdot \left(1 - (1 - 10^{-L_s})^t\right) = 1, \tag{12}$$

where $t = \sum_{b \in G_a} D(a,b)$. From Eqs. (9) and (10), both L_s and t should be enlarged as much as possible for minimizing both P_{error} and P_{forge}. From Eq. (12), the L_s and t at the same time can't be reduced. If $t = 20$, from Eq. (12), $L_s \approx 7$. In this case, from Eqs. (9) and (10), $P_{error} \approx 10^{-7}$ and $P_{forge} \approx 10^{-6}$ (i.e. the security strength corresponds to a six-digit code). Hereinafter, $L_s = 7$ is used.

4. Signature Analysis

4.1. Four Machines

This section discusses whether traffic signatures satisfy inequality (7) by using machine set $G = \{a,b,c,d\}$, which consists of four machines listed in Table 1, under the condition that fluctuation $\Delta L_i = 0$. Figure 5 exhibits distances d_i derived from user signatures of machines in G. It can be seen from Figs. 3 and 5 that d_i, $1 \leq i \leq 14$, correctly quantify signature differences. Let us see the distance between

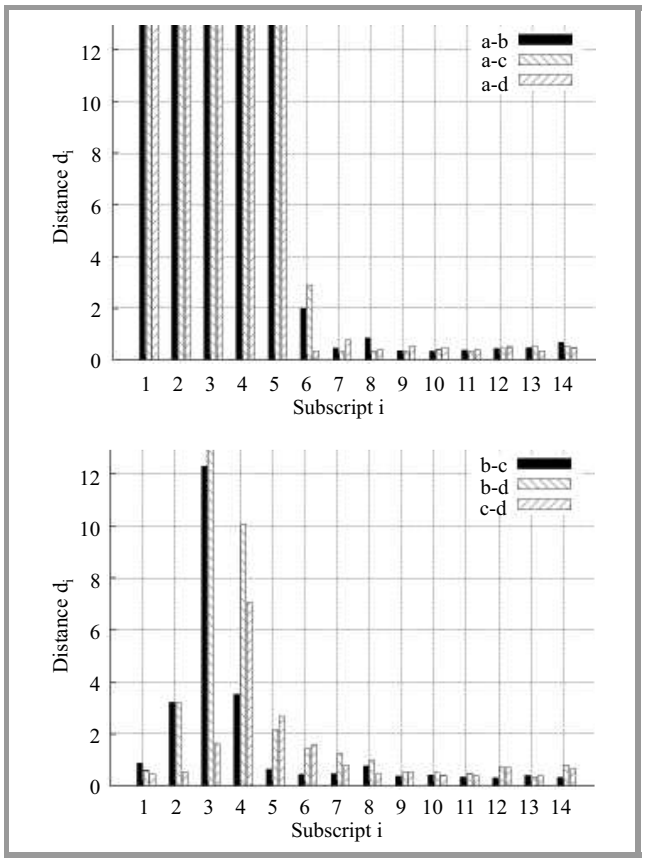


Fig. 5. Distances d_i for all machine pairs in Table 1 – "*a-b*" indicates $d_i(a,b)$.

signatures of machines a and b ("*a-b*" in Fig. 5). From the figure, $d_i(a,b) > L_i(=7)$ at $i = 1,2,3,4,5$. Therefore, $D(a,b) = 5$. Similarly, $D(a,c) = D(a,d) = 5$, $D(b,d) = 2$, and $D(b,c) = D(c,d) = 1$. Accordingly, (7) holds for all sets G_x, $x = a,b,c,d$ when $G_x = G \setminus \{x\}$. Figure 5 also shows that for any $x,y \neq a$, $D(a,y) \geq D(x,y)$. Namely, machine a creates the most characteristic signature. This may be because from Table 1, machine a is the oldest PC (therefore, it may be composed of many unique devices) and its Windows version is different from those of the others.

4.2. Fluctuation Range

Next the fluctuation range ΔL_i discussed in Subsection 3.2 is estimated. Let Δd_i be the decrease in the distance $d_i(x,y)$ caused by the increase in the load of machine x, i.e.

$$\Delta d_i = d_i(x,y) - d_i(\tilde{x},y), \tag{13}$$

where $\tilde{x}$ denotes machine x whose load has been raised. The author estimates the distribution of Δd_i through experiments with various machines x, and then determine ΔL_i such that Δd_i is not greater than ΔL_i with probability 0.95 ($\Pr(\Delta d_i \leq \Delta L_i) = 0.95$).

From Eq. (13), mean μ_i^y and variance $(\sigma_i^y)^2$ of machine y are necessary to obtain Δd_i. These values are independent of β_i^x and $\beta_i^{\tilde{x}}$. However, y should be as normal as possible. Therefore, one can assume that variance $(\sigma_i^y)^2$ is equal to the mean of variances $(\sigma_i^x)^2$ of various machines x, i.e.,

$$(\sigma_i^y)^2 = m\left((\sigma_i^x)^2\right), \tag{14}$$

where this paper uses $m(Z_i)$ and $s(Z_i)$ to indicate the mean and standard deviation of samples $\{Z_i\}$. On the other hand, μ_i^y can be determined by assuming that due to the load increase, the distance decreases to L_s, i.e.

$$d_i(\tilde{x},y) = L_s. \tag{15}$$

From Eqs. (13) and (15), we have

$$d_i(x,y) = \Delta d_i + L_s. \tag{16}$$

In short, the load increase lowers the distance from $L_s + \Delta d_i$ to L_s. Using Eqs. (5) and Eqs. (14), (15) and (16) are rewritten as

$$-\log F\left(0; |\mu_i^{\tilde{x}} - \mu_i^y|, (\sigma_i^{\tilde{x}})^2 + m(\sigma_i^x)^2\right) = L_s \tag{17}$$

$$-\log F\left(0; |\mu_i^x - \mu_i^y|, (\sigma_i^x)^2 + m(\sigma_i^x)^2\right) = \Delta d_i + L_s. \tag{18}$$

Given $\mu_i^{\tilde{x}}$ and $(\sigma_i^{\tilde{x}})^2 + m(\sigma_i^x)^2$, (17) has two solutions μ_i^y, so that (17) and (18) yield two Δd_i values for each x. The 24 Δd_i values are obtained using various desktop and laptop PCs x. For measuring signatures of $\tilde{x}$, the CPU, memory, and hard disk utilization rates are raised by playing a video (whose bitrate is 2.4 Mb/s) stored on the hard disk. The Shapiro-Wilk and Anderson-Darling tests do not reject the normality of 24 Δd_i samples for all $i = 1,2,\ldots,14$. Thus, this paper assumes that Δd_i has a normal distribution. Using mean $m(\Delta d_i)$ and standard deviation $s(\Delta d_i)$, ΔL_i is given by

$$\Delta L_i = m(\Delta d_i) + z_{0.95} s(\Delta d_i), \tag{19}$$

where $z_{0.95}$ satisfies $F(z_{0.95};0,1) = 0.95$; i.e., $\Pr(\Delta d_i \leq \Delta L_i) = 0.95$. Table 2 shows $m(\Delta d_i)$, $s(\Delta d_i)$, and ΔL_i derived from Δd_i samples. From the table, fluctuation ranges ΔL_i depend on i and are between 2.5 and 5. If ΔL_i values in Table 2 are used for calculating L_i in Eq. (11), we have $D(x,y) \geq 1$ for all $x,y \in G$ in Subsection 4.1 except for $D(c,d)$. Therefore, in equality (7) does not hold. In other words, the four machine models do not satisfy requirement R1 in Subsection 3.2. In Section 5, author improves signature uniqueness to fulfill the requirement.

Table 2
Means $m(\Delta d_i)$, standard deviations $s(\Delta d_i)$, and fluctuation ranges ΔL_i for $i = 1,2,\ldots,14$

i	1	2	3	4	5	6	7
$m(\Delta d_i)$	–0.1	–0.2	0.3	–0.1	0.3	0.2	0.1
$s(\Delta d_i)$	2.6	1.9	2.3	1.9	1.9	2.1	2.2
ΔL_i	4.2	2.9	4.1	2.9	3.3	3.7	3.7

i	8	9	10	11	12	13	14
$m(\Delta d_i)$	0.3	–0.1	0.0	0.9	1.3	–0.4	0.0
$s(\Delta d_i)$	2.4	1.9	2.1	2.5	2.3	1.7	2.3
ΔL_i	4.2	3.0	3.4	5.0	5.0	2.5	3.8

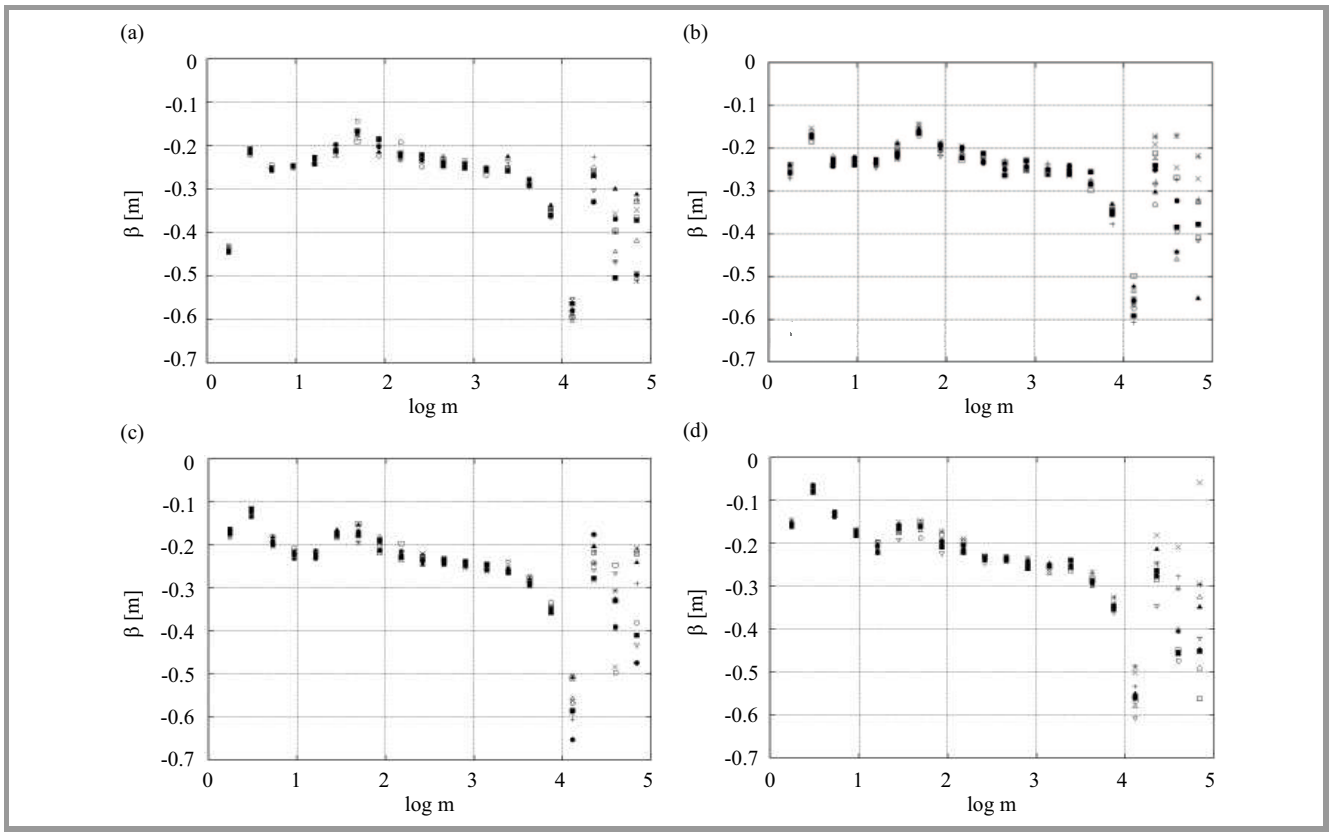


Fig. 6. User signatures for (a) g^1, (b) g^8, (c) g^{16}, and (d) g^{32}.

5. Uniqueness Improvement

So far the case where different machine models yield different signatures was considered. This result suggests that replacing some hardware or software components of C-PC might produce signatures that satisfy requirements R1 and R2. This section discusses the way how to change user signatures without adding a hardware device to C-PC. Note that as mentioned in the previous section, changing user signatures is often necessary for security.

5.1. Process Interference

One approach for uniqueness improvement is to increase the number of WinDump processes on C-PC. Let g^k denote machine g on which k WinDump processes are running. Figure 6 shows user signatures for g^1, g^8, g^{16}, and g^{32}. As shown in the figure, each number k creates a unique signature.

Figure 7 exhibits distances d_i between signatures of g^1 and g^k, $k > 1$. The figure demonstrates that the process interference-based approach generates many distinguishable signatures since $D(g^1, g^k) \geq 1$ for $k \in \{4, 8, 16, 32\}$. Figure 7 also provides the following attractive facts:

- distances $d_i(g^1, g^k)$ at $1 \leq i \leq 4$ increase with k,
- numbers i that satisfy $d_i(g^1, g^k) \geq L_i$ increase with k.

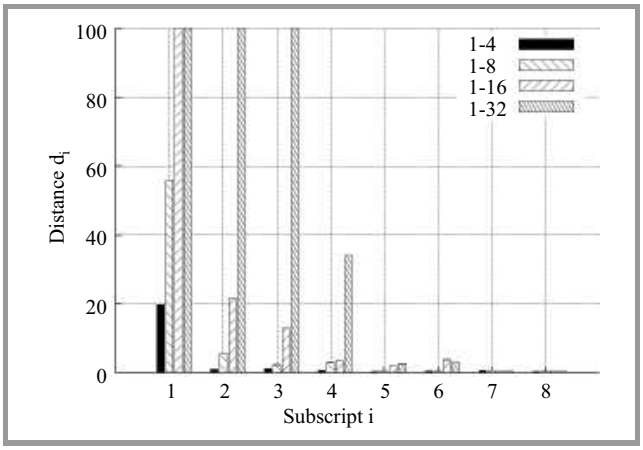


Fig. 7. The impact of the number of WinDump processes on distances d_i – "1-4" indicates $d_i(g^1, g^4)$.

The author conjectures that these phenomena arise due to process interference. Every time a packet arrives at C-PC, a packet processing process and k WinDump processes all start at once, so that they severely compete for CPU and buffer resources if k is large. Number k represents the degree of competition. As k increases, packet arrival timestamps become more inaccurate, and this inaccuracy result in the emergence of unique user signatures.

The process interference approach is neither CPU nor memory intensive. Therefore this approach can be considered

as a key technology for traffic signature-based user authentication.

5.2. Other Approaches

This subsection explores other possible approaches for producing unique signatures and investigates whether they are available under various hardware and software configurations. Experiments were made with 14 machines *e*-*r*, whose system information is listed in Table 3. The research is focused on Windows machines because most of personal desktop and laptop PCs use Windows OS. For comparison, Mac and Linux machines are included in the table. On Mac and Linux machines, tcpdump [15] runs instead of WinDump. Since their packet analysis mechanisms are different [16], experimental results may depend on which of them is used. The table also shows the maximum distances ($\max_i d_i$) between signatures measured before and after each of the following three operations:

Snaplen: The snapshot length of each packet collected by WinDump or tcpdump is increased to 4096 bytes (the default is 68 bytes). As a result of this, a larger data amount are stored in the hard disk.

Interfer: Eight WinDump (or tcpdump) processes are executed so that they severely compete for CPU and memory resources. Figure 6 is the result obtained by this operation with machine *g* in Table 3.

Load: The machine workload is raised by executing eight VLC media players [17], all of which play a video file on the hard disk.

In Table 3, symbol ◎ implies that the operation has an ability to create distinguishable signatures, and ○ indicates that distinguishable signatures may be obtained if the operation is adequately tuned (e.g. the snapshot length further increases). The table shows the following three results:

- the snapshot length-based approach may not yield long distances,
- the load-based approach is not suited to laptop PCs since the online news may abnormally terminate,
- the process interference-based approach is the most effective and stable approach.

However, this approach should be adequately tuned since $\max_i d_i$ depends on the machine configuration. Some machines require a large number of WinDump (or tcpdump) processes.

An advantage of the process interference-based approach is that a variety of recent and ongoing computer technologies keep producing unique and unpredictable signatures. Even if detail hardware and software specification of C-PC is given, obtaining user signatures through computation must be a difficult task. At the same time, however, new computer technologies make $\max_i d_i$ variable, so that the number of WinDump processes may need to be revised. The author considers the following recent technologies must have impacts on d_i:

Table 3
Three operations are performed to see whether they can yield distinguishable signatures under various machine configurations of C-PC. Symbols ◎ and ○ indicate $\max_i d_i \geq L_i$ and $2 < \max_i d_i < L_i$, respectively. Numbers in parentheses denote $\max_i d_i$

PC	OS	CPU	Snaplen	Interfer	Load
e	Vista	Q9450		○ (5)	
f	Win 7	i5-2400S		(2)	◎ (54)
g		i7-930		◎ (18)	◎ (51)
h		i7-960	○ (5)	◎ (104)	○ (3)
i		i3-2120	○ (6)	◎ (69)	○ (3)
j		i7-2600		◎ (62)	◎ (90)
k		i3-2130		○ (3)	◎ (37)
l	Win 8	i5-3350P	◎ (16)	◎ (56)	
m		i7-4770	◎ (15)	◎ (37)	◎ (13)
n (1)	Win 7	AMD	◎ (12)	◎ (66)	(4)
o (1)		Atom		◎ (38)	(4)
p (1)	Win 8	i5-3317U	◎ (24)	◎ (59)	
q	Linux	i7-3770K		(3)	
r	MAC	i7-2630		○ (5)	

Notes:
(1) *n*, *o*, *p* – laptop computers.
(2) $\max_i d_i = 2$ at 16 processes.
(3) $\max_i d_i = 2$ at 64 processes.
(4) The online news abnormally terminates.

Timestamp precision: WinPcap (a Windows library used by WinDump) by default obtains the timestamp through kernel function KeQueryPerformanceCounter(), which provides a time reference with microsecond precision. By modifying a registry key, timestamps are generated through faster i386 instruction RDTSC, which accesses TimeStamp Counter (TSC), whose precision is equivalent to the CPU frequency. RDTSC works only on Intel CPUs and is expected to provide nanosecond time resolution [18], [19].

Multiprocessing: A packet received by a NIC is stored in the NIC driver buffer. The timestamp of the packet is measured after the capture driver is invoked through a hardware interrupt. If there are pending interrupts, the driver is executed after all these interrupts are served. Therefore, the timestamp is significantly inaccurate if there are a large number of pending interrupts [18]. WinPcap works on symmetric multi-processing (SMP) machines. Multiple processors concurrently execute the same instance of the

capture driver, so that each processor handles a different packet stored in the NIC driver buffer. Accordingly, the delay of the driver execution depends on the number of processors [20].

Turbo Boost: Turbo Boost is a technology that enables the processor to run above its base operating frequency when workload on the processor calls for faster performance. The timestamp accuracy is affected by the technology since it dynamical changes processing capability.

5.3. Traffic Control

Some traffic control software tools effectively create unique signatures (probably because they frequently consult the current time). Traffic control is performed to reduce congestion, latency and packet loss by prioritizing, controlling, or reducing the network traffic. One of the traffic control tools is dummynet [21]. It emulates a network link that consists of a transmission link with fixed bandwidth B and propagation delay t_D and a finite FIFO queue with tail-drop. For link emulation, dummynet delays each packet i by $(\ell_i + Q_i)/B + t_D$, where ℓ_i is the length of packet i and Q_i is queue occupation when packet i was queued.

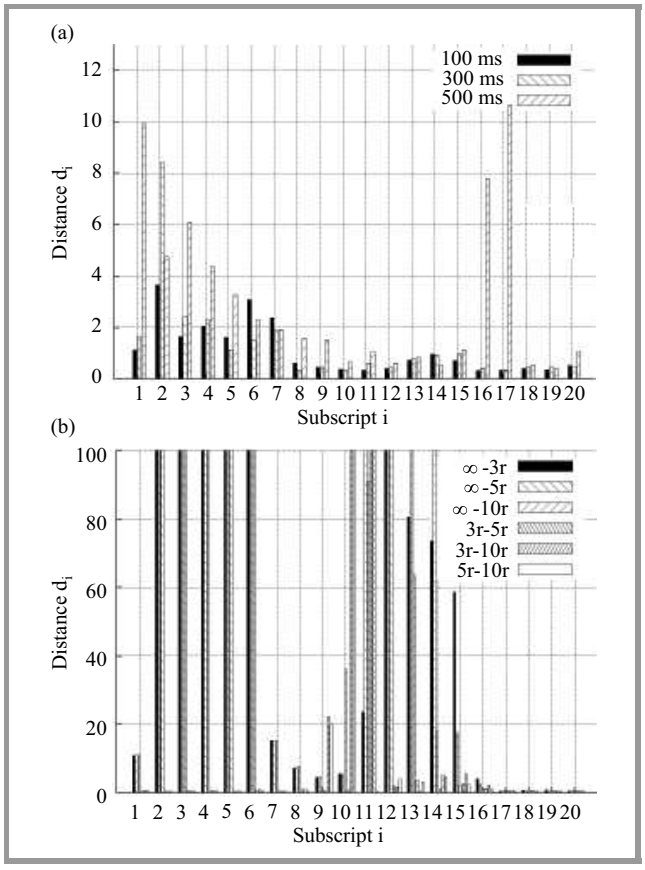


Fig. 8. The impact of: (a) propagation delay t_D and (b) bandwidth B on the distance. "100 ms" indicates the distance between two signatures measured at $t_D = 0$ and $t_D = 100$ ms. "$3r-5r$" indicates the distance between two signatures measured at $B = 3r$ and $B = 5r$.

Figure 8a shows distances between signatures measured at $t_D = 0$ and $t_D > 0$, where dummynet on C-PC delays every incoming and outgoing packets by t_D. Since video quality deteriorates greatly when $t_D = 500$ ms, the TCP window scale option [22], which allows larger windows to be used, is set to work when $t_D = 500$ ms. The figure demonstrates that $\max_i d_i$ is too small to distinguish signatures for all t_D values. On the other hand, as shown in Fig. 8b, dummynet bandwidth B is useful in raising $\max_i d_i$ greatly. From the figure, $\max_i d_i$ exceeds 100 at multiple values of i. By looking closely at Fig. 8b, it can be seen that in the case of "$\infty - 3r$", $\max_i d_i > 100$ at $i \in \{2,3,4,5,6,12\}$, where "$\infty - 3r$" denotes the distance between two signatures measured at $B = \infty$ (i.e. the bandwidth is unlimited) and $B = 3r$, where r is the average rate of the video stream. Note that perceived video quality is not degraded as long as B is at least three times greater than r. All approaches in Subsection 5.2. change user signature $\{\beta_i\}$ only at small time scales i (e.g. $i \leq 5$), whereas by changing bandwidth B, $\{\beta_i\}$ varies at large i (e.g. $9 \leq i \leq 14$). This is an important advantage of this approach.

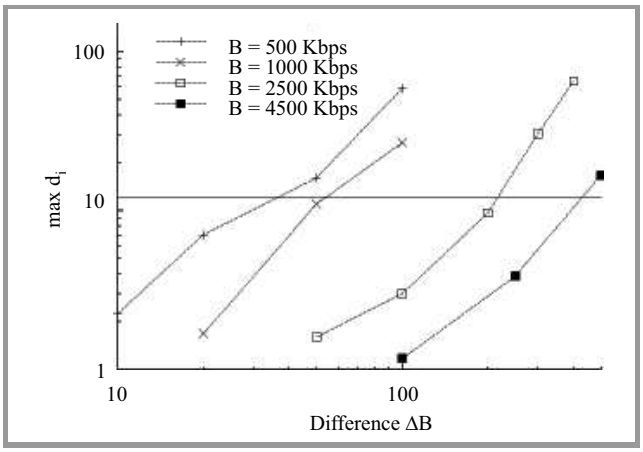


Fig. 9. A larger B requires a larger ΔB to satisfy $D \geq 1$. The solid line denotes $\max_{1 \leq i \leq 14} L_i$.

Figure 9 shows $\max_i d_i$ between two signatures obtained when the dummynet bandwidths are B and $B + \Delta B$ Kb/s. The solid line in the figure denotes $\max_{1 \leq i \leq 14} L_i (= 12)$, so that $D \geq 1$ if $\max_i d_i$ is above the line. From the figure, one can roughly estimate how many distinguishable signatures one can be obtained by changing the bandwidth, because the figure explains how $\max_i d_i$ increases with ΔB and how the minimum ΔB that satisfies $D \geq 1$ grows with B. For example, one can get roughly ten distinguishable signatures in the range of $500 \leq B \leq 1000$ Kb/s since from the figure, the smallest ΔB that satisfies $D \geq 1$ is approximately 50 Kb/s in the range.

6. Conclusions

For protecting users who place high value on their accounts, various what you have authentication technologies

have been proposed. However, they are not widely used today mainly because security hardware added to the user machines poses other new problems. Additional hardware is not necessary if the user machine itself is identified. In this paper, the feasibility of applying the traffic signature to the user machine identification has been discussed, where the signature is calculated from HTTP-based video traffic transmitted by the authentication server. This paper focused on uniqueness and reproducibility of the signature based on the distance function defined in this paper and obtained the following results:

Uniqueness was verified based on a criterion, which requires that the distance between any two signatures is not less than $L_s + \Delta L_i$, where $L_s = 7$ and ΔL_i is the fluctuation range of the i-th decay rate. The security strength corresponds to a six-digit code when $L_s = 7$. Although different machine configuration models tended to provide different signatures, these signatures did not always meet the criterion. However, the process interference approach, in which the number of executing packet-capture processes is used as a parameter for controlling the accuracy of packet arrival timestamps, was shown to be effective for producing signatures that meet the criterion.

Reproducibility was verified by calculating signatures from real Internet traffic delivered by France 24 live. Although the traffic traversed 30 routers and experienced a long propagation delay, signatures measured on various machines were stable especially over small time scales. Sample signatures showed that the fluctuation ranges ΔL_i were between 2.5 and 5. Therefore, $9.5 \leq L_s + \Delta L_i \leq 12$. When the machine load is highly increased by playing eight video files in parallel, five machines out of fourteen generated signatures that exceed $L_s + \Delta L_i$. Therefore, some machines need to reduce their loads before performing the authentication. However, dummynet, a traffic control tool, is expected to mitigate the impact of the load because dummynet generated many signatures whose distances from the original signature were significantly large (more than 100).

References

[1] R. P. Guidorizzi, "Security: Active authentication", *IT Professional*, vol. 15, no. 4, pp. 4–7, 2013.

[2] Federal Financial Institutions Examination Council, "Authentication in an internet banking environment", 2005 [Online]. Available: http://www.ffiec.gov/pdf/authentication_guidance.pdf

[3] M. Jakobsson, R. Chow, and J. Molina, "Authentication – are we doing well enough? [guest editors' introduction]", *IEEE Secur. & Priv.*, vol. 10, no. 1, pp. 19–21, 2012.

[4] D. DeFigueiredo, "The case for mobile two-factor authentication", *IEEE Securi. & Priv.*, vol. 9, no. 5, pp. 81–85, 2011.

[5] M. Sasse and C. C. Palmer, "Protecting you" [guest editors' introduction], *IEEE Secur. & Priv.*, vol. 12, no. 1, pp. 11–13, 2014.

[6] C. Herley, "More is not the answer", *IEEE Secur. & Priv.*, vol. 12, no. 1, pp. 14–19, 2014.

[7] K. Oida, "A traffic signature sensitive to client machines", in *Proc. Int. Conf. Adv. Comp. Inform. Technol. ACIT'13*, Kuala Lumpur, Malaysia, 2013.

[8] E. De Cristofaro, H. Du, J. Freudiger, and G. Norcie, "Two-factor or not two-factor? A comparative usability study of two-factor authentication", Computing Research Repository, 2013.

[9] K. Oida, "Video traffic attributes for end host identification", *Int. J. Comp. Commun. Engin.*, vol. 1, no. 4, pp. 396–401, 2012.

[10] Trusted Computing Group, "Trusted platform module (TPM) summary", July 2009 [Online]. Available: http://www.trustedcomputinggroup.org/resources/trusted_platform_module_tpm_summary

[11] O. Oyman and S. Singh, "Quality of experience for http adaptive streaming services", *IEEE Commun. Mag.*, vol. 50, no. 4, pp. 20–27, 2012.

[12] J. Beran, *Statistics for Long-Memory Processes*. Chapman & Hall/CRC Press, 1994, vol. 61.

[13] WinDump [Online]. Available: http://www.winpcap.org/windump/

[14] T. Weigold, T. Kramp, and M. Baentsch, "Remote client authentication", *IEEE Secur. & Priv.*, vol. 6, no. 4, pp. 0036–43, 2008.

[15] tcpdump [Online]. Available: http://www.tcpdump.org/

[16] F. Risso and L. Degioanni, "An architecture for high performance network analysis", in *Proc. 6th IEEE Symp. Comp. & Commun. ISCC 2001*, Hammamet, Tunisia, 2001, pp. 686–693.

[17] VLC media player [Online]. Available: http://www.videolan.org/vlc/

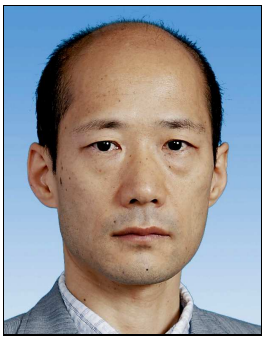
[18] L. Degioanni, M. Baldi, F. Risso, and G. Varenni, "Profiling and optimization of software-based network-analysis applications", in *Proc. 15th IEEE Symp. Comp. Architec. & High Perform. Comput. SBAC-PAD'03*, São Paulo, Brazil, 2003, pp. 226–234.

[19] P. Orosz and T. Skopko, "Performance evaluation of a high precision software-based timestamping solution for network monitoring", *Int. J. Adv. Softw.*, vol. 4, no. 1 and 2, pp. 181–188, 2011.

[20] G. Varenni, M. Baldi, L. Degioanni, and F. Risso, "Optimizing packet capture on symmetric multiprocessing machines", in *Proc. 15th IEEE Symp. Comp. Architec. & High Perform. Comput. SBAC-PAD'03*, São Paulo, Brazil, 2003, pp. 108–115.

[21] M. Carbone and L. Rizzo, "Dummynet revisited", *ACM SIGCOMM Comp. Commun. Rev.*, vol. 40, no. 2, pp. 12–20, 2010.

[22] V. Jacobson *et al.*, "TCP extensions for high performance", IETF, RFC 1323, May 1992 [Online]. Available: www.ietf.org/rfc/rfc1323.txt



Kazumasa Oida received the Bachelor of Information Science, Master of Engineering, and Doctor of Informatics degrees from the University of Tsukuba in 1983, Hokkaido University in 1985, and Kyoto University in 2002, respectively. He is currently a Professor in the Department of Computer Science and Engineering, Fukuoka Institute of Technology, Japan. His main interests include analysis and modeling of packet traffic and the origin of adaptive behavior in complex dynamic systems.
E-mail: oida@fit.ac.jp
Department of Computer Science and Engineering
Fukuoka Institute of Technology
Fukuoka, 811-0295 Japan

Paper

On Providing Cloud-awareness to Client's DASH Application by Using DASH over HTTP/2

Jordi Mongay Batalla[1], Piotr Krawiec[1], Daniel Negru[2], Joachim Bruneau-Queyreix[2], Eugen Borcoci[3], Andrzej Bęben[4], and Piotr Wiśniewski[4]

[1] *National Institute of Telecommunications, Warsaw, Poland*
[2] *CNRS-LaBRI, Bordeaux, France*
[3] *University Politehnica Bucharest, Bucharest, Romania*
[4] *Institute of Telecommunication, Warsaw University of Technology, Warsaw, Poland*

Abstract—Mobile Cloud Networks group together mobile users and clouds containing content servers. Hence, they are an ideal framework for media content delivery. Stream-switching adaptive video players cope well with some limitations of Mobile Cloud Networks as low bandwidth and bandwidth variability in access network. Nonetheless, other limitations, as cloud congestion, are difficult to be managed by the video players. This paper presents a system for discovering fault situations at the cloud (e.g., cloud congestion) and notifying to the video player, which will take appropriate actions for saving the quality of media transmission. In proposed implementation the video application is DASH-capable and adaptation action may be both stream rate adaptation and content server adaptation. The communication between client and server uses "bidirectional" communication feature of HTTP/2 thanks to the new deployed modules running DASH over HTTP/2 in both client's and server's applications.

Keywords—adaptive video streaming, DASH, HTTP/2.

1. Introduction

Mobile Cloud Networks (MCN) are a great opportunity for media delivery since they group together mobile users and media content servers (in cloud networks) under the same umbrella. However, the challenging issue still remains the quality of the delivery coming up to the user's expectations, without escalating the cost [1]. In fact, most of the recognized problems of MCN [2]–[5] directly affect the media streaming and should be solved for offering integrated solutions of media delivery. Specifically, the limitations of MCNs that mainly affect media delivery are:

- low bandwidth in the wireless access networks [2]. Even when 4G technology increased the bandwidth in mobile devices, the continuous raise of demand for mobile applications presents bandwidth limitations at users' disposal;
- excessive latency in wireless access networks [3], due to physical media delay caused by the low-quality connectivity with cell tower, and the latency of the protocol for access to the physical media;
- resource limitations of mobile devices [3], i.e., electric power, processing power and storage capacity. In the case of video applications, an added resource limitation in mobile devices is the poor display in comparison to laptops or TV;
- non-optimal management of access network resources [4], which influences negatively in the aforementioned points causing inappropriate access to the available bandwidth and increasing unnecessarily the latency in access networks;
- congestion of cloud networks [5], defined as the overload of any of the cloud resources (e.g., processing capacity, uplink bandwidth), which causes delay in the services damaging the delivery of media content and, as a consequence of this, impoverishing the quality of the media event.

Aforementioned limitations require extra-management mechanisms into the cloud directed to ensure appropriate user's satisfaction of the media event [6]. This paper centers on one of the mentioned limitations. It presents a management framework for discovering Cloud congestion situations and informing the user's video player about the predicted state of the cloud (avoiding to keep per-connection information in the cloud). The user's video player has then information for interpreting the cause of the reduction of bandwidth measured at the player. If the bandwidth reduction is caused by the cloud congestion, then the video player could adapt the video content server by switching the streaming to another server located in a different cloud. If the bandwidth reduction is caused by congestion in the user's mobile network, then the video player could adapt the streaming by reducing the media bitrate.
Even when the presented framework is independent of the video player (client application), the solution fits well in the case of stream-switching adaptive video players. Stream-switching adaptive players (e.g., Dynamic Adaptive Streaming over HTTP-DASH, Akamai HD Video Streaming – AHDVS, Adobe Dynamic Streaming and Apple HTTP Adaptive Live Streaming – HLS) request consecutive small portions of video (called chunks or segments), each one with the appropriate media rate (called representation rate).

The player decides both the representation rate and the content server from where the next segment will be downloaded, so it may adapt to the current state of the network and of the content server.

The paper is organized as follows. Section 2 gathers the state of the art of the proposed management framework and stream-switching adaptive protocols, whereas Section 3 provides some exemplary simulation results that prove the gain of considering the information about the state of the cloud into adaptation decisions. Section 4 presents the end-to-end framework considering both architecture and communication between involved entities and Section 5 shows implementation details of client's and server's applications deploying the presented framework. The proposed implementation has been optimized from the point of view of cloud management since authors avoid a separate channel for signaling between client's and server's applications. The communication uses the streaming channel instead, and it is based on DASH over HTTP/2, which makes bidirectional-like (initiated from the client or from the server) communication feasible.

An undoubted added value of this paper is the extension of *QTSamplePlayer* [7] for interworking with HTTP/2, which is available to the researchers in the web page: http://www.nit.eu/offer/research-projects-products/http2dash.

Results of the tests performed on the implemented software are presented in Section 6. At last, Section 7 summarizes the paper.

2. Background

Between all the limitations of MCN, stream-switching adaptive protocols cope pretty well with low bandwidth and bandwidth variability thanks to the adaptation mechanism that dynamically selects the media bitrate that fits better with the current download rate. Mobile networks with unmanaged and variable resources require from the client's application to constantly control the occupancy of its buffer [8] in order to avoid, from one side, re-buffering and, from the other side, suboptimal use of resources in the wireless link.

Even if stream-switching adaptive protocols are relatively new, the papers on this issue are countless. Until the publication of some open standards, the stream-switching adaptive protocols were, mainly, close solutions. Therefore, many of the first papers were based on inverse engineering, i.e., the analysis of existing solutions by testing them in different scenarios. A good recompilation of such analyses was made by Akhshabi *et al.* in [9], where the authors compared the adaptation mechanisms used by the most important service providers from the point of view of how aggressive/conservative they are in different scenarios. The authors of [10] compared the Smooth Streaming protocol which downloads video chunks periodically with traditional technique of continuous consecutive download. With the publication of the open MPEG standard Dynamic Adaptive Streaming over HTTP (DASH), the research centered on improving the adaptation algorithm in order to balance the stability of video playout and the stability of buffer occupancy. This trade raises due to the variability of download bandwidth. Since the instability of buffer occupancy is less harmful from the end user's point of view, the applications try, as the main goal, to trade off oscillations in video playout. In order to avoid such oscillations, Dobrian *et al.* [11] outlined the importance of the variability of the size of consecutive segments when estimating the download rate and Seo and Zimmermann [12] proposed to estimate the download rate from the rate measurements of a number of downloaded segments. The number of segments to be considered should be dependent of the state of the network. Other approaches for consolidating video playout consisted of monitoring the connections at the TCP level [8] or using not only measurements of download rate but also other measurements, such as packet losses, delay and TCP throughput [13], [14].

In MCNs the conditions, which the client's application should adapt to, are not controllable at the client's terminal. In other words, it is unlikely to differentiate the increase of the end-to-end delay from the decrease of bandwidth [11], so the congestion in the cloud is difficultly recognizable at the client's side. A more general vision about the situation of the server and network could improve the adaptation decisions. Some researchers have proposed to place the adaptation logic out of the end user's video player. For example Liu et al. proposed to deploy centralized video controllers with a global view of network and server conditions that may take decisions about adaptation for the clients [15]. Rejaie and Kangasharju proposed to introduce proxies for managing the quality of the media streaming at the network level [16]. At last, some papers showed several benefits of controlling the adaptation logic at the server side. For example, Sodagar [17] proposed to base adaptation decisions on the state of the sending buffer instead of the receiving one. Anyway, it seems that the adaptation logic should be located in the end user's video application, mainly by two reasons: the client application is in the best position to detect and respond to the dynamics on time and, on the other hand, there is a strong need for keeping minimal per-connection information in the cloud and servers [15].

In this paper a mechanism for informing the client about cloud congestion situations is proposed. With this information, the client may optimize the adaptation decisions by differentiating mobile access congestion from cloud congestion distinguishing two adaptation actions: media bitrate adaptation when there is mobile access congestion, and content source switching when there is cloud congestion. Moreover, the content server keeps no information about situation of the clients in presented solution.

3. Rationale

The simulation tests provided in this section aim to show that the adaptation decisions based only on user's side

(video player) measurements are more error prone than in-cloud decisions assuming that further are supported by measurements at the network level (as it occurs in commercial cloud systems). Let us remark that the current simulation analysis does not aim to cover a wide range of cases and scenarios for offering exhaustive results about the necessity of considering the state of the cloud into adaptation decisions, but they only aim to confirm (in an example scenario) that more information (about the state of the cloud) is useful for taking better decisions about the media streaming.

With this scope, the downloading of segments and calculate the time of download of each segment (segment download time) in presence of background traffic is modeled. A posteriori (when the simulations are finished), the behavior of user- and cloud-based adaptation algorithms for the values of segment download rate obtained in the simulations will be compared.

The simulation scenario is shown in Fig. 1. The cloud uplink is modelled by a single server with infinite FIFO queue that serves (service time equal to 120 μs) the stream under test, which is composed of "segments" of 333 packets (10^5 of such segments), and the background traffic, which is Poisson traffic with a rate that varies for different tests: $R_{bg} = \{2,\ 3,\ 4,\ 5 \text{ and } 6\} \times 10^3$ packet/s. The packets of each segment of the stream under test arrive to the queue with an interarrival time equal to 12 μs (shaped packets arrival without considering TCP effects) and the first packet of each segment is sent only when the last packet of the previous segment finished its service in the queue.

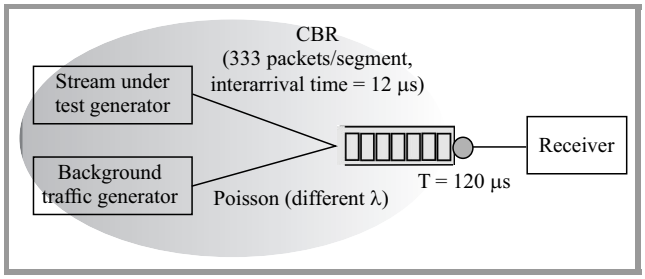


Fig. 1. Simulation model for rationale.

Note that if no background traffic were in the queue, then the complete segment would be downloaded in around 40 ms and the next segment could be sent immediately after that. Whereas, if there were 6×10^3 packet/s background traffic, then the segment would be downloaded in around 0.22 s, which means 7.5×10^3 packet/s of total traffic (stream under test and background) and server utilization $\rho \approx 0.9$.

The test is repeated for five values of background traffic rate indicated above. The mean rate of the stream under test (called *base rate*) equals $\{3.9,\ 3.3,\ 2.8,\ 2.2 \text{ and } 1.6\} \times 10^3$ packet/s, respectively for each aforementioned value of background traffic rate, which means queue utilization $\rho \approx \{0.7, 0.75, 0.8, 0.85 \text{ and } 0.9\}$, respectively. Let us remark that for each segment of the stream, the download rate is slightly different to the *base rate* due to the unpredictability of Poisson traffic.

When the simulations are finished, the download rate of each segment (10^5 segments in each one of the five tests), r^i, as the number of packets in the segment (333 packets) divided by the segment download time are calculated.

With the obtained values of segment download rate, the authors try to understand how the adaptation algorithm works in the case when it has no information about the state of the cloud (the server and the queue) and when it has this information. During adaptation decisions, three potential representations with rates: R_1, R_2 and R_3, where R_2 is the base rate minus 5% (of the base rate) are assumed. R_1 is the base rate minus 15% and R_3 is the base rate plus 5%, as indicated in Table 1. Note that R_2 is the reference representation, i.e., the representation that should be selected for all segments if the background traffic were Constant Bit Rate (CBR) instead of Poisson.

Table 1
Probability of erroneous decision

P	Representation (packets $\times 10^3$/s)	User-based P_{err}	Cloud-based P_{err}
0.70	$R_1 = 3.3$, $R_2 = 3.7$, $R_3 = 4.1$	9.7×10^{-3}	3.8×10^{-3}
0.75	$R_1 = 2.8$, $R_2 = 3.1$, $R_3 = 3.5$	1.1×10^{-2}	4.5×10^{-3}
0.80	$R_1 = 2.4$, $R_2 = 2.7$, $R_3 = 2.9$	1.2×10^{-2}	5.2×10^{-3}
0.85	$R_1 = 1.9$, $R_2 = 2.1$, $R_3 = 2.3$	1.4×10^{-2}	5.9×10^{-3}
0.90	$R_1 = 1.4$, $R_2 = 1.5$, $R_3 = 1.7$	1.5×10^{-2}	6.5×10^{-3}

For the adaptation algorithm which does not consider information about the state of the cloud (called user-based adaptation algorithm), authors assume an algorithm that adapts the representation bitrate of the next segment to the download rate of the last segment. This is, the representation selected for segment i, R_i, is calculated as:

$$R^i = \max_n \{R_n | R_n < r^{i-1}\}, \tag{1}$$

where r^{i-1} is the download rate of segment $i-1$.

The authors are conscious that the assumed adaptation algorithm is too simple (compared to commercial ones), but it is enough to compare user- and cloud-based adaptations. This adaptation algorithm is applied to the 10^5 ordered values of segment download rate r^i, obtained in the simulations. In this case, the representation R_1 (simulation results for total load in the server $\rho = 0.7$) will be selected for 472 segments (since 472 segments were downloaded with rate lower than R_2). The representation R_3 will be selected in 502 segments (since 502 segments were downloaded with rate higher than R_3). All the other segments will be requested with representation rate equal to R_2.

For analyzing the adaptation decisions, authors consider that the adaptation decision is erroneous when the selected representation for segment i, R^i, is higher than the download rate of segment i: r^i (resulted from the simulations). This is, the decision is incorrect if $R^i > r^i$. Let us remark that $R^i > r^i$ could cause image freezing in video players with short playback buffer. Note that R^i is a function of r^{i-1}, so the erroneous decision rate is closely related to the correlation of the segment download rate. The probability of erroneous decision for different values of ρ is presented in Table 1 (user-based column).

The second algorithm analyzed is the so-called cloud-based algorithm. It considers the information about the current state of the cloud and (based on historical data) the average conditions of the cloud. In the presented simulations, the average traffic is the same during all the simulations, so the average conditions (load, available bandwidth) of the cloud does not vary. Therefore, we assume that the cloud-based algorithm selects the same representation (R_2) for all the segments: $R^i = R_2$, $i = 1 \ldots 10^5$. Also in this case, we consider that taken decision for segment i was erroneous when $R^i > r^i$.

The probability of error (erroneous decisions divided by total decisions, i.e. 10^5) is presented in Table 1 for each of the five tests (different values of ρ), together with the values of the representation rates. As we can observe, the error for user-based decision is always much higher than in the case of cloud-based decisions since cloud-based decisions are based on the knowledge of the situation of the cloud bottleneck, unlike user-based decisions. For increasing values of ρ, the probability of erroneous decisions raises, which is explained by the higher variability of the state of the queue, which causes higher variability into the download time of the consecutive segments.

The authors are aware that the presented results are very dependent on the assumptions (especially on the assumed adaptation algorithms), but the aim of the simulation-based comparison was only to show that user-based decisions are less reliable since the user's video application does not have information about the bottleneck. A lector could find other algorithms that provide better results, but the conclusion would be the same. The cloud has information about the cloud's bottleneck that the user does not own, and such information may be useful for taking right adaptation decisions.

4. End-to-end Framework for Cloud-aware Adaptation

In order to provide awareness about the state of the cloud into adaptation decisions, authors propose that cloud system performs measurements at the cloud premises (generally, in the cloud access), which will be used for predicting the state of the cloud for the next few seconds (in proposed implementation the state of the cloud is predicted for the next 5 s). The information about potential restrictions of the cloud is then passed to the content server, which is responsible for sending it to the user's DASH application (client DASH application) by using the push function of HTTP/2. The DASH application, on its turn, will request the next segment by considering the state of the user's mobile access network (as measured by the DASH application) and the state of the cloud (as indicated by the information received from the content server). The decisions taken by the client application can be to perform the media bitrate adaptation or the content server adaptation (switching to another cloud for serving the request). Media bitrate adaptation is efficient if the congestion is in the wireless access since the unique possibility is to reduce bitrate for reducing congestion but, in the case of cloud congestion, better results by switching the content server while maintaining the previous media bitrate (saving the quality of the future streaming) can be obtained. Let us remark that both media bitrate and content server adaptation decisions should be in accordance with the original Media Presentation Description (MPD) file managed by the DASH application (different representations for media bitrate adaptation and different BaseURL tags for content server adaptation).

In order to obtain reliable information about cloud congestion, the system performs the next operations:

- The Monitoring Resource Mediator (MRM) (see Fig. 2) collects (in some time windows before a current instant of time) bandwidth information available at uplink of the cloud and processor load in the servers. Bandwidth information about aggregate traffic avoids potential problems of scalability during the measurements. This information is stored in the monitoring database at the Cloud Manager (CM). Let us remark that commercial clouds actually contain MRMs that monitor the state of the links.

- At time t, the Traffic Forecast within the CM (using the collected data read from the monitoring database) provides small-term bandwidth forecast for the next T seconds, i.e. $[t,\ t+T]$, and decides whether the cloud will experience in that time an over-load (congestion) that will be able to impoverish the quality of the media transmission. In that case, the CM informs the content server and the latter sends information to the client DASH application. The system finishes the above steps before time t, so that an over-load alert can be sent to the DASH application at time t for the period $[t,\ t+T]$, after which the above process is repeated for the next period $[t+T,\ t+2T]$. Let us remark that the time T is not synchronized with the segment duration or segment download duration but the two operations (DASH streaming and cloud congestion control) work independently.

Figure 2 shows the entities involved in providing cloud-awareness to the user's terminal.

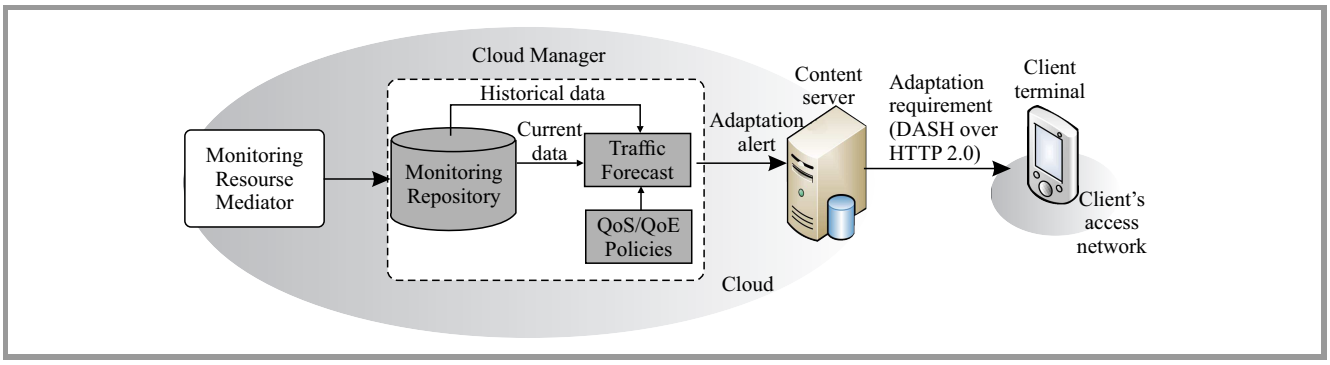


Fig. 2. Framework for cloud-aware adaptation.

The MRM is the monitoring tool that provides information about the traffic in selected points of the cloud (especially in the uplink of the cloud) and about the state of the processors in the servers. Many enterprize clouds, open-source clouds and CDNs provide a multimodal selection of possible monitoring metrics that span beyond the CPU utilization and Bulk Transport Capacity up to more detailed per service or user-defined metrics (e.g. [18]). In this case, the measurements used for predicting over-load of the cloud are the bandwidth of the aggregate (at cloud uplink) and the number of new requests arrived to the cloud during the last 30 s. The commercial Riverbed Stringray Traffic Manager tool [18] provides the requested measurements. Two different kinds of measurements are necessary for the traffic forecast algorithm: the current bandwidth measurements and the historical data (stored into the Monitoring Repository), as shown in Fig. 2.

The Traffic Forecast module has to efficiently predict the needed bandwidth capacity in the cloud based on observed fluctuations of cloud resources and to conclude which situations may lead to cloud congestion. For the implementation of the traffic forecast algorithm, a similar approach as in [19] is proposed but considering each http request (during one video session) as a separate video channel. The algorithm estimates the bandwidth required to the server (b_T) during the next instant of time, T, according to estimated values of active population (N_T) and target download rate which, on mean, each user will require (R), as indicated in Eq. (2).

$$b_T = R \times N_T \,. \tag{2}$$

Generally, the estimation of the population is based on the past measurements of population that are downloading content (active population) during a long period of time. These population time series are processed by using different mathematical techniques (e.g. Box-Jenkins) in order to eliminate periodicity and trends related to specific periods of time (e.g. daily periods) [19]. In this way, the data can be used independently of the moment when they were taken. The output of these operations can be characterized by autoregressive moving-average (ARMA) model. The characteristics of the past active population define the population during the next instant of time.

In presented implementation the active population that will download a content (N_T) to the server as the average of the measurements of last 30 s (which is the reference value of playback buffer for many video players) is calculated, and it is assumed that during this short period there is no periodicity trend that could have negative influence into the prediction. Note that the estimations presented are very sensitive to error since they use short-term measurements, but let us remember that the implemented system is a proof of concept. Its deployment in commercial networks should consider more sophisticated (by using long-term processed measurements) forecast algorithms.

The target download rate R is calculated as follows. Let us assume that the server has c different contents with the same popularity. Each content c contains i different representations with rate equal to R_{ic}. Then the average of the target download rate is:

$$R = \frac{1}{C} \times \sum_{c=1}^{C} \frac{\sum_{i=1}^{I_c} R_{ic}}{I_c} \,. \tag{3}$$

If the predicted bandwidth goes beyond a given threshold (90% of the uplink bandwidth of the server in author's implementation), then the traffic forecast algorithm predicts cloud congestion in the next time slot. In this case, the cloud manager contacts the content servers in order to inform about the situation. The interface implementation between CM and content server is based on JSON/RPC protocol. The content server contacts the users' terminals, which the server is actually serving in order to inform about the congestion situation. The server could decide to send such information only to a number of clients (e.g. one of five clients) in order to avoid avalanche situations (all the clients switch to another cloud). The communication between server and clients is based on the push functionality of HTTP/2. Details of such a communication are given in the Section 5.

At last, the client DASH application is responsible for taking the final decision about adaptation. Such a decision may include:

- media adaptation, i.e. switching media stream to lower representation without changing content server, as it is illustrated in Fig. 3,

- content server adaptation, i.e. downloading further segments from another content server (from the set of available servers specified in *MultiBaseURL* element into MPD) leaving representation unchanged.

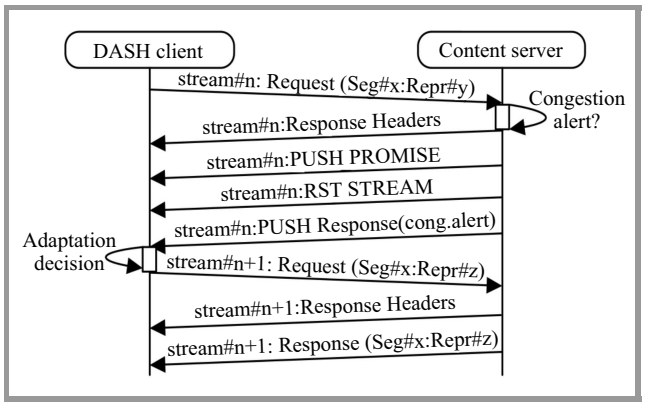


Fig. 3. Sequence diagram for push-based congestion alert.

For this, the DASH application should implement an enhanced adaptation algorithm that may adapt not only Media bitrate but also content server (e.g. to select another cloud for streaming the content) considering the information gathered by the same application and the information arrived from the content server. An example algorithm fulfilling such functionalities was presented in [20]. The selection of the server is, generally, a blind decision in the DASH application. It is responsibility of the service provider to inform the application about the characteristics of the new servers, but this is out of the scope of this paper. In case when DASH application switches the content server (content server adaptation), it should avoid to come back to previously used server, from which the client downloaded a previous segment, in order to avoid ping pong effect of endless switching between two overloaded clouds.
The scalability of the solution is ensured due to the fact that the unique interchanged information is about the cloud congestion, which is not specific per video session. cloud congestion can be predicted only by taking measurements of the aggregate traffic and marked flows, which saves most of the potential scalability issues in the monitoring tool. Moreover, the network measurements are taken in specific nodes, which are the bottlenecks of the cloud system (e.g. uplink). These points are well-known to the cloud provider and are constantly controlled by traffic manager tools.

5. Implementation Details of DASH over HTTP/2 Module

HTTP/2, which is still under development within the IETF HTTPbis Working Group, is a binary protocol that aims at better utilization of network capacity than previous versions while preserving compatibility with the transaction semantics of HTTP 1.1. HTTP/2 introduces a framing layer between HTTP and TCP used for multiplexing several HTTP requests into one TCP connection. HTTP/2 provides efficient header compression in order to reduce the protocol overhead [21] and also proposes server push mechanism, which allows a server to send a response without an explicit request from the client. In [22], the authors employ the server push to decrease media delivery latency in DASH live video streaming. In presented approach, authors apply the push feature to transfer information about cloud congestion from server to client without the need for client's request.
Figure 3 presents the communication between DASH client and content server assuming that the server received a congestion alert from the CM. The client DASH application downloads media segments in separate streams. The content server, after receiving a request for a segment, checks if during the time, which elapsed from the previous request, the CM signaled a congestion. If no, the server returns requested segment. If yes (what is depicted in Fig. 3), the server responses to the request with Headers frame and next sends Push_Promise frame to notify that the server intends to initiate new stream for "pushed" data. Then, the server sends RST_Stream header to reset current stream, followed by Push Response, which carries information about congestion. Canceling the current stream allows for faster reaction to congestion alert since the client does not need to wait with adaptation process until the whole required segment (Seg#x:Repr#y in Fig. 3) will be downloaded. Such process delays downloading of the current segment by Round Trip Delay required for transferring, one by one, Push_Promise, RST_Stream, Push Response frames and a new request for the segment. On the other hand, the client may ask for the new adapted segment (Seg#x:Repr#z in Fig. 3) just after receiving Push Response. Seg#x:Repr#z is the result of adaptation decisions after receiving cloud congestion alert (media bitrate adaptation or content server switching).
Let us remark that server push action can be executed only when server receives a segment request from the client (as the "supplementary response" to this request) since in HTTP client-server scheme an exchange of messages is initiated solely by the client.

5.1. DASH HTTP/2 Client Implementation

In the above-mentioned papers [21] and [22], the authors used implementations of SPDY protocol to develop HTTP/2-compilant DASH applications. Although HTTP/2 and SPDY have equivalent functionalities (Google's SPDYv2 protocol was chosen as the basis for HTTP/2), they are incompatible due to, for example, different header compression mechanisms (GZIP in SPDY, whereas HTTP/2 uses dedicated HPACK scheme). Therefore, authors have implemented own version of HTTP/2 DASH client using *nghttp2* library [23], which is compliant with IETF HTTP 2.0 Draft v13 [24].
For this purpose, the SPDY-based QTSamplePlayer, an open-source DASH application provided by Bitmovin [7] (which bases on *libdash3* library) has been extended, by implementing new class HTTP2Connection responsible for

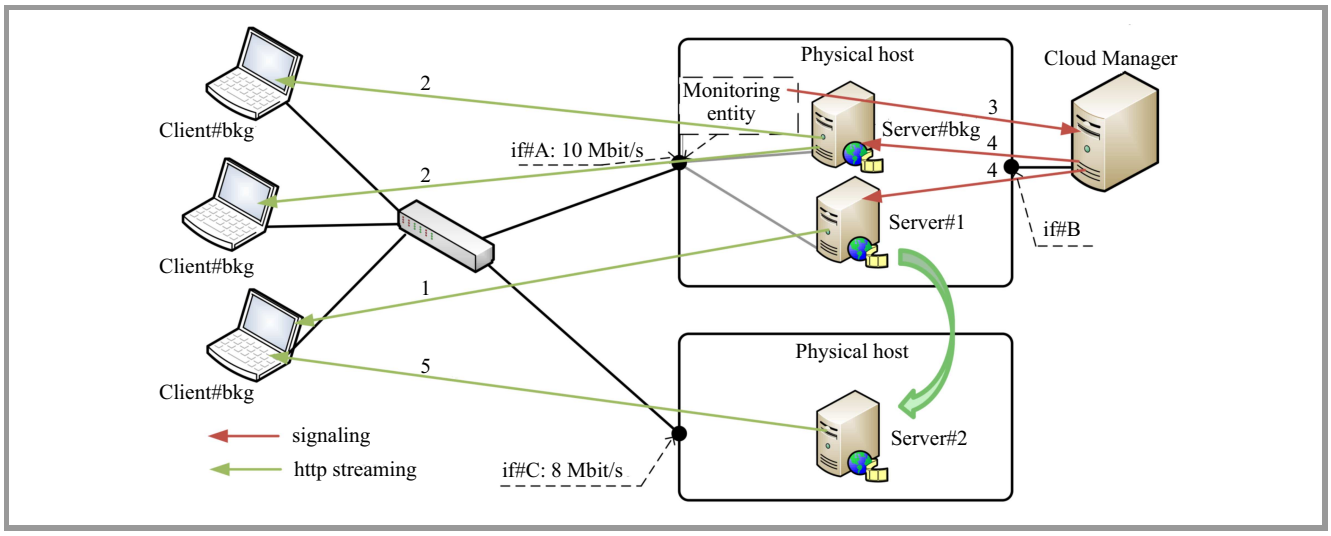


Fig. 4. Experimental setup for evaluation tests.

establishing and handling HTTP/2 connections. Moreover, we modified *DASHReceiver* module of the QTSample-Player to include *adaptationAlert()* method for interrupting the current segment downloading process whenever a server push occurs.

5.2. Content Server Implementation

The HTTP/2 content server was implemented based on *nghttpd* server implementation provided by *nghttp2* [23]. The authors deployed *AlertModule*, which contains JSON-RPC server for receiving congestion alerts from the CM. The *havePushData()* method of the *AlertModule* is called by main thread of content server whenever a new request for media segment arrives (which means opening a new stream). This method checks if a congestion alert from the CM exists and, in positive case, it records the label of congestion alert in order not to propagate the same alert in the future and returns data which should be pushed to the client. When the response of the *havePushData()* is positive, the server triggers push procedure and, at the same time, cancels the stream related with media segment request. Let us remark that additional functionality results in a very low overhead comparing to standard HTTP/2 server. This overhead is related with receiving a JSON message (of small size) from CM and performing one extra step in client's request handling flow to check if there is a message to be pushed.

Source code for both implementations (client and server), are available on web page http://www.nit.eu/offer/research-projects-products/http2dash.

6. Test Results

The experimental setup, presented in Fig. 4, includes two physical hosts with three virtual machines (labeled as server#bkg, server#1 and server#2) containing our aforementioned implementation of content streaming server. Each physical host emulates one separate cloud domain characterized by own IP prefix.

The servers in the first cloud are connected through a link that is constantly monitored (monitoring entity is located on physical host output interface if#A, see Fig. 4). The available bandwidth in the output interfaces if#A and if#C was restricted to 10 Mb/s and 8 Mb/s, respectively, by using the Linux Traffic Control system (*tc* command). The servers provide media content (Big Buck Bunny movie [25]) with 15 different representations, from 100 Kb/s up to 6 Mb/s, divided into segments of two seconds duration.

The client applications, client#1 and two client#bkg, run under Linux Ubuntu 14.04. The adaptation algorithm in the DASH client applications is based on mean download rate but it has been modified in order to switch the content server whenever congestion information arrived from the server (see [20]). Just after connecting with the server, the clients increased its HTTP/2 flow control window from default value equal to 64 KB up to 1048 KB using a Settings frame. In this way, streaming stop is avoid due to exhaustion of client's window space, and also we limit the number of Windows_Update frames generated by the clients, which indicate how many bytes the server is permitted to transmit. The server#1 starts streaming the content to client#1 (arrow no. 1 in Fig. 4) with the highest representation. At second 60, both client#bgk start downloading the same content from server#bgk located at the same cloud domain as server#1 (arrows no. 2 in Fig. 4), so the uplink of the cloud becomes overloaded. The monitoring information of if#A arrives to the CM (arrow no. 3 in Fig. 4), which determines that there is bandwidth congestion since the occupancy of the interface is higher than 0.9 (simple prediction algorithm created for testing purposes). Then, the CM sends an alert about cloud congestion to the content servers: server#1 and server#bgk (arrow no. 4 in Fig. 4). The next request for media segment arrived to server#1 and server#bgk are used to perform server push to the clients. The DASH adaptation

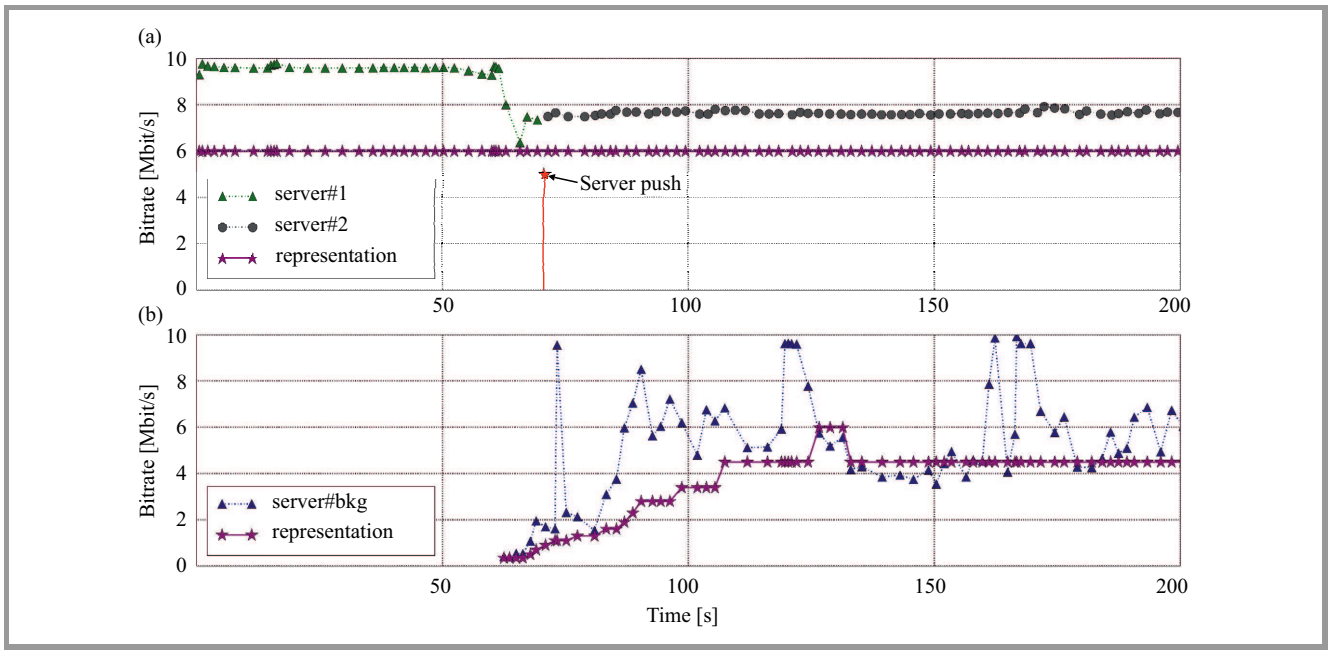


Fig. 5. Segment download rate with cloud-awareness.

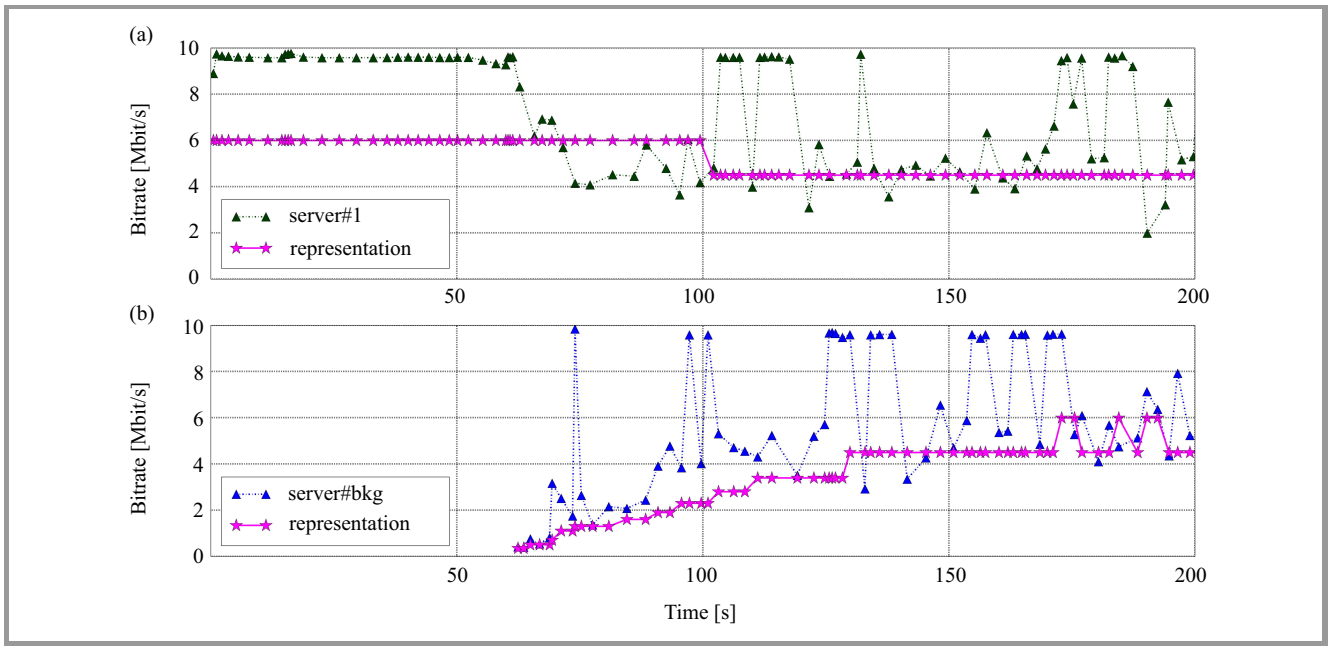


Fig. 6. Segment download rate without cloud-awareness.

mechanism at the client#1 terminal receives information about cloud congestion and performs content server adaptation by switching the streaming to server#2, as it is signaled by arrow no. 5 in Fig. 4. However, for demonstration purposes both client#bkg applications received MPDs without *MultiBaseURL* option, therefore they performed media adaptation only (without content server adaptation).

Figure 5a shows download rate of each segment received by client#1 from server#1 and server#2, as well as the representation rate selected by the client for each segment. The moment when server push occurred is indicated by a red vertical line in Fig. 5a (pointed also by an arrow).

Figure 5b shows the segment download rate and selected representation for one of client#bkg.

As may be observed in the results presented in Fig. 5a, thanks to the cloud congestion notification, client#1 could switch the content server while maintaining the same (highest) representation during the whole downloading process. Moreover, thanks to prediction algorithm, the adaptation algorithm is able to react fast to the congestion situation (only 3 segments from the moment when server#bkg started streaming to client#bkg). Both client#bkg compete for the if#A bandwidth, so they download the content with lower representation (Fig. 5b).

Figure 6 shows the same scenario but, in this case, the cloud congestion information is not sent to the server, so DASH application of client#1 performs media bitrate adaptation with a delay of 14 segments from the moment when server#bkg started streaming (client#1 had to wait for collecting enough measurement data to discover bandwidth decrease). This delay may result in image freezing during video playout, if only the client buffer size is not enough to compensate degradation of downloading conditions.
By comparing Figs. 5 and 6, it may be concluded that cloud-awareness improves the performance of the system enhancing QoE due to the predictive feature that allows to fast reaction from the DASH application. Moreover, content server adaptation allows maintaining higher media bitrate by switching the transmission to another (non-overloaded) cloud.

7. Conclusions

The system presented in this paper allows for communication between Cloud Manager and video player in the end user's terminal by means of the content server. Such a communication is used for notifying situations of cloud congestion foreseen for close time. This way, the client video application may take proper decisions about adaptation taking into account both bandwidth limitations in the mobile access and congestion situations in the cloud. The results obtained by means of the system implemented on DASH-capable video player and DASH-capable server, present the applicability of the proposed system in situations of congestion in the cloud and compare the same situation when the DASH application does not own information about congestion. In the latter case, the application might not adapt bitrate in time, which would cause frozen image (in the case of strong degradation in the cloud). Moreover, making a distinction between cloud congestion and mobile access congestion allows for dual adaptation (media bitrate and content server), which may improve the quality of the media event experience.
Two issues will be addressed in planned future work: the use of bidirectional feature of HTTP/2 communication for sending dynamic MPD from the server to the client's DASH application and new dual adaptation algorithms in DASH application that integrate together rate measurements at the video player and information about congestion arrived from the Cloud Manager.

Acknowledgements

This work is supported by the European research project DISEDAN (http://wp2.tele.pw.edu.pl/disedan/) under the CHIST-ERA framework program.

References

[1] T. Jursonovics and S. Imre, "Quality-based charging solutions for wireless multimedia services", *Int. J. Netw. Manag.*, vol. 24, no. 5, pp. 357–401, 2014.

[2] M. Mehta, I. Ajmera, and R. Jondhale, "Mobile cloud computing", *Int. J. Elec. Commun. Engin. & Technol.*, vol. 4, no. 5, pp. 152–160, 2013.

[3] S. Qureshi *et al.*, "Mobile cloud computing as future for mobile applications – Implementation methods and challenging issues", in *Proc. IEEE Int. Conf. Cloud Comput. & Intell. Syst. CCIS 2011*, Beijing, China, 2011, doi: 10.1109/CCIS.2011.6045111.

[4] N. Fernando, W. L. Seng, and W. Rahayu, "Mobile cloud computing: A survey", *Future Gener. Comp. Sys.*, vol. 29, no. 1, pp. 84–106, 2013.

[5] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: Architecture, applications, and approaches", *Wirel. Commun. & Mob. Comput.*, vol. 13, no. 18, pp. 1587–1611, 2013.

[6] J. Famaey and F. De Turck, "Federated management of the Future Internet: Status and challenges", *Int. J. Netw. Manag.*, vol. 22, no. 6, pp. 508–528, 2012.

[7] GitHub repository for bitmovin libdash library [Online]. Available: https://github.com/bitmovin/libdash/tree/http2 (last access: Aug. 2015).

[8] C. Dovrolis, M. Jain, and R. Prasad, "Measurement tools for the capacity and load of Internet paths" [Online]. Available: http://www.cc.gatech.edu/fac/Constantinos.Dovrolis/bw-est/ (last access: Aug. 2015).

[9] S. Akhshabi, A. C. Begen, and C. Dovrolis, "An experimental evaluation of rate-adaptation algorithms in adaptive streaming over http", in *Proc. 2nd Ann. ACM Conf. Multimed. Syst. MMSys 2011*, San Jose, CA, USA, 2011, doi: 10.1145/1943552.1943574.

[10] S. Akhshabi, L. Anantakrishnan, C. Dovrolis, and A. C. Begen, "What happens when http adaptive streaming players compete for bandwidth?, in *Proc. 22nd Int. Worksh. Netw. Oper. Sys. Supp. for Digit. Audio Video NOSSDAV'12*, Toronto, Ontario, Canada, 2012, doi: 10.1145/2229087.2229092.

[11] F. Dobrian *et al.*, "Understanding the impact of video quality on user engagement", in *Proc. ACM SIGCOMM 2011 Conf.*, Toronto, ON, Canada, 2011, doi: 10.1145/2043164.2018478.

[12] W. C. B. Seo and R. Zimmermann, "Efficient video uploading from mobile devices in support of http streaming", in *Proc. 3rd Ann. ACM Conf. Multimed. Syst. MMSys 2012*, Chapel Hill, NC, USA, 2012, doi: 10.1145/2155555.2155589.

[13] M. Mirza, J. Sommers, P. Barford, and X. Zhu, "A machine learning approach to tcp throughput prediction", *IEEE/ACM Trans. Netw.*, vol. 18, no. 4, 2010, doi: 10.1109/TNET.2009.2037812.

[14] Q. He, C. Dovrolis, and M. Ammar, "On the predictability of large transfer tcp throughput", *ACM SIGCOMM Comp. Commun. Rev.*, vol. 35, no. 4, pp. 145–156, 2005.

[15] X. Liu *et al.*, "A case for a coordinated internet video control plane", in *Proc. ACM SIGCOMM 2012 Conf.*, Helsinki, Finland, 2012, doi: 10.1145/2342356.2342431.

[16] R. Rejaie and J. Kangasharju, "Mocha: A quality adaptive multimedia proxy cache for internet streaming", in *Proc. 21st Int. Worksh. Netw. Operat. Syst. Support for Digit. Audio and Video NOSSDAV'11*, Vancouver, BC, Canada, 2011, doi: 10.1145/378344.378345.

[17] I. Sodagar, "The MPEG-DASH standard for multimedia streaming over the Internet", *IEEE MultiMedia*, vol. 18, no. 4, pp. 62–67, 2011.

[18] Riverbed Stingray Traffic Manager [Online]. Available: http://www.riverbed.com/

[19] D. Niu, Z. Liu, B. Li, and S. Zhao, "Demand forecast and performance prediction in peer-assisted on-demand streaming systems", in *Proc. 30th IEEE Int. Conf. Comp. Commun. IEEE INFOCOM'11*, Shanghai, China, 2011, doi: 10.1109/INFCOM.2011.5935196.

[20] J. Mongay Batalla and S. Janikowski, "In-segment content server adaptation for dual adaptation mechanism in DASH", in *Proc. 5th Int. Conf. Comput. Intell., Commun. Syst. Netw. IEEE CICSyN 2013*, Madrid, Spain, 2013, doi: 10.1109/CICSYN.2013.16.

[21] C. Mueller, S. Lederer, C. Timmerer, and H. Hellwagner, "Dynamic adaptive streaming over HTTP/2.0", in *Proc. IEEE Int. Conf. Multim. & Expo ICME 2013*, San Jose, CA, USA, 2013, doi: 10.1109/ICME.2013.6607498.

[22] W. Sheng and V. Swaminathan, "Low latency live video streaming over HTTP 2.0", in *Proc. 24st Int. Worksh. Netw. Operat. Syst. Support for Digit. Audio and Video NOSSDAV'14*, Singapore, 2014, doi: 10.1145/2578260.2578277.

[23] nghttp2 – HTTP/2 C Library. Project webpage [Online]. Available: http://nghttp2.org (last access: Aug. 2015).

[24] M. Belshe *et al*., "Hypertext Transfer Protocol version 2", IETF HTTPbis Working Group Internet-Draft, 2014 [Online]. Available: http://tools.ietf.org/html/draft-ietf-httpbis-http2-13

[25] S. Lederer, C. Müller, and C. Timmerer, "Dynamic adaptive streaming over HTTP Dataset", in *Proc. 3rd Ann. ACM Conf. Multimed. Syst. MMSys 2012*, Chapel Hill, NC, USA, 2012, doi: 10.1145/2155555.2155570.

Jordi Mongay Batalla received his M.Sc. degree from Universitat Politecnica de Valencia (Spain) in 2000 and Ph.D. degree from Warsaw University of Technology (WUT) in 2009. His work experience includes jobs in Centro Nazionale di Astrofisica in Bologna, Italy as well as Telcordia Poland. Currently, he is with National Institute of Telecommunications as Head of Internet Architectures and Applications Department. He has also an Associate Professor position at WUT. His research interest focus mainly on quality of service in both IPv4 and IPv6 infrastructures, Future Internet architectures, as well as applications for Future Internet (Internet of Things, Smart Cities, IPTV).
E-mail: jordim@tele.pw.edu.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Piotr Krawiec received his M.Sc. and Ph.D. degrees in Telecommunications from Warsaw University of Technology, in 2005 and 2011, respectively. Since 2012 he is an Assistant Professor at the Department of Internet Architectures and Applications, National Institute of Telecommunications, and Institute of Telecommunications, Warsaw University of Technology. His research areas include IP networks (fixed and wireless), Future Internet architectures and applications, prototyping and testbeds.
E-mail: P.Krawiec@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Daniel Negru received his Ph.D. from the University of Versailles Saint Quentin en Yvelines in 2006 in the field of Broadcast and Internet convergence solutions at the network and service levels. In 2007, he became Associate Professor at ENSEIRB School of Engineers/University of Bordeaux, specializing in multimedia and networking. From 2010 to 2014, he has been coordinating the ICT FP7 ALICANTE IP project that tackles networking and multimedia research fields. He has participated to more than 10 collaborative research projects at the national or European level, published more than 50 papers, such IEEE Communication Magazine, IEEE Multimedia, Globecom, ISCC, FIA. In 2013, he received his Habilitation à Diriger des Recherches (HDR) from the University of Bordeaux.
E-mail: daniel.negru@labri.fr
CNRS-LaBRI, University of Bordeaux
351 cours de la Libération
33 405 Talence CEDEX, France

Joachim Bruneau-Queyreix received his M.Sc. in Telecommunications at ENSEIRB-MATMECA graduate school of engineering, Bordeaux, in 2014. Since 2014, he is pursing his Ph.D. at LaBRI/Bordeaux Computer Science Laboratory in the field of multi-criteria optimization for content delivery within the Future Media Internet. His research areas include video codecs, streaming protocols, Future Internet streaming systems and architectures as well as multimedia streaming application prototyping.
E-mail: jbruneau@labri.fr
CNRS-LaBRI, University of Bordeaux
351 cours de la Libération
33 405 Talence CEDEX, France

Eugen Borcoci, Ph.D, is full professor at University "Politehnica" of Bucharest (UPB), Electronics, Telecommunications and Information Technology Faculty. His expertise has been oriented to specific domains of telecommunications and computer networks architectures, technologies and services. Recently, his research

interest and activities are on new architectural approaches: Future Internet, SDN/NFV, Content Aware/Centric Networking. He has published 5 books, 4 textbooks and over 130 scientific or technical papers and scientific reports. He has been UPB team leader in several European research projects. He is member of several International Conferences Committees and member of the Technical Sciences Academy of Romania.

E-mail: eugen.borcoci@elcom.pub.ro
University Politehnica Bucharest
Splaiul Indeependenei
Bucharest 5, 060042 Romania

Andrzej Bęben received his M.Sc. and Ph.D. degrees in Telecommunications from Warsaw University of Technology (WUT), Poland, in 1998 and 2001, respectively. Since 2001 he has been Assistant Professor at WUT, where he is a member of the Internet Architectures and Applications research group. His research areas cover Future Internet, IP networks, information centric networks, network virtualisation, traffic engineering, multi-criteria decision theory, simulation techniques, measurement methods, and testbeds.

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

Piotr Wiśniewski is a Ph.D. student at the Institute of Telecommunications at the Warsaw University of Technology, where he received his M.Sc. (2010) and B.Sc. (2009) degrees in Telecommunications. He is a specialist at the National Institute of Telecommunications and a Research and Teaching Assistant at the Warsaw University of Technology. His research interests include quality of service, Information Centric Networks, media streaming solutions, Future Internet architectures and applications.

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

Paper

Analysis of Burst Ratio in Concatenated Channels

Jakub Rachwalski and Zdzisław Papir

AGH University of Science and Technology, Krakow, Poland

Abstract—Burst ratio is a parameter that quantifies packet loss patterns in transmission networks. It has been defined for an end-to-end scenario, therefore burst ratio can be determined only if the characteristics of the whole transmission path are known. In this paper, the burst ratio parameter applicability to cases when the transmission path consists of a series of transmission channels with known packet loss rate and burst ratio values is extended. The paper also presents the results of simulations performed with NS2 software, demonstrating the validity of the burst ratio analysis. Consequently, the research makes it possible to determine the value of the burst ratio parameter in concatenated packet networks, which in turn supports delivering higher quality VoIP services.

Keywords—bursty packet loss, E-model, quality of experience, voice over IP.

1. Introduction

Voice over Internet Protocol (VoIP) applications play a crucial role in connecting people and businesses around the world. It is a huge business for hardware manufacturers, network operators and service providers. In order to assure end customer satisfaction, the transmission networks must be designed well, and the quality of the provided VoIP service must be constantly monitored and maintained. In order to achieve this, all factors that affect the application quality of experience (QoE) [1] must be recognized.

The quality of VoIP carried over packet networks is influenced by multiple factors [2]. They include user-dependent aspects (e.g. user expectations), terminal quality (e.g. microphone sensitivity) and application settings (e.g. audio codec). The quality is also affected by transmission network-dependent factors, which include throughput, round-trip time and packet loss. To some extent, they can be controlled by network design and maintenance.

One of the transmission network-dependent factors that influences the perceived quality of VoIP transmissions is the burst ratio parameter [3]. It quantifies the packet loss pattern by describing the extent to which the packets were lost in bursts. The burstiness of packet loss affects the perceived media quality. If the number of audio packets lost sequentially is low enough not to be noticed by the human cognitive system, or it can be concealed by the packet loss concealment (PLC) technique [4], then the event has no impact on the perceived quality. In contrast, long sequences of lost packets can be easily perceived as an annoying quality deterioration. Therefore, the burstiness (burst ratio) of packet loss can be correlated with the perceived quality of VoIP service [5].

In order to provide a VoIP service of the best possible quality, the burst ratio parameter needs to be well recognized and analyzed. Thus far, it has only been defined for end-to-end transmission scenarios. In this case, in order to calculate the burst ratio of a transmission, the characteristics of the complete, end-to-end transmission path must be measured. This article describes the research into defining the end-to-end value of the burst ratio parameter, when the transmission is carried over multiple concatenated transmission channels and only the characteristics of each individual intermediate channels are determined.

Although extensive research on the influence of bursty packet loss on the QoE of VoIP has been carried out [6], [7], the authors are the first to analyze burst ratio in concatenated channels. In work [8], the results of theoretical studies are presented in which the formula for burst ratio in the concatenated scenario is derived. This article presents results of NS2 simulations [9] performed in order to validate the equations in a real environment. The results demonstrate the validity of the aforementioned theoretical considerations.

The results help control the burst ratio parameter by describing the impact of individual transmission channels on the burst ratio of the complete transmission path. The results will improve the quality and reliability of VoIP applications, thus improving end user satisfaction.

The remainder of this paper is structured as follows. In Section 2 the burst ratio parameter is presented and described in detail. In Section 3 we describe the methodology and features of the simulations that were carried out to validate the theoretical studies. Section 4 presents the results of the validation of the equation for Burst Ratio in concatenated channels. In Section 5 the verification of the simplified form of the equation is presented. Potential applications of the results are presented in Section 6. Finally, the conclusions are given in Section 7.

2. Burst Ratio Overview

This section presents the definition and application of burst ratio. It also contains results of our previous studies in the field of extending the burst ratio parameter applicability to multi-channel scenarios.

In order to describe packet loss of a communication channel, the packet loss rate Ppl is used. It indicates the probability of losing a packet during transmission over the channel. However, it is not a complete channel description as it does not capture packet loss patterns. Under the same packet loss rate, the loss can be evenly distributed over the whole transmission, or take place in bursts if multiple consecutive packets are lost.

The parameter that describes the packet loss pattern is burst ratio (denoted as $BurstR$). It is defined in [3] as the average length of observed bursts in a packet arrival sequence (average burst length) normalized over the length of burst expected for purely random loss (μ):

$$BurstR = \frac{\text{Average measured burst length}}{\mu}. \qquad (1)$$

Burst ratio describes the packet loss pattern by expressing how much longer or shorter the measured bursts were than in the hypothetical case when all the packets were lost randomly under the same packet loss rate. Therefore, the burst ratio quantifies the observed packet loss as:

- bursty if $BurstR > 1$,
- random if $BurstR = 1$,
- scattered if $BurstR < 1$.

The length of packet loss burst expected for purely random loss (μ) is given as [10]:

$$\mu = \frac{1}{1 - Ppl}, \qquad (2)$$

where Ppl stands for the probability of packet loss. The formula shows that even for purely random loss the observed burst length increases with higher packet loss, in the multiplicative inverse way. This is why the $BurstR$ value can differ dramatically for the same observed packet loss burst length, depending on the packet loss rate μ.

Generally speaking, for the same packet loss rate, higher values of burst ratio indicate that the packets are being lost in series. Conversely, lower values of the parameter mean that the packet loss was distributed more evenly over the transmission.

It is common to model packet loss in digital transmission channels with time-discrete state models, Markov chains [11], [12]. The approaches include two-state Markov chain, Gilbert or Gilbert-Elliot models. When examining the lossy transmission, authors are focusing on two-state Markov chain due to its simplicity and flexibility. In two-state Markov chain the successful transmission of a packet over a channel and losing a packet are marked with two different transmission channel states (Markov chain states). An example of the chain is shown in Fig. 1. In this case, if the channel successfully transmits a packet, it is in the F (found) state. If the packet is lost, the channel is in the L (lost) state. At any given time, the channel can only be in one of these two states.

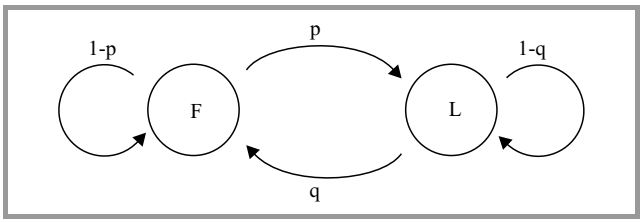


Fig. 1. In two-state Markov loss model F and L represent the found and lost states of a channel, while p and q describe the probabilities of switching the F and L states.

The two-state Markov chain is described with two parameters: p and q probabilities. The probability of losing a packet if the previous packet was successfully transmitted (transition from F to L) is described by p. Similarly, q defines the probability of successfully transmitting a packet if the previous one was lost (transition from L to F). Consequently, probability $1-p$ describes the probability of losing packets in series.

In two-state Markov chains a packet may be lost if the previous packet was successfully transmitted (with probability p) or if the previous packet was lost (with probability $1-q$). Therefore, for two-state Markov chains the probability of losing a packet is determined as:

$$Ppl = \frac{p}{p+q}. \qquad (3)$$

For random loss, $q = 1-p$, the probability of losing a packet is equal to p:

$$Ppl = p. \qquad (4)$$

A transmission channel modeled with the two-state Markov chain exhibits the burst ratio following the formula [13]:

$$BurstR = \frac{1}{p+q}. \qquad (5)$$

Burst ratio is used in E-model [13], a commonly used analytical method of voice quality assessment. E-model uses numerous transmission parameters in order to calculate the transmission ratio factor R, which can then be used to obtain an estimated mean opinion score for the conversational scenario.

Figure 2 presents how the estimated mean opinion score value changes when the burst ratio parameter value varies between 1 and 4. The figure was created with an assumption that the G.711 codec without packet loss concealment (PLC) was used, a 1% packet loss rate was observed and other E-model parameters were used at their default values [14].

Figure 2 shows that there is a clear correlation between the application quality and the burst ratio value. Therefore, in order to calculate estimated mean opinion scores using the E-model, the burst ratio parameter must be accurately determined.

Originally, burst ratio was defined only for scenarios where the transmission is monitored and analyzed end-to-end. In [8] authors studied the burst ratio in a situation

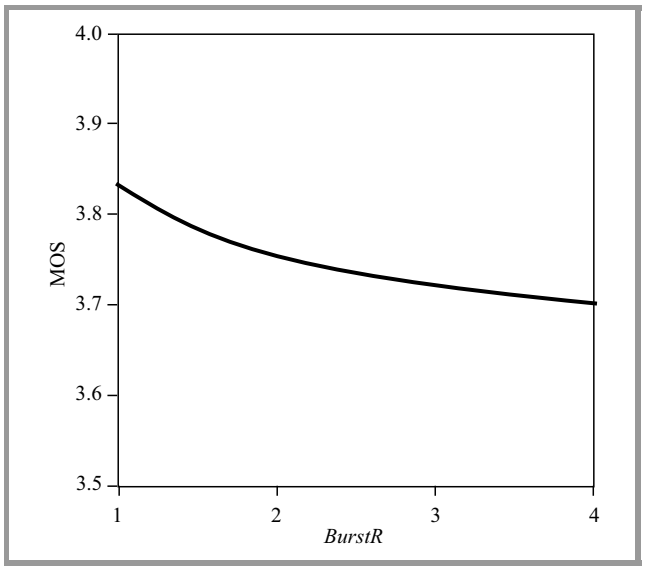


Fig. 2. Based on the E-model relationship between the estimated mean opinion score (MOS) and burst ratio parameter (*BurstR*) for 1% packet loss and the G.711 codec without PLC.

where the transmission path consists of a series of channels, and each is monitored separately. In this case, the burst ratio of the complete path must be calculated using the measured characteristics of separate channels, as presented in Fig. 3.

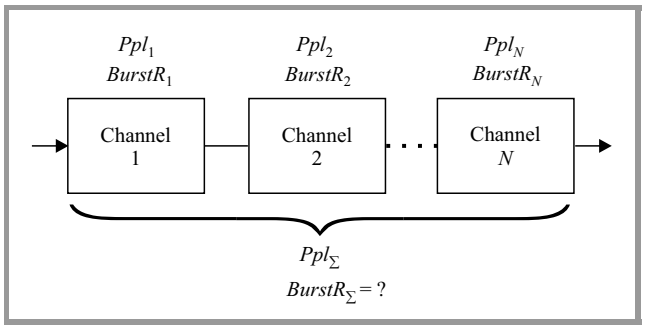


Fig. 3. The problem of burst ratio in concatenated channels network.

It was shown in [8] that if each channel can be modeled with a two-state Markov chain, the burst ratio of the complete transmission path consisting of N channels is described by the formula:

$$BurstR_\Sigma = \frac{1 - \prod_{n=1}^{N} (1 - Ppl_n)}{1 - \prod_{n=1}^{N} \left(1 - \frac{Ppl_n}{BurstR_n}\right)}, \tag{6}$$

where Ppl_n and $BurstR_n$ are the parameters of the n-th channel.

The exact value of the burst ratio can be determined with the regular burst ratio equation. However, for channels characterized by low packet loss the following formula can be assumed:

$$\prod_{n=1}^{N} Ppl_n = 0\,. \tag{7}$$

In this case the packet loss of multiple concatenated channels is as follows:

$$Ppl_\Sigma = \sum_{n=1}^{N} Ppl_n\,. \tag{8}$$

Based on this assumption, the burst ratio value of concatenated channels can be presented with the following, simpler equation.

$$BurstR'_\Sigma = \frac{\sum_{n=1}^{N} Ppl_n}{\sum_{n=1}^{N} \frac{Ppl_n}{BurstR_n}} \tag{9}$$

Analysis performed in [8] shows that this simplification is a reliable approximation of Eq. (6). The error introduced by the simplification depends on the characteristics of each channel and increases with increasing packet loss rate and burst ratio values.

As the assumption of modeling the channels with two-state Markov chains is a simplification, the authors verified the formula in a simulated network using Network Simulator 2 (NS2). The results of this verification are shown below.

3. Simulation Environment

In this section the methodology used to verify the accuracy of Eqs. (6) and (9) is described. The verification has been performed by running extensive simulations in NS2 [9].

The fundamental part of the simulation environment was designed during a seminar in Telekom Innovation Laboratories [15], which is a recognized research and develop-

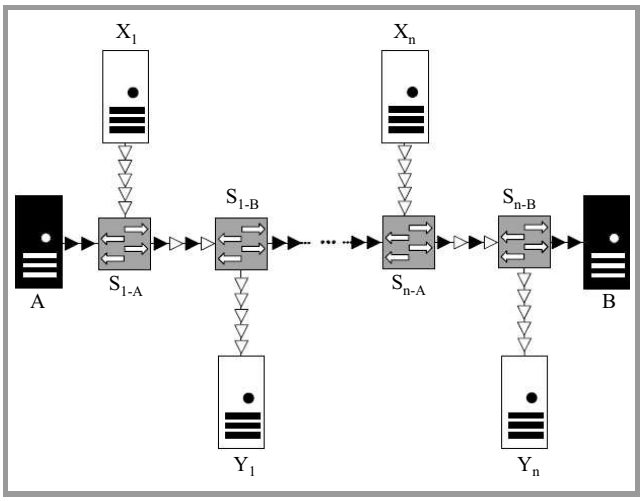


Fig. 4. The generic topology used in the simulations.

Table 1
Simulations parameters

Object	Parameter	Value	Comment
VoIP traffic	Transport protocol	UDP	
	Traffic generator	CBR	
	Packet size	50–1500 bytes	Value selected randomly (uniform distribution)
	Inter-packet interval	0.002–0.06 s	
	Start time delay	0.5–1 s	
Backgroud traffic	Number of streams transported by a single switch	1–10	Value selected randomly (uniform distribution)
	Transport protocol of a stream	TCP, UDP	
	TCP packet size	1000 bytes	
	TCP window size	2–20	
	TCP congestion control algorithm	Tahoe	
	TCP application	FTP	
	UDP traffic generator	Pareto	
	UDP Pareto shape parameter	1.4	
	UDP Pareto burst time	50–5000 ms	Value selected randomly (uniform distribution)
	UDP Pareto idle time	30000–375000 ms	
	UDP Pareto sending rate in burst	400–700 kb/s	
	UDP Pareto packet size	50–1500 bytes	
	Start time delay for each stream	0.5–1 s	
Switches	Number of intermediate switches	2–10	Each simulation repeated for every value
	Queuing scheme of each switch	DropTail, RED, FQ, SFQ	Value selected randomly (uniform distribution)
	Buffer size of each switch	2–20 packets	
Links	Capacity	500–1000 kb/s	Value selected randomly (uniform distribution)
	Propagation delay	0–200 ms	
Simulation	Duration	10, 100, 1000 s	Each simulation repeated for every value

ment institute in the field of quality of audio and multimedia applications.

NS2 is a commonly used [16] simulation environment for testing and studying communication protocols and networks. It can be used to simulate TCP/IP protocol stacks, traffic sources of various distributions and packet queuing and dropping mechanisms.

The release NS2 2.35 was used in this research in order to simulate packet transmission over a series of switches and to analyze packet loss. Each switch serves a number of packet streams and drops packets in case of a buffer overflow. After each simulation the burst ratio calculated at the end of the transmission path using Eq. (1) is compared with the burst ratio value calculated from the transmission parameters of each intermediate switch using Eq. (6). The calculations are performed by analyzing the NAM trace files generated by each NS2 simulation.

The topology used in the simulations is a path presented in Fig. 4. It contains two endpoints (A and B) responsible for a VoIP transmission, n pairs of background traffic servers $(X_1, Y_1, \ldots, X_n, Y_n)$ and n pairs of switches $(S_{1-A}, S_{1-B}, \ldots, S_{n-A}, S_{n-B})$. VoIP traffic, marked with black arrows, is sent from server A to server B. n background traffic streams, marked with white arrows, are sent between servers X_1 and $Y_1, \ldots, X_n$ and Y_n. VoIP traffic and background traffic compete for resources of shared links, which are built up by pairs of switches $S_{1-A} \longleftrightarrow S_{1-B}, \ldots, S_{n-A} \longleftrightarrow S_{n-B}$. Consequently, at switches $S_{1-A}, \ldots, S_{n-A}$ the VoIP packets and the background transmission compete for access to the shared links. If not enough bandwidth is available to serve both streams, the switches drop packets. Therefore, in the simulation the transmission path of the VoIP application consists of a series of links. However, packets may be dropped at shared links only. Other links do not drops packets because they always have enough bandwidth due to transmitting either VoIP or background traffic only. At the end of the simulation, the packet loss analysis of each switch which drops packets is performed.

During the analysis the VoIP application packet loss rate and the burst ratio value are calculated. Using these values and Eq. (6), the burst ratio of the whole transmission (from node A to B) is calculated. The calculated value is compared with the value calculated at node B based on the analysis of VoIP stream packets that were not successfully delivered, Eq. (1). The result of the comparison quantifies the accuracy of Eq. (6).

It should be noted that in the simulations the packet loss takes place in shared links only. Therefore, in the remaining sections the terms "channel" and "shared link" are used interchangeably.

The results of the simulations may depend on the topology as well as transmission and network parameters. The complete list of parameters identified and analyzed during the simulations is presented in Table 1. The parameters were randomly altered within a range of values during each simulation in order to reduce the influence of a specific parameter value on the results. The parameter values and ranges of values were adjusted so the results of the simulations were relevant for the study of burst ratio parameter.

In order to obtain meaningful results it was important that the VoIP traffic was constantly generating packets. Therefore VoIP traffic utilized the user datagram protocol (UDP) with a constant bit rate. Additionally, the randomization of the background traffic was of crucial importance in order to assure a full spectrum of simulation conditions. Therefore, the background traffic used UDP (with the Pareto distribution) and TCP protocols, both selected randomly for each simulation. Moreover, the start time and the total number of transmitted packets within each transmission were also randomized. As a result the VoIP traffic faced different conditions in each simulation run. The wide spectrum of conditions meant the VoIP traffic was characterized by a wide range of parameters values *BurstR* and packet loss rate Ppl.

This paper presents the results of a total 250,000 simulations, each representing different network conditions. They were carried out in order to demonstrate the validity of the equations. As a result, the validation contains relevant and fully conclusive results.

4. Accuracy of Burst Ratio Calculation

In this section the simulation results run in order to validate Eq. (6) are presented. The equation was numerically verified by the authors in [8], where a transmission channel was modeled by a two-state Markov channel. This section contains simulations results, where the transmission environment was modeled with real networks characteristics, simulated using NS2.

The verification has been performed by comparing two burst ratio values:

- the *BurstR* value measured at the end of the transmission path using Eq. (1),
- the value calculated using Eq. (6), which incorporates the characteristics of each intermediate transmission channel, denoted below as $BurstR_\Sigma$.

The comparison is presented as relative error δ_Σ, defined as follows:

$$\delta_\Sigma = \frac{BurstR_\Sigma - BurstR}{BurstR}. \tag{10}$$

If δ_Σ is equal to 0, Eq. (6) is perfectly accurate. A positive value of δ_Σ means that the experienced packet loss is less bursty than that estimated using Eq. (6). A negative value of δ_Σ means that the burst ratio value calculated with Eq. (6) underestimated the burstiness of the analyzed traffic.

The number of shared links may have an impact on the final results, because the VoIP traffic needs to compete for resources in each link. The more shared links, the more VoIP packets may be lost. In order to study this impact, each simulation was rerun with two, six and ten shared links.

The results published in this section present the relationship between relative error δ_Σ (in %) and packet loss Ppl, number of transmitted packets or *BurstR* of the complete transmission. The error is analyzed in the form of a mean and its confidence intervals. The mean value of the relative error is shown using black lines. The 95% confidence intervals of the mean are marked with gray areas.

Figure 5 presents the relationship between the relative error δ_Σ of the burst ratio calculation using Eq. (6) and packet

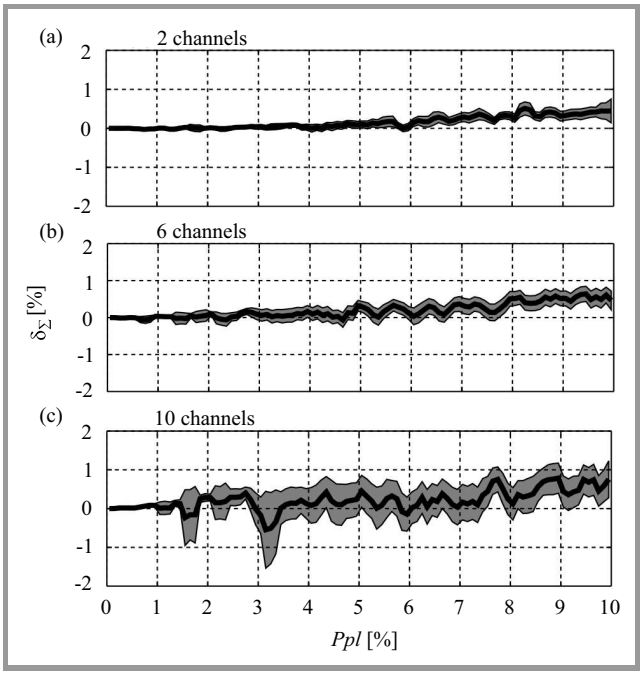


Fig. 5. Relationship between the burst ratio calculation error δ_Σ and the packet loss rate Ppl of the whole transmission. The solid line represents mean relative error while the gray areas present the 95% confidence intervals of the mean. The figures were created with a packet loss range of 0–10%. The subplots presents results for simulations of two, six and ten intermediate channels.

loss *Ppl* of the whole transmission. It can be observed that for values of packet loss lower than 1%, the relative error is negligible, regardless of how many intermediate channels the transmission contains. As the packet loss increases, the mean error and its confidence interval increase slightly as well. The observed increase is dependent on the number of intermediate channels. The higher the number of channels, the higher the error for the same value of packet loss. However, the relative error never reaches 2%, which indicates a high accuracy of the equation.

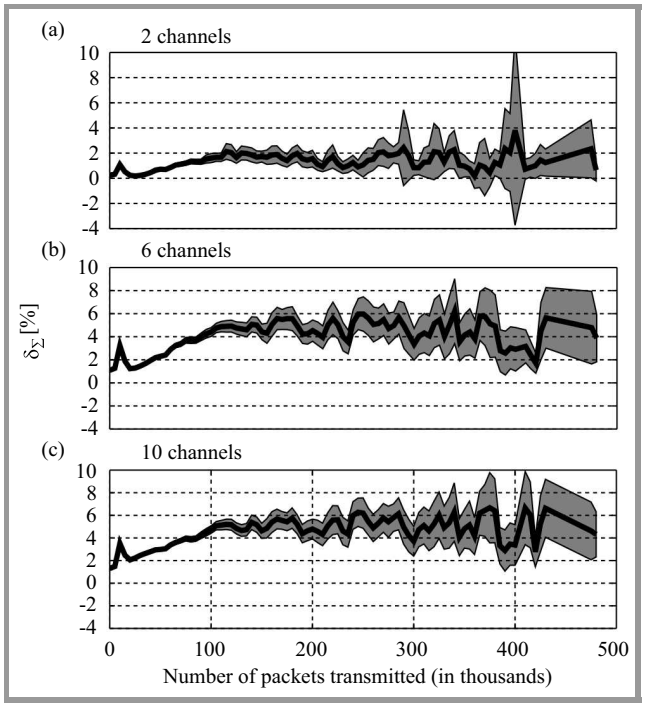


Fig. 6. Relationship between the burst ratio calculation error δ_Σ and the number of transmitted packets. The solid lines represent mean relative error while the gray areas present the 95% confidence intervals of the mean. The subplots presents results for simulations of two, six and ten intermediate channels.

Figure 6 presents the relationship between the burst ratio calculation error δ_Σ and the number of transmitted packets during measurement. The figure shows that the mean error initially slightly increases for the shorter observations and then stabilizes at a level of 2% for two intermediate channels or 5% for ten channels. Figure 6 presents results for up to 500,000 transmitted packets, which corresponds to approximately 2 hours 45 minutes observation of a transmission. Such a long observation is unrealistic and its results are presented only for reference. More reasonable duration of observation is up to 5 minutes, which corresponds to 0–15,000 of transmitted packets. In this range the error never exceeds 4%, regardless of the number of intermediate channels.

Figure 7 presents the relationship of the relative error δ_Σ of the burst ratio calculation using Eq. (6) and burst ratio value *BurstR* of the complete transmission. It can be seen that regardless how many intermediate channels are used the relative error is low around $BurstR = 1$. For two channels, the error value is negligible, regardless of the burst ratio value. In the case of several intermediate channels, as the burst ratio increases, the error decreases and for $BurstR > 1.5$ the error becomes negative. In the worst case, for the scenario of ten intermediate channels the error reaches -9%. It can be seen that for fewer channels, *BurstR* of the complete path reaches higher values. For ten intermediate channels the highest value of *BurstR* slightly exceeds 2.5, while for two channels it is over 3.5. This effect can be explained by analyzing Eq. (9). The formula shows that *BurstR* value of the complete path is approximately equal to the weighted harmonic mean of all intermediate channels' *BurstR* values. As the result, the more channels are involved in the transmission, the lower probability that end-to-end burst ratio reaches high values.

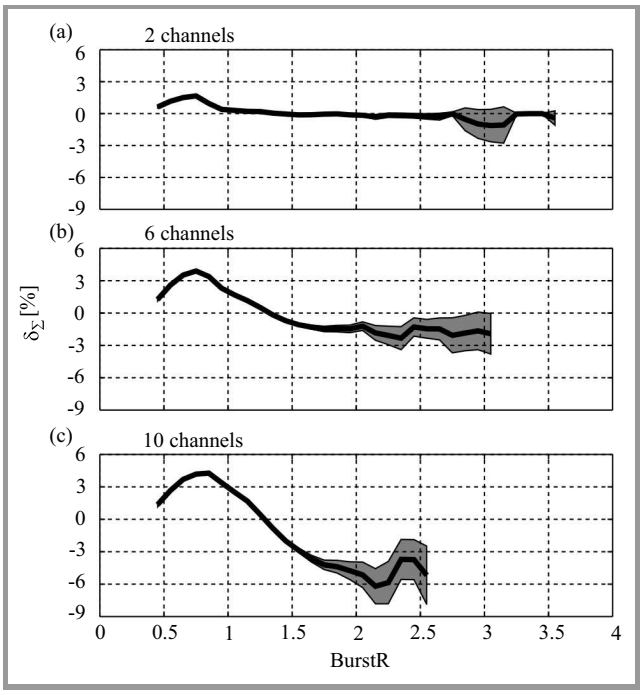


Fig. 7. Dependency of the burst ratio calculation error δ_Σ on the *BurstR* value of the complete transmission. The solid lines represent mean relative error while the gray areas present the 95% confidence intervals of the mean. The subplots presents results for simulations of two, six and ten intermediate channels.

All these results show that when Eq. (6) is used it provides reliable results and a high precision of the measurement. The accuracy of the calculation is always very high, but the most precise results are achieved in the two-channel scenario, when packet loss of the complete transmission path is limited or the burst ratio of the complete transmission path is not higher than $BurstR = 1.5$.

5. Accuracy of the Simplified Equation

As well as the regular burst ratio equation, validated above, we also show a simplified version of the equation, Eq. (9).

This simplification reveals that the burst ratio of the whole transmission path can be approximated with a weighted harmonic mean of properties of individual channels. This equation was verified numerically in [8]. The results indicate that the simplified equation's inaccuracy increases with higher values of packet loss and burst ratio of the whole transmission. However, the verification was performed with the assumption that the transmission channels can be modeled with two-state Markov chains, which is a form of simplification. This section presents the results of equation validation performed in an environment that simulates real characteristics of transmission channels.

The verification of the simplified burst ratio equation – Eq. (9) is performed by calculating the simplification error Δ_{BurstR}. It expresses the difference between the error of the simplified equation and the error of the regular burst ratio equation – Eq. (6). The values that are compared are mean relative error (in %) and the 95% confidence interval of the mean. Both were introduced in Section 4. The comparison of mean error is performed by calculating the difference between absolute values of mean error δ'_{Σ} of the simplified burst ratio equation (Eq. 9) and mean error of the regular burst ratio equation δ_{Σ}, as described in Eq. 10. The comparison is presented below:

$$\Delta_{BurstR} = \left|\delta'_{\Sigma}\right| - \left|\delta_{\Sigma}\right| . \tag{11}$$

If the calculated difference of the mean error is equal to 0, both Eq. (6) and Eq. (9) are equally accurate. When Δ_{BurstR} is positive, Eq. (6) is more accurate, while if Δ_{BurstR} is negative, the simplified equation is more accurate.

The comparison of the 95% confidence intervals of the mean is performed in a similar way, by subtracting the value of the confidence interval for the regular equation from the value of the confidence interval for the simplified equation.

The figures published in this section present the calculated differences of mean error using black lines. The gray areas in the figures correspond to the confidence interval differences of the means.

Figure 8 presents the differences of mean errors and confidence intervals in the domain of packet loss in the range of 0–10%. It can be seen that regardless how many intermediate channels are used, the difference is negligible in that it never exceeds 0.5%. However, it should be noted that there is almost no difference in the confidence interval width (marked with gray fields).

The results clearly show the validity of the simplified equation. The difference in performance, compared with the regular equation, is almost indistinguishable. However, the regular equation almost always performs slightly better than the simplified formula. Therefore, when the highest accuracy of the measurements is required, the regular equation is used. However, when the top priority is ease of calculation, the simplified equation is applied.

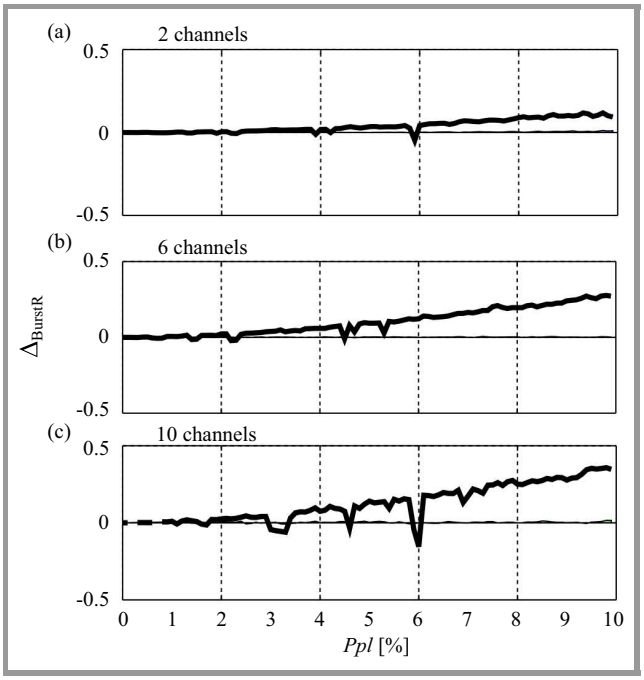


Fig. 8. Difference Δ_{BurstR} between the calculation errors of the regular equation δ_{Σ} and the simplified equation δ'_{Σ} in the domain of packet loss *Ppl* of the whole transmission. The solid lines represent the difference between mean relative errors while the gray areas represent the difference between the 95% confidence intervals. The subplots presents results for simulations of two, six and ten intermediate channels.

6. Applications

As mentioned above, burst ratio is one of the parameters used in the ITU-T E-model, which is used to assess the quality of VoIP. Therefore, the formula presented has a wide spectrum of potential applications, mainly facilitating the VoIP MOS level assessment.

The formulas can be used during network planning. When a network is being designed, a set of technical requirements is specified for the network. They include packet loss, round trip time and mean opinion score (MOS) of VoIP transmission. When network topology is defined, the characteristics of all the network elements are assumed. Even if the topology is complex and the network contains hundreds of elements, the VoIP transmission MOS assessment between any endpoints may be required. Without proper calculation of the burst ratio value between the endpoints, a precise assessment of application quality is not possible. Using the formulas presented and the E-model, MOS can be easily and precisely assessed between any endpoints of the designed network. Therefore, during the network design phase, corrections may be applied to the network topology to help provide the best quality of the VoIP service.

Another application of the formula is when a network is already operating and a re-design of the topology or routing is required. In this case the formula may help assess the impact of the changes on the quality of the VoIP transmission. A good example would be a network that contains multiple elements which introduce packet loss. If only one

of them could be upgraded, it would be important to select the optimal element to upgrade. By using the formula, the network administrator can easily assess how end-to-end VoIP quality would be affected, depending on which elements are upgraded.
The formulas can also be successfully used during monitoring of networks. The measurements, as described in [17], need specially configured environments. Therefore they can only be performed within a single network, owned by a single company. If a VoIP transmission path is established via several different networks, which are administered by different companies, the complete path monitoring is not possible. In this case, the formulas can be used in order to calculate the VoIP transmission MOS using monitoring logs of the individual networks.

7. Conclusions

The results clearly show that the equations presented can be successfully used to calculate the burst ratio parameter, when the complete transmission path consists of multiple concatenated channels. Although the equation has been derived theoretically using two-state Markov models, in real-life scenarios, simulated here using NS2, the equation is still valid. Its accuracy is the highest when the number of concatenated channels is limited to two, when the packet loss of the complete transmission path is low, or the burst ratio of the complete transmission path is not higher than $BurstR = 1.5$.
Moreover, the results show that the simplified version of the equation is almost as accurate as the regular equation, therefore it can be used as an engineering tool. The simplified formula reveals that the burst ratio value of the complete transmission path can be regarded as a harmonic mean of the individual channels burst ratio values, weighted with the their packet loss probabilities.
The results also demonstrate that the equation is valid and therefore can be used in QoE measurements and network performance assessment. Moreover, the formula has a wide spectrum of potential application. As such, it would be useful in improving the quality of VoIP applications.

Acknowledgments

This work has been supported by the AGH University of Science and Technology under contract no. 11.11.230.018. The authors would like to thank Prof. Sebastian Moeller, Prof. Alexander Raake, Dr. Błażej Lewcio and Dennis Guse for consultations and assistance in designing the validation methodology presented in this paper.

References

[1] "Vocabulary for Performance and Quality of Service", ITU-T Rec. P.10/G.100 (incl. Amendment 2), Tech. Rep., 2008.

[2] A. Raake, *Speech Quality of VoIP: Assessment and Prediction*. Wiley, 2006.

[3] J. W. McGowan, "Burst ratio: a measure of bursty loss on packet-based networks", 16 2005, US Patent 6,931,017.

[4] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet loss recovery techniques for streaming audio", *IEEE Network*, vol. 12, no. 5, pp. 40–48, 1998.

[5] A. Raake, "Short-and long-term packet loss behavior: towards speech quality prediction for arbitrary loss distributions", *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 14, no. 6, pp. 1957–1968, 2006.

[6] A. D. Clark, P. D. F. Iee, and M. Ieee, "Modeling the effects of burst packet loss and recency on subjective voice quality", in *Proc. IP Telephony Worksh.*, New York, 2001.

[7] W. Jiang and H. Schulzrinne, "Perceived quality of packet audio under bursty losses", in *Proc. 21st Ann. Joint Conf. IEEE Comp. Commun. Soc. InfoCom 2002*, New York, USA, 2002.

[8] J. Rachwalski and Z. Papir, "Burst ratio in concatenated markovbased channels", *J. Telecommun. Inform. Technol.*, no. 1, pp. 84–90, 2014.

[9] "The Network Simulator NS-2" [Online]. Available: http://www.isi.edu/nsnam/ns/

[10] S. M. Ross, *Introduction to Probability Models*, 11th ed. Academic Press, 2014.

[11] J.-C. Bolot, "Characterizing end-to-end packet delay and loss in the interne", *J. of High Speed Netw.*, vol. 2, no. 3, pp. 305–323, 1993.

[12] H. A. Sanneck and G. Carle, "Framework model for packet loss metrics based on loss runlengths", *Proc. of SPIE*, vol. 3969, pp. 177–187, 1999.

[13] "The E-Model, a computational model for use in transmission planning", "ITU-T Rec. G.107, Tech. Rep., 2014.

[14] "Transmission impairments due to speech processing", ITU-T Rec. G.113, Tech. Rep., 2007.

[15] J. Rachwalski, "Bursty loss modelling in E-model", Semminar in Telekom Innovation Laboratories, Berlin, 25 Apr. 2013.

[16] S. Kurkowski, T. Camp, and M. Colagrosso, "MANET simulation studies: the incredibles", *ACM SIGMOBILE Mob. Comput. and Commun. Rev.*, vol. 9, no. 4, pp. 50–61, 2005.

[17] "Cisco IOS IP SLAs Configuration Guide", Cisco Systems, Inc., Tech. Rep., 2008.

Jakub Rachwalski obtained his M.Sc. from the University of Science and Technology in Krakow, Poland in 2009. He is currently a Ph.D. student under the supervision of Prof. Zdzisław Papir. His current research focuses on the Quality of Experience in VoIP applications, with special interest in the influence of the packet loss distribution on the perceived quality.
jrachwal@agh.edu.pl
Department of Telecommunications
Faculty of Computer Science, Electronics and Telecommunications
AGH University of Science and Technology
Mickiewicza av. 30
30-059 Krakow, Poland

Zdzisław Papir is Professor and a deputy chair at Department of Telecommunications, AGH University of Science and Technology in Krakow, Poland. During 1994–1995 he was serving for the Polish Cable Television as a Network Design Department Manager. Between 1999–2006 he was a guest co-editor for IEEE Communications Magazine responsible for the Broadband Access Series. He has been participating in several R&D IST European projects being responsible for Network performance evaluation and quality assessment of communication services. He has also been appointed as an ICT expert by the European Commission. His current research interests include modeling of telecommunication networks/services and measuring quality of experience.

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

Paper

Measured Interference of LTE Uplink Signals on DVB-T Channels

Massimo Celidonio[1], Pier Giorgio Masullo[1], Lorenzo Pulcini[1], and Manuela Vaser[2]

[1] *Fondazione Ugo Bordoni, Rome, Italy*
[2] *University of Rome Tor Vergata, Rome, Italy*

Abstract—Because of the decision, taken during the ITU WRC-07, to allocate the upper part of the so-called digital dividend spectrum for mobile services on a co-primary basis with TV broadcast services, the involved stakeholders have a great interest in avoiding any interference caused by signals transmitted in adjacent bands. In this context the paper presents some experimental results of a study addressed to assess the effects produced by an interferential LTE signal transmitted from a user terminal when it is in proximity of a television antenna that receives DVB-T signals. The study has been conducted in the context of collaboration between Fondazione Ugo Bordoni and ISCTI, the scientific and technical body of the Italian Ministry of Economic Development, using high professional laboratory equipments and considering different experimental simulation test setups. Several simulation scenarios have been analyzed and results in terms of protection ratio and protection distance have been carried out.

Keywords—DVB-T, LTE, TV Interference, UHF Measurements, uplink.

1. Introduction

Mobile telecommunications are currently experiencing the fourth generation technology, whose base is the standard 3GPP Long Term Evolution (LTE), and this is leading to large economical investments as well as innovative technological developments such as new high-speed and high-quality services, convergence towards all-IP networks and very effective radio network performances in terms of spectrum efficiency, peak data rate and latency.

In December 2008, the LTE specification was published as part of Release 8 and the first implementation of the standard was deployed in 2009. The first release of LTE supported radio network delay less than 5 ms and multiple input multiple output (MIMO) antenna techniques achieving rather high data rates. Later on, in December 2009, Release 9 has been introduced improving several functional features that were already present in Release 8.

As a result of these technological developments, improvements in speed and capacity have been made available from telco operators to customers, boosting the market to propose services and applications more and more hungry for bandwidth. On the other hand, the device proliferation like smartphones and tablets, as well as the globally interconnection of devices, envisaged by Internet of Things (IoT), are going to cause, in the next future, a significant data traffic explosion with the consequence that demand could outstrip supply. This consideration is feeding the debate among regulators on how timely cope with technical and regulatory developments which might influencing the allocation, management and use of the radio-frequency spectrum, as witnessed in the last World Radiocommunication Conference held in Geneva, Switzerland, in 2012 (WRC-12).

According to the spectrum reforming process, currently in progress all around the world, the frequency bands designated for the new services in Europe are the ones at present used by UMTS and GSM, as well as new bands at 2.6 GHz. However, the best frequency band option to open up these services has been, definitely, the digital dividend portion of the UHF spectrum (from 790 to 862 MHz in ITU Region 1, consisting of Europe, Africa and parts of the Middle East), resulting from the digital switchover (DSO) process from analog to digital television (DTV) that was concluded at the end of the year 2013 in almost all European countries.

Furthermore, during WRC-12 in order to meet the demand of additional bandwidth, it was decided to allocate further UHF spectrum to mobile services. The planned new spectrum allocation involves the frequency band from 694 to 790 MHz, and is proposed to come into force in 2015, in order to enable the conclusion of the necessary technical studies regarding the availability and assignment of the new band, before allowing its use.

As a consequence of these decisions LTE will operate alongside broadcast applications (UHF TV channels) and, for this reason, potential coexistence issues might arise.

The digital TV antenna, as well as LTE user terminals, always receives the wanted signal in the presence of unwanted signals generated by other radio systems. On the basis of the outlined spectrum allocation, a certain number of coexistence scenarios have been identified. Depending on the severity of the interference, this may cause degradation of receiver quality, e.g. artifacts in case of digital TV reception or loss of throughput in case of mobile radio, up to complete failure in receiving the wanted signal. It is a matter of fact that the most important factors, which characterize the influence of this interference, depend on the frequency offset between interferer and victim signals as well as on the power level of both wanted and unwanted signals.

In this paper the considered coexistence scenario involves an LTE transmitter, which acts as interferer on the digi-

tal TV receiver. In particular, the adopted LTE equipment is a mobile terminal transmitting an uplink signal in the 832-862 MHz band, where, as described in ECC Decision 09-03 [1], the harmonized ITU Region 1 plan for the digital dividend band has allocated the new mobile services adopting the FDD duplex technique for downlink and uplink communication services, with the uplink one located in the upper part of this band.

Experimental tests have been carried out in the Istituto Superiore delle Comunicazioni e delle Tecnologie dell'Informazione (ISCTI), sound and television broadcasting laboratory of Ministry of Italian Economic Development (MiSE) through high-professional devices, considering several operative setup configurations. The paper will provide, in Section 2, an overview of the uplink LTE transmission and potential influences it could cause on signals transmitted on DVB-T channels, as well as the performance parameters that should be considered for assessing the corresponding interfering effects (protection ratio and protection distance). In Section 3 are illustrated the considered experimental setups, including a short description of interfering and victim signals, the measurement methodology and the features of the most relevant equipment used (DVB-T receivers and MATV masthead amplifiers). Successively in Sections 4 and 5, are reported the resulting measurements carried out during the experimental activity as well as relative considerations about the protection distance, which guarantees the quality of service (QoS) levels required by operators. The paper ends with some conclusive remarks reported in Section 6.

2. LTE Influence on DVB-T Reception: Uplink Analysis

2.1. Overview

Since spectrum-sharing criteria normally depend upon parameters of both interfering and interfered systems, as well as their operating environments, the identification of the appropriate interference protection criteria to be adopted is often a confusing, time-consuming process.

This is due to several regulatory and technical terms used to identify the potential interference, including: permissible interference, accepted interference, harmful interference, protection ratios, allowable performance degradation, interference protection criteria, spectrum sharing criteria, and so on.

In 1990, the ITU-T SG1, with the adoption of Rec. ITU-R SM.669 [2], made an attempt to reduce this complexity by the definition of a matrix of protection ratios for various combinations of interfering and wanted signal modulation types, but the interference protection criteria presented in this Recommendation have become largely obsolete. In fact it is mainly addressed to analog and early digital modulation techniques which are in many cases being supplanted by more complex, digital modulations. Furthermore, efforts are focused primarily on modulation with little regard to radio service requirements or factors linked to the operating frequency.

On the contrary, most spectrum sharing studies today focus primarily on radio services and specific frequency bands. As a result of various compatibility studies conducted within the European Conference of Postal and Telecommunications Administrations (CEPT), the frequency plan, resulting from the decisions approved by the WRC-07, employs a duplex direction that is reversed when compared to the normal European convention. Normally, mobile bands are planned with the uplink (base stations receive/mobile terminals transmit) in the lower band and the downlink (base stations transmit/mobile terminals receive) in the upper band. However, due to concerns regarding interference from mobile transmission to Digital Terrestrial Television (DTT) services operating below 790 MHz, it was decided for the 800 MHz band to reverse the duplex direction, so that the downlink is in the lower band. The reason of this choice is to protect the DVB-T broadcast signal transmitted on UHF channels (in particular on channel 60), from interference produced by the uplink transmission of mobile Electronic Communication Networks (ECN) terminals, which is the most critical aspect for DVB-T receiving systems, due to the unpredictable position of the LTE terminal. Thanks to the adoption of the above mentioned frequency allocation a virtual guard band of 42 MHz is provided. In addition, the remaining 11 MHz duplex gap could be exploited by other unspecified services (see Fig. 1).

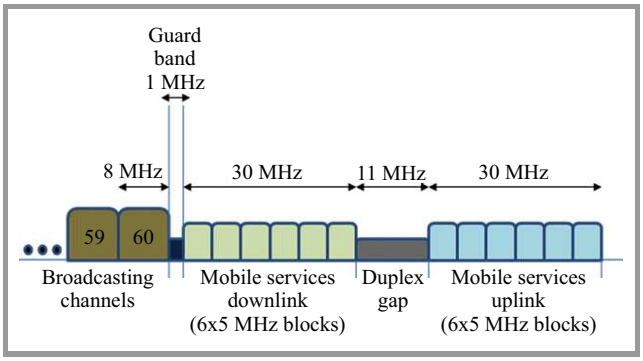


Fig. 1. CEPT channelization proposal for the digital dividend band in case of FDD duplexing technique implementation.

Many studies have been performed to verify the coexistence conditions between LTE downlink and DVB-T signals [3]–[5]. On the contrary, many less studies have been conducted with regard to the interferential effects produced by the LTE uplink signal [4], [5] also because of the higher frequency offset that should theoretically ensure a greater protection. Most of them provide results carried out using computer simulations [6] others consider a limited number of experimental setup configurations [7], [8].

The present work focuses on results obtained through laboratory measurements, referred to the coexistence of DVB-T services, operating in the upper part of the TV UHF band (470–790 MHz), with uplink LTE mobile communication services, operating in the digital dividend band, including

an extended set of experimental configurations involving a MATV masthead amplifier. The analysis has been conducted for a selection of reference scenarios, evaluating the Protection Ratio (PR) and the Protection Distance (PD) parameters, with the aim to provide an overview of potential drawbacks resulting from the usage of LTE mobile equipments nearby TV antennas, in order to give an idea of appropriate countermeasures to be adopted to mitigate these undesired effects.

2.2. Protection Ratio Concept

As defined in Report ITU-R BT.2215-4 [9], the PR is the minimum value of the signal-to-interference ratio, usually expressed in decibels, required to obtain a specified reception quality under specified conditions at the receiver input. In this work, the wanted signal is a DVB-T one whilst the unwanted signal is a LTE uplink one.

PR measurements have been carried out for different power levels of the DVB-T signal at the DTT receiver antenna. In this way, it has been possible to simulate the magnitude of the interference in different operating conditions, trying to cover the widest range normally present at the input of the TV sets.

Usually, PR is conditioned by different parameters of the involved signals, i.e. the frequency offset between the wanted and the interfering signals, the power of the victim signal, as well as by the specific receiver features because usually it has different ability to discriminate interfering signals operating on frequencies adjacent to the ones used by the wanted signal. For this reason, the results of the laboratory measurements have been repeated for different typologies of DVB-T receivers, by configuring the LTE signal at different frequency offset from the carrier of the DVB-T channel under test, evaluating, for each of them, the minimum power level of the LTE signal which affects the quality of the DVB-T video signal or, alternatively, a critical value for the BER of the victim signal higher than $2 \cdot 10^{-4}$ after Viterbi decoding or higher than 10^{-11} after Reed-Solomon decoding.

Another relevant parameter to determine the signal QoS is the overloading threshold corresponding to the carrier-to-interference (C/I) ratio value where the TV signal is lost. This parameter was not deeply investigated in this work however some minimal considerations are reported in the conclusions.

2.3. Protection Distance Concept

From a technical point of view, the PD is defined as the minimum spatial distance, expressed in meters, between an interfering device (LTE user equipment, for the uplink analysis conducted in this work) and the receiving antenna of the system to be protected (DVB-T receiver) in order to ensure that the interfering signal power levels at the front-end antenna of the receiver are still low enough to guarantee an acceptable QoS of the wanted signal.

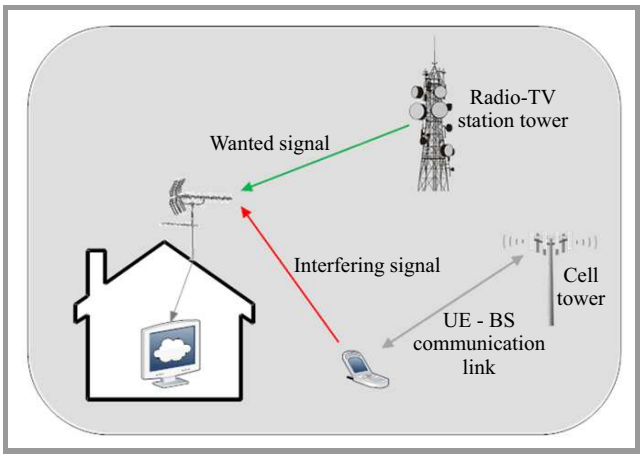


Fig. 2. Reference scenario for evaluating the PD parameter.

Recommendation ITU-R SM.337-5 [10] provides a general method to evaluate this parameter, as well as the frequency separation between the interfering and the victim antennas in order to bind the interfering signal under an acceptable radiated power level. According to this method, the PD can be evaluated under the simplified hypothesis that the LTE mobile terminal is in line-of-sight with the DTT receiving antenna (see Fig. 2) and that the path attenuation of the radio signals can be calculated using the free space propagation model, as it is also suggested in Rec. ITU-R P.525.2 [11].

3. Experimental Scenario

3.1. Experimental Setup

The measurements have been carried out using the experimental setup shown in Fig. 3, in accordance with the method suggested in ECC Report 148 [12]. The DVB-T signal is generated by the Rohde-Schwarz SFU broadcast test system and is modulated by the MPEG-2 transport stream provided by the Rohde-Schwarz DVG generator.

On the other side, the LTE uplink signal is generated by the Rohde-Schwarz SMBV100A vector signal generator. The DVB-T and LTE signals are combined using a directional coupler (Agilent model 775D) operating in the 450–940 MHz frequency band (see Fig. 4), which is able to ensure an effective decoupling of about 60 dB between the two signal generators, if it is used as shown in Fig. 3.

The output of the coupler has an impedance of 50 Ω and needs a 50–75 Ω impedance adapter in order to avoid mismatch reflection to the input of the TV receiver or to the masthead amplifier, in case it is present in the MATV chain.

In the experimental setup option including a masthead amplifier, the output of this device is connected to a variable attenuator in order to simulate the losses of a real TV distribution system in order to guarantee a good DVB-T signal power level (–50 dBm) at the input of the TV receiver.

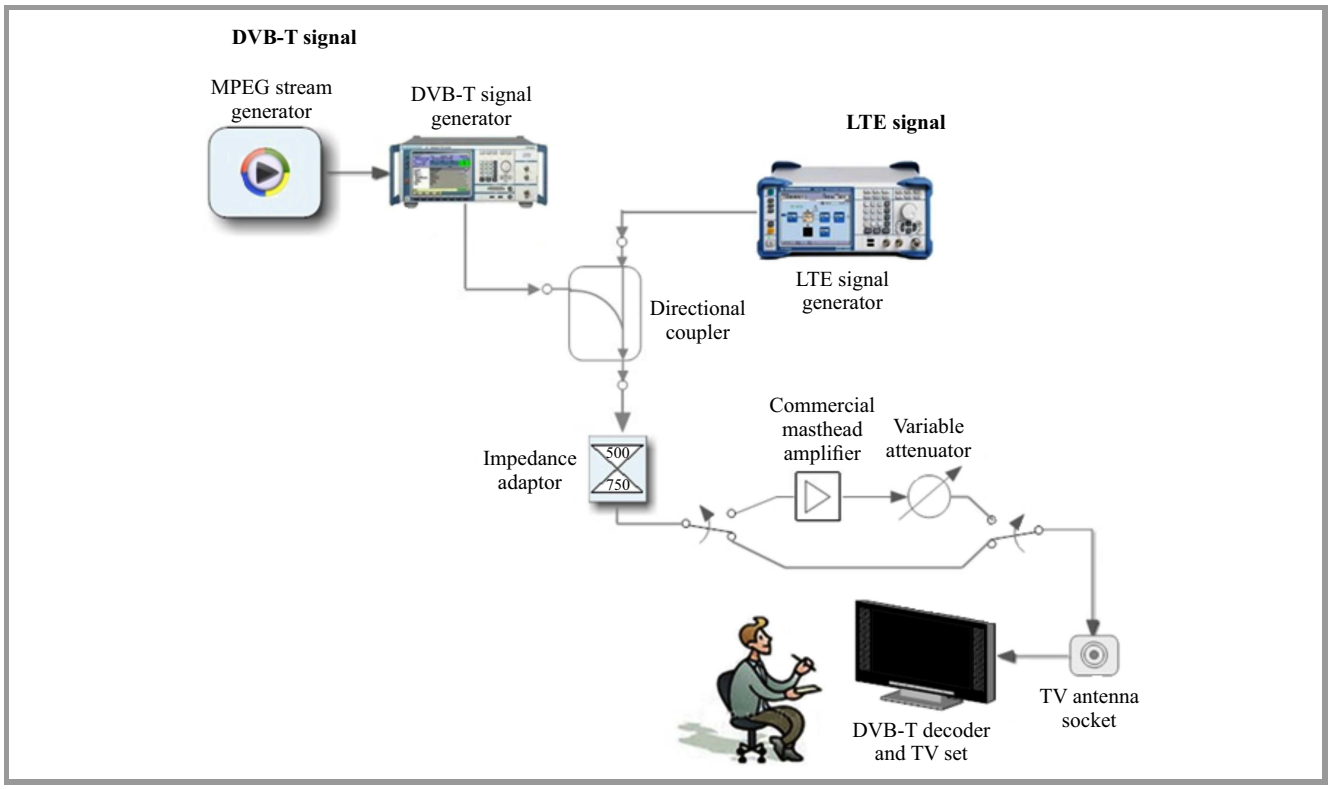


Fig. 3. Functional schematic of the test bed.

Fig. 4. Directional coupler.

3.2. System Parameters Adopted: Interfering and Victim Signals

As above mentioned, in this experimental context the LTE uplink signal acts as the interferer. It has been configured in accordance with the provisions of the WRC-07, and with reference to the regulation fixed by European authorities. According to 3GPP TS 36.211 [13], the LTE signal at 800 MHz is characterized by FDD duplexing technique and the spectrum dedicated to uplink transmissions ranges from 832 to 862 MHz (see Fig. 1).

The adopted frame structure is Type-1, and the transmission scheme for LTE uplink signals is the Single Carrier Frequency Division Multiple Access (SC-FDMA). The main LTE uplink signal frame specifications for the PHY layer are reported in Table 1.

In addition, taking into account of a typical scenario, three different LTE uplink carrier frequencies have been considered: 837 MHz (from 832 to 842), 847 MHz (from 842 to 852) and 857 MHz (from 852 to 862). In any case, for the LTE signal, a traffic load level of 100% has been configured.

Table 1
LTE uplink signal frame specifications for the PHY layer

LTE uplink signal frame parameter	Value
Duplex scheme	FDD
Radio frame structure	Type-1
Duration of radio frame	10 ms
Number of subframes (TTI)	10
Duration of a subframe	1 ms
Number of time-slot per subframe (Resource block or RB)	2
Duration of a time-slot	0.5 ms
Bandwidth of a RB	180 kHz
Number of subcarrier per RB	12
Bandwidth of a subcarrier	15 kHz
Bandwidth for each operator	10 MHz

The victim signal considered in this study is a reference DVB-T one compliant with the ETSI Rec. EN 300 744 [14] whose main features are summarized in Table 2.

Measurements have been carried out for different signal levels at DTT receiver antenna input or at the input of a masthead amplifier, if it is present. QoS of five DVB-T

Table 2
Main settings of the victim signal

Mod-ulation	OFDM sub-carriers	Code rate	Guard interval	Band-width	Data rate
64QAM	8 K	3/4	1/4	8 MHz	22.39 Mb/s

Table 3
TV channel frequencies considered in the experimental setup

DVB-T channel	Central frequency [MHz]	Start frequency [MHz]	End frequency [MHz]
60	786	782	790
59	778	774	782
58	770	766	774
57	762	758	766
56	754	750	758

multiplex channels have been tested spanning from channel 56 up to channel 60 (carrier spacing 8 MHz), as reported in Table 3.

3.3. DVB-T Receivers Typology

During the experimental study, two DVB-T receivers, implementing different tuning techniques (CAN and silicon tuner), have been tested:

- CAN tuner (also known as superheterodyne) is placed in a metal enclosure to prevent interferences and the RF input signal is mixed with a Local Oscillator to obtain an Intermediate Frequency (IF) signal. The superheterodyne architecture used in this kind of tuner leads to an image frequency 72 MHz above the tuned signal;
- Silicon tuner uses Large Scale of Integration (LSI) chips. The receivers using this kind of tuner are not affected by the problem of image frequency, because the input signal is directly converted in base band or in a very low IF.

3.4. MATV Masthead Amplifier Features

Masthead amplifiers are used in television receiving systems to increase the level of signal received at a television set for single or multi user systems. They are mounted on the roof mast or on the roof space, very close to the aerial, and are connected to the TV set by a coaxial cable or a more complex distribution network.
The amplifiers commercially available before the migration of mobile services in the UHF television band show amplification features almost constant up to the end of the digital dividend band (862 MHz), today unnecessary and even harmful, because the presence of unwanted signals could force it to operate in nonlinear mode.

Usually, in order to prevent the amplifier to operate in the nonlinear area, its gain is fixed to a value which allows, at its output and for each DVB-T channel, a signal level adequately lower than the maximum value provided by the manufacturer. This is necessary in order to reduce the potential signal degradation caused by saturation distortion and intermodulation signals production. To this aim, international regulation bodies suggest to use, for each DVB-T channel, the following formula [15], [16]:

$$P_{out} = P_{outmax-ampli}(\text{dBm}) - M(\text{dB}) \tag{1}$$

where: $M(\text{dB}) = 10\log(N-1)$, N = number of received DVB-T channels (2 or more), $P_{outmax-ampli}$ = maximum output power indicated by the manufacturer (usually expressed in dBμV over 75 Ω).

For example, in case of $N = 30$ DVB-T channels, M is equal to 14.6 dB and P_{out} measured for the DVB-T channel having the highest received power level should be fixed at $P_{out} = P_{outmax-ampli}(\text{dBm}) - 14.6$.
In this experimental study, measurements have been carried out using two typical commercial MATV masthead amplifiers, which will be indicated in this work as AMP1 and AMP2, designed to cover UHF band up to 862 MHz. In the following are reported some features of these amplifiers.
AMP1 is a broadband amplifier. Its technical specifications are reported in Table 4. Figure 5 shows the measurement

Table 4
Main parameters of AMP1

Parameter	Value
Noise figure	7.5 dB
Max output	125 dBμV
Minimum gain	10 dB $\pm$2
Maximum gain	30 dB $\pm$2
Power supply	12 V_{DC}
Impedance	75 Ω

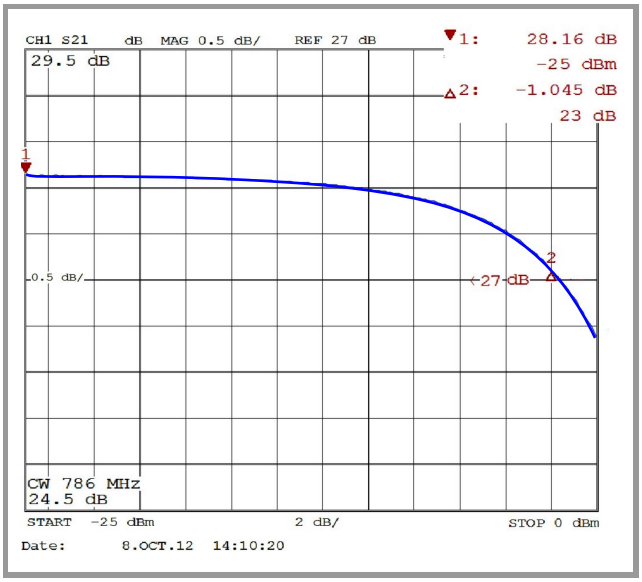


Fig. 5. 1 dB compression point for AMP1.

of 1 dB compression point for AMP1, which occurs in correspondence to an input level of –16.0 dBm and an output level of 11.2 dBm (measurement carried out for a frequency value of 786 MHz and a gain value set to the maximum).

Table 5
Main parameters of AMP2

Parameter	Value
Noise figure	8 dB
Max output	121 dBμV
Minimum gain	30 dB$\pm$2
Maximum gain	53 dB$\pm$2
Power supply	12 V_{DC}
Impedance	75 Ω

AMP2 is a high gain broadband amplifier. Its relevant technical specifications are reported in Table 5. Figure 6 shows the measurement of 1 dB compression point for AMP2. It occurs in correspondence of an input level of –21.5 dBm and an output level of 29.7 dBm (measurement carried out for a frequency value of 786 MHz).

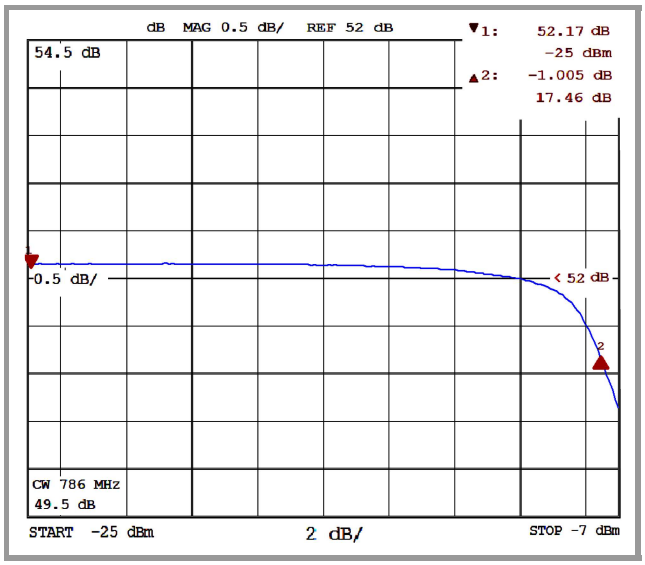


Fig. 6. 1 dB compression point for AMP2.

In order to take into account, as much as possible, of the potential effects of signal degradation due to intermodulation distortion, during the measurements, AMP2 has been set out not only near the maximum gain, but also at 10 dB below the maximum (in practice the signal at the input of the amplifier is first regulated with an attenuation of 10 dB and then amplified with its maximum gain of 52 dB).

3.5. Measurement Methodology

In order to determine the power levels of LTE uplink signal causing quality degradation of DVB-T channels, a proper procedure has been adopted and repeated for each experimental measurement, based on the assessment of the subjective quality of the received TV signal in terms of picture failure (PF). As a preliminary step, the attenuation in dB of all passive elements in the test bed chain has been determined, in order to take them into account during the evaluation of the power levels, of the desired and interfering signals, at the input of the TV receivers or masthead amplifiers as well. Furthermore, a suitable MPEG stream feeding the DVB-T generator and transmitted to the TV receiver has been chosen from the ones available on the MPEG stream generator. After selecting the DVB-T channel to be tested, its power level, the LTE uplink signal carrier frequency and, in case of using the masthead amplifier, its gain value, the power level of the LTE signal has been progressively increased until picture failures of video signal within an interval of 30 s of observation occurred. When this effect appeared, the resulting LTE power level shown on the generator, reduced of the attenuation due to transmission chain passive elements, has been recorded. Note that the reported results depend on the spectral characteristics of the adopted LTE and DVB-T signals as well as on the noise platform introduced by LTE signal generator, which represents a co-channel interference for the DVB-T signal under observation. Any variant of these experimental configurations (e.g. DVB-T signal degradations due to propagation with multiple paths or absence of additional noise), could lead to conclusions different from those obtained in this paper.

4. Experimental Results for the Protection Ratio

As above mentioned, on the basis of the described experimental scenario, two different TV receivers have been used for carrying out the laboratory measurements, the first one operating with a CAN tuner (TV RX1) and the second one with a silicon tuner (TV RX2). These measurements have been performed in absence and presence of two different masthead amplifiers, with the three following configurations:

- configuration A1 – amplifier 1 (AMP1) with a gain of 30 dB (max. allowed),
- configuration A2 – amplifier 2 (AMP2) with a gain of 52 dB (max. allowed),
- configuration A3 – amplifier 2 (AMP2) with a gain of 42 dB.

The measurements have been repeated for each TV receiver with DVB-T power level settings specified in Table 6.
The larger number of measurements carried out in presence of the masthead amplifier was necessary to have a better discrimination of the intermodulation effects due to the amplifier nonlinearities when the DVB-T signal power level is stepping up.
For each configuration, the protection ratio and the protection distance have been evaluated.

Table 6
DVB-T Power levels fixed in correspondence of measurements points

Case	DVB–T Power levels [dBm]	Measurement point
Without masthead amplifier	–35 (max. permitted) –50 –65 (min. permitted)	TV receiver input
With masthead amplifier	–35 (max. permitted) –45 –55 –65 –75 (min. permitted)	Masthead amplifier input

4.1. Measurements in Absence of Masthead Amplifier

In absence of the masthead amplifier, the results obtained with TV RX 1 (equipped with CAN tuner) highlight the presence of an image frequency signal effect 72 MHz below the LTE signal carrier. This impairment may significantly affect the quality of the tuned DVB-T signal, causing the above mentioned picture failures on the video image.
The degradation of DVB-T signal QoS, for the channel influenced by this effect, occurs in presence of an LTE signal power level about 20 dB below the ones noticed for channels not affected by the image frequency impairment. Specifically, for DVB-T signal power levels close to the sensitivity threshold of the receiver, the PR parameter assumes a constant value around –55 dB.

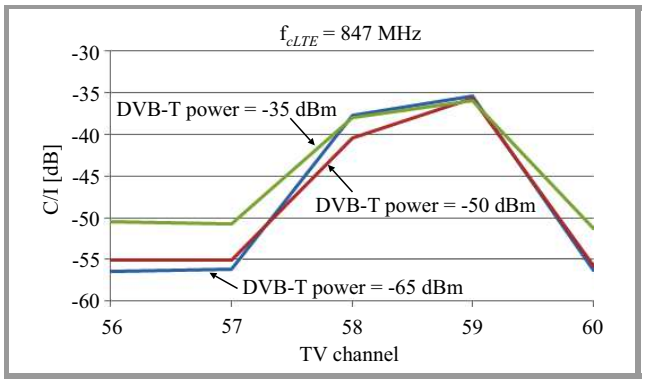


Fig. 7. PR measurements using TV RX1 and LTE signal at $f_{cLTE} = 847$ MHz, for different values of DVB-T power, without masthead amplifier.

Depending on the carrier frequency of LTE uplink interferer signal, the effects of the image channel moves from DVB-T channels 57 (central frequency 762 MHz) and 58 (central frequency 770 MHz), in correspondence of an LTE carrier frequency of 837 MHz, to channels 58 and 59 (central frequency respectively at 770 and 778 MHz, see Fig. 7), with an LTE carrier frequency of 847 MHz, up to channels 59 and 60 (central frequency respectively at 778 and 786 MHz) with an LTE carrier frequency of 857 MHz. These effects are shown in Fig. 8 for a DVB-T signal power level fixed at –50 dBm and become even more evident for decreasing values of the DVB-T signal power level (see Fig. 7).

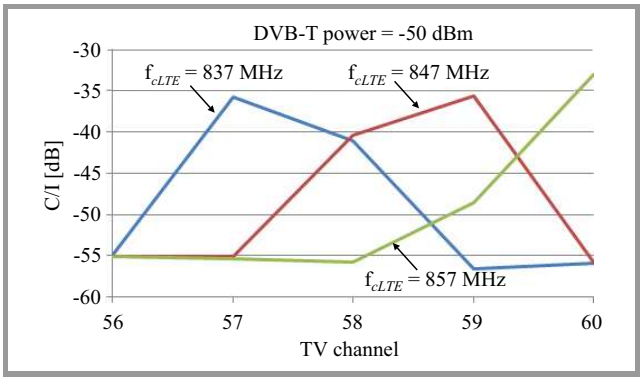


Fig. 8. PR measurements using TV RX1 and –50 dBm DVB-T power level, for different values of f_{cLTE}, without masthead amplifier.

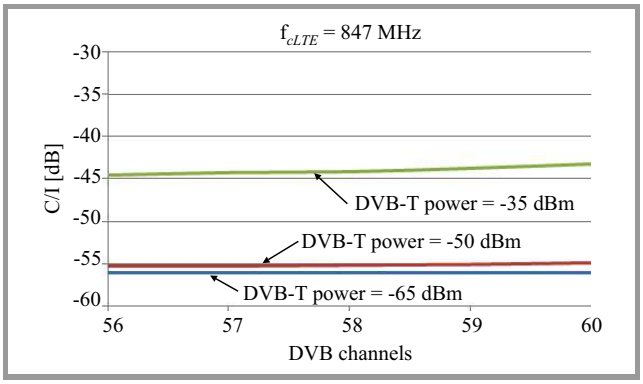


Fig. 9. PR measurements using TV RX2 and LTE signal at $f_{cLTE} = 847$ MHz, for different values of DVB-T power levels, without masthead amplifier.

On the contrary, experimental results obtained with the TV RX 2 receiver (equipped with silicon tuner) show that the interference effects are negligible. In particular, the LTE signal power levels that cause impairments on the DVB-T signal are quite high, as it is evident in Fig. 9, and therefore the reduction of DVB-T signal QoS is evident only in specific environmental conditions, for example when the LTE user terminal is very close to the aerial.

4.2. Measurements with TV RX1 and Masthead Amplifiers

In this experimental setup a masthead amplifier is introduced in the test bed chain after LTE and DVB-T signals have been combined. Two different amplifiers (AMP1 and AMP2) have been used, operating in the above mentioned three different experimental configurations (A1, A2 and A3). Results have been carried out for both TV RX1 and TV RX2. In this section the results with TV RX1 are illustrated.
In configuration A1, the effects of the image frequency are more evident for low power levels of the DVB-T signal

(less than –50 dBm), with an increase of the protection ratio of about 20 dB. For higher power levels of the victim signal (i.e. –35 or –45 dBm), because of approaching the nonlinear behavior (saturation) of the amplifier, the effects of the image channel are less noticeable (Fig. 10).

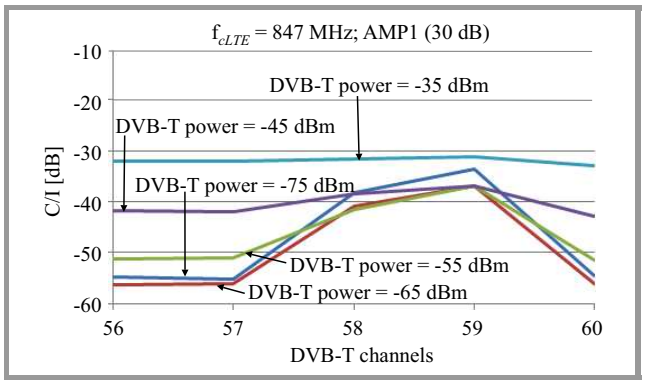


Fig. 10. PR measurements using TV RX1, LTE signal at f_{cLTE} = 847 MHz, for different values of DVB-T power levels, with AMP1 (gain = 30 dB).

In configuration A2 the effects of the image frequency are still evident. In this case, however, the impact is limited to very low values of DVB-T signal power levels (around –70 dBm), since for higher values the consequences of the saturation condition of the amplifier are very considerable. In fact, the behavior is strongly influenced by the amplifier: as soon as the LTE signal at its input exceeds a specific threshold power level (around –35 dBm), the amplifier starts operating in the nonlinear area, consequently resulting in a significant reduction of the video signal quality (Fig. 11).

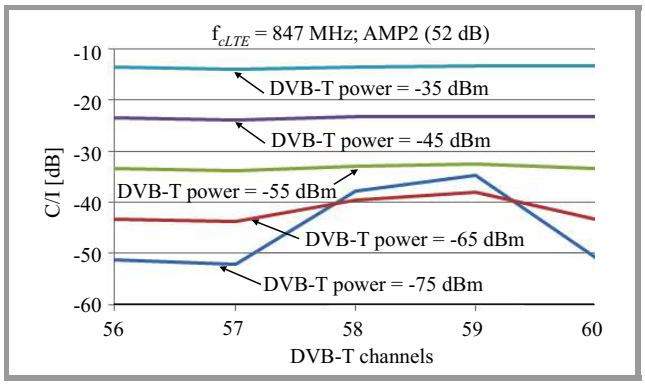


Fig. 11. PR measurements using TV RX1, LTE signal at f_{cLTE} = 847 MHz, for different values of DVB-T power levels, with AMP2 (gain = 52 dB).

Finally, using the A3 configuration, the result is similar to the one obtained with the amplifier at the maximum gain. However, in this case, the DVB-T signal is more protected because, for same power levels of a DVB-T signal at the input of the amplifier, the protection ratios are 10 dB lower of one's obtained in configuration A2. In this case it was not possible to perform measurements with the DVB-T signal power equal to −75 dBm, because at the output of the amplifier the signal resulted severely corrupted even in absence of the interfering LTE signal due to the presence of comparable power levels of device noise (see Fig. 12).

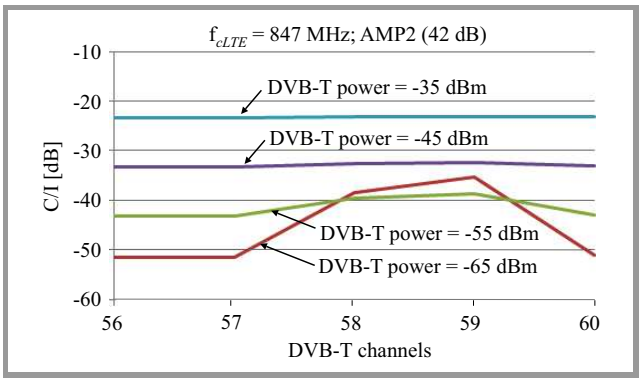


Fig. 12. PR measurements using TV RX1, LTE signal at f_{cLTE} = 847 MHz, for different values of DVB-T power levels, with AMP2 (gain = 42 dB).

As a conclusive analysis concerning the TV RX1 receiver, Figs. 13 and 14 illustrate a comparison between LTE and DVB-T signal power level measurements in presence and absence of masthead amplifiers. The analysis is referred to an LTE signal centered at 847 MHz and considering the DVB-T signals received on channels 59 and 60.

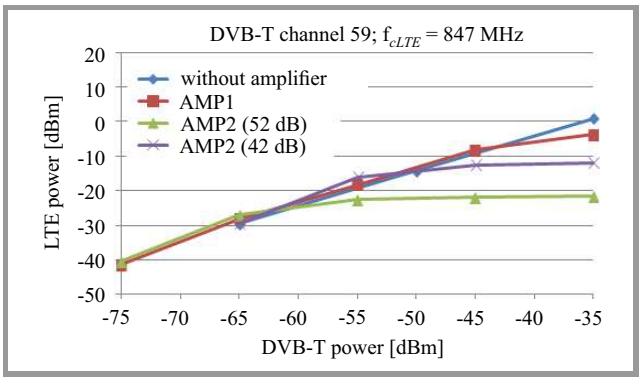


Fig. 13. LTE power level causing interference vs. DVB-T power level in presence and absence of the masthead amplifier, using TV RX1, for DVB-T channel 59 and f_{cLTE} = 847 MHz.

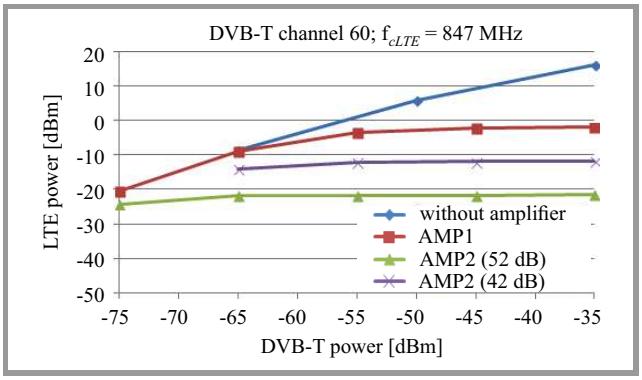


Fig. 14. LTE power level causing interference vs. DVB-T power level in presence and absence of the masthead amplifier, using TV RX1, for DVB-T channel 60 and f_{cLTE} = 847 MHz.

The resulting curves provide a clear evidence of the non-linear effects of the amplifiers and the corresponding values of LTE signal power levels causing this harmful condition.

4.3. Measurements with TV RX2 and Masthead Amplifiers

Also in this case, the analysis has been conducted adopting the above mentioned three different experimental configurations (A1, A2, A3) for AMP1 and AMP2 in combination with the TV RX2 receiver. The test results illustrated in the graphs reported in Figs. 15, 16 and 17 show a uniform behavior in terms of protection ratio for any considered DVB-T signal power level, regardless of the considered TV channel.

However, the LTE signal power level that causes QoS degradation increases with the power level of the DVB-T received signal, reaching the highest levels for DVB-T signal powers of –35 and –45 dBm. The latter two experimental configurations lead to poorly significant results. In fact, as shown in the next section, for such values of the interferer signal the mobile terminal should be located very close to the antenna TV (few centimeters, not in far field condition) in order to originate a perceptible interference.

As for TV RX1, even in this case an analysis illustrating

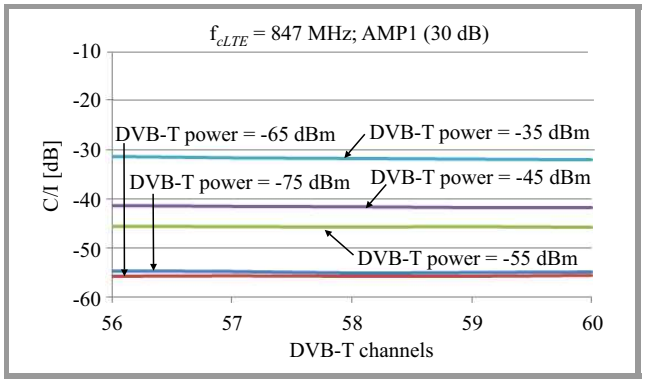


Fig. 15. PR measurements using TV RX2, LTE signal at f_{cLTE} = 847 MHz, for different values of DVB-T power levels, with AMP1 (gain = 30 dB).

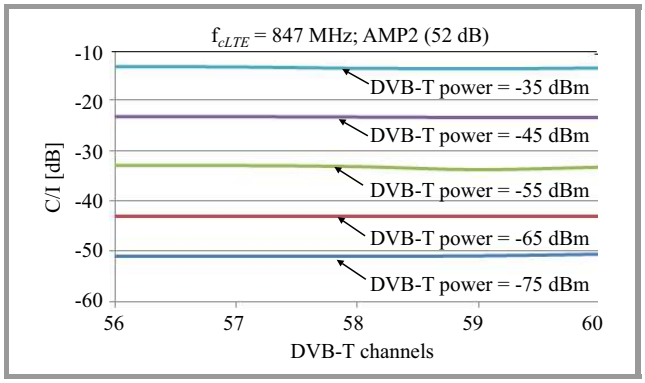


Fig. 16. PR measurements using TV RX2, LTE signal at f_{cLTE} = 847 MHz, for different values of DVB-T power levels, with AMP2 (gain = 52 dB).

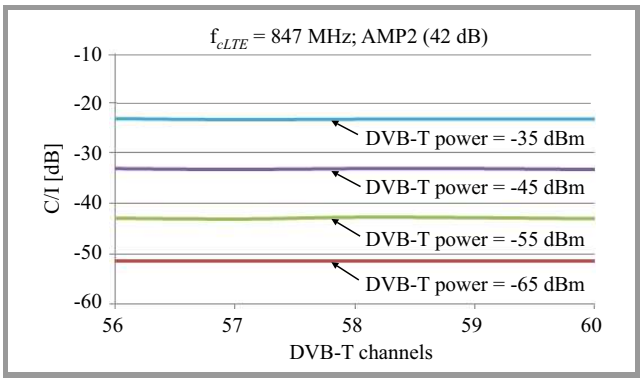


Fig. 17. PR measurements using TV RX2, LTE signal at f_{cLTE} = 847 MHz, for different values of DVB-T power levels, with AMP2 (gain = 42 dB).

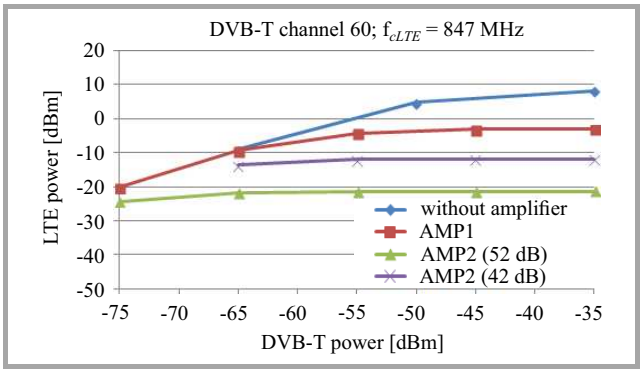


Fig. 18. LTE power level causing interference vs. DVB-T power level in presence and absence of the masthead amplifier, using TV RX2, for DVB-T channel 60 and f_{cLTE} = 847 MHz.

the lowest LTE power levels causing interferential effects at different power levels of DVB-T signal (channel 60), with and without masthead amplifiers, has been provided (see Fig. 18). From this graph nonlinear effects of the amplifiers and saturation points are evident and it is clear how the presence of an amplifier affects the QoS of a DVB-T system in presence of LTE uplink signals. As a consequence of the C/I flat trend highlighted in Figs. 15, 16 and 17, a very similar behavior has been noticed also for all the others considered TV channels.

5. Protection Distances Based on Experimental Results

With the goal to assess the impact of the mutual position of the mobile terminal with respect to the receiving antenna of a TV set, the protection distance parameter between an LTE mobile terminal and a DVB-T antenna has been measured in some of the above considered experimental scenarios.

For this analysis it has been assumed the experimental conditions mentioned in Subsection 2.2 (the mobile user equipment is in Line-Of-Sight with the receiving DTT antenna and the path attenuation can be calculated using the free space propagation model [11]), adopting the LTE

mobile terminal and the DVB-T receiver antenna parameters reported in Table 7 (as suggested in the Rec. ITU-R BT.419-3 [17]).

Table 7
LTE UE and DVB-T antenna parameters

Parameter	Value
Max TX power	23 dBm
UE antenna gain	–3 dBi
RX antenna gain (DVB-T)	9.15 dBi
UE antenna pattern	Omnidirectional

In this context, three case studies have been analyzed:

Case 1 – TV RX1 without amplifier and LTE carrier frequency at 837 MHz.

In this case, as illustrated in Fig. 19 the protection distance ranges from less than one meter up to about 20 m. In particular, the worst condition is detected in correspondence of the image frequency, with a DVB-T signal power level equal to –65 dBm. It should be taken into account that values less than one meter are considered as approximated extrapolation of free space far field behavior.

Case 2 – TV RX1 using AMP2 with a gain fixed at 52 dB and LTE carrier frequency at 847 MHz.

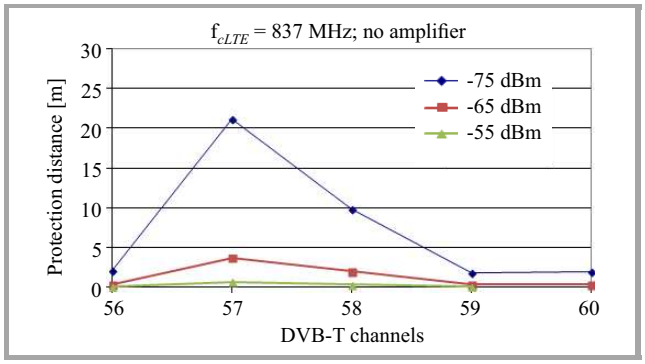


Fig. 19. Protection distance for different DVB-T channels and power levels, at $f_{cLTE} = 837$ MHz, using CAN TV receiver, without masthead amplifier.

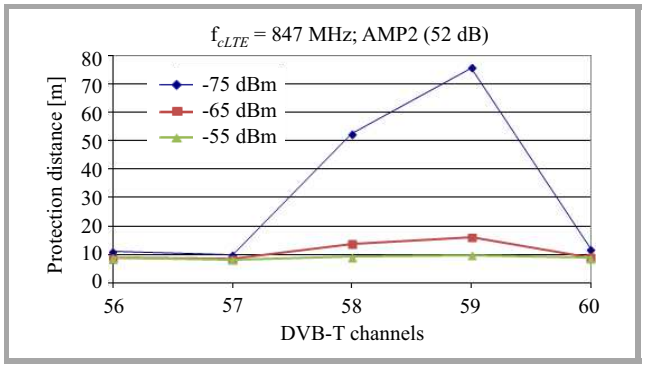


Fig. 20. Protection distance for different DVB-T channels and power levels, at $f_{cLTE} = 847$ MHz, using CAN TV receiver, with AMP2.

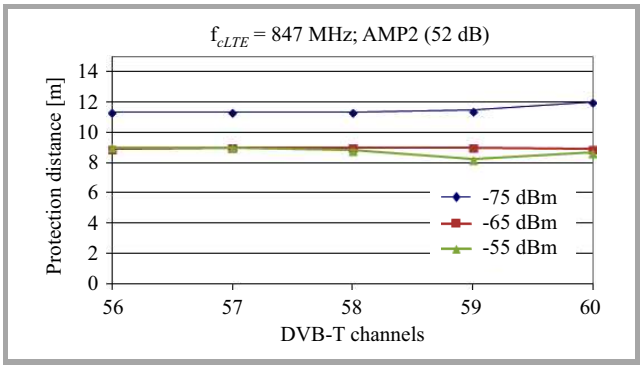


Fig. 21. Protection distance for different DVB-T channels and power levels, at $f_{cLTE} = 847$ MHz, using silicon TV receiver, with AMP2.

In this case, as shown in Fig. 20, the protection distance ranges from less than 10 m up to about 75 meters. The worst condition is still detected in correspondence of the image frequency and for DVB-T signal power level equal to –75 dBm.

Case 3 – TV RX2 using AMP2 with a gain fixed at 52 dB and LTE carrier frequency at 847 MHz.

In this case, as illustrated in Fig. 21, the protection distance is around 10 meters for all channels and all considered DVB-T power levels. Similar results have been assessed for 837 and 857 MHz LTE carriers.

6. Conclusions

In the present study the problem of coexistence of DVB-T and LTE mobile communication systems operating in contiguous UHF frequency bands has been analyzed.

In particular, it has been investigated the interference effect that an LTE uplink transmission might produce on the TV broadcasting service in specific scenarios, by means of the evaluation of protection ratio and protection distance parameters referred to experimental simulation laboratory setups representing real operative situations.

The obtained results highlighted that the TV receiver features, as well as the presence or absence of a masthead amplifier in the receiving chain, may heavily influence, in specific circumstances, the DVB-T signal QoS perceived by the user. More in detail, for the protection ratio, the results can be synthesized as follows.

In absence of the masthead amplifier, the CAN TV receiver determines the presence of an image frequency contribution on DVB-T channels 57–60, depending on the LTE signal carrier ranging from 837 to 857 MHz, with a consequential increasing of the PR of about 20 dB for the channel influenced by this effect. On the contrary, this result does not occur for the silicon TV receiver where, only in presence of DVB-T signal power levels close to the minimum allowed or in presence of quite high LTE signal power levels, reachable in extremely rare conditions

(LTE mobile terminal located very close to the UHF TV antenna), the reduction of DVB-T signal QoS is relevant, independently from the considered TV channel;

In presence of the masthead amplifier, it has been observed a behavior similar to the one described in the previous case, until the power levels of the victim signal are under a threshold value, inversely proportional to the gain of the amplifier. Once this value is exceeded, the nonlinearity characteristic effects of the amplifier become predominant and the picture failure of the video signal appears even for very low power levels of the interfering LTE signal gathered by TV antenna. It must be underlined that in this case the signal degradation affects almost all DVB-T channels and not only a specific one. Furthermore the increase of the power level of the LTE interference signal of a few decimal of dB (0.2 to 0.5) leads to completely break down the TV signal, reaching the condition of overloading threshold.

About the protection distance parameter, it can be deducted that the lower is the received DVB-T signal power level the higher should be the distance of the LTE transmitter from the receiving UHF TV antenna in order to avoid interference, mainly in correspondence of the channel that is influenced by the image frequency, if the TV set equipped with a CAN tuner is considered. This effect is even more emphasized in presence of a masthead amplifier when, the previous distance, is in the order of tens of meters, depending on the gain of the amplifier. Obviously, in case of a TV set equipped with a silicon tuner, due to the absence of image frequency effect, the protection distance behavior is almost constant for all considered DVB-T channels, with a slight worsening in case of DVB-T power levels close to the minimum allowed values at the input of masthead amplifier. From these results it is clear that some interference problems exist, which could be mitigated by adopting appropriate countermeasures including the installation of a low pass filter in the head-end of the MATV plant.

Acknowledgements

The authors wish to thank Dr. Elio Restuccia, Dr. Gianmarco Fusco, Mr. Massimo Ferrante and Mr. Roberto Dal Molin of Istituto Superiore delle Comunicazioni e delle Tecnologie dell'Infomazione (ISCOM) for their valuable support in planning and carrying out the laboratory activities.

References

[1] "ECC Decision of 30 October 2009 on harmonised conditions for mobile/fixed communications networks (MFCN) operating in the band 790-862 MHz", Electronic Communications Committee, Oct. 2009.

[2] Rec. ITU-R SM.669, Protection Ratios for Spectrum Sharing Investigations, 1990.

[3] I. Parker and S. Munday, "Assessment of LTE 800 MHz Base Station Interference into DTT Receivers", ERA Tech. Rep. no. 2011-0351, Survey, UK, Jul. 2011.

[4] N. Cardona, "Coexistence of broadcast and mobile technologies in digital dividend band", in *IEEE BTS Gold Worksh.*, Brasov, Romania, 2014.

[5] B. Randhawa, J. Parker, and S. Antwi, "LTE Interference into Domestic Digital Television Systems", Cobham Technical Services, ERA Technology RF and EMC Group, Rep. no. 2010-0026, Jan. 2010.

[6] I. Cho, I. Lee, and Y. Park, "Study on coexistence between long term evolution and digital broadcasting services", *Int. J. Adv. Sci. Technol.*, vol. 38, Jan. 2012.

[7] K. Sakic and M. Gosta, "Measurements on the influence of the LTE System on DVB-T reception", in *Proc. 54th Int. Symp. ELMAR 2012*, Zadar, Croatia, 2012.

[8] "Dynamics of 3GPP LTE uplink: 800 MHz DTT and LTE coexistence", Real Wireless Tech. Rep., ver. 5.0, Feb. 2012.

[9] Rep. ITU-R BT.2215-4, Measurement of protection ratios and overload thresholds for broadcast TV receivers, Nov. 2014.

[10] Rec. ITU-R SM.337-5, Frequency and distance separations, 2007.

[11] Rec. ITU-R P.525.2, Calculation of Free-Space Attenuation, 1994.

[12] ECC Report 148, "Measurements on the performance of DVB-T receivers in the presence of interference from the mobile service (especially from LTE)", Jun. 2010.

[13] 3GPP TS 36.211 Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation (Release 12).

[14] ETSI EN 300 744 Digital Video Broadcasting (DVB); Framing structure, channel coding and modulation for digital terrestrial television, 2015.

[15] EN 50083-8:2013 Cable networks for television signals, sound signals and interactive services – Part 8: Electromagnetic Compatibility for Networks.

[16] EN 50083-5:2002 Cable networks for television signals, sound signals and interactive services – PART 5: Headend Equipment.

[17] Rec. ITU-R BT.419-3, Directivity and Polarization of Antennas in the Reception of Television Broadcasting, 1992.

Massimo Celidonio received the Electronic Engineering degree in 1991. In the same year he started working at Ugo Bordoni Foundation as a researcher and successively, in 1996, became an IEEE member joining the IEEE Communication Society. At the beginning he was involved in studies on channel coding techniques and digital signal quality. Actually he is working on the subject of broadband wireless access (BWA) in the "next generation networks technologies" area. In this context he studied original network architectures and coordinated experimental trials employing Wi-Fi, HiperLan and WiMAX technologies. Recently, as a consequence of his participation at the ECC/TG4 CEPT workgroup, in cooperation with the Italian Administration, he managed a project on the subject of interference between broadcast TV and LTE signals operating in the UHF digital dividend band.
E-mail: celi@fub.it
Fondazione Ugo Bordoni
via del Policlinico 147
00161 Rome, Italy

Pier Giorgio Masullo received the Electronic Engineering degree at University of Rome La Sapienza. In 1984 he started working at Fondazione Ugo Bordoni as a researcher. He was involved in studies on the influence of hydrometeors on radiopropagation in SHF band, on spatial diversity at 12.5 GHz, on radiometric techniques for remote sensing applications. He cooperated with the Italian Administration participating to national and international activity of ITU-R SG6 and ITU-T SG9 (2001–2005) and to the activity of Tetra Mou Certification Body (2007–2010). Recently he was involved in studies about the influence of LTE systems on DVB-T signals.

E-mail: giorgio@fub.it
Fondazione Ugo Bordoni
via del Policlinico 147
00161 Rome, Italy

Lorenzo Pulcini received the Telecommunication Engineering degree from La Sapienza University of Rome in 2001. He joined Ugo Bordoni Foundation (FUB) in July 2002 as a researcher. Since the beginning of his activities in FUB he studied and tested the modern broadband wireless access network technologies and architectures based on the IEEE 802.11, IEEE 802.16 and ETSI HiperLAN standards, collaborating to the coordination of technological WiMAX trials in Italy, in cooperation with the Italian Ministry of Communication. At present, he is involved in the "next generation networks technologies" area, working on projects concerning 5G mobile communications. Recently, he has been involved in EC projects in the context of the Seventh Framework Programme. His current research interests are focused on studies relying on interference between DVB-T and LTE signals operating in the UHF digital dividend band.

E-mail: lpulcini@fub.it
Fondazione Ugo Bordoni
via del Policlinico 147
00161 Rome, Italy

Manuela Vaser received M.Sc. in Telecommunication Engineering in 2011 at University of Rome Tor Vergata, and she is Ph.D. student in Telecommunication and Microeletronics Engineering in collaboration with Telecom Italia SpA. Her research activity relies on LTE/SAE systems, User Experience and QoS end-to-end characterization. She has collaborated with Fondazione Ugo Bordoni with the assessment of LTE physical signal and frame characteristics in a study about coexistence between DVB-T and LTE systems. Actually, she is also in collaboration with ETSI TC INT group, about Voice over LTE (VoLTE) quality tests and LTE/IMS interoperability.

E-mail: manuelavaser@gmail.com
University of Rome Tor Vergata
DIE – Via Politecnico
1-00133 Rome, Italy

Paper

The Integration, Analysis and Visualization of Sensor Data from Dispersed Wireless Sensor Network Systems Using the SWE Framework

Yong Jin Lee[1], Jarrod Trevathan[2], Ian Atkinson[1], and Wayne Read[3]

[1] *eResearch Centre, James Cook University, Townsville, Australia*
[2] *School of Information and Communication Technology Griffith University, Brisbane, Australia*
[3] *School of Engineering and Physical Sciences, James Cook University, Townsville, Australia*

Abstract—Wireless Sensor Networks (WSNs) have been used in numerous applications to remotely gather real-time data on important environmental parameters. There are several projects where WSNs are deployed in different locations and operate independently. Each deployment has its own models, encodings, and services for sensor data, and are integrated with different types of visualization/analysis tools based on individual project requirements. This makes it difficult to reuse these services for other WSN applications. A user/system is impeded by having to learn the models, encodings, and services of each system, and also must integrate/interoperate data from different data sources. Sensor Web Enablement (SWE) provides a set of standards (web service interfaces and data encoding/model specifications) to make sensor data publicly available on the web. This paper describes how the SWE framework can be extended to integrate disparate WSN systems and to support standardized access to sensor data. The proposed system also introduces a web-based data visualization and statistical analysis service for data stored in the Sensor Observation Service (SOS) by integrating open source technologies. A performance analysis is presented to show that the additional features have minimal impact on the system. Also some lessons learned through implementing SWE are discussed.

Keywords—environmental data, environmental monitoring, sensor technologies, standardization, web-based visualization.

1. Introduction

In recent decades Wireless Sensor Networks (WSNs) have been dramatically advanced and adopted by many domains to remotely monitor environments [1]. The development of low-cost sensor technologies that are capable of capturing various properties of physical phenomena has led to the growing popularity of WSNs. This has made it easier to observe many environmental aspects [2]. WSNs can also reduce the time needed for collecting large amounts of data on key environmental factors. Furthermore, WSNs provide access to the collected data via the Internet, thereby allowing environmental scientists and decision makers to gain a better real-time understanding about the observed environment.

Analyzing sensed data requires a significant amount of time and effort. Such analysis involves the discovery and integration of data from multiple sources (e.g. various and different types of sensors), assessing quality issues (e.g. missing/suspicious data), hypothesis testing, and visualizing the test results to support decision making. Manual analysis of large amounts of heterogeneous and spatiotemporal data is difficult and complicated. Automatic integration, analysis, and visualization of sensed data from multiple sources can reduce the workload needed for addressing data quality issues and understanding environmental conditions. Such automation can also minimize human mistakes during the analysis phase. However, different WSN systems provide different encodings, models, and services for their sensor data. This makes the integration of differing sensor technologies and network systems problematic. Furthermore, the encodings, models, and services are typically designed for a particular application, which makes it difficult to reuse these services for other WSN applications. A standardized model, encoding, and service for WSN data would avoid the constant and inefficient need to "reinvent the wheel", and can facilitate the discovery and exploitation of sensor data.

This paper describes a system architecture based on the Open Geospatial Consortium (OGC) Sensor Web Enablement (SWE) framework [3], [4]. A middleware integration platform has been designed to collect and integrate sensor data from disparate WSN systems, referred to as the James Cook University (JCU) Sensor Federation (JSF). JSF provides a flexible solution for automating the process of transforming sensor data into the corresponding SWE encoding and storing the data in the Sensor Observation Service (SOS) via the web service interface. Furthermore, additional features have been added to the existing SOS web service interface to provide web-based access to the data and statistical analysis tools. Several real world WSN projects of varying scales and complexities have been integrated into one SOS using JSF to demonstrate the system's versatility. A performance analysis indicates that the additional features have minimal impact on the system.

This paper is organized as follows. Section 2 describes the SWE framework, provides a brief overview of the WSN projects that authors have been involved with, and presents the motivation for the work presented in this paper. Section 3 describes the JSF system architecture, and proposes a middleware integration platform for the automation of transforming data from multiple WSNs into the SWE encodings. Section 3 also shows how SOS is integrated with open source freely available technologies to support web-based data visualization and statistical analysis of the data stored in the SOS. Section 4 analyses the performance of the enhanced SOS and discusses some of the issues authors had when implementing a SWE system. Section 5 provides some concluding remarks and avenues for future work.

2. Related Work and Problem Motivation

2.1. Sensor Web Enablement

Historically, WSN applications have been completely proprietary. A specific vendor would provide all of the sensor technologies, hardware, software, and network infrastructure. This predicament meant that WSNs were very technical, application-specific, inflexible, and expensive to purchase and maintain. There was limited scope to integrate heterogeneous sensor technologies (i.e. sensors from different vendors). Furthermore, the sensed data was formatted/encoded according to the vendor's own standards, which restricted data sharing and reuse.

In recent years, the concept of the *Sensor Web* has gained momentum [3]–[10]. The Sensor Web's aim is to make all sensors interoperable (regardless of the vendor) so that heterogeneous sensor technologies can be combined to create low-cost, non-proprietary WSNs. Furthermore, collected data becomes available to the Sensor Web which promotes data sharing and reuse. The data can be reused by other consumers for purposes that may be unrelated to, or extend upon the original motivation for collecting the data. This is possible as the WSNs and the data they collect adhere to a set of mutually accepted standards.

The OGC is made up of representatives from academia, industry, and enthusiasts to develop the standards behind the Sensor Web. The OGC SWE framework provides a set of standards that enables all types of sensors, transducers and sensor data repositories to be discoverable, accessible and usable via the Web [3], [4]. The SWE framework consists of following standards and services:

- Observations and Measurements (O&M) – defines XML schemas for accessing and exchanging observations, measurements, procedures, and metadata of sensor systems;
- Sensor Model Language (SensorML) – defines standard models and XML schemas for describing the processes within sensor and observation processing systems. SensorML provides a functional model of the sensor system, where all components including sensors, transducers, actuators, and processors are modeled as processes;
- Sensor Observation Service (SOS) – enables the querying of observations, sensor metadata and representations of observed features, registration/deletion of sensors, and inserting new observations of a registered sensor. SOS is essentially a data repository at the heart of an SWE WSN;
- Sensor Planning Service (SPS) – defines interfaces for queries that provide information about the capabilities of a sensor and how to task the sensor;
- Sensor Alert Service (SAS) – provides a standard web service interface for publishing and subscribing to alerts from sensors; and
- Web Notification Services (WNS) – provides a standard web service interface for asynchronous delivery of messages or alerts from SAS and SPS web services.

The SWE architecture has reached broad acceptance by sensor network application developers. Schade *et al*. [11] applied the SWE framework to volunteered geographic information sensing and event detection techniques. Shafi *et al*. [12] introduced an automated detection/alert system based on the SWE framework (SOS, SAS and WNS) that detects radiation leakage and sends a notification to its subscribed users. Hu *et al*. [13] extended the SensorML model to support sensor observation capability information, i.e. depth, quality, frequency, and range, that enables the accurate discovery of qualified sensors. Srimathi *et al*. [2] proposed a sensor grid architecture that combines a metamodeling tool, the SWE framework, and sensor grid (Hadoop framework). Back *et al*. [14] presented a conceptual design for bridging two domains: a supervisory control and data acquisition system and a *Geographic Information System* (GIS), where the SOS is used to provide a standardized service model for GIS.

Churcher *et al*. [15], [16] describe their experiences with applying SWE to a telecare application involving a number of projects using bespoke sensor hardware, interfaces, and communications. Guru [17] show how they are using the a river catchment WSN to evaluate specifications for SWE in terms of its ability to facilitate water resource management tools. Markovic *et al*. [8] also describe a system for river pollution monitoring and alerts using architecture based on SWE. Lee and Reichardt [18] discuss how open standards for sensor interfaces and data formats can aid in speeding up the identification of threats to homeland security. Samadzadegan *et al*. [19] developed a system architecture for monitoring air quality observations using SWE standards (i.e., SOS, SAS, SPS and WNS) for integrating/interoperating heterogeneous sensors and discovering air pollution to send a notification.

2.2. Proposed Sensor Network Projects

The authors have been involved in several projects where WSNs were deployed in different locations and operate independently. The WSN projects differ in size, complexity, and application. These WSN projects include:

Smart Environmental Analysis and Technologies (SEMAT) [20]: The SEMAT project revolved around constructing smart sensor networks that can be deployed in aquatic settings for the purposes of conducting marine studies. Authors have undertaken SEMAT deployments at Deception Bay and Heron Island in Queensland Australia. The system was designed to take a heterogeneous, low-cost approach, which allowed for near real-time access to data. In each deployment, five buoys containing on-board electronics (Gumstix Computer-On-Module) equipped with various sensors from Dataflow Systems (temperature, light, water pressure, and salinity) were positioned in shallow water environments. The buoys communicated sensor data back to the end user via a base station located near-by on land. This project grappled with significant WSN issues including limited power supply, communications over and underwater, and problems with marine fouling and water ingress.

Digital Homestead Project: This project involved building a low-cost and smart WSN suitable for applications in a digital homestead (i.e. remote farming properties) and urban environments. The project's initial WSN deployment at Rowes Bay in Queensland Australia used and Seeeduino Stalker with eight DS18B20 temperature sensors, DHT22 humidity sensors, and analogue light sensors placed under different types of roofing materials for observing the energy efficiency measures. This study is being used to explore renewable energy solutions that can benefit biodiversity maintenance through planned urban landscapes.

Greening Federation Place: This project's goal was to demonstrate how heritage buildings in tropical environments can evolve into sustainable buildings while retaining cultural significance. Federation Place is a heritage listed building located in Townsville Australia. A WSN containing DS18B20 temperature sensors was deployed at Federation Place to examine the thermal properties of the building and identify fine scale sources of temperature variation.

Over time, difficulties arose as a result of each of the individual WSN deployments using different types of sensors (with different capabilities) and requiring different setup configurations. From a software perspective, each of the projects contained its own models, encodings, and services for sensor data. Also, differing amounts and types of data were available to describe each deployment's characteristics, e.g. the positioning of nodes and sensors are available for the SEMAT deployments, but not for the Rowes Bay deployment. Furthermore, each deployment was initially integrated with different types of visualization and analysis tools based on individual project requirements [21].

As the number of projects grew and their complexity increased, the need for standardization of sensor configuration, data, storage, communication, and a generic web-based user interface for data visualization and analysis became apparent. The solution required was more comprehensive than the existing solutions proposed by the literature in Subsection 2.1 due to a number of factors. The existing proposals from the literature were either for a specific project, or proposed frameworks that were too broad to be applied in practice. A system that could be used over multiple disparate WSN projects with completely different applications was desired. The system also needed to remove the manual process of generating documents that adhere to SWE standards (i.e., SensorML, O&M). When performed manually, this process is tedious, repetitious of work conducted in other WSNs, and is often error-prone. Automating this process would increase the speed of setting up a WSN and reduce the possibility of errors in the SWE documents. Furthermore, to authors' knowledge little or no literature exists on providing a general web-based interface and statistical analysis features that can interact with the SWE framework.

3. JSF System Architecture

In order to integrate and interoperate sensor data from the disparate WSNs described in Subsection 2.2, the SWE framework was extended by creating a *middleware integration platform*. The intention is to provide an interface (referred to as the SWE API) for each WSN deployment that facilitates interoperability according to SWE standards. The SWE API provides common encoding/decoding functions that can be used by any WSN. Encoding functions specific to a particular WSN are abstracted from the SWE layer and are implemented in an extension level. For example, with the SEMAT Heron Island deployment, the HeronSOSEncoder implements functions specific to the Heron Island deployment, by extending the SWE SOSEncoder. The point of this approach is that changes to any particular WSN do not affect how any other WSN application interacts with the SWE framework. Furthermore, the system automates the generation of the SWE documentation (i.e., SensorML, O&M) to ease WSN set-up time or changes in the WSN configuration, thereby reducing the potential for errors in adhering to SWE standards.

The authors also decided to extend SWE's SOS standard by providing a generic web-based user interface and statistical analysis functionality. The combination of the WSN projects and our extended SWE functionality is referred to as the JCU Sensor Federation (JSF). Figure 1 presents the proposed JSF system architecture. The system is comprised of:

- A Data User – this is the individual user/stakeholder who is interested in accessing and viewing the data from any WSN connected to the system;
- A SOS with support for web-based data visualization and statistical analysis – this provides storage of sensor metadata and sensor observations. The extended

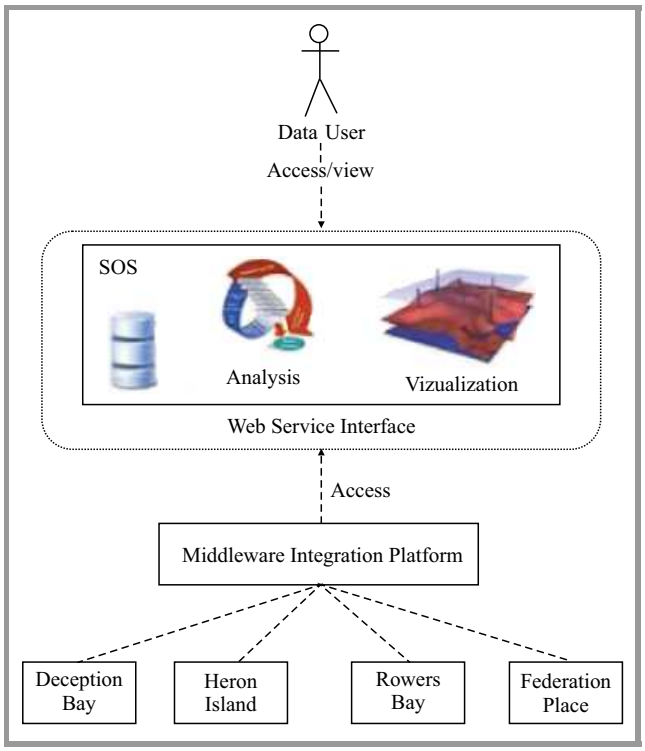


Fig. 1. The JSF system architecture for the integration and interoperability of sensor data from multiple data sources.

functionality allows for automated input, integration, visualization, and analysis of sensed data. Environmental scientists can use these features to enhance the decision making process and/or the discovery of new information from the environment under observation;

- A Middleware Integration Platform – this hides the heterogeneity of models, encodings, and web service interfaces for sensor data. This enables a user to access the sensor data via the Internet without having to learn individual models and service interfaces, and also facilitates the integration and interoperability of the heterogeneous data through automating the process of generating SWE-compliant documents;

- Individual WSN Projects – the projects undertaken as part of the work presented in this paper. These include the SEMAT deployments at Deception Bay and Heron Island, the Digital Homestead deployment at Rowes Bay, and the Greening Federation Place deployment (refer to Subsection 2.2). Conceivably the number of projects can scale with the system.

A major benefit of employing the SWE framework is for easy discovery and use of the data. The decision was made to not utilize different SOSs for each deployment as a distributed approach reduces this benefit. For example, the distributed approach would require a user to send requests to different SOS URLs. Furthermore, users would need to be informed every time a new SOS is added to the system. Having a single SOS provides a single service interface to multiple WSN deployments' data/metadata. Therefore, the proposed architecture brings all of the WSN deployments together in one SOS.

3.1. The JSF Middleware Integration Platform

Sensor data stored in a SOS must be available via the SOS web service interface. However, this process requires significant effort in practice to achieve. For example, registering a sensor via the SOS web service interface requires three steps:

- the sensor data need to be mapped into the respective message encoding (i.e., the InsertSensorDocument),

- the document needs to be formatted based on a SOS protocol binding, e.g., Simple Object Access Protocol (SOAP),

- the formatted document is then transferred to the SOS via its web service interface.

Furthermore, it is necessary to have a common agreement on how to apply SWE within a specific domain. This is because there are different SWE specifications available, where the encodings, models, and services are different. For example, SOS version 2.0 provides the ability to store observation metadata and data through different transactions, whereas this ability is not available in SOS version 1.0. That is, SOS version 1.0 requires metadata to be transferred every time an observation is to be stored. However, SOS version 2.0 only requires the metadata to be stored once. Therefore, only the observation data is required, which reduces the amount of data transferred and makes transfers faster. Also, the SOS implementations can vary based on the needs of a particular domain (e.g. protocol bindings). Due to the aforementioned reasons, and that SOS version 2.0 has a richer array of functionality, for JSF the SOS version 2.0 is used.
JSF was developed to integrate the sensor data from multiple WSNs using the SWE architecture. JSF manages adapted SWE specifications of an individual domain application to map, format, and store its sensor and sensed data. JSF provides the ability to extract sensor data from a domain WSN application, and transforms the data into the corresponding request document using its SWE API implementation. Then, the encoded document is formatted and transferred using the Transaction API, which provides binding protocols (i.e. XML binding or SOAP binding) and the HTTP functions (request/response). This abstracts the underlying WSN projects from SWE and also automates the process of generating SWE-compliant documents.
Figure 2 shows the design of the JSF middleware integration platform. This platform consists of three layers:

- Extraction layer – extracts sensor network data from web services or databases and converts the data into a corresponding SWE API function,

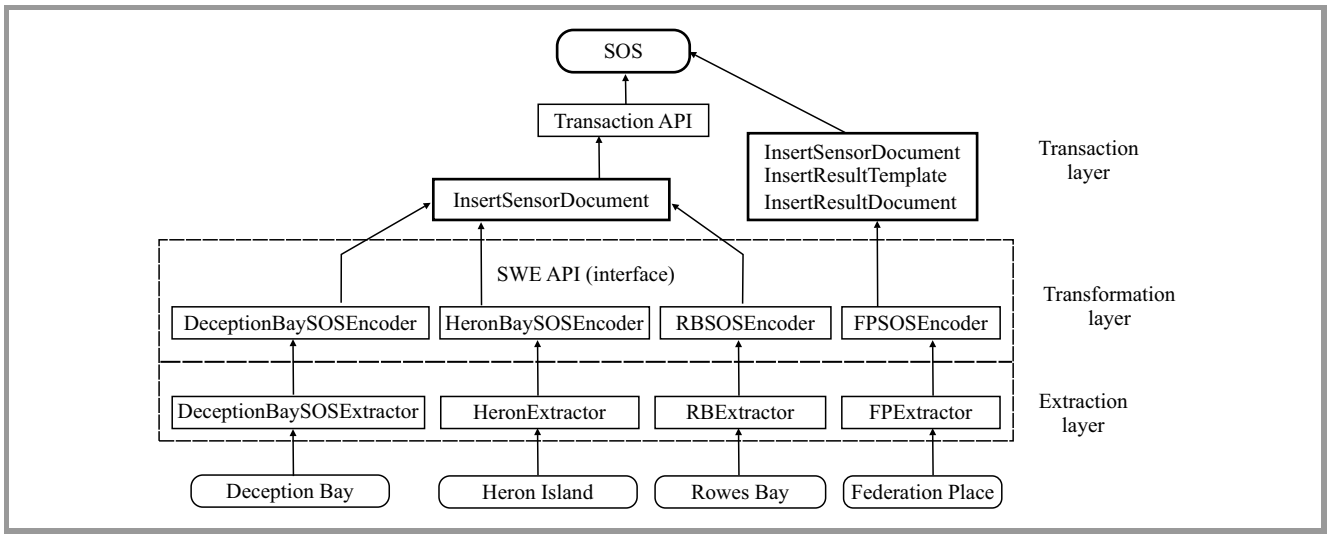


Fig. 2. JSF middleware integration platform design (the arrows denote the direction of the flow of information).

- Transformation layer (SWE API) – maps data into a corresponding SOS request document (i.e., InsertSensorDocument, InsertObservationDocument, InsertResultTemplateDocument, or InsertResult),
- Transaction layer – binds/transfers the encoded documents to a corresponding SOS web service interface.

Consider the SEMAT Deception Bay deployment. The data is first extracted by the DeceptionBayExtractor (i.e. the Extraction layer) and appropriate SWE API function is called. The data is then encoded into a SOS-compliant document by the DeceptionBaySOSEncoder (i.e. the Transformation layer). Finally, the SOS request document is transferred and stored in the SOS according to the SOS web service (i.e. the Transaction layer). A similar process occurs for data from any of the other WSN deployments, i.e. Heron Island, Rowes Bay, and Federation Place.

3.1.1. The SWE API

The SWE Application Programming Interface (API) provides standardized and portable system abstractions that allows JSF to transform data into SWE encodings (i.e. the Transformation layer). The SWE API consists of the following components:

- Interface layer – provides five interfaces for the SOS API. Each interface corresponds to the individual SWE standards (SWEFrame_SOSEncoder, SWEFrame_SMLEncoder, SWEFrame_OMOLEncoder, SWEFrame_SASEncoder, and SWEFrame_SPSEncoder);
- Abstract layer – is associated with a particular SWE implementation that provides a list of commonly used functionality for the particular SWE implementation, i.e. encoding data, query request documents, and binding and transmission operations;
- Implementation layer – is an extension of the Abstract layer that maps sensor network data from a particular system type into the SWE documents.

A deployment must first be registered with JSF. During registration, an API is created for the deployment. A deployment's API implementation contains two properties "identifier" and "version". The combined value must be unique in order to store it into the API container. The registered implementation can be retrieved from the container by passing the respective identifier and version onto the container interface. The Abstract layer can simply be extended to add a new encoding, binding, or transmission process.

Figure 3 presents an example of an SWE API implementation for the data from the SEMAT Heron Island deployment. It shows the identifier, version, and methods for each class, and relationships between classes. The HeronSOSEncoder class uses all of the other classes. Individual classes represent each corresponding SWE component, e.g. HeronSOSEncoder for SOS. These classes transform sensor data into their respective SWE encodings (e.g., the getInsertSensorDocument functionality maps sensor data into the SOS InsertSensorDocument).

The HeronSensorMLEncoder implements the AbstractSensorMLEncoder that maps the sensor data into an InsertSensorDocument. The HeronOMLEncoder implements the AbstractOMLEncoder that maps the observation data into an InsertObservationDocument, where its observation metadata and data can be mapped in separate documents by the HeronResultTemplateEncoder and HeronResultEncoder respectively. The HeronSosEncoder provides a single interface to access to the aforementioned implementations.

To describe a sensor in a SWE-compliant way requires the following attributes:

- Identification – this requires the user to supply uniqueID ("urn:ogc:def:identifier:OGC:uniqueID") and offeringID (urn:ogc;def:identifier:OGC:offeringID), where an uniqueID attribute is used for

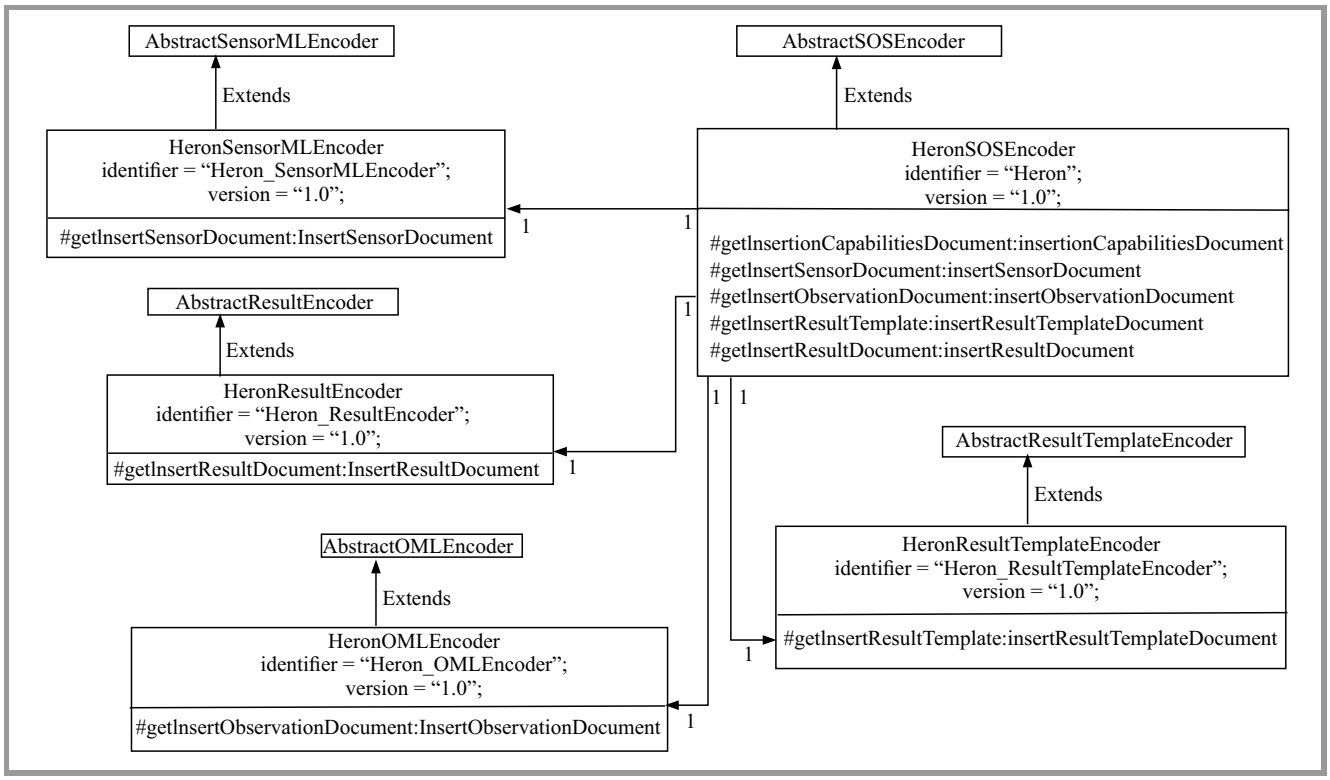


Fig. 3. Class diagram of the SWE API implementation for the SEMAT Heron Island deployment.

Table 1
The integration of the heterogeneous sensor data models/encodings using the SWE encodings

WSN	Sensor Data Model/Encoding	SWE Encoding
Deception Bay, Heron Island and Rowes Bay	**Node model:** name, latitude, longitude, buoy model, power log and sensors	**SensorML:** identifier, position, outputs and components
	Sensor model: serial number, manufacturer, description, type, parameter number and unit of measurement	**SensorML:** identifier, position and outputs
	SensedData model: type, unit of measurement, position, time, raw_Data, calibrated_Data and power log	**O&M:** field and values
Federation Place	**Node model:** name, latitude, longitude and sensors	**SensorML:** identifier, position and components
	Sensor model: name, type, observedProperty, code, altitude and definition	**SensorML:** identifier, position, outputs and observableProperty
	Observation model: definition, code and type	ResultTemplate: field
	Data model: value and timestamp	**Result:** value

querying sensor (system) metadata, and offeringID is used for inserting the sensor's sensed data. This is the same value that must be used within O&M in order to link the observation and sensor metadata/data;

- Capabilities– this attribute is used to describe the feature that the sensor is measuring (e.g. Ocean);
- Location – describes the sensor's (system) geographical location, its format is defined by its "reference-frame" definition;
- Inputs – describes the sensor's (system) process input;
- Output – describes the sensor's (system) process output;
- Components – other systems that is included within the system.

A WSN deployment typically consists of sensor nodes, where sensors are attached/installed. So proposed WSNs are described as a sensor node (identification, capabilities, location, inputs, outputs, and components) that combines

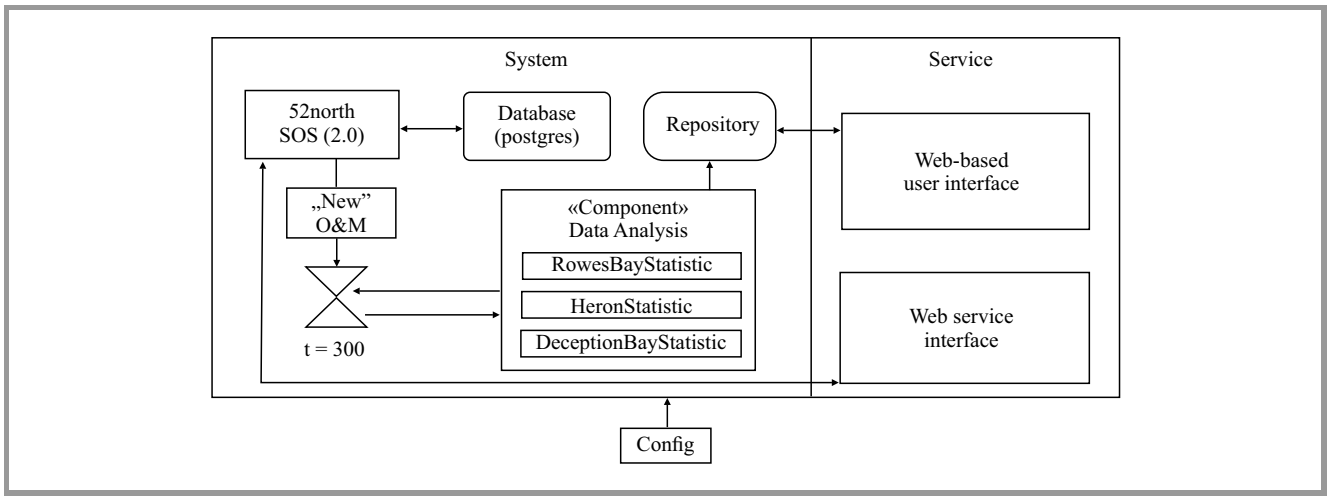


Fig. 4. The enhanced SOS architecture with web-based data visualization and statistical analysis functionality.

all the sensors under the components attributes, where individual sensors are described (identification, capabilities, location, inputs, outputs). The SEMAT Heron Island deployment is described under the sensor node's capabilities attribute and individual sensor's capabilities describe what actual environmental factor the sensor is measuring, e.g. temperature, light. Note that SensorML documents can be quite large. The InsertSensor document for the Heron Island deployment contains over 800 lines of code.

Table 1 shows how the SWE API integrates the heterogeneous sensor data models/encodings using the SWE encodings. The SWE API implementation for the Deception Bay, Heron Island and Rowes Bay deployments converts the Node and Sensor model into SensorML, and the SensedData model into O&M. The API implementation for the Federation Place deployment transforms the Node and Sensor model into SensorML, and the Observation and Data model into InsertResultTemplate and InsertResult. The purpose of this table is to show how each deployments own characteristics can be maintained, while the SWE API maps these characteristics to the appropriate corresponding attributes in SWE.

3.2. SOS Design and Implementation

As previously mentioned, the authors required a generic web-based user interface to operate across all the WSN deployments using JSF. Therefore, an existing SOS implementation (52° North SOS version 2.0) is extended to support web-based data visualization and statistical analysis of the collected data. This was achieved using the following open source technologies and APIs:

- Apache Common Math API – provides mathematics and statistics, i.e. descriptive statistics, simple/multiple regression, rank transformation, covariance, correlation, and statistical tests;
- Weka API – enables Java to support several data mining tasks including data pre-processing, clustering, classification, regression, and feature selection;
- Highstock library – provides general timeline charts with navigation options, e.g. scrolling and date picker;
- Google Maps API – allows for the embedding of Google Maps on a web page.

The system provides additional features to the existing SOS web service interfaces (i.e., InsertObservation, InsertResultTemplate and InsertResult) that analyses and visualizes the data based on the system configuration. The SWE encodings/documents are provided by the OGC (net.opengis.* package). 52° North SOS does not provide support for all the semantic definitions to describe attributes such as FeatureOfInterest, Location, Format, etc. These attributes are described using the 52° North API (ogc.n52.sos.ogc.om.features*).

Figure 4 illustrates the enhanced SOS architecture and how the system operates. At the heart of the architecture is the 52° North SOS. SWE compliant documents are pushed to, or retrieved from the SOS via SWE's web service interface. When a sensor transfers sensed data, the system stores the data in a temporary location. This sensed data is formatted according to the O&M SWE standard.

The system runs a batch-process at predefined intervals to iterate through new data to generate and store a JavaScript Object Notation (JSON) file for each O&M document. The interval is determined by a configuration file. For example, in this instance the configuration file contains $t = 300$ which means that batch process runs every 300 s.

Individual statistical analysis implementations can integrate data from multiple sensors by providing a list of sensor identifiers supplied by the offering parameter. Note that the offering parameter is O&M's equivalent of the sensor identifier in sensorML. This parameter can be accessed by invoking the isOfferingListed function in O&M. The system also checks for any statistical analysis implementations that use the sensed data through the offering property within the O&M document during the iteration. The result of each

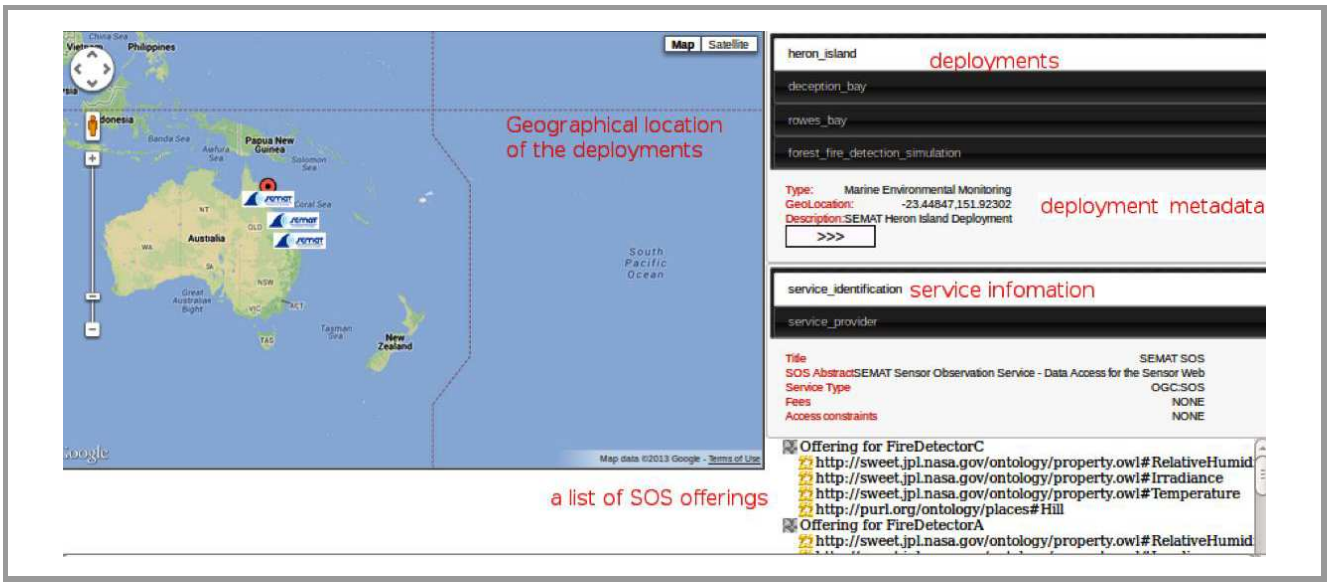


Fig. 5. The main web page shows the geographical location of the WSN deployments on the Google Maps, and sensor/observation data provided by the SOS.

Fig. 6. An example deployment page for the web-based user interface.

analysis process is mapped with the Java Map interface, which is transformed into a JSON file and stored in the repository. The JSON files provide a light-weight data interchange format to facilitate asynchronous browser/server communication. This technology is used for mapping deployment data via the Java Map interface.

The web-based user interface provides graphical visualization of the sensors, observation data, and the analysis results. When the user interface requires information, it can access the data from the 52° North SOS using SWE's web service interface. Alternately, when statistical analysis/aggregation data is required, the user interface can access the repository.

The user interface consists of two primary sections:

- Main web page – provides an overview of registered deployments, presents the deployment locations on a Google Map, and allows the user to select and navigate to a specific deployments web page,
- Deployment web page – provides access to all details specific to an individual deployment, and presents

the user with options to graphically visualize sensed data and analysis results, or to export the data to a file.

Figure 5 shows the user interface's main web page. The main page of the interface consists of four frames:

- Map – illustrates the geographical location of various WSN deployments and sensor nodes as markers on Google Maps. A user can view a deployment's parametric information by clicking on the corresponding marker in the map;
- Deployment – displays a list of WSN deployments associated with a particular project (e.g. SEMAT) and shows each deployment's name, geographical location, type, and description. A user can also view information specific to a WSN deployment by clicking the corresponding deployment button;
- Service – provides the SWE service information;
- SOS offering – shows a list of the sensed data provided by the SOS.

Figure 6 shows an example of what is shown in the user interface when a specific deployment has been clicked on. This page consists of the following components:

- Visualization metadata – provides a brief description about the deployment's purpose;
- Sensors – shows a list of sensors associated with the deployment and allows a user to click on a specific sensor to view its metadata and sensed data;
- Sensor metadata – displays sensor metadata information (e.g., the sensor type, description);
- Analysis – lists any analysis processes being conducted on the collected data, and provides the user with several options regarding the types of statistical analysis that can be performed;
- Analysis result – provides graphical illustration of the analysis results;
- Data graph – provides graphical illustration of the sensed data. Note that numerous sensor data sets can be overlayed on the same graph;
- Time control bar – allows the user to control the analysis result by varying the time span.

Figure 6 shows the deployment page for the Rowes Bay WSN. The interface is showing the data graphs corresponding to the temperature sensors and the sensor metadata, i.e. where each sensor is located in relation to the roofing material. The user can graph each sensor's data individually, or overlay all sensor data. The deployment page also illustrates the statistical analysis results conducted on the temperature data. The analysis section is charting the minimum, maximum, variance, and average temperature of each temperature sensor at hourly intervals. An user has the ability to export the collected (and analyzed) data in a series of formats for use in other software packages.
The web-based user interface is written in JavaScript and HTML to increase flexibility and reusability. The interface also provides the ability to add, modify, or remove a data source (i.e. sensor and analysis results), and to configure the visualization type (i.e. charting type) from the deployment web page using a system configuration file. A new deployment web page can be added by creating a web page and modifying the configuration file (i.e. setting its data sources and visualization types). This is automatically converted into the corresponding JavaScript functions to generate charts/graphs and Google Maps. In this manner, the web-based interface essentially becomes "generic" in that it can display data from any WSN regardless of its application.

4. Performance and Lessons Learned

4.1. Performance of the Enhanced SOS System

To evaluate the integration of the statistical analysis and visualization services with the SOS, the 52° North SOS is compared with presented enhanced SOS. The test environment was setup with a virtual machine and a laptop. Both SOS implementations were installed on the virtual machine (Linux kernel 2.6.32, Red Hat 4.4.7, 2.93 GHz CPU and 1 GB RAM) with Java 1.7 and Tomcat 7. JSF was installed on the laptop (Mac OS X 10.7.5 2.8 GHz CPU, 4 GB RAM) to simulate test data sets (InsertObservationDocument).
The effects of the additional features in the system performance were observed. 100 simulated sensors were registered on the both SOS implementations, and the enhanced SOS was configured to provide descriptive statistics on the sensed data transferred by these sensors every minute. A test function was implemented in the middleware to encode and transfer the InsertObservationDocument with 100 data points for each sensor.
Figure 7 presents the results of the performance comparison between the 52° North SOS and the enhanced SOS. The X-axis and Y-axis gives the number of sensors and the time (in seconds) respectively. The test results show that the InsertObservation, InsertResultTemplate and InsertResult service time of the enhanced SOS are approximately 0.5, 0.1, and 0.3 seconds slower respectively than the original SOS for every transaction. This indicates that the additional features of proposed enhanced SOS have minimal impact on the original service performance of the 52° North SOS.

4.2. Lessons Learned from Implementing SWE

In this section some of the practical challenges the authors faced when using SWE are briefly described and how they managed to overcome these issues.
A major hurdle for implementing a SWE-compliant WSN is that SWE takes time to learn (i.e. a few months). The SWE documentation is complex. An SWE implementer

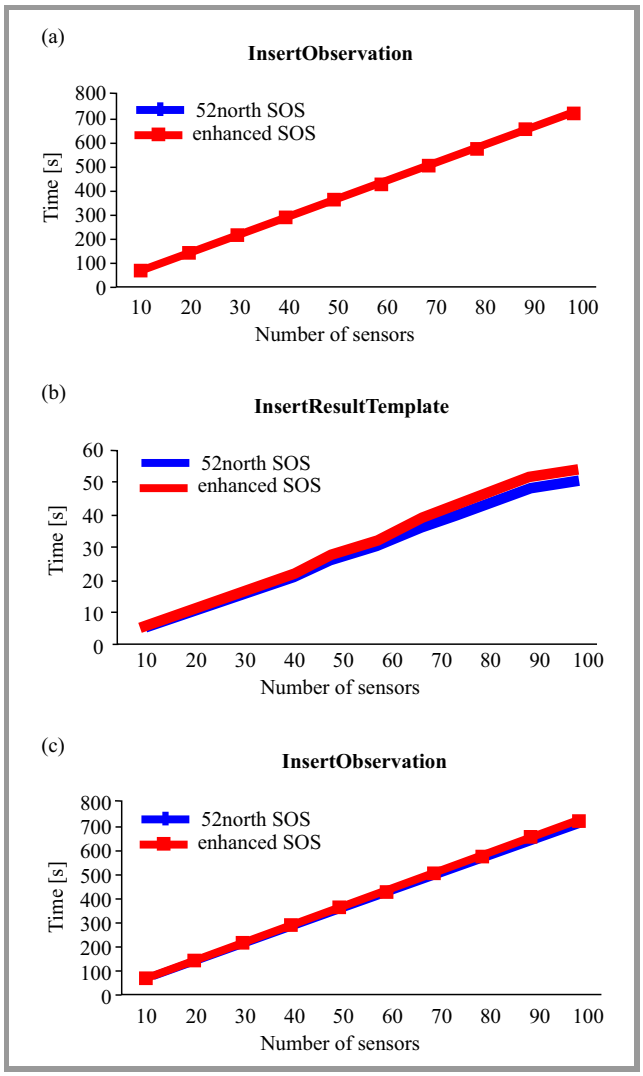


Fig. 7. A performance comparison between the 52° North SOS and the enhanced SOS.

has to go through every part of the documentation or a least the relevant sections in order to successfully get a SWE-compliant WSN operational.

The process with moving a WSN towards SWE is found to first focus on getting a SOS running. The next step is understanding how to push a SensorML document into the SOS using the sensor's identifier (and also usually a semantic web resource). Once a sensor is registered with SOS, the sensor's identifier is used to push a sensor observation into the SOS. SOS will then use the sensor's identifier to automatically link observation data with the corresponding SensorML document.

A major issue that hinders an SWE implementation is that SWE is very pedantic with its expectations about data and message formats. SWE requires specific data to be provided in a particular order with an explicit structure, and also with an exact semantic resource. If any data or message does not strictly adhere to these formatting rules, the information is not recognized and causes errors. This lack of flexibility can make implementing SWE a challenge. In some instances SWE will raise the error to the user's attention. On other occasions it may appear that data has been stored correctly (as no errors are flagged). However, as the format is incorrect, the data is unable to be retrieved. This lack of storage safeguard can cause frustration if a significant quantity of data has been incorrectly stored and can no longer be recovered. Therefore, extensive testing of formatting is required before employing the system for use with real data.

A further frustration with SWE is that SOS asks you to describe a large number of characteristics. None of these characteristics are optional. If you do not have this information, you need to put in dummy values as placeholders. Furthermore, SOS version 2.0 has a semantic web component that requires you to define what the metadata is for (semantic web resource). This may not be relevant in the context of your system, or if you do not desire to use the semantic web components. Additionally, it can be difficult to locate some of these semantic resources.

JSF attempts to alleviate some of these issues as the SWE API transforms the sensor data context of a deployment into the SWE encoding (SensorML, O&M, etc.). As this process is fully automated, it makes it easier to generate the encodings. All you have to do is pass the parameters (system name, process name, type of process) and the SWE API creates the respective SWE-compliant document. This ensures that data is properly formatted, data is not missing (or dummy values are created), and semantic resources are in place. This reduces the amount of work required and the number of potential errors when operating the system.

5. Conclusion

The authors have been involved with multiple real-world WSN projects including the SEMAT deployments at Deception Bay and Heron Island, the Digital Homestead deployment at Rowes Bay, and the Federation Place WSN. Each project differed in size, complexity, and application. Different sensors were involved (i.e., temperature, humidity, light, water pressure, and salinity) which provided different data streams. Furthermore, each project used different technologies to collect, log, transfer, and store the data. A system has been required to speed up the deployment process by automating common tasks, and a way to integrate, visualize, and analyze the data under a common web-based user interface to support the decision making processes of the end user.

This paper presented a system architecture based on the SWE framework that facilitates the integration and interoperability of sensor data from dispersed WSN systems. The authors proposed a middleware integration platform JSF, which provides integration between the WSN systems and the SWE framework. JSF manages adapted SWE specifications of an individual domain application to map, format, and store its sensor and sensed data, and provides the ability to extract sensor data from a domain WSN application. JSF transforms the data into the corresponding request doc-

ument using its SWE API implementation, formats the data (via a protocol binding), and then transfers the requested document to the SOS. JSF facilitates the discovery and exploitation of sensor data from dispersed WSNs, and reduces the amount of effort needed for developing a new WSN application. As presented solution provides a standardized framework, JSF can be reused for any type of WSN application.

The paper also presented an enhanced SOS implementation for JSF that provides support for web-based data visualization and statistical analysis using open source freely available technologies. Additional features were added to the existing SOS web service interfaces that integrates data from multiple sensors. The enhanced SOS provides functionality for analyzing the data and allows the data and the analysis result to be visualized graphically via a web-based user interface. A performance analysis was conducted to compare the impact that the additional functionality of the enhanced SOS has on the system compared to a regular SOS implementation. The performance analysis showed that the additional features have minimal impact on the system performance, where each additional transaction with 100 sensed data points increased its service response time by approximately 0.5 s. Also some lessons learned for implementing a WSN systems using SWE are briefly described. Some impediments/difficulties for integrating SWE are the initial learning curve, having to strictly adhere to SWE document formats, having to supply extraneous or unnecessary information, dealing with subtle errors in SOS due to incorrect message formats or missing data, and the tedious process of generating lengthy SWE-compliant documents.

Future work involves extending the SOS server with a semantically-enabled SOS server [22]. A significant issue with the SWE architecture is the lack of semantically rich discovery mechanisms. This makes it hard to explore related concepts, subgroups of sensor types, or other dependencies between the sensors and the data they collect. Integrating SOS with semantic technologies will enable the SOS server to query high-level knowledge of the environment as well as the raw sensor data. This can facilitate knowledge sharing and exchange, and automated processing of web resources.

References

[1] F. L. Lewis, "Wireless sensor networks", in *Smart Environments: Technology, Protocols and Applications*, D. Cook and S. Das, Eds. New York: Wiley, 2004, pp. 11–46.

[2] C. Srimathi, C. Eunmi, H. A. Jemal, and N. Rajesh, "Sensor grid middleware metamodeling and analysis", *Int. J. Distrib. Sensor Netw.*, pp. 1–12, 2014.

[3] M. Botts, G. Percivall, C. Reed, and J. Davidson, "OGC Sensor Web Enablement: Overview and High Level Architecture", in *GeoSensor Networks – 2nd Int. Conf. GSN 2006, Boston, MA, USA*, S. Nittel, A. Labrinidis, and A. Stefanidis, Eds. *LNCS*, vol. 4540, pp. 175–190. Springer, 2008.

[4] M. Botts and A. Robin, "Bringing the sensor web together", *Geosciences*, pp. 46–53, 2007.

[5] K. Delin and S. Jackson, "The sensor web: a new instrument concept", in *SPIE's Symposium on Integrated Optics*, San Jose, CA, USA, 2001.

[6] K. Delin, "Sensor webs in the wild", in *Wireless Sensor Networks: A Systems Perspective*, N. Bulusu and S. Jha, Eds. London: Artech House, 2005.

[7] P. Gibbons, "Irisnet: An architecture for a worldwide sensor web", *IEEE Pervasive Comput.*, vol. 2, no. 4, pp. 22–33, 2003.

[8] N. Markovic, A. Stanimirovic, and L. Stoimenov, "Sensor web for river water pollution monitoring and alert system", in *Proc. 12th AGILE Int. Conf. on Geographic Inform. Sci. AGILE 2009*, Hannover, Germany, 2009.

[9] D. Moodley and S. Ingo, "A new architecture for the sensor web: the SWAP framework", in *Semantic Sensor Netw. Worksh. A workshop of the 5th Int. Semantic Web Conf. ISWC 2006*, Athens, Georgia, USA, 2006.

[10] E. Torres-Martinez, P. Granville, M. Schoeberl, and M. Kalb, "A web of sensors: enabling the earth science vision", *Acta Astronautica*, vol. 53, no. 4–10, pp. 423–428, 2003.

[11] S. Schade *et al.*, "Citizen-based sensing of crisis events: sensor web enablement for volunteered geographic information", *Appl. Geomat.*, vol. 5, no. 1, pp. 3–18, 2013.

[12] S. Shafi, A. A. Reshi and A. Kumaravel, "Wireless sensor network based early warning and alert system for radioactive radiation leakage", *Middle-East J. Scient. Res.*, vol. 19, no. 12, pp. 1602–1608, 2014.

[13] S. M. Guru, P. Taylor, H. Neuhaus, Y. Shu, D. Smith, and A. Terhorst, "Hydrological sensor web for the South Esk catchment in the Tasmanian state of Australia", in *Proc. 4th Int. Conf. on eScience, eScience'08*, Indianapolis, IN, USA, 2008, pp. 432–433.

[14] C. Hu, Q. Guan, N. Chen, J. Li, X. Zhong, and Y. Han, "An observation capability metadata model for EO sensor discovery in sensor web enablement environments", *Remote Sens*, vol. 6, no. 11, pp. 10546–10570, 2014.

[15] S. Back, S. B. Kranzer, T. J. Heistracher, and T. J. Lampoltshammer, "Bridging SCADA systems and GI systems", in *Proc. IEEE World Forum on Internet of Things WF-IoT 2014*, Seoul, Korea (South), 2014, pp. 41–44.

[16] G. E. Churcher, G. Bilchev, J. Foley, R. Gedge, and T. Mizutani, "Experiences applying sensor web enablement to a practical telecare application", in *Proc. 3rd Int. Symp. on Wirel. Pervasive Comput. ISWPC 2008*, Santorini, Greece, 2008, pp. 138–142.

[17] G. E. Churcher and J. Foley, "Applying and extending sensor web enablement to a telecare sensor network architecture", in *Proc. 4th International ICST Conf. on COMmun. Syst. softWAre and middleWaRE*, Dublin, Ireland, 2009, p. 6, 2009.

[18] K. B. Lee and M. Reichardt, "Open standards for homeland security sensor networks", *Instrument. & Measur. Mag.*, vol. 8, no. 5, pp. 14–21, 2005.

[19] F. Samadzadegan, H. Zahmatkesh, M. Saber, and H. J. Ghazi khanlou, "An interoperable architecture for air pollution early warning system based on sensor web", *ISPRS – Int. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 1, no. 3, pp. 459–462, 2013.

[20] J. Trevathan, R. Johnstone, T. Chiffings, I. Atkinson, N. Bergmann, W. Read, S. Theiss, and T. Stevens, "SEMAT – The next generation of inexpensive marine environmental monitoring and measurement systems", *Sensors*, vol. 12, no. 7, pp. 9711–9748, 2012.

[21] Y. J. Lee, J. Trevathan, I. Atkinson, W. Read, and R. Johnstone, "The evolution of the SEMAT sensor network management system", in *Proc. 7th Int. Conf. on Intell. Sensors, Sensor Netw. and Inform. Process. ISSNIP 2011*, Adelaide, Australia, 2011, pp. 229–234.

[22] C. A. Henson, J. K. Pschorr, A. P. Sheth, and K. Thirunarayan, "SemSOS: Semantic Sensor Observation Service", in *Int. Symp. on Collabor. Technol. and Syst.*, Baltimore, MA, USA, 2009.

Yong Jin Lee completed his Ph.D. in Information Technology in 2015 at James Cook University. His research interests include wireless sensor networks, datamining, ecommerce security and fraud, and big data. He worked on the Smart Environmental Monitoring and Analysis Technologies project in conjunction with the University of Queensland. Dr. Lee now works in industry in South Korea as an analyst/programmer. His current projects involves big data analysis of sensor data.

E-mail: yong.lee1@jcu.edu.au
eResearch Centre
James Cook University
Townsville, Queensland, Australia

Jarrod Trevathan is a lecturer/researcher for the School of Information and Communication Technology at Griffith University. His research interests include ecommerce security and fraud, dynamic handwritten signature verification, phishing scams, and wireless sensor networks. He has worked in industry as an analyst/programmer and has an extensive history working with James Cook University and the University of Queensland. He took a lead role in the Smart Environmental Monitoring and Analysis Technologies project. Dr. Trevathan has published over 70 papers in conferences, journals and book chapters.

E-mail: j.trevathan@griffith.edu.au
School of Information and Communication Technology
Griffith University
Nathan, Queensland, Australia

Ian Atkinson is a Professor and the director of the eResearch Centre at James Cook University. His research interests include environmental monitoring, sensors, data management and supramolecular chemistry. He has worked in various professional/corporate IT roles and is the recipient of several major research grants, including the Smart Environmental Monitoring and Analysis Technologies and CSIRO Digital Homestead projects.

E-mail: ian.atkinson@jcu.edu.au
eResearch Centre
James Cook University
Townsville, Queensland, Australia

Wayne Read is an applied mathematician Associate Professor. He is a lecturer/researcher for the School of Engineering and Physical Sciences at James Cook University. His research interests include numerical and computational mathematics, mathematical modeling (particularly for ground water), and ecommerce security and fraud. He has worked in various roles at James Cook University, including Head of School for over five years. He has a long standing relationship with IT/Computer Science researchers in mentoring and providing applied mathematical solutions for algorithms and data structures.

E-mail: wayne.read@jcu.edu.au
School of Engineering and Physical Sciences
James Cook University
Townsville, Queensland, Australia

Paper

Lorentzian Operator for Angular Source Localization with Large Array

Youssef Khmou [1], Said Safi [1], and Miloud Frikel [2]

[1] *Department of Mathematics and Informatics, Beni Mellal, Morocco*
[2] *Greyc UMR 6072 CNRS, ENSICAEN, Caen, France*

Abstract—Source localization problem consists of an ensemble of techniques that are used to obtain spatial information of present radiation in given medium of propagation, with a constraint of the antenna geometry and the characteristics of radiating sources. This condition gives multitude of cases to study, hence several methods were proposed in the literature. In this paper, a new algorithm for estimating the Direction of Arrival (DoA) of narrowband and far field punctual sources is introduced. By exploiting the spectrum of covariance matrix of received data, the Lorentzian function on spectral matrix to filter the eigenvalues is applied. This filtering process eliminates the eigenvalues belonging to signal subspace. Parameters of Lorentz function are adjusted using first and second statistics of eigenvalues. The algorithm requires the knowledge of minimum eigenvalue and is performing when the dimension of antenna is relatively large which is confirmed by several Monte Carlo simulations.

Keywords—array processing, Direction of Arrival, narrowband, operator.

1. Introduction

In the context of array signal processing, source localization [1] refers to the techniques implemented to detect the location of present radiation in space. The origin of these radiations is often considered to be punctual sources due to far field assumption. The radius of propagation is larger than the maximum antenna dimension [2]. These techniques are valid for both electromagnetic and acoustic waves, thus a source can be cosmic, a cell phone, a seismic wave, sound in underwater and so on. Each source is characterized by its frequency, for example it can be narrowband [2] or wideband [3]. Due to this diversity, this field of research has attracted more interest due to its usefulness in many applications including radioastronomy [4], geolocalization such as Global Positioning System (GPS) [5], localization of mobile stations [6], radar and sonar [7] in both civilian and military applications, underwater acoustics [8], medical signal processing and seismology.

Most of localization techniques exploit the space-time diversity, some methods are based on time delay, known as Time of Arrival (ToA) [1]. This concept requires synchronization between the transmitters and the receivers. Other methods use the properties of propagating wavefront along the antenna to calculate the Angle of Arrival (AoA) of radiating sources [2]. This mechanism has the advantage of no requirement of synchronization. In fact, to compute the angle of single source, at least two aerials are needed and the distance between them is a function of wavelength of incoming wave. In case of multiple sources, the resulting wavefield is a superposition of each radiation. In this situation, an antenna with larger number of sensors must be used, thus the antenna beamwidth becomes narrower, which gives the ability to separate two sources with small angular difference.

In some cases, the problem of localization becomes difficult, for example, when some sources with different signal power are present, or the propagating signals have different carrier frequencies or when the source signals are highly correlated [9]. The preliminary solutions are based on preprocessing techniques [2] to decorrelate the waveforms. Additionally, some of these problems are caused by the transmission channel. During the propagation many phenomena can occur [10], for example a wave can be scattered, when it hits objects having dimensions smaller than the wavelength. This condition is known as Rayleigh scattering [10]. When a wavefield enters a medium with different electrical properties than the previous one, a refraction occurs [10]. Another type of deviation can happen when wave faces a smooth surface like metal, a reflection takes place with the Angle of Incidence (AoI) equals the angle of reflection in this situation.

The problem of localization depends on the environment and search dimensions, one dimensional scan focuses only on azimuth angle, this type requires only one dimensional arrays geometry. For two dimensional localization, it is mandatory to use two dimensional arrays such as circular [11], rectangular [2], L-shaped [12] and fractal arrays.

Concerning the mathematical aspects, some Direction of Arrival (DoA) techniques are based on extracting signals information from second order statistics [2]–[4], in the other hand alternative approaches use high order statistics [13]. Covariance based methods (also called spectral matrix or cross correlation matrix) can be divided into beamforming and eigendecomposition techniques [2], subspace based techniques have resolution power that is able to locate sources under angular limit resolution of array. These techniques use several spectral decomposition which are eigen-

decomposition [2], QR or LU factorization and Singular Value Decomposition (SVD) [14].

Recent researches are focused on enhancing the eigen-based approaches, when external factors impact the electrical properties of sensors, such as temperature, humidity, pressure and vibration. These variables generate coupling effects between sensors, which degrade the performance of DoA methods [15]. Another type of ongoing researches offer a consistency of spectral techniques when the system dimensions become larger [16], precisely number of sensors and number of acquired samples. In the other hand, as the dimensions tend to infinity, a computational complexity increases, recent solution implements a non-regular sampling of impinging signals, this method is known as compressed/compressive sensing [17].

In this paper, authors introduce a new algorithm for localizing narrowband sources. Using the order of the covariance matrix spectrum, a new operator that performs a filtering operation on eigenvalues to isolate the noise subspace is introduced which is orthogonal to signal subspace. From mathematical definition this operator is the Lorentz function of Hermitian matrix, also known as Cauchy distribution [18]. This function needs adjustment of two parameters, the index of the function's peak, which corresponds to the smallest eigenvalue and the width that is related to a threshold between signal and noise eigenvalues. The authors use a theorem that offers a bound of minimum eigenvalue using first and second order statistics of spectral eigenvalues. The obtained bound is efficient, when the antenna contains large number of sensors comparatively to number of sources.

The proposed operator is validated through several Monte Carlo simulations along with other techniques.

In the Section 2 of this paper, the statistical signal model for DoA problem is described. In the Section 3, the author's contribution is presented and in the Section 4 some computer simulation results for performance analysis is shown.

2. Statistical Data Model

Let us consider a geometry given Fig. 1 which consists of one punctual radiating source and a uniform linear array of N sensors placed along y axis, the system source antennas are placed in the same horizontal plane $(x, y, z = cte)$.

The sensors are located in the farfield region relative to the source where the wavefront arriving are considered plane waves. The uniform distance between the sensors is half the wavelength of the emitting source $d = \lambda/2$ and the farfield condition implies that the Line of Sight (LoS) r_0 is much larger than the length of the array $r_0 \gg L_\lambda = (N-1)d$. The propagation model [21] is given by the equation:

$$\vec{\nabla}^2 \vec{E}_i(\vec{r},t) = \frac{1}{c^2}\frac{\partial^2 \vec{E}_i(\vec{r},t)}{\partial t^2}, \tag{1}$$

where c denotes the velocity of propagation $c = (\mu\varepsilon)^{\frac{-1}{2}}$ and $\vec{\nabla}^2$ is Laplace operator $\vec{\nabla}^2 = \Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$. By

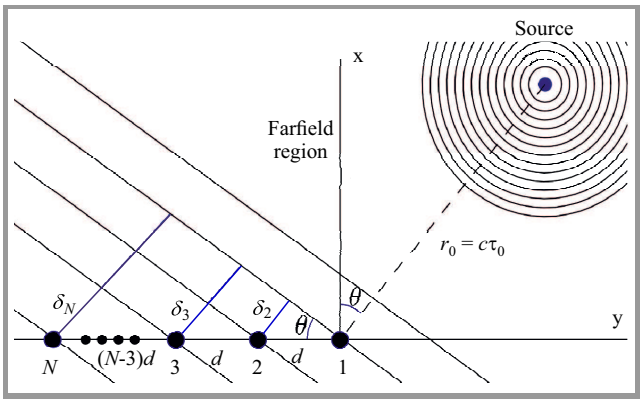


Fig. 1. A farfield punctual source emitting radiations received by a uniform linear array of sensors with azimuth angle θ relative to the reference.

assuming the transverse mode of propagation by considering the z component of the wave vector $\vec{E}_i = (E_x = 0, E_y = 0, E_z \neq 0)$, the solution of the i-th source is given by:

$$E_z(\vec{r},t) \simeq s_i(t) e^{j(\omega t - \vec{k}_i.\vec{r})}. \tag{2}$$

The solution is only an approximation because the variation of the function $s_i(t)$ is temporally negligible than the oscillation of the carrier wave with frequency $\omega = 2\pi f_c$ where $c = \lambda f_c$ [21]. This is also known as Slowly Varying Envelope Approximation (SVEA) in other fields. The $\vec{k}_i$ is the wave vector having the components in spherical coordinates as:

$$\vec{k}_i = \frac{-2\pi}{\lambda_i}\begin{pmatrix} \sin\varphi_i \cos\theta_i \\ \sin\varphi_i \sin\theta_i \\ \cos\varphi_i \end{pmatrix}, \tag{3}$$

where (θ_i, φ_i) are the azimuth and elevation of the i-th source respectively. The received wavefront is a superposition of all existing sources and the magnitudes of the collected signals are proportional to $\sum_{i=1}^{P} \vec{E}_i$. During the acquisition time $T = T_s K$, where T_s is the sampling period and K is the number of measurements, the carrier frequencies [21] terms are removed from the signals $e^{j\omega t}$, so for any measurement at instant $t \in \{1, \ldots, K\}$, the signal at the m-th sensor with position $\vec{r}_m$, is given by:

$$x_m(t) = \sum_{i=1}^{P} s_i(t) e^{-j\vec{k}_i.\vec{r}_m} + n_m(t), \tag{4}$$

where $n_m(t)$ is the additive noise at the m-th sensor considered complex and random process with zero mean. While considering the uniform linear array (ULA), the complex vector of signals at instant t is:

$$x(t) = A(\theta)s(t) + n(t), \tag{5}$$

with $s(t) \in \mathbb{C}^{P\times 1}$ is the source waveforms, $n(t) \in \mathbb{C}^{N\times 1}$ is the noise waveforms and $A(\theta) \in \mathbb{C}^{N\times P}$ is the steering matrix given by:

$$A = \begin{pmatrix} 1 & \dots & 1 \\ e^{-j\mu_1} & \dots & e^{-j\mu_P} \\ \dots & \dots & \dots \\ e^{-j(N-1)\mu_1} & \dots & e^{-j(N-1)\mu_P} \end{pmatrix}, \quad (6)$$

where $\mu_i = 2\pi d\lambda^{-1}\sin(\theta_i)$ and θ_i is the AoA of the i-th punctual source. The rank of the steering matrix A in Eq. (6) is P such that the P sources are located in different angular positions θ_i. In a compact form, the matrix of received signals is $X(t) = A(\theta)S(t) + N(t)$ with the dimensions $X(t) \in \mathbb{C}^{N\times K}$, $S(t) \in \mathbb{C}^{P\times K}$ and $N(t) \in \mathbb{C}^{N\times K}$. The objective is to calculate a localization function $f(\theta)$ from which positions θ_i can be derived. Most of high resolution DoA techniques are based on second order statistics of $X(t)$, the spectral matrix $< x(t)x^+(t) >$ has the following theoretical expression:

$$\Gamma = \Gamma_s + \Gamma_n, \quad (7)$$

where $\Gamma_s = A\Gamma_{ss}A^+$, $\Gamma_n = \sigma^2 I_N$ and $\Gamma_{ss} =< s(t)s^+(t) >$ is the correlation matrix of sources, if the waveforms are not correlated, then $[\Gamma_{ss}]_{ij} = \delta_{ij}\sigma_i^2$, where σ_i^2 is the power of the i-th signal. In decreasing order, the spectrum of matrix Γ is given as:

$$\sigma_\Gamma = \left\{\lambda_1 \geqslant \lambda_2 \geqslant \dots \geqslant \lambda_P > \lambda_{P+1} \simeq \dots \simeq \lambda_N = \sigma^2\right\}.$$

For example one of the high resolution DoA operators [19] exploits the threshold between the smallest signal eigenvalue λ_P and the largest noise eigenvalue λ_{P+1} to calculate the projector in the noise subspace. This latter is obtained by spectral decomposition, indeed the spectral matrix is decomposed as $\Gamma = U\Lambda U^+$, where Λ is a diagonal matrix of eigenvalues and $U \in \mathbb{C}^{N\times N}$ is the orthonormal matrix such as $\|U\|_F = \sqrt{Tr\{UU^+\}} = \sqrt{N}$. The first P columns of U correspond to the largest eigenvalues to form a base of signal subspace $U_s \in \mathbb{C}^{N\times P}$ and the remaining $N-P$ columns form a noise subspace $U_n = [u_{P+1}, \dots, u_N]$. The projector into the noise subspace $P_n \in \mathbb{C}^{N\times N}$ is defined by the relation $P_n = U_n U_n^+$. For given steering vector $a(\theta)$ with testing angle $\theta \in \Omega = [\theta_{min}, \theta_{max}]$, the localization function verifies

$$f(\theta) = a^+(\theta)P_n a(\theta) = \begin{cases} 0 & \text{if } \theta \text{ is DoA,} \\ \neq 0 & \text{otherwise.} \end{cases} \quad (8)$$

After performing an angular scan in the region Ω, the indexes of the peaks of $f(\theta)$ indicate the angles of arrival of radiating sources.

3. Lorentzian DoA Algorithm

For real variable $x \in \mathbb{R}$, the single peak normalized Lorentz function centered at x_0 is defined by [18]:

$$f(x) = \frac{1}{\pi}\frac{\beta}{(x-x_0)^2 + \beta^2}, \quad (9)$$

with parameters (x_0, β) such that:

$$\int_{\mathbb{R}} f(x)dx = 1.$$

At the abscissa x_0, the function has maximum value of $f(x_0) = \frac{1}{\beta\pi}$ and equals half maximum at $x = x_0 \pm \beta$, which makes the Full Width at Half Maximum (FWHM) to be 2β, the inflection points x_c occur when the second derivative is null:

$$\frac{\partial^2 f(x)}{\partial x^2} = \frac{-2\beta}{\pi((x-x_0)^2+\beta^2)^2} + \frac{8\beta(x-x_0)^2}{\pi((x-x_0)^2+\beta^2)^3} = 0. \quad (10)$$

This equation has a solution of $x_c = x_0 \pm \frac{\beta}{\sqrt{3}}$, and $f(x_c) = \frac{3}{4\pi\beta}$. Remark that at the inflection points the magnitude is reduced, comparatively to the maximum value, by a factor of 0.75.

Given the condition that the spectral matrix Γ is positive definite, then $\sigma_\Gamma \in \mathbb{R}^+$. Let us denote $\lambda \in \mathbb{R}^+$ the scalar function representing the eigenvalues and $\lambda_{\min}$ its lowest value, the spectrum is considered to be binary $\{\lambda_n \simeq \lambda_{\min}, \lambda_s\}$. We search for function that normalizes the eigenvalue $\lambda_{\min}$ and forces any signal eigenvalue λ_s to zero, for this purpose the following version of Lorentz function is used:

$$f(\lambda) = \frac{1}{\alpha(\lambda - \lambda_{\min})^2 + 1}, \quad (11)$$

where α is the scaling parameter of the width, this principle is illustrated in Fig. 2. From the Eq. (11), we need to

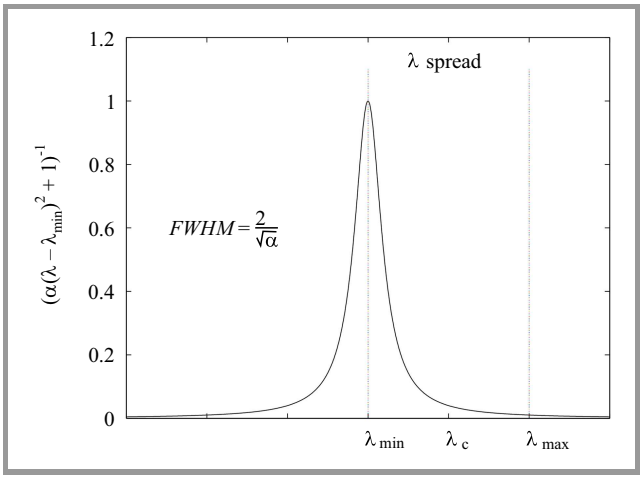


Fig. 2. Lorentzian function with parameters $\{\alpha, \lambda_{\min}\}$ applied to the spectrum of operator Γ.

calculate two parameters. The minimum eigenvalue can be estimated using the power method, first the largest eigenvalue $\lambda_{\max} = \lambda_1$ is computed, next the $\lambda_{\min}$ is calculated using the condition number $\tau = \lambda_{\max}/\lambda_{\min}$. The random vector $\phi \in \mathbb{C}^{N\times 1}$ with norm $||\phi||_\infty = 1$ is chosen, and for $m \geq 2$ the following iterations are performed:

$$\begin{aligned} \phi_{m+1} &= \Gamma\phi_m, \\ \mu_m &= \phi_m^+\phi_{m+1}, \\ \phi_m &= \frac{\phi_{m+1}}{\phi_{m+1}^+\phi_{m+1}}. \end{aligned} \quad (12)$$

When $\mu_m \to \lambda_{\max}$, the minimum eigenvalue is calculated by the following equation:

$$\lambda_{\min} = \frac{\lambda_{\max}}{\tau} = \frac{\lambda_{\max}}{||\Gamma||_2 ||\Gamma^{-1}||_2}. \tag{13}$$

The inversion of spectral matrix is required, the scaling parameter α is necessarily related to a threshold λ_c that differentiates the two subsets $\{\lambda_s, \lambda_n\}$. The authors impose the condition that at abscissa λ_c, the function equals the value $\varepsilon = 10^{-3}$, this is equivalent to $\alpha(\lambda_c - \lambda_{min})^2 \simeq \varepsilon^{-1}$, consequently the chosen scaling parameter is given by:

$$\alpha = \frac{\varepsilon^{-1}}{(\lambda_c - \lambda_{min})^2} = \frac{10^3}{(\lambda_c - \lambda_{min})^2}. \tag{14}$$

The threshold λ_c is proposed as the bound of minimum eigenvalue $\lambda_{\min}$, this theoretical bound can be calculated using only the trace of spectral matrix. The theorem of the smallest eigenvalue bounds [20] is based on mean and standard deviation of σ_Γ, before announcing the theorem, the following variables are defined:

$$< \lambda >= \frac{tr(\Gamma)}{N} = \frac{1}{N}\sum_{i=1}^{N} \Gamma_{ii}, \tag{15}$$

$$\Delta\lambda = \sqrt{< \lambda^2 > - < \lambda >^2} = \sqrt{\frac{tr(\Gamma^2)}{N} - \left(\frac{tr(\Gamma)}{N}\right)^2}. \tag{16}$$

Using these two statistics and for matrix with real eigenvalues, the bounds for smallest and largest eigenvalues are given by the Theorem 1.

Theorem 1: For Hermitian matrix $\Gamma \in \mathbb{C}^{N\times N}$ ($\Gamma^+ = \Gamma$), the extremum eigenvalues are bounded by:

$$\left(< \lambda > - \Delta\lambda\sqrt{N-1}\right) \le \lambda_{\min} \le \left(< \lambda > - \frac{\Delta\lambda}{\sqrt{N-1}}\right), \tag{17}$$

$$\left(< \lambda > + \frac{\Delta\lambda}{\sqrt{N-1}}\right) \le \lambda_{\max} \le \left(< \lambda > + \Delta\lambda\sqrt{N-1}\right). \tag{18}$$

The proof is presented in [20]. The smallest signal eigenvalue satisfies $\lambda_s > \lambda_{\min}$, for a relatively large array ($N > 2P$). The proposed threshold is given by:

$$\lambda_c = < \lambda > - \frac{\Delta\lambda}{\sqrt{N-1}}. \tag{19}$$

Note: While having only two sensors $\Gamma \in \mathbb{C}^{2\times 2}$ and eventually single source the eigenvalues are exactly $\lambda_{\min} = < \lambda > - \Delta\lambda$, $\lambda_{\max} = < \lambda > + \Delta\lambda$. This special case can be used in the presence of single source $P = 1$ and N sensors where the $N/2$ spectral matrices Γ_i for $i = 1, \ldots, N/2$ can be calculated, and theirs eigenvalues using the above equations are computed.

After describing the theoretical expressions for the couple $\{\alpha, \lambda_{min}\}$, the Lorentz function is now adaptive to the variation of parameters describing the physical system. The application of the proposed function on self adjoint operator Γ acts on its eigenvalues $f(\Gamma) = Uf(\Lambda)U^+$, then we have the following result.

Proposition: Given Hermitian matrix $\Gamma = K^{-1}XX^+$, The Lorentz operator defined by:

$$f(\Gamma) = \left(\alpha(\Gamma - H)^2 + I_N\right)^{-1} \tag{20}$$

is an approximation to the projector into the noise subspace, with $H = \lambda_{\min} I_N$, $\alpha = \frac{10^3}{(\lambda_c - \lambda_{\min})^2}$ and $\lambda_c = < \lambda > - \frac{\Delta\lambda}{\sqrt{N-1}}$. Indeed, developing the above equation, based on the relation $f(\Gamma) = Uf(\Lambda)U^+$, yields to the decomposition:

$$f = U\frac{I_N}{\alpha(\Lambda - \Gamma_0)^2 + I_N}U^+ = \sum_{g=1}^{N} \frac{u_g u_g^+}{\alpha(\lambda_g - \lambda_{\min})^2 + 1}$$

$$= \sum_{i=1}^{P} \frac{u_i u_i^+}{\alpha(\lambda_i - \lambda_{\min})^2 + 1} + \sum_{j=P+1}^{N} \frac{u_j u_j^+}{\alpha(\lambda_j - \lambda_{\min})^2 + 1}$$

$$\simeq \sum_{j=P+1}^{N} u_j u_j^+ \simeq P_n. \tag{21}$$

From numerical experiments, the obtained operator is not an exact a projector because either the noise eigenvalues are not normalized and have some fluctuating errors (example $f(\lambda_n) = 0.98$), or the signal eigenvalues are not totally annihilated (e.g. $f(\lambda_s) = 10^{-3}$). The Algorithm 1 summarizes the proposed method.

Algorithm 1: Lorentzian operator algorithm

Input: $\Gamma \in \mathbb{C}^{N\times N}$ ($N > P$).

1. Compute $\lambda_{\min}$ using power method for example.
2. Compute statistics of operator Γ
   $m = \frac{Tr(\Gamma)}{N}$ and $s = \sqrt{\frac{Tr(\Gamma^2)}{N} - m^2}$.
3. Compute parameters $\lambda_c = m - \frac{s}{\sqrt{N-1}}$,
   $\alpha = \frac{10^3}{(\lambda_c - \lambda_{\min})^2}$ and $H = \lambda_{\min} I_N$.
4. Calculate $P_n = (\alpha(\Gamma - H)^2 + I_N)^{-1}$.

4. Results and Discussion

4.1. Simulation Results

In this section, some computer simulations using a configuration of Uniform Linear Array of $N = 11$ sensors considered isotropic and identical are performed. The available range for this type of array is $\Omega = [\frac{-\pi}{2}, \frac{\pi}{2}]$.

The distance between the sensors is half the wavelength of the carrier waves. It is assumed the presence of $P = 4$ narrowband and far field sources impinging on array from directions $-80°$, $15°$, $20°$ and $56°$, the number of samples is set to $K = 200$. The signals are chosen to be ergodic

complex and zero mean random processes with uniform power of 1 W. In all the tests, the perturbative noise is additive and also complex zero mean random process uncorrelated between the sensors and independent of $s(t)$. The noise power is derived from $SNR = 20\log(\frac{1}{\sigma})$. Figure 3 represents an average of $L = 100$ Monte Carlo trials of Lorentzian localization function, with $N = 11$, $P = 4$, $\theta = [-80^\circ, 15^\circ, 20^\circ, 56^\circ]$, $K = 200$, $d = \lambda/2$, $s(t) \sim \mathscr{CN}(0, I_4)$ and $SNR = 5$ dB.

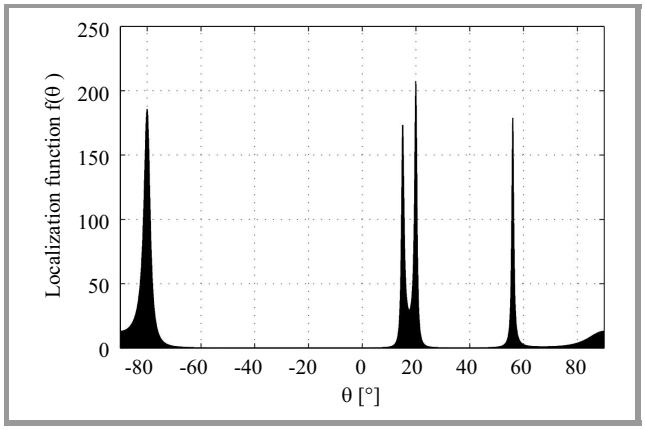


Fig. 3. Average of 100 trials of the proposed operator.

The obtained result proves that the function has the ability to separate the sources. In the second test, the proposed operator with several spectra is compared. The authors choose a critical situation where the transmitted signals and noise signals are equipowered $SNR = 0$ dB. Figure 4 presents an average of $L = 100$ trials of Lorentzian function against the first three spectra, which are based on subspace computation with $L = 100$ trials, $N = 11$, $P = 4$, $\theta = [-80^\circ, 15^\circ, 20^\circ, 56^\circ]$, $K = 200$, $d = \lambda/2$ and $s(t) \sim \mathscr{CN}(0, I_4)$ and $SNR = 0$ dB.

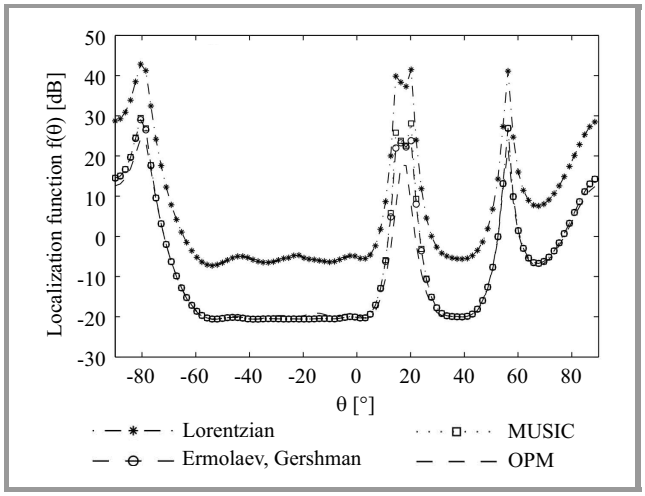


Fig. 4. Lorentzian operator against three different spectra.

Lorentzian spectrum is compared with Multiple Signal Classification (MUSIC) method [23], Ermolaev and Gershman subspace [19] with parameter $m = 10$, and orthonormalized propagator [24]. Schmidt's method and Ermolaev subspace are identical in this case. They are successful in locating all angle indexes, the OPM could locate the farthest source at 80° but did not separate sources at 15° and 20°. Lorentzian spectrum identifies all the AoAs with higher magnitudes.

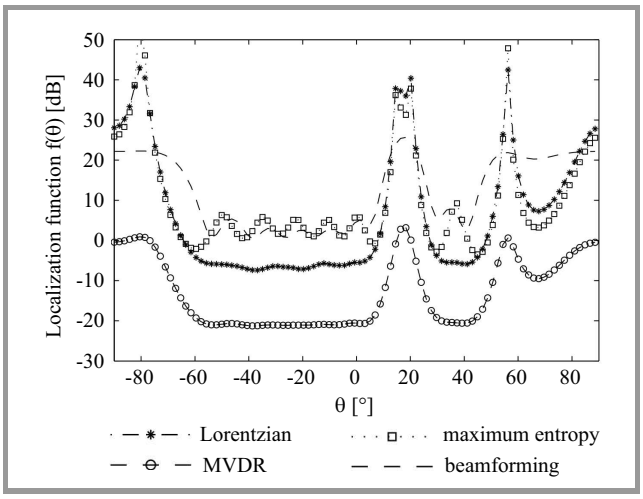


Fig. 5. Lorentzian operator against three different spectra.

In Fig. 5 the presented function is compared with three other techniques (with the same conditions as in Fig. 4). The authors realize that Capon's method (MVDR) [2]–[10] failed to separate sources at 15°, 20° and the shape of its localization function is similar to the beamforming [2] in these conditions. Maximum entropy method [10] with parameter $l = 1$ and Lorentzian are approximately the same.

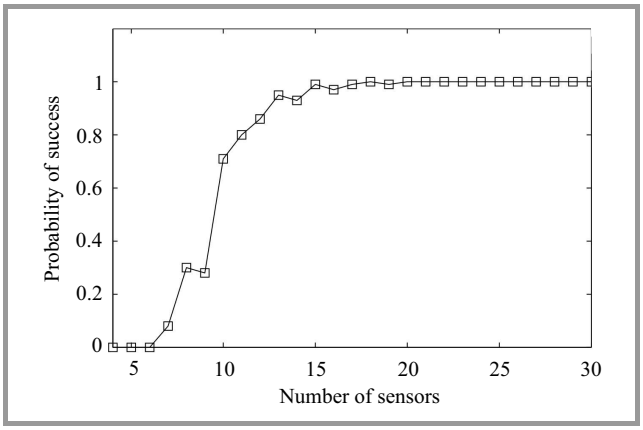


Fig. 6. Probability of success of Lorentzian operator.

To evaluate the performance with respect to the dimension of the antenna in Fig. 6 the probability of success over $L = 100$ trials with varying number of sensors starting from $N = 4$ is presented, with $L = 100$ trials, $P = 4$, $\theta = [-80^\circ, 15^\circ, 20^\circ, 56^\circ]$, $K = 200$, $d = \lambda/2$, $s(t) \sim \mathscr{CN}(0, I_4)$ and $SNR = 5$ dB. The result informs that the bound λ_c in Eq. (27) is not valid unless the number of sensors N is about triple the number of sources P, in fact the probability of detection reaches 90% when $N = 12$.

4.2. Experimental Results

In this second part of performance evaluation, the resolution power of the proposed operator using underwater acoustical data obtained from linear array of hydrophones [25] is tested. The received echoes are generated by two acoustic sources. The Table 1 summarizes the data description.

Table 1
Description of experimental underwater acoustic data

Data	Value
Number of hydrophones	$N = 6$
Interelement spacing	$d = 0.9$ m
Length of the array	$L_\lambda = 4.5$ m
Number of samples	$K = 4096$
Sources wavelength	$\lambda = 5.32$ m
Average power of data $X(t)$	$Tr(\Gamma)/N = 0.99$ W
Eigenvalues of Γ	[4.3648, 1.4835, 0.1225, 0.0220, 0.0051, 0.0007]
Number of sources	$P = 2$
Angular step	$d\theta = 0.1^\circ$ in the range $[-\pi/2, \pi/2]$
Estimated noise power	$\sigma^2 \simeq 0.0376$ W
Estimated powers of sources	$\sigma_1^2 \simeq 0.7188$ W and $\sigma_2^2 \simeq 0.3092$ W
Estimated signal to noise ratios	$SNR1 \simeq 25.62$ dB and $SNR2 \simeq 18.30$ dB

The noise power or minimum eigenvalue is computed using the equation:

$$\sigma^2 = \frac{1}{4}\sum_{j=3}^{6} \lambda_j \,. \tag{22}$$

The powers of sources are calculated using the beamforming as:

$$f_{BF}(\theta) = \frac{1}{N^2} a^+(\theta)\Gamma a(\theta)\,, \tag{23}$$

where the values of two largest peaks are approximately equal to the powers of sources σ_1^2 and σ_2^2. The implemented steering vector $a(\theta) \in \mathbb{C}^{6\times 1}$ is defined by the relations:

$$\begin{cases} a(\theta) = e^{-2\pi j r^T \lambda^{-1}\sin(\theta)} \\ r \quad\;\; = [0.00, 0.90, 1.80, 2.70, 3.60, 4.50] \end{cases}.$$

The seven DoA spectral techniques are applied to identify the locations of the acoustic sources which are the Lorentzian operator, MUSIC projector [23], Orthonormal Propagator (OPM) [24], Ermolaev and Gershman operator [19], Maximum Entropy Method (MEM) [10], where the operator is computed using the fourth column of Γ^{-1}, Minimum Variance Distortionless Response operator (MVDR) [2]–[10] and partial propagator method (PAR) [26]. Figure 7 presents the obtained results [25].

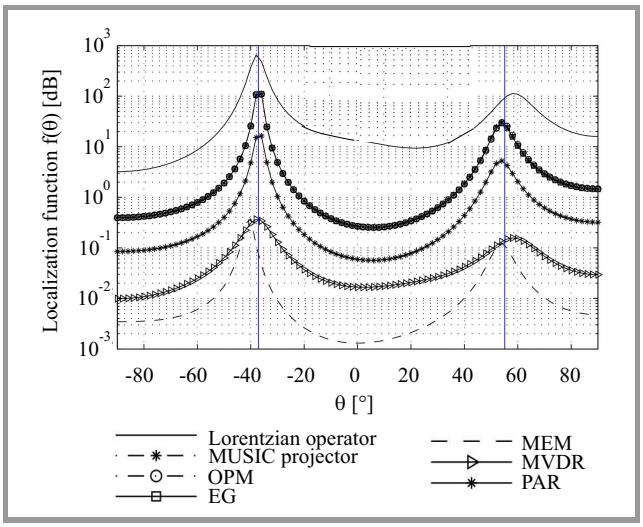


Fig. 7. DoA localization functions $f(\theta)$ of acoustic sources from experimental data.

The Lorentzian localization function has the highest values of peaks and the majority of the spectra present some deviations of locations. To quantify these fluctuations, in Table 2 the estimated DoAs using peak detection algorithm are presented.

Table 2
Estimated DoAs of acoustic sources

Spectral technique	Source 1 θ_1 [°]	Source 2 θ_2 [°]
Lorentzian	–37.70	58.60
MUSIC	–37.00	54.00
OPM	–37.00	54.20
EG	–37.00	54.00
MEM	–40.80	53.80
MVDR	–37.50	58.20
PAR	–36.90	53.60
Mean values	–37.70	55.20

The three subspace techniques MUSIC, OPM and EG operators identify the acoustic sources with same values of θ_1 and θ_2, the MVDR and Lorentzian functions present the same result where the angular position of the second source is different than the result of the first three subspace techniques by 4°. This difference is reduced for the partial propagator method where $\theta_2 = 53.60^\circ$. The Maximum Entropy Method is efficient if the fourth column is chosen as reference, however the peak of the first source is deviated by approximately 3°.

5. Research Perspectives

The proposed approach for DoA problem is based on bandpass filter using single shaped Lorentzian function. The similar solution was proposed using Gaussian function with

exponential operator [27]. Another solution consists of using Heaviside function as low pass filter of eigenvalues, one of the approximations is given by the equation:

$$f(\lambda) = 1 - \frac{1}{e^{-\alpha(\lambda-\lambda_c)}+1}, \tag{24}$$

where λ_c is the threshold such as $\lambda_{P+1} < \lambda_c < \lambda_P$ and α is a constant that controls the rate of decay as illustrated in Fig. 8 where larger value gives fast transition.

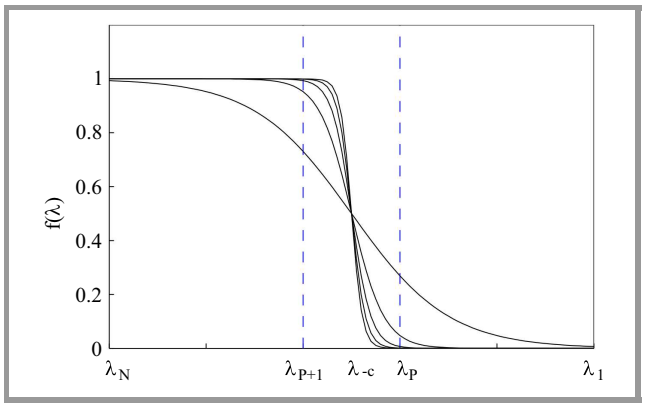


Fig. 8. Approximation of rectangular function $f(\lambda)$ applied to eigenvalues of spectral matrix Γ with parameters λ_c, $\alpha = \{1, 3, 5, 7, 9\}$.

As perspective, a theoretical value of threshold λ_c and fast approximation of exponential operator $e^{-\alpha(\Gamma-\lambda_c I_N)}$ may provide accurate results comparatively to conventional DoA spectra.

6. Conclusion

In this paper, a new high resolution algorithm for narrowband source localization problem using large array is proposed. The main idea consisted of applying Lorentz function on spectral matrix of received data such as low pass filter, where the cut-off value is the threshold between signal and noise eigenvalues, this mechanism requires a priori knowledge of minimum eigenvalue, which is the index of function's peak. Theoretical threshold and scaling parameter of Lorentz function were derived using first and second order statistics of eigenvalues using only the trace.
Several computer simulations demonstrated the resolution power of the proposed algorithm when the dimension of the antenna is relatively large.

References

[1] E. Xu, Z. Ding, and S. Dasgupta, "Source localization in wireless sensor networks from signal time-of-arrival measurements", *IEEE Trans. Sig. Process.*, vol. 59, no. 6, pp. 2887–2897, 2011.

[2] Z. Chen, G. Gokeda, and Y. Yu, *Introduction to Direction-of-Arrival Estimation*. Norwood, MA, USA: Artech House, 2010.

[3] Y.-S. Yoon, L. M. Kaplan, and J. H. McClellan, "TOPS: new DOA estimator for wideband signals", Signal Processing, *IEEE Trans. Sig. Process.*, vol. 54, no. 6, pp. 1977–1989, 2006.

[4] R. Feliachi, "Spatial processing of cyclostationary interferers for phased array radio telescopes", Ph.D. thesis, Université d'Orléans, Orléans, France, 2010.

[5] B. Yang, F. He, J. Jin, H. Xiong, and G. Xu, "DOA estimation for attitude determination on communication satellites", *Chinese J. Aeronautics*, vol. 27, no. 3, pp. 670–677, 2014.

[6] Y.-H. Ko, Y.-J. Kim, H.-I. Yoo, W.-Y. Yang, and Y.-S. Cho, "DoA estimation with cell searching for mobile relay stations with uniform circular array", in *Proc. IEEE 20th Int. Symp. Person., Indoor Mob. Radio Commun.*, Tokio, Japan, 2009, pp. 993–997.

[7] M. Jiang, J. Huang, W. Han, and F. Chu, "Research on target DOA estimation method using MIMO sonar", in *Proc. 4th IEEE Conf. on Indust. Elec. Appl. ICIEA 2009*, Xi'an, China, 2009, pp. 1982–1984.

[8] C. Shao-hua, Z. Wei, and L. Hui-bin, "Improved DOA estimation of underwater target with acoustic cross array", in *Proc. IEEE 11th Int. Conf. Sig. Process. ICSP 2012*, Beijing, China, 2012, vol. 3, pp. 2071–2074.

[9] I.-K. Rhee and H.-S. Kim, "Improved DOA estimation of correlated signals in correlated antenna noises environment", in *Proc. Int. Conf. Inform. Netw. ICOIN 2013*, Bangkok, Thailand, 2013, pp. 66–70.

[10] F. B. Gross, *Smart Antennas for Wireless Communications with Matlab*. New York, NY, USA: McGraw-Hill Professional, 2005.

[11] P. Tan, P. Wang, Y. Luo, Y. Zhang, and H. Ma, "Study of 2D DOA estimation for uniform circular array in wireless location system", *Int. J. Comp. Netw. Inform. Secur.* (*IJCNIS*), vol. 2, no. 2, pp. 54–60, 2010.

[12] L. Liu, Q. Ji, and Y. Jiang, "Improved Fast DOA Estimation Based on Propagator Method", in *Proc. APSIPA Ann. Summit and Conf. APSIPA ASC 2011*, Xi'an, China, 2011.

[13] Y.-H. Chen and Y.-S. Lin, "Fourth-order cumulant matrices for DOA estimation", *IEE Proc. Radar, Sonar and Navig.*, vol. 141, no. 3, pp. 144–148, 1994.

[14] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques", *IEEE Trans. Acoust., Speech and Sig. Process.*, vol. 37, no. 7, pp. 984–995, 1989.

[15] J. Dai, W. Xu, and D. Zhao, "Real-valued DOA estimation for uniform linear array with unknown mutual coupling", *Sig. Process.*, vol. 92, no. 9, 2012.

[16] X. Mestre and M.-A. Lagunas,"Modified subspace algorithms for DoA Estimation with large arrays", *IEEE Trans. Sig. Process.*, vol. 56, no. 2, pp. 598–614, 2008. doi: 10.1109/TSP.2007.907884.

[17] Y. Wang, G. Leus, and A. Pandharipande, "Direction estimation using compressive sampling array processing", in *Proc. IEEE/SP 15th Worksh. on Statis. Sig. Process. SSP 2009*, Cardiff, UK, 2009, pp. 626–629. doi: 10.1109/SSP.2009.5278497.

[18] S. A. Clough and F. X. Kneizys, "Convolution algorithm for the Lorentz function", *Applied Optics*, vol. 18, no. 13, pp. 2329–2333, 1979.

[19] V. T. Ermolaev and A. B. Gershman, "Fast algorithm for minimum-norm direction-of-arrival estimation", *IEEE Trans. Sig. Process.*, vol. 42, no. 9, pp. 2389–2394, 1994. doi: 10.1109/78.317860.

[20] H. Wolkowicz and G. P. H. Styan, "Bounds for eigenvalues using traces", *Linear Algebra and its Applications*, vol. 29, pp. 471–506, 1980.

[21] H. Krim and M. Viberg,, "Two decades of array signal processing research: the parametric approach", *IEEE Sig. Process. Mag.*, vol. 13, no. 4, pp. 67–94, 1996.

[22] Y. Khmou, S. Safi, and M. Frikel, "Comparative study between several direction of arrival estimation methods", *J. Telecommun. Inform. Technol.*, no. 1, pp. 41–48, 2014.

[23] R. O. Schmidt, "Multiple emitter location and signal parameter estimation", *IEEE Trans. Antenn. Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[24] S. Marcos, A. Marsal, and M. Benidir, "The propagator method for source bearing estimation", *Sig. Process.*, vol. 42, no. 2, pp. 121–138, 1995.

[25] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Upper Saddle River, NY, USA: Prentice-Hall, 2005.

[26] J. Chen, Y. Wu, H. Cao, and H. Wang, "Fast algorithm for DOA estimation with partial covariance matrix and without eigendecomposition", *J. Sig. Inform. Process.*, vol. 2 no. 4, pp. 266–269, 2011.

[27] Y. Khmou, S. Safi, and M. Frikel, "Exponential operator for bearing estimation", *Int. J. Adv. Sci. Technol.* (*IJAST*), vol. 74, pp. 1–10, 2015.

Youssef Khmou obtained the B.Sc. degree in Physics and M.Sc. degree from polydisciplinary faculty, in 2010 and from Faculty of Science and Technics Beni Mellal, Morocco, in 2012, respectively. Now he is Ph.D. student and his research interests include statistical signal and array processing and statistical physics.
Email: khmou.y@gmail.com
Department of Mathematics and Informatics
Beni Mellal, Morocco

Said Safi received the B.Sc. degree in Physics (option Electronics) from Cadi Ayyad University, Marrakech, Morocco in 1995, M.Sc. degree from Chouaib Doukkali University and Cadi Ayyad Universrsity, in 1997 and 2002, respectively. He has been a Professor of information theory and telecommunication systems at the National School for applied Sciences, Tangier, Morocco, from 2003 to 2005. Since 2006, he is a Professor of applied mathematics and programming at the Faculty of Science and Technics, Beni Mellal, Morocco. In 2008 he received the Ph.D. degree in Telecommunication and Informatics from the Cadi Ayyad University. His general interests span the areas of communications and signal processing, estimation, time-series analysis, and system identification – subjects on which he has published 35 journal papers and more than 70 conference papers. Current research topics focus on transmitter and receiver diversity techniques for single- and multi-user fading communication channels, and wide-band wireless communication systems.
E-mail: safi.said@gmail.com
Department of Mathematics and Informatics
Beni Mellal, Morocco

Miloud Frikel received his Ph.D. degree from the center of mathematics and scientific computation CNRS URA 2053, France, in array processing. Currently, he is with the GREYC laboratory (CNRS URA 6072) and the ENSICAEN as Assistant Professor. From 1998 to 2003, Dr. Frikel was with the Signal Processing Lab, Institute for Systems and Robotics, Institute Superior Tecnico, Lisbon, as a researcher in the field of wireless location and statistical array processing, after been a research engineer in a software company in Munich, Germany. He worked in the Institute for Circuit and Signal Processing of the Technical University of Munich. His research interests span several areas, including statistical signal and array processing, cellular geolocation (wireless location), space-time coding, direction finding and source localization, blind channel identification for wireless communication systems, and MC-CDMA systems.
E-mail: mfrikel@greyc.ensicaen.fr
GREYC UMR 6072 CNRS
Ecole Nationale Supérieure d'Ingénieurs
de Caen (ENSICAEN)
6, B. Maréchal Juin, 14050 Caen, France

Paper

Maintenance of Lead-acid Batteries Used in Telecommunications Systems

Ryszard Kobus, Paweł Kliś, and Paweł Godlewski

National Institute of Telecommunications, Warsaw, Poland

Abstract—The article presents numerous problems with standby batteries used in telecommunications systems, with a particular emphasis placed on the assessment of their real capacity. The methods used to evaluate the technical condition of batteries and to measure their real capacity are presented. Also, the a new test device which measures the actual battery capacity is presented. The said measurement is based on the discharge test method and is performed with the use of a new TBA-A automated test unit. The article is targeted for electronic designers, managers and telecommunications hardware maintenance personnel, as well as for other telecommunications systems experts.

Keywords—battery capacity measurements, maintenance, telecommunications systems.

1. Introduction

Nowadays, a high degree of reliability is an aspect of key significance in the delivery of telecommunications services. This means that telecommunications systems should remain powered even if a mains failure occurs. Lead-acid batteries are the most popular back-up energy source and it is expected that such batteries will remain in use for a long time to come, in spite of introduction, to the market, of new battery types and new reserve power source chemistries. The above means that the problem of maintenance of good lead-acid batteries still remains an issue of high importance.

2. Batteries Used in Telecommunications Systems

Telecommunications systems should ensure continuous availability of services. This applies both to commercial services offered to the general public, and to emergency services supplied over critical infrastructure networks. That means that telecommunications systems should be powered without any interruptions.

Telecommunications systems are powered by installations relying on rectifier-based power systems (PS) and a number of batteries connected in parallel. The batteries should be able to provide backup for a given telecommunications system for a few hours or more. When the mains voltage is present, PS supply energy to the telecommunications equipment and to the batteries associated therewith. Under such conditions, the rectifier provides float voltage (about 54 V) to the batteries, preventing their self-discharge.

Figure 1 shows three basic configurations of power systems dedicated to use on telecommunications sites. The simplest structure, and thus the least reliable, is presented in Fig. 1a. In the case of a mains failure, the powered equipment (PE) is supplied from the battery until either the battery discharges or the mains voltage is restored. Additionally, it should be borne in mind that if rectifier or battery maintenance is performed, an additional, transportable backup power source has to be connected. The configuration shown in Fig. 1b is more reliable due to the added redundancy. It allows to disconnect one rectifier unit or one battery without any disturbances to the PE supply. The configuration shown in Fig. 1c is the most reliable, but at the same time the most expensive. It relies on two independent power systems and two independent mains networks.

In order to increase the level of AC voltage supply reliability even further a backup diesel generator may be connected to the system via an automatic switch [1], [2]. Batteries are the source of power during mains failures. Therefore, their key features should include long battery life, low overall costs of purchase and operation, as well as safety of use.

It should be noted that battery weight is not an important factor in this particular case. Hence, lead-acid batteries fulfill all the requirements mentioned above. They are characterized by high power density of up to 0.1 kWh/kg and by low internal resistance. Despite of advanced technologies relying on other battery chemistries, i.e. NiCd, NiMH, Li-Ion and Li-Po, lead-acid batteries remain the primary standby source of energy in telecommunications power supply systems.

3. VLRA Batteries

Flooded lead-acid batteries have been used in the telecommunications sector for about 100 years now. Because of their open design, they must be installed in separate, ventilated and secured rooms. The first leak-proof, valve-regulated lead-acid (VRLA) batteries first appeared in the 1960s in the USA, but they only began to be used on a wider scale in Europe in the 1990s. It is difficult to say if VRLA batteries are significantly superior to the flooded variety, but they offer certain advantages which have contributed to their widespread use. VRLA batteries have a shorter lifetime, but their maintenance cost is lower. They do not require separate, special rooms, but there is a need to provide float voltage thermal compensation. VRLA units can be installed in rooms used by staff or other electronic equipment, but in a designated area. Adequate room must be

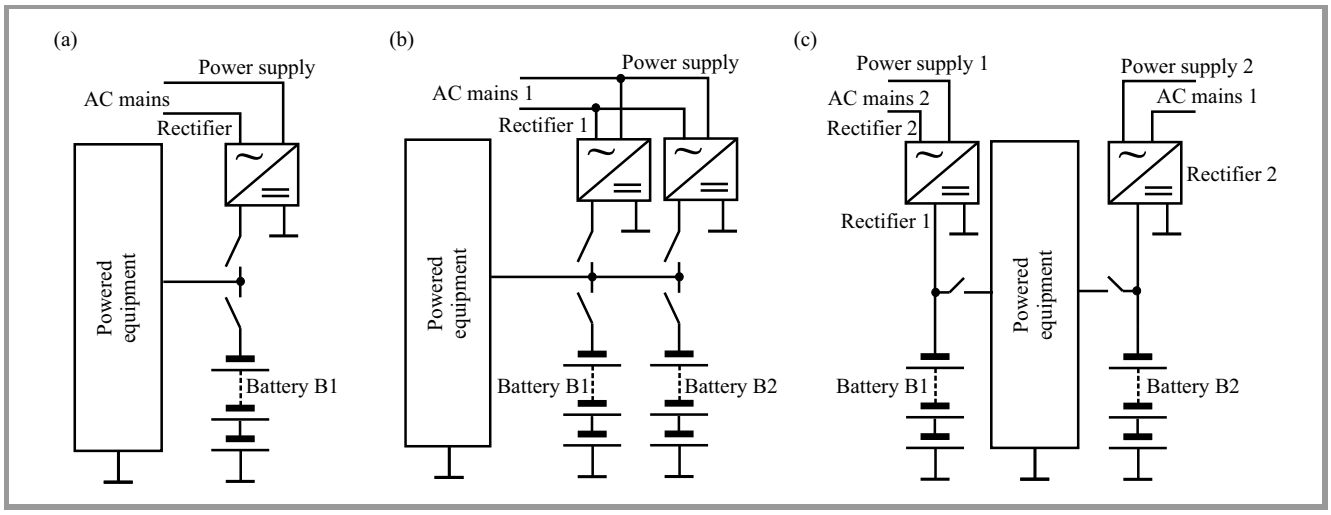


Fig. 1. Examples of power system configurations.

provided around the battery system to allow maintenance, including the exchange of blocks/cells. The area required for installation of VRLA batteries can be smaller than in the case of their flooded counterparts. The battery can be placed vertically or horizontally and can be stacked with the use of a dedicated rack enclosure.

All lead-acid batteries have a defined maximum storage time of six months at the temperature of 18–30°C. By the end of that period, batteries should be either installed or charged. Therefore, it is important to schedule the delivery of batteries to the site and their date of production as close as possible to the date of actual installation.

Two primary types of VRLA batteries exist, relying on gel and AGM technologies. In the case of the gel technology, silica dust is added to the electrolyte, forming a thick putty-like gel. The Absorbed Glass Mat (AGM) technology employs a fiberglass mesh between the individual battery plates. The mesh absorbs and retains the electrolyte. Both technologies offer similar advantages and disadvantages in comparison to conventional battery types.

4. Battery Condition Monitoring

To improve operational functionality of batteries and to protect them against damage, the use dedicated equipment is required. A relevant device can be integrated within the power supply system, may constitute a part of the battery itself, or may be installed as additional, stand-alone test equipment [3], [4]. The following parameters can be monitored with use of this equipment:

- battery voltage,
- charging/discharging current,
- ambient temperature,
- all cell/block voltages and temperatures,
- AC ripple current and voltage.

Voltage and temperature measurements pertaining to all cells/blocks enhance assessment of battery cell balancing and help detect damaged cells. Comparisons of battery string temperatures, in turn, allow for detection of thermal runaways.

Specialized circuits are used in order to improve cell voltage balancing. They reduce cell voltage if it is higher than the prescribed limit value while the battery is charging. Balancers are also used in which the cells with the lowest voltage levels are charged with higher current in order to manage cell voltage more effectively [3], [5], [6].

Generally speaking, monitoring systems are capable of indicating the actual condition of the batteries, but do not reflect their actual capacity.

5. Key Parameter Measurements

All battery manufacturers recommend periodic check of batteries condition including:

- leaks,
- verification of cell interconnection resistance,
- battery capacity measurements.

Battery maintenance always requires that periodic site visits be paid (even on unmanned sites), but attempts are made to minimize the maintenance time. It is recommended that only a few measurements be made to evaluate the condition of a battery, with a particular focus on its capacity and the remaining lifetime. With the accuracy of all crucial parameter measurements, the time and cost of tests, as well as the need to mitigate test-related risks taken into consideration, one may conclude that no single method meeting all the requirements exists [4], [7], [8]. Therefore, internal resistance measurements and discharge tests are among the most commonly used procedures. The properties of such methods are described in detail below.

6. Internal Resistance Measurement

The common internal resistance measurement procedure is cheap, fast and safe, and usually does not require that the battery undergoing the test be disconnected from the power system. It relies either on the analysis of DC pulses or on resistance measurements performed with the use of AC signals [5], [9]. The resistance reflects not only the battery capacity, but also:

- grid corrosion,
- loss of active material from electrodes,
- possible sulfation,
- temperature increase,
- internal short circuit,
- other cell failures.

Measurement equipment manufacturers recommend that a principle be adopted in line with which a 20% loss of battery capacity is related to a 25% increase in the resistance of each cell. It is also estimated that the loss of battery capacity is related to only 40% of the total internal resistance (for the entire battery). Additionally, internal battery resistance may vary by approximately $\pm 10\%$ for the same type. Therefore, it is recommended to measure each cell and the entire block separately, and directly at battery terminals. If the measurements are performed periodically under the same conditions, i.e. temperature and charge level, it is possible to identify deterioration of the cells based on historical data analysis. Unfortunately, research fails to prove that the internal resistance test may be considered an equivalent of the battery capacity measurement that relies on the discharge test. Hence, it is not commonly used to actual battery capacity assessment.

7. Discharge Test

The discharge test is the only reliable method used to evaluate actual battery capacity with a high degree of accuracy. It takes a long time to perform – even up to 20 hours. While the measurements are performed, the battery needs to be disconnected, which results in a considerable depletion of the amount of reserve energy available on site. A few test procedures and equipment setups may be employed, which can provide results characterized by a varying degree of accuracy. The cost of the measurements performed may vary as well. Examples of the test procedures are presented below.

7.1. Discharge Test Built into the Power System

Modern DC power systems offer an advanced functionality enabling the efficient use of energy from VRLA batteries, referred to as the "battery test". This function is capable of controlling the powered equipment based on priority levels assigned (e.g. critical equipment and non-critical equipment). The test may be run periodically, e.g. after a prolonged mains failure, or on-demand. Charging voltage may be boosted or reduced.

The test is based on a simultaneous, partial discharge of all batteries (up to 50% of the batteries' design capacity). During the test, the output voltage of the power system's rectifier is temporarily reduced to the pre-programmed value, e.g. 44 V. If the batteries manage to keep the telecommunications equipment powered up, over a pre-defined period of time, with the voltage remaining higher than the rectifier-provided value, the test result is deemed positive. If the battery voltage drops below the rectifier-fed value, over a period of time that is shorter than specified, the test result is considered negative.

Power consumption of modern telecommunications systems remains constant. Therefore, the amount of energy drained from batteries can be measured quite easily. Interpretation of test results is much easier when cell voltage of all batteries is monitored. This solution is simple and cheap to implement, but the capacity of batteries available at the final stages of the test is unpredictable. Therefore, the real battery capacity is unknown.

7.2. Discharge Test Using Battery Discharger

In order to determine the real capacity of a battery, a discharge with the current of 0.1 C is usually performed [10]–[12]. There are many types of battery dischargers, but in general, all of them are passive and rely on the transformation of power into heat. The majority of modern battery dischargers are equipped with monitoring circuits that measure the following parameters: battery voltage, individual cell voltage, discharging current and battery temperature. Hence, they are capable of working out the battery capacity. It is possible to set threshold values for the parameters referred to above, and to program the discharger to discontinue the test if one of them is reached. This simplifies the entire test procedure and allows to protect the battery from damage caused by excessive discharge. An example of the discharger unit that can sink up to 120 A

Fig. 2. Battery discharger with nominal current of 120 A.

Fig. 3. Stand-alone battery ATE: (a) up to 160 A and (b) up to 50 A.

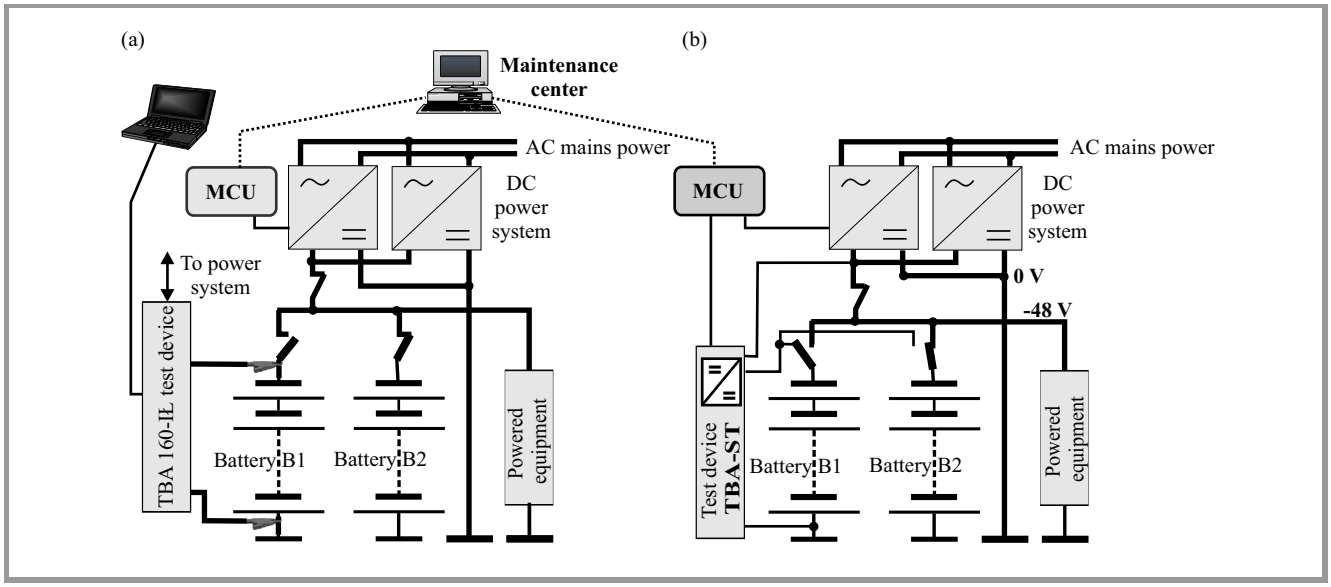


Fig. 4. Power system with ATE (a) stand-alone TBA160-IŁ and (b) built-in test functionality (TBA-ST).

from 48 V batteries is presented in Fig. 2 [13]. However, tests performed with the use of such devices are time consuming, as the full test procedure requires that several steps be completed:

- disconnecting the battery to be tested from the power system,
- charging the battery to its full capacity,
- discharging the battery,
- re-charging the battery to restore its operational parameters,
- reconnecting the battery to the power system.

Unfortunately, the return charging process is not monitored and battery energy efficiency cannot be assessed. Each stage of the process requires that battery connections be altered, and that the measurements be activated manually. One full battery test cycle takes approximately one day to complete, which means that in the case of sites with two batteries, the power system operates with a reduced energy capacity for two days. Therefore, often only partial discharges are performed.

It needs to be added that large amounts of heat are dissipated in the course of the test, which increases ambient temperature in the room and, of course, the battery temperature. The above means that not only all discharge energy is lost, but that air conditioning systems in use on the site consume more power as well.

7.3. Battery Test Automation

The entire battery test cycle can be automated, thanks to the use of sophisticated testers, either of the stand-alone variety, or ones that are built-in to the power supply system. TBA-IL is an example of a stand-alone portable device designed to measure real capacity of batteries at telecommunications sites (Fig. 3). The device can be connected to the battery and the power system via universal flexible cables, or with the use of a dedicated terminal box. As mentioned above, the battery undergoing the test needs to be disconnected from the power system. TBA160-IŁ was developed within the framework of a project titled "The new generation of VRLA battery control devices for telecommunications power systems", and was subsidized by the European Union under the Innovative Economy Operating Program [14].

The test unit can operate in a full automatic mode (Automated Test Equipment) – Fig. 4a. All input parameters and measurement results are stored in its memory and may be transferred to a local or remote PC by means of the LAN-WAN interface. The unit presented in Fig. 3 is very efficient – energy discharged from the battery is returned to the power system and less than 5% energy is dissipated in the form of heat. The enclosure of the device is also smaller than that of a typical resistive discharger.

The ATE test device may be integrated with the power system, as shown in Fig. 4b. In this case, it is supervised by a power system controller and managed by the maintenance center. In such a case, the ATE comprises only a bidirectional power converter and relevant sensors. Therefore, the built-in test device can be a few times cheaper than the stand-alone version. Automated battery switch-off functionality is another of the advantages of this particular configuration. No manual operation is required, and the switching-off process is initiated by the power system controller. However, there are certain restrictions inherent in this solution. The only drawback is the fact that the built-in unit is capable of controlling batteries with the maximum capacity of 1000 Ah.

8. Universal Module for Charging/Discharging Batteries

The National Institute of Telecommunications (ITL) boasts extensive experience in designing devices for testing batteries used at telecommunications sites. ITL cooperates with the Electronic Power and Market (EP&M) company. A consortium led by ITL won a contract from the National Centre for Research and Development for designing a "Control systems for telecommunications site energy reserve solution - SKOT". TBA-A with the module shown in Fig. 5 was developed within the framework of this project, which can serve as a TBA-ST device integrated with the power system. There are similar solutions available on the market, i.e. [15], but ITL TBA-ST ATE offers optimized functionality. The TBA-ST is dedicated for medium size telco sites and it is capable of driving/sinking current of up to 50 A. When combined with the TBA-W control unit, the TBA-A module forms another ATE unit. Its firmware was also developed under the project in question.

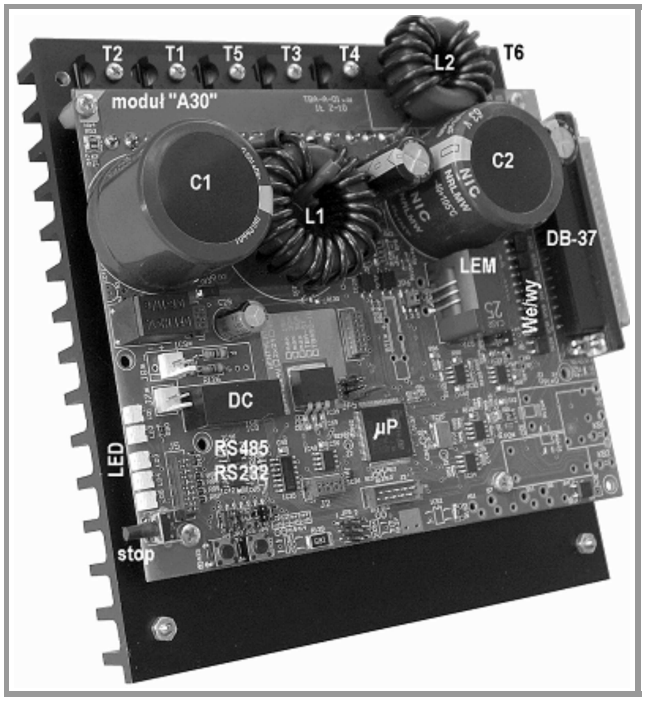


Fig. 5. Universal battery charging/discharging module.

The core of the TBA-A has the form of a bidirectional additive/subtractive power converter based on T1–T4 switching transistors, L1 inductor and C1–C2 capacitors. The switching process is controlled by the PWM circuit at the fixed frequency of 35 kHz, enabling output voltage to be regulated, and the energy from the tested battery (either Battery 1 or Battery 2) to be transferred to the power system or in the reverse direction. The charging and discharging current is stabilized by using a high accuracy LEM current sensor. The input and output voltage is monitored for exceeding threshold values. The power conversion is controlled by the STM32F103VE 32 bit microcontroller. It generates PWM waveforms, reads battery voltage, power system voltage, each cell/block voltage and charging/discharging current from the LEM transducer. It also offers an external communications interface. All parameters and operational modes can be transferred remotely via the RS232/485 interface. The serial port is used also for downloading the measurement results. The TBA-A is also equipped with an additional RS232 port used for servicing. More details are presented in Fig. 6 and in [16].

The firmware of the device presented above constitutes is core component, as it controls the bidirectional converter. It was developed based on the authors' extensive experience. The first power converter dedicated to charging/discharging batteries was developed by ITL 15 years ago and weighed approximately 10 times more than the current solution [6], [17]–[19]. The use of fast MOSFET transistors with internal diodes, the 32-bit ARM-based microcontroller and sophisticated firmware has enabled to develop a very small, light and powerful unit.

Voltages of the batteries (B1 and B2 in Fig. 6) and cells (a1…a4, b1…b4) are measured with the accuracy better than 1%. The real capacity calculated in relation to the capacity at 20°C is saved with the accuracy 2%. In addition, the device calculates works out the energy of the discharged battery. The test device offers high energy efficiency. About 95% of the discharged energy is returned to the power system to supply telecommunications equipment. As no heat is generated, the measurement conditions remain very stable. The room itself and especially the battery are not exposed to any additional heat, which means that the air conditioner operates under stable ambient conditions. The TBA-ST ATE device was developed under the "Monitoring system for telecommunications site

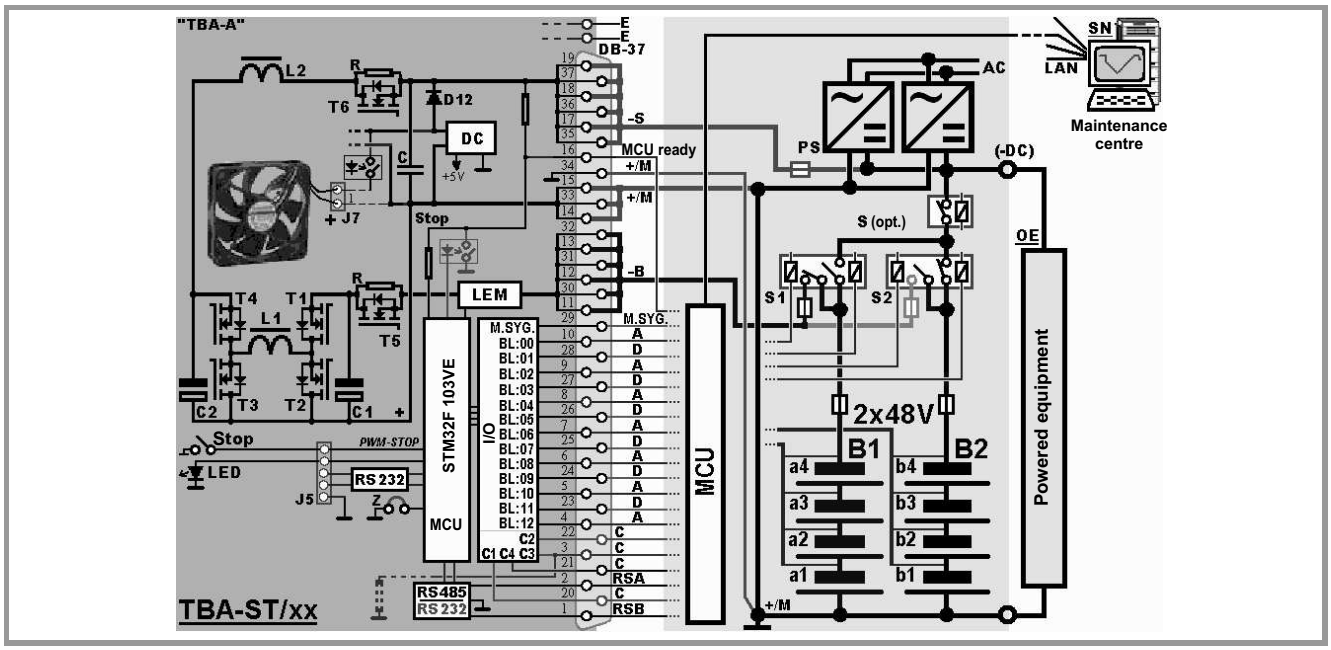


Fig. 6. TBA-A block diagram and its connection to the power system.

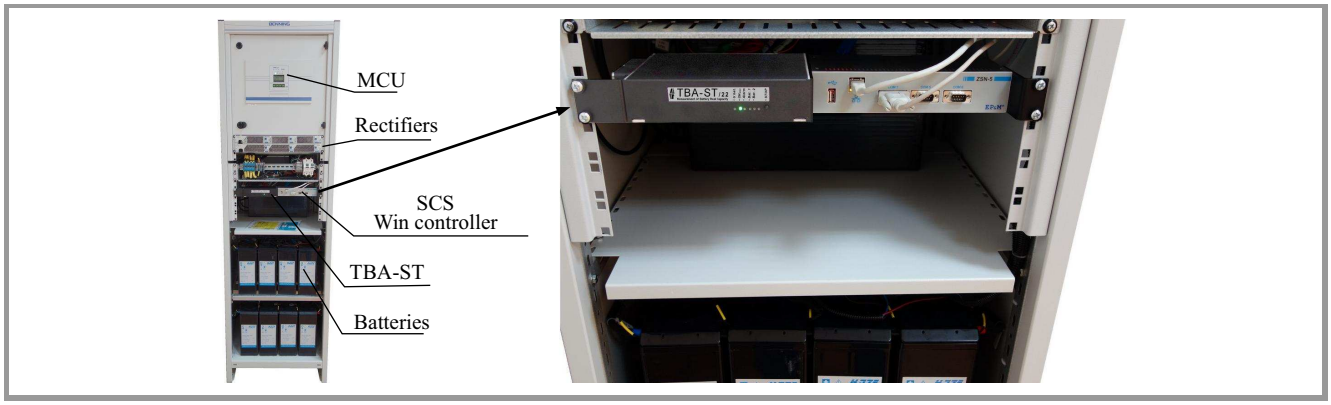


Fig. 7. Power system with a built-in universal battery testing module.

energy reserve solutions – SKOT" project. The project was implemented by the National Institute of Telecommunications and the EP&M company, and was co-financed by the European Regional Development Fund under the Innovative Economy Operational Program.

9. Deep or Partly Battery Discharge

The primary objective of the study is to evaluate the energy reserves stored in the battery. It is usually assumed, in the case of telecommunications power systems, that the battery remains operational if its capacity (Q) is not lower than 80% of rated value at the discharge current of 0.1 C. That is why the designed battery capacity is 20% greater. It enables to achieve the required capacity within the power system at the end of the battery's lifetime declared by its manufacturer. Due to the adverse operating conditions, some batteries fail to achieve the average declared life expectancy, but a significant portion of them remain operational until the end of the specified period. It should be noted that the efficiency of each battery is determined by the condition of its weakest cell.

Figure 8 shows the results of checks performed on various batteries rated at 48 V/1000 Ah after operation lead times. The drawings present cell voltages during the discharge and charge test under the same conditions. The cell discharge cut-off voltage was set at 1.80 V. If the voltage of any of the tested cells drops below that value, the battery discharge stops.

The initial charging current was set at 0.1 C (100 A), the final charging battery voltage was 56.00 V, and the highest cell voltage was set at 2.38 V. If either the voltage of the battery reaches 56.00 V or any the voltage of any of the cells is equal to 2.38 V, the charging current is decreased and the charging process is stopped.

The discrepancies between cell voltage characteristics shown in Fig. 8 tend to increase with time of use, and with the reduced battery capacity. It can also be noted that

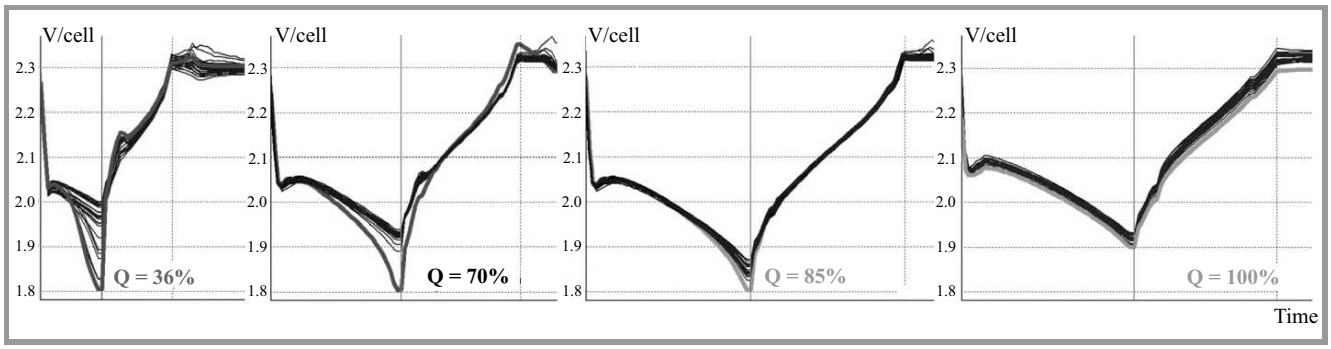


Fig. 8. The results of inspection of various 48 V/1000 Ah batteries.

in the first stage of the discharging process, cell voltages are usually similar and do not suggest a failure of any of the cells. Moreover, the voltage of smaller capacity cells recorded during the first stage of discharge process may be higher than that of higher capacity cells. Such a case is presented in Fig. 9. The cell with the lowest voltage in the final stage of the discharge phase (after 360 minutes) has the smallest capacity. But in the first stage of the discharging process, the voltage of this particular cell was good, and no evidence enabling to predict its reduced capacity existed. This means that is not easy, or even impossible, to evaluate battery parameters, especially its capacity, based on the cell voltage chart during the first stage of the discharge process. The National Institute of Telecommunications [20] has performed research focusing on this particular issue, but no effective algorithm to predict the battery capacity based on short discharge results only has been developed yet.

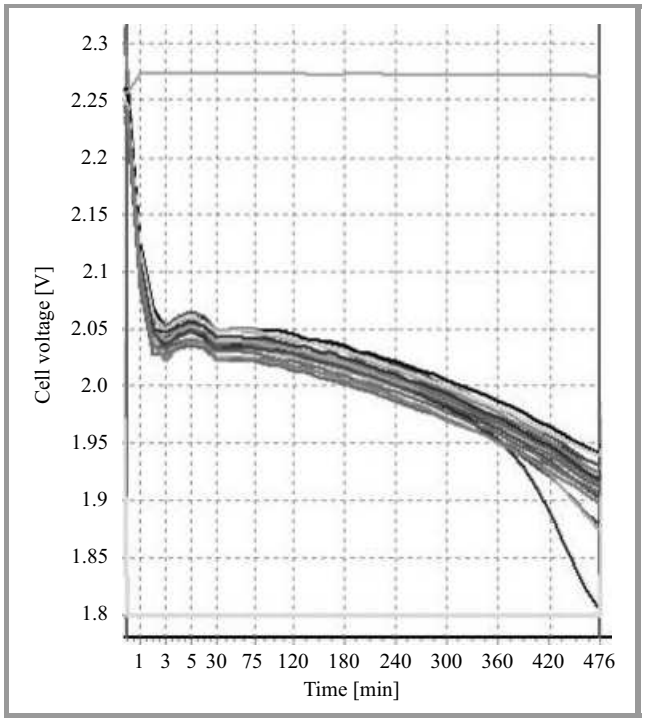


Fig. 9. Detailed discharge characteristics of a 700 Ah battery.

10. Conclusions

Currently, the discharge test method remains the only reliable way to evaluate the real capacity of batteries. Such a measurement method renders results with the accuracy of $\pm 2\%$, a level that is unattainable in the case of remaining methods. Unfortunately, measurements made with dischargers or stand-alone testers are expensive and time consuming. The use of a measurement module that is integrated with the power supply system can significantly reduce the cost of testing batteries to the level that is competitive with alternative solutions.

The method presented and the ATE testers do not reduce the measurement lead time, but offer the opportunity to stop the test at any given moment, e.g. if the continuity of power supply is jeopardized. Once the test is completed, the battery is reconnected to the power system and the reserve power is increased.

The measurement module enables also to disconnect the battery remotely, should a need arise.

These benefits make the application of the ATE system very profitable in the case of remote telecommunications facilities.

References

[1] W. D. Reeve, *DC Power System Design for Telecommunications*. Hoboken, NJ, USA: Wiley, 2007.

[2] S. Fischer, "Cell Site Power System Management, Including Battery Circuit Management", U.S. Patent Application Publication, US 2011/0227414A1.

[3] "The changing face of backup", Bestmagazine, Energy Storage Publishing, no. 39, Winter 2013 [Online]. Available: www.bestmag.co.uk

[4] "Implementing proactive battery management. Strategies to protect your critical power system", white paper from the experts in business-critical continuity, Emerson Network Power, Columbus, OH, USA, 2008.

[5] G. Hedlund, "Monitoring of lead-acid batteries", an Eltek white paper. Different Methods of Predicting Premature Capacity Loss, 2013 [Online]. Available: www.eltek.com

[6] P. Godlewski and B. Regulska, "Automatyzacja oraz zdalne badania baterii akumulatorów w obiekcie telekomunikacyjnym", *Przegląd Telekomunikacyjny*, vol. 8–9, p. 1260, XXVIII Krajowe Sympozjum Telekomunikacji i Teleinformatyki KSTiT, Warszawa, 2012 (in Polish).

[7] P. Shore and G. May, Battery Optimization Services. A guide to optimize your battery maintenance. Emerson Network Power and FOCUS Consulting, 2013.

[8] "Handbook for Gel-VRLA-Batteries, part 2: Installation, Commissioning and Operation", Industrial Energy, Technical Support, Rev. 5, Dec. 2003, Deutsche Exide GmbH, Büdingen, Germany.

[9] E. Davis, D. Funk, and W. Johnson, "Internal ohmic measurements and their relationship to battery capacity – EPRI's Ongoing Technology Evaluation" [Online]. Available: http://www.battcon.com/papersfinal2002/davispaper2002.pdf

[10] B. Blohm, "The keys to extended battery life", Electrical Construction & Maintenance (EC&M), 2002 [Online]. Available: http://ecmweb.com/content/keys-extended-battery-life

[11] Panasonic, Katalog akumulatorów kwasowo-ołowiowych serii LC-R i LC-X. Wamtechnik, Piaseczno, Poland, 2004 (in Polish).

[12] Operating instructions – Valve regulated stationary lead-acid batteries, Hoppecke Batterien, Brilon, Germany [Online]. Available: http://www.hoppecke.com

[13] Battery Discharger & Capacity Tester, Specifications, Megger, Dallas, TX, USA [Online]. Available: http://www.megger.com

[14] Operating manual TBA160-IŁ and TBA150-IŁ, Instytut Łączności, 2010–2012 [Online]. Available: http://www.itl.waw.pl/tba

[15] M. Zakrzewski, "Sposób i układ do diagnostyki baterii akumulatorów, zwłaszcza w systemach teleinformatycznych", Patent RP, B1 218334 (in Polish).

[16] P. Godlewski, K. Niechoda, K. Olechowski, and B. Regulska, "Stacjonarne urządzenia TBA-ST – do pomiaru dysponowanej pojemności akumulatorów siłowni telekomunikacyjnych – projekt SKOT", *Telekomunikacja i Techniki Informacyjne*, no. 3–4, pp. 14–23, 2014 (in Polish).

[17] P. Godlewski *et al.*, "Sposób i układ do zdalnej kontroli dysponowanej pojemności akumulatorów w siłowni telekomunikacyjnej", Patent RP, PL 219471 (in Polish).

[18] P. Godlewski, B. Chojnacki, and R. Kobus, "Funkcjonowanie i budowa urządzeń TBA160-IŁ", Biblioteka Infotela. Szerokie Pasmo – Rozwiązania technologiczne i usługi, p. 38, 2012 (in Polish).

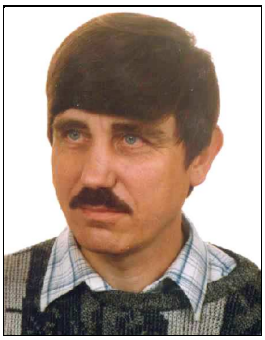
[19] P. Godlewski and T. Kunert, "Przekształtnik TBA2-IŁ", *Telekomunikacja i Techniki Informacyjne*, no. 3–4, pp. 117–121, 2003 (in Polish).

[20] A. Binkiewicz, "Uniwersalny test baterii akumulatorów kwasowo-ołowiowych", *Wiadomości Elektrotechniczne*, vol. 6, no. 10–11, 2014 (in Polish).

[21] W. Fechalos, "Battery System and Management Method", U.S. Patent Application Publication, US 2011/0298626A1

Ryszard Kobus received his B.Sc. and M.Sc. degrees from the Faculty of Electronics of the Warsaw University of Technology in 1975. Kobus has been working at the National Institute of Telecommunications since 1975. He is a member of the Expert Technical Committee CEN/TC 331 specializing in postal services, and the deputy chairman of the Postal Service Committee PKN/TC 259. He is a co-author of many patented telecommunications solutions. Research interests: telecommunications, measurements and evaluation of quality of telecommunications services, quality surveys, evaluation the quality of postal services, standardization.
E-mail: R.Kobus@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Paweł Kliś received his B.Sc. degree from the Faculty of Electrical Engineering of the Opole School of Engineering in 1976. He has been working at the National Institute of Telecommunications since 1976, formerly in the Power Systems Department, currently in the Central Chamber for Telecommunications Metrology. He is a co-designer of numerous telecommunications power systems and devices. He is a co-author of several scientific publications and co-author of several patents. His research interests include: telecommunications power systems, metrology of basic electrical parameters.
E-mail: P.Klis@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland



Paweł Godlewski received his B.Sc. degree from the Faculty of Electronics of the Warsaw University of Technology in 1973. He has been working at the National Institute of Telecommunications since 1973. He is the designer of many devices, and co-author of a system for the assessment of quality of telecommunications services, a well as AWP-IL and TBA-IL ATE equipment. He is the author of numerous scientific publications and co-author of patented solutions. His research interests include: visualization systems used in the telecommunications sector, programmable measurement devices for telecommunications.
E-mail: P.Godlewski@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

JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY 4/2015