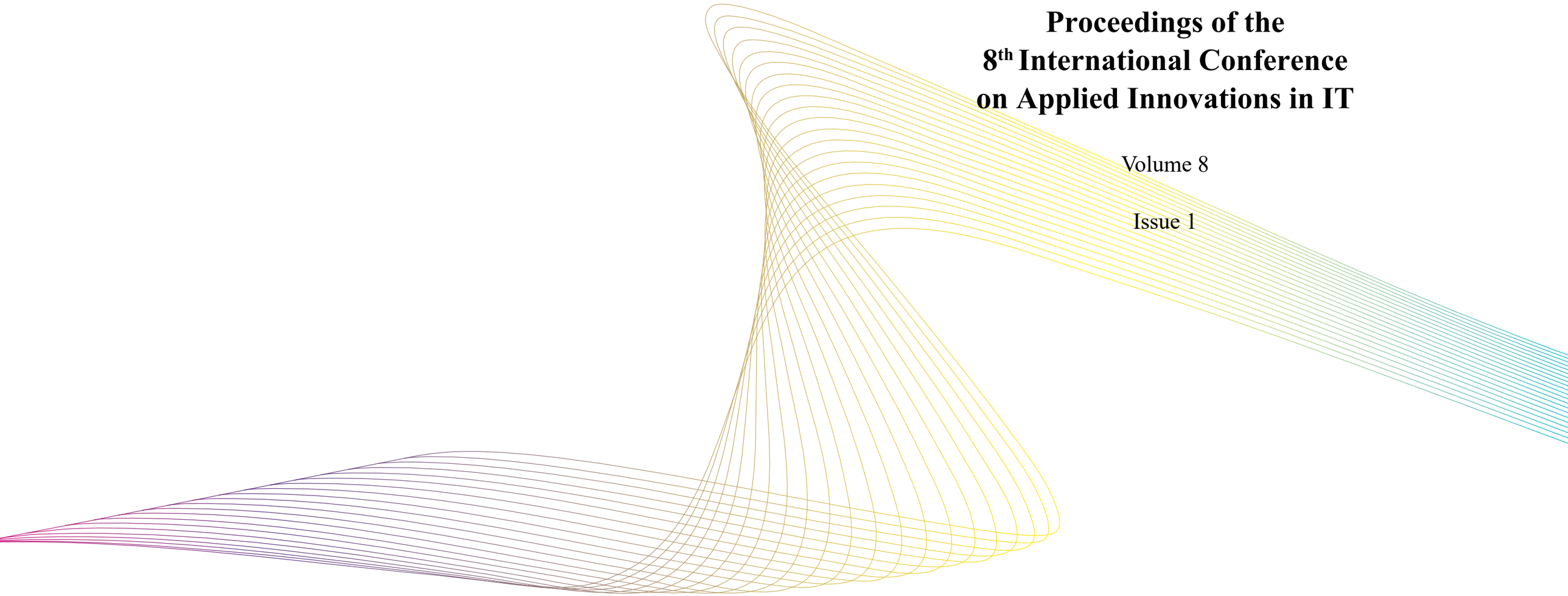
by Anhalt University of Applied Sciences and Perm National Research Polytechnic University

Proceedings of the 8th International Conference on Applied Innovations in IT

Volume 8

Issue 1



EDITION
Hochschule Anhalt

Proceedings of the 8th International Conference on Applied Innovations in IT

Volume 8 | Issue 1

Koethen , Germany
10 March 2020

Editors:

Prof. Dr. Eduard Siemens* (editor in chief),
Dr. Leonid Mylnikov**

(*Anhalt University of Applied Sciences,
** Perm National Research Polytechnic University)

This volume contains publications of the International Conference on Applied Innova-tions in IT (ICAIIT), which took place in Koethen March 10th 2020. The conference is devoted to problems of applied research in the fields of automation and communication technologies. The research results can be of interest for researchers and development engineers, who deal with theoretical base and the application of the knowledge in the respective areas.

ISBN: 978-3-96057-118-6 (Online)
ISSN: 2199-8876

Additional copies of this publication are available from:
FB6 Anhalt University of Applied Sciences
Postfach 1458
D-06354 Koethen, Germany
Phone: +49 3496 67 2327
Email: eduard.siemens@hs-anhalt.de
Web: http://icaiit.org

Content

Section 1. Communication technologies

Section 2. Data Analysis and Management

Section 3. Control and Automation

Available Bandwidth Metrics for Application-Layer Reliable Multicast in Global Multi-Gigabit Networks

Kirill Karpov[1,2], Maksim Iushchenko[1,2], Nikolai Mareev[1,2], Dmytro Syzov[1], Eduard Siemens[1] and Viatcheslav Shuvalov[2]

[1] *Future Internet Lab Anhalt, Anhalt University of Applied Sciences, Bernburger Str. 57, Köthen, Germany*

[2] *Department of Transmission of Discrete Data and Metrology, Siberian State University of Telecommunications and Information Sciences, Kirova Str. 86, Novosibirsk, Russia*

{*kirill.karpov, maksim.iushchenko, nikolai.mareev, dmytro.syzov, eduard.siemens*}*@hs-anhalt.de, shvp04@mail.ru*

Keywords: Application Layer Multicast, Point-to-Multipoint, Available Bandwidth, Minimum Spanning Tree, Networking, High Bandwidth.

Abstract: Application layer multicast shows its efficiency when it is necessary to transmit enormous amount of data to many nodes. One of the most important issues for such kind of transmission is "what is the criteria for path construction?". In the global networks, when the distance between nodes becomes a significant factor, the time delay between nodes seems self-evident. However, in the presence of cross-traffic in the channel, the minimum spanning tree based on the delay might construct the tree which provides the lower data rate than possible alternative. The point of this work is to study how available bandwidth estimation techniques might solve such kind of challenges, how to adopt available bandwidth as a metric to construct data distribution paths and the cases when it gives higher performance in comparison with delay as a metric.

1 INTRODUCTION

In the current time, there is a possibility to create a global high bandwidth network infrastructure using cloud services for a relatively low cost. Cloud services provide the possibility to anyone to deploy own intercontinental network on top of the Internet or create own Content Delivery Network (CDN). However, dealing with the global network, it is necessary to deal with its properties, such as the high delay between nodes, the possibility of packet loss, interfering traffic (so-called cross-traffic) in the links, forbidden network features such as multicast. To make high-speed data transmissions over WAN possible, the RMDT [1] transport protocol was used, since it can overcome challenges described above: it protocol provides WAN acceleration service, which makes network packet losses and latency up to 1 second nearly negligible. It can serve up to 10 receivers in parallel within a single session without fairness issues meaning that available bandwidth will be shared evenly. It has a centralized congestion control, which allows the coexistence with the cross-traffic in IP WANs. The multicast functionality has been implemented and studied during the previous work [2] using the Minimal Spanning Tree (MST) algorithm. Previous work shows the efficiency of RMDT in conjunction with the MST algorithm using delay or RTT values between hosts as a metric in the global network. However, the presence of the cross-traffic in the global links is more than possible. Nevertheless that with the delay metric, the MST algorithm can construct the optimal tree, cross-traffic can negatively affect the bandwidth along a constructed path. This means that cross-traffic as a significant factor should be taken into account.

This paper studies the performance of application layer multicast in combination with high-bandwidth data transport applications using available bandwidth (AvB) metric and where and how it can outperform the delay metric.

The remainder of this paper is structured as follows. Section 2 provides an overview of related researches and methods. Section 3 gives the AvB metric description, how it can be obtained and adopted to the MST algorithm. Section 4 reviews the testing environment and software equipment that was used for the experimental setup. Section 5 is devoted to the retrieved results of experiments with application layer multicast implementation using RTT and AvB metrics. Interpretation and discussion of the results and future work can be found in Section 6.

2 RELATED WORK

An analysis of application layer multicast in the Wide Area Network has been made in the previous work [2]. It shows the efficiency of ALM in conjunction with Reliable Multi-Destination Transport Protocol (RMDT) in the global networks between hosts spaced across continents. In the research, the delay was used as a metric for the tree-first application layer multicast. The system has been tested in the AWS infrastructure.

The available bandwidth estimation research has been provided by Kirova, et. al [3, 4]. These researches introduce the Kite2 application for available bandwidth estimation, which is used in the given work.

The ALM model as a service implemented on the top of Hierarchical Peer to Peer Architecture, in order to give media streaming based applications or conferencing applications [5].

A multicast framework for point-to-multipoint and multipoint-to-point-to-multipointvideo streaming from drones presented in the work [6]. The proposed rate-adaptive approach outperforms legacy multicast in terms of goodput, delay, and packet loss.

Adaptable, ISP-friendly multicast overlay network proposed in [7]. The system is adaptable to different conditions and easily reconfigurable in the construction and management of the multicast overlay distribution tree. The ALM tested has been tested in a network emulation tool.

There is an ALM based on an encoding-free non-dominated sorting genetic algorithm (EF-NSGA) [8]. The approach aimed to construct a tree based on multi-objective criteria: minimization transmission delay and instability simultaneously.

3 METRIC DESCRIPTION

The used AvB metric is based on Probing Rate Model (*PRM*) of the active probing measurement and uses analysis of inter-packet interval deviation on the receiver side (*IpIr*) in order to detect whether the sending rate of probe packets meets the available bandwidth limits of the link between the sender and the receiver.

Figure 1 illustrates the active probing model. S_n is an n-th probe packet, i_{sn} is the sending inter-packet interval and i_{rn} is the inter-packet interval on the receiver side. The main principle of this method is to find the inter-packet interval, which is dependent on timing operations precision. The accurately chosen inter-packet interval on the sender side allows achieving the actual available bandwidth of the network path. The goal of the AvB algorithm is to find the minimal value i_s after which the difference $i_r - i_s$ is minimal.

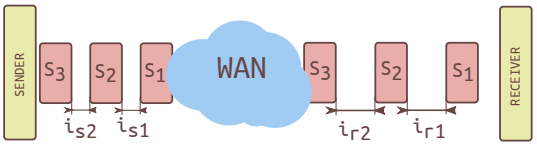


Figure 1: AvB active probing model [4].

Since the given ALM system uses the tree-first approach, the available bandwidth estimation values are collected only once as a first step. Collected metrics are formed into adjacency matrix A and inverted by Hadamard inverse rule as $A^{\circ -1}$ because the MST algorithm calculates the shortest tree where the minimal weight of an edge is preferable. The $A^{\circ -1}$ matrix passed to the MST or DCMST algorithm then. The resulted tree is treated as optimal from the perspective of the available bandwidth of the link.

4 EXPERIMENTAL SETUP

The section provides a detailed description of the testing infrastructure and software, which is used during the experiments.

4.1 Testing Environment

As an experimental environment, *Amazon AWS* has been chosen. It provides the virtual infrastructure in selected continents and regions. Cascade network transmission infrastructure based on **c5.xlarge** virtual instances, the configuration provided in Table 1.

Table 1: c5.xlarge host configuration.

Name	Parameters
Operating system	Ubuntu 18.04
CPU model	Intel Xeon Platinum 8000 (Cascade Lake)
CPU Frequency	3.6 GHz
Number of vCPUs	4
RAM	8 GB
Bandwidth	up to 10 Gbps

The instances are distributed all over the world in the AWS regions, which shown in Figure 2. The regions are highlighted by colors here and further. **US West (Oregon)** - orange, **EU (Frankfurt)** - blue, **EU (London)** - red, **Asia Pacific (Singapore)** - cyan, **Canada (Central)** - green.

In each region, three **c5.xlarge** virtual instances have been deployed. The regions have been chosen to get different variations of network conditions, such as long and short distances, international and intercontinental links. The RTT probing results shown in Figure 3. The index rows and columns of the matrix are rep-resented each host by the first letter of its region and the number of the virtual machine in the region. Left indexes are output nodes, top indexes are input nodes. The matrix colored as a heatmap, where the cells with the highest values have more saturated color.

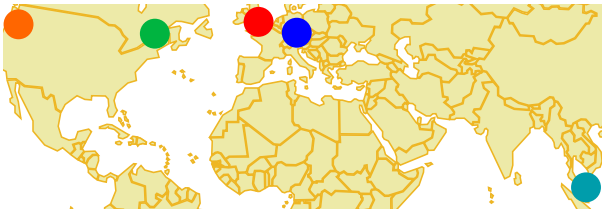


Figure 2: The map of AWS hosts location.

	l1	l2	l3	f1	f2	f3	c1	c2	c3	o1	o2	o3	s1	s2	s3
l1	0	0.1	0.1	13.4	14.3	14.2	89.1	85.6	88.8	134	137.4	133.7	184.8	181.2	171.4
l2	0.1	0	0.1	13.3	15.3	13	85.6	86.9	87.3	139	133.7	137.8	185.2	181.2	184.7
l3	0.1	0.1	0	13.1	13.7	13.6	88.8	88.8	86.8	137	139	139.9	180.7	181.2	171.9
f1	13.4	13.3	13.1	0	0.1	0.1	100.4	100.4	98.9	159	161.1	164.3	177.7	169.6	173.7
f2	14.3	15.3	13.7	0.1	0	0.1	99.3	99.7	99.2	159	162.6	162.7	173.3	173.4	173.7
f3	14.2	13	13.6	0.1	0.1	0	100.4	99.3	98.9	163	158.3	163.9	173.6	174	173
c1	89.1	85.6	88.8	100.4	99.3	100.4	0	0.1	0.1	66.5	66.6	65.1	219.8	219.9	219.8
c2	85.6	86.9	88.8	100.4	99.7	99.4	0.1	0	0.1	65.6	66.6	67.1	219.9	219.7	219.7
c3	88.8	87.3	86.8	98.9	99.2	98.9	0.1	0.1	0	67.4	66.6	66.7	219.9	219.6	219.6
o1	133.6	138.6	137.1	158.5	159.1	163.2	66.5	65.6	67.4	0	0.3	0.3	163.2	162.5	162.6
o2	137.4	133.7	139	161.2	162.6	158.2	66.6	66.6	66.6	0.3	0	0.3	163.1	162.9	162.8
o3	133.7	137.8	139.9	164.3	162.7	163.8	65.1	67.1	66.6	0.3	0.3	0	163.5	163.1	162.6
s1	184.8	185.2	180.7	177.7	173.3	173.6	219.7	219.9	219.9	163	163.1	163.5	0	0.1	0.1
s2	181.2	181.2	181.2	169.6	173.4	174	219.9	219.7	219.6	163	162.9	163.1	0.2	0	0.1
s3	171.4	184.7	171.9	173.7	173.7	173	219.8	219.7	219.6	163	162.8	162.6	0.1	0.1	0

Figure 3: Matrix of RTT metric values between hosts in milliseconds.

The experiments with additional cross-traffic are provided with the network load which can be seen in the matrix in Figure 4. The matrix is generated in the way to let MST algorithm construct the path, based on the AvB metric, which is similar to the RTT-based tree, but with two swapped branches: *frankfurt* and *singapore* hosts. Such configuration aimed to confuse RTT probing and tree construction process, which re-turn the non-optimal tree from the AvB perspective.

	l1	l2	l3	f1	f2	f3	c1	c2	c3	o1	o2	o3	s1	s2	s3
l1	0	0	0	11,21	10,99	10,4	5,41	5,19	5,38	8,1	8,33	8,11	0	0	0
l2	0	0	0	11,23	10,99	11,2	0	5,27	5,29	8,4	8,11	8,35	0	0	0
l3	0	0	0	10,95	10,99	10,42	5,39	5,38	5,26	8,32	8,43	8,49	0	0	0
f1	11,21	11,23	10,95	0	0	0	13,33	13,33	13,33	9,89	9,89	9,91	10,77	10,51	10,53
f2	10,99	10,99	10,99	0	0	0	13,33	13,32	13,32	9,85	9,88	9,89	10,29	10,52	10,55
f3	10,4	11,2	10,42	0	0	0	13,33	13,32	13,32	9,86	9,87	9,86	10,53	10,53	10,49
c1	5,41	5,19	5,39	13,33	13,33	13,33	0	0	0	4,04	4,04	0	6,09	6,02	6,09
c2	5,19	5,27	5,38	13,33	13,33	13,32	0	0	0	3,98	4,04	4,07	6,09	6,05	6,02
c3	5,38	5,29	5,26	13,33	13,32	13,32	0	0	0	4,09	4,04	4,04	6	6,02	6
o1	8,1	8,4	8,32	9,9	9,85	9,86	4,04	3,98	4,09	0	0	0	9,61	9,65	9,89
o2	8,33	8,11	8,43	9,89	9,88	9,88	4,04	4,04	4,04	0	0	0	9,77	9,86	9,59
o3	8,11	8,35	8,48	9,91	9,89	9,86	3,95	4,07	4,04	0	0	0	9,96	9,87	9,94
s1	0	0	0	10,77	10,28	10,53	6,09	6,09	6	9,61	9,77	9,96	0	0	0
s2	0	0	0	10,51	10,52	10,53	6,02	6,05	6,02	9,64	9,86	9,87	0	0	0
s3	0	0	0	10,53	10,55	10,49	6,09	6,02	6	9,9	9,6	9,94	0	0	0

Figure 4: Matrix of the amount of cross-traffic values between hosts in Mbps.

4.2 Software Equipment

For the experiments, the following software and technologies have been used:

1) **RMDT** — Reliable Multi-Destination Transport Protocol is a C++ software library, developed at the Future Internet Lab Anhalt at Anhalt University of Applied Sciences. It provides point to multipoint data transport functionality [1] using UDP. It uses **BQL** congestion control [9] which is tolerant of big delays and dramatic packet loss rates.
2) **Dataclone** — is the transfer application based on RMDT. In the experiments, it is configured to allocate 100 MB of RAM for both send and receive buffers. For the experiments, Dataclone has been set to the third CPU core. The most actual **v1.0.5** version has been used.
3) **Kite2** — is a software application, written in C++, developed at the Future Internet Lab Anhalt at Anhalt University of Applied Sciences. It is an implementation of the modified *AvB* active probing algorithm which can work in 10 Gbps links [4, 3].
4) **Cascade** — is a client-server application, written in *Python*, developed at the Future Internet Lab Anhalt at Anhalt University of Applied Sciences.. Cascade provides ALM functionality, it collects metrics such as RTT, using ICMP packets and AvB, using Kite2. Based on collected metrics it calculates transmission routes using MST or DCMST. In the client mode, it orchestrates Dataclones as transport software and communicates between other server instances.
5) **tc** — an open-source utility, which is used to configure Traffic Control in the Linux kernel. With this tool, the bandwidth of the interface is shaped down to **100 Mbps**. Traffic Control configured as follows:

```
qqdisc cbq 1: root refcnt 2 rate 10Mbit \
(bounded,isolated) prio no-transmit
qdisc sfq 30: parent 1:30 limit 127p quantum \
1514b depth 127 divisor 1024 perturb 10sec
qdisc sfq 10: parent 1:10 limit 127p quantum \
1514b depth 127 divisor 1024 perturb 10sec
qdisc sfq 20: parent 1:20 limit 127p quantum \
1514b depth 127 divisor 1024 perturb 10sec
qdisc ingress ffff: parent ffff:fff1 ----------------

class cbq 1: root rate 10Mbit (bounded,isolated) \
 prio no-transmit
class cbq 1:1 parent 1: rate 100Mbit (bounded,isolated) \
 prio 5
class cbq 1:10 parent 1:1 leaf 10: rate 100Mbit prio 1
class cbq 1:20 parent 1:1 leaf 20: rate 90Mbit prio 2
class cbq 1:30 parent 1:1 leaf 30: rate 80Mbit prio 2
```

6) **iperf** – is an open-source utility for performing network throughputmeasurements. For the experiments, iperf has been set to the second CPU core. The cross-traffic between hosts has been created with the following command:

```
// the process attached to core 2 with taskset 0x1
// the client started as
taskset 0x1 iperf -c DESTINATION_IP -b 10M -u -t 1000000
 // the server started as
taskset 0x1 iperf -s
```

The software testbed configuration is shown in Figure 5. **Cascade** represents the ALM system. It orchestrates the probing software, such as **Kite2** in case of AvB metric, or sends **ICMP** packets in case of RTT. It configures Dataclone as a sender, receiver, or relay node. It defines the source and destination for the data, which needs to be transmitted. The **iperf** utility is used as a cross-traffic generator. The **tc** util-ity restricts the datarates within 100 Mbps, to neglect the CPU and NIC as the bottleneck and to reduce the spendings for AWS traffic.

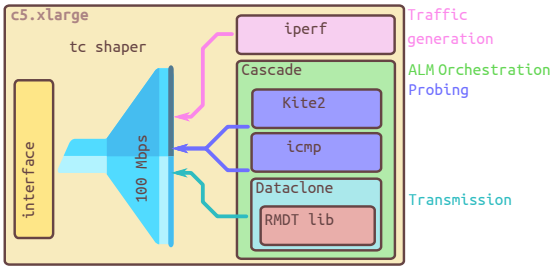


Figure 5: Software equipment scheme.

5 EXPERIMENTAL RESULTS

Experiments have been made in two scenarios: network with generated **cross-traffic**, described in section 4.1, and without interfering traffic, this scenario called **pure network**. Both RTT and AvB metrics have been tested in these conditions. Each test case has been repeated in 10 trials.

5.1 RTT Metric

In both cases, with and without cross-traffic, the **cascade** has built the same data distribution tree, which is shown in Figure 6. The tree constructed based on the metrics collected during the probing state of the **cascade** work. The delay matrix is shown in Figure 3. Cells with red bold values are related to paths built with RTT metrics.

The experiment without additional cross-traffic in the links given the transmission paths configuration provides **45.81** Mbps.

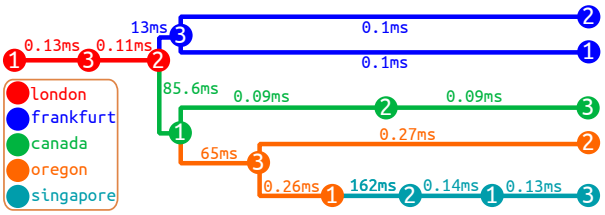


Figure 6: The tree generated by MST algorithm in the AWS network cloud infrastructure using RTT as a metric (in milliseconds).

In the presence of cross-traffic in the links, the same tree provides **34.89** Mbps. The cross-traffic generated for the test is shown in Figure 4. Cells with red bold values are related to paths built with RTT met-rics.

5.2 AvB Metric

In the case of links without additional traffic, AvB metric shows it's inefficiency. Since the links in the network are relatively pure from the interfering traffic, the AvB estimation probing returns similar met-rics, e.g 100 Mbps for all edges. This leads to uncertainty in the tree construction. The trees obtained during all 10 trials of an experiment with a "pure" network, were completely different and gained from 15 - 24 Mbps, the average data rate of the experiments is **23.65** Mpbs. The Figure 7 demonstrates one of the given trees.

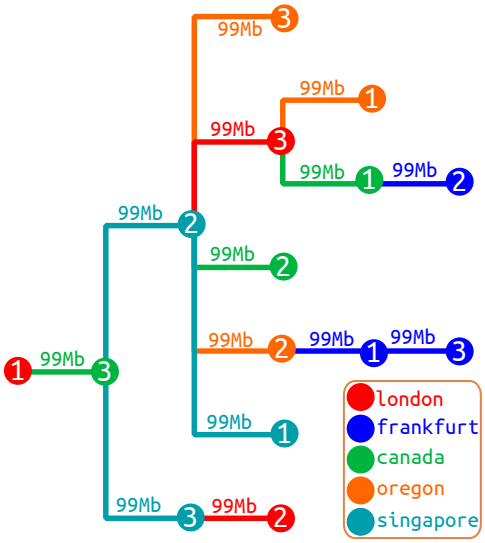


Figure 7: The tree generated by MST algorithm in the pure network cloud infrastructure using AvB as a metric (in Mbps).

However, in the presence of cross-traffic, shown in Figure 4, the edges of the tree, obtained with MST

highlighted with red borders. The tree constructed us-ing the AvB metric is shown in Figure 8. The data rate achieved in the test is **44.65** Mbps, which is similar to the result of an experiment with the RTT metric with-out cross-traffic.

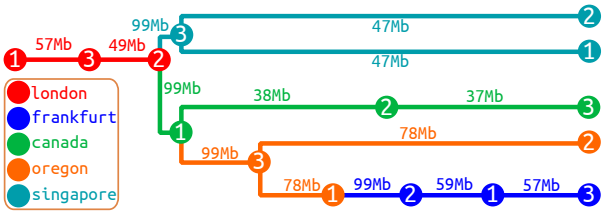


Figure 8: The tree generated by MST algorithm in the AWS network cloud infrastructure in the presence of cross-traffic using AvB as a metric (in Mbps).

6 CONCLUSIONS

The results of experiments allow to conclude the fol-lowing:

1) There are cases when available bandwidth as a metric for ALM can achieve higher performance for data transmission than using delay between hosts as a metric.

2) AvB gives insufficient results in the case of a ”pure” network.

3) Even in the modest presence of cross-traffic, the given transport system based on the RTT metric loses about 20 % of the data rate. This can be seen in the resulting Table 2.

4) For the AvB, as well as for delay metric it’s not necessary to be as precise as possible, it’s only necessary to establish the right relation between all edges of the network graph.

Table 2: Results of the data transmission experiments in Mbps.

	AvB	RTT
Cross Traffic	44.65	34.89
Pure Network	23.65	45.81

Since there are cases where one metrics is prefer-able than others to get the higher performance of the system, it makes sense to find criteria for the decision which metric should be used in the given situation. Finding such criteria is a preferable approach for the ongoing work. These criteria will be helpful in the process of developing the metric which will take into account both factors at once RTT and AvB.

ACKNOWLEDGMENTS

This work has been funded by Volkswagen Foundation for trilateral partnership between scholars and scientists from Ukraine, Russia and Germany within the CloudBDT project: “Algorithms and Methods for Big Data Transport in Cloud Environments”.

REFERENCES

[1] A. V. Bakharev, E. Siemens and V. P. Shuvalov, “Analysis of performance issues in point-to-multipoint data transport for big data,” in 2014 12th International Conference on Actual Problems of Electronics Instrument Engineering (APEIE). IEEE, 2014, pp. 431-441.

[2] K. Karpov, D. Kachan, N. Mareev, V. Kirova, D. Syzov, Siemens and V. Shuvalov, “Adopting Minimum Spanning Tree Algorithm for Application-Layer Reli-able Mutlicast in Global Mutli-Gigabit Networks,” in Proceedings of the 7th International Conference on Ap-plied Innovations in IT, 2019.

[3] V. Kirova, E. Siemens, D. Kachan, O. Vasylenko and K. Karpov, “Optimization of Probe Train Size for Available Bandwidth Estimation in High-speed Net-works,” in MATEC Web of Conferences, vol. 208. EDP Sciences, 2018, p. 02001.

[4] V. Kirova, K. Karpov, E. Siemens, I. Zander, O. Va-sylenko, D. Kachan and S. Maksymov, “Impact of Modern Virtualization Methods on Timing Precision and Performance of High-Speed Applications,” Future Internet, vol. 11, no. 8, p. 179, 2019.

[5] M. Amad, A. Boudries and L. Badis, “Application Layer Multicast Based Services on Hierarchical Peer to Peer Architecture,” Applied Mechanics and Materials, vol. 892, pp. 64-71, June 2019.

[6] R. Muzaffar, E. Yanmaz, C. Raffelsberger, C. Bettstet-ter and A. Cavallaro, “Live multicast video stream-ing from drones: an experimental study,” Autonomous Robots, vol. 44, no. 1, pp. 75–91, January 2020.

[7] A. Sampaio and P. Sousa, “An adaptable and ISP-friendly multicast overlay network,” Peer-to-Peer Networking and Applications, vol. 12, no. 4, pp. 809-829, July 2019.

[8] Q. Liu, R. Tang, H. Ren and Y. Pei, “Optimizing multicast routing tree on application layer via an encoding-free non-dominated sorting genetic algorithm,” Applied Intelligence, vol. 50, no. 3, pp. 759-777, March 2020.

[9] N. Mareev, D. Kachan, K. Karpov, D. Syzov and E. Siemens, “Efficiency of BQL Congestion Control under High Bandwidth-Delay Product Network Conditions,” in Proceedings of the 7th International Conference on Applied Innovations in IT, 2019.

Dynamic Real-Time Stream Reservation for IEEE 802.1 Time-Sensitive Networks with OpenFlow

Martin Böhm, Jannis Ohms, Manish Kumar, Olaf Gebauer and Diederich Wermser
Research Group Communication Systems, Ostfalia University, Salzdahlumer Str. 46/48, D-38302 Wolfenbüttel
{*ma.boehm, jannis.ohms2, m.kumar, ola.gebauer, d.wermser*}*@ostfalia.de*

Keywords: Time-Sensitive Networking, Software-Defined Networking, Network Management, Industry 4.0.

Abstract: Industrial network communication requirements are changing within Industry 4.0. Current static industrial networks will require flexibility and need on demand stream reservation with real-time capabilities. Time-sensitive Networking (TSN) offers real-time communication for Ethernet while also providing a mechanism to dynamically request streams (IEEE 802.1Qcc). The standard does not provide concrete specifications for the implementation. This paper evaluates the OpenFlow protocol known from Software-Defined Networking (SDN) for network management in TSN-networks. Requirements for a centralized TSN-controller were identified and OpenFlow has been evaluated if it can fulfill these requirements. An architecture for a TSN-controller has been presented. A proof-of-concept has been implemented and evaluated.

1 INTRODUCTION

Future production facilities are changing within Industry 4.0. New needs emerge for self configuration of network devices. This flexibility allows devices to request their own communication streams. Furthermore, networks have to react on changed configuration while monitoring its health e.g. link failures. These features are not new but getting more demanded when including control and field levels. Current Ethernet-based real-time solutions are often incompatible to standard Ethernet [1] and are also proprietary. As an open standards solution, Time-Sensitive Networking (TSN) guarantees "packet transport with bounded latency, low packet delay variation, and low packet loss" [2] in IEEE 802 networks. The IEEE 802.1Qcc standard [3] specifies on demand TSN stream reservation which is visualized in Figure 1. In combination with the machine to machine communication protocol OPC Unified Architecture (OPC UA), a standard for OPC UA over TSN [4] is currently in standardization. Within the OPC UA PubSub [5] architecture, a centralized communication broker handles the registration of TSN-streams for all communication partners.

The IEEE 802.1Qcc standard does not provide concrete specifications regarding implementation which raises questions for the selection of a User/Network Interface (UNI) protocol for user requests, algorithms for schedule calculation and the selection for a protocol for deploying configurations.

This paper proposes the OpenFlow [6] protocol for the configuration deployment. OpenFlow is commonly used for Software-Defined Networks [7] and already offers a high degree of flexibility. Based on a requirements analysis, an architecture is presented which integrates OpenFlow in the context of TSN. A working prototype has been implemented using an existing open source SDN-controller in accordance with IEEE 802.1Qcc.

This paper is structured as follows. First, the basics of TSN and SDN are presented. Section 3 discusses related work. In Section 4 features and requirements for the in Section 5 presented architecture are described. A proof-of-concept implementation, a testbed and an evaluation is presented in Section 6. Finally, Section 7 concludes and presents future work.

2 BASICS

This chapter gives an overview of the basic functions of TSN. A more detailed view of the IEEE 802.1Qcc standard is given. Later, information about SDN and the OpenFlow protocol are provided.

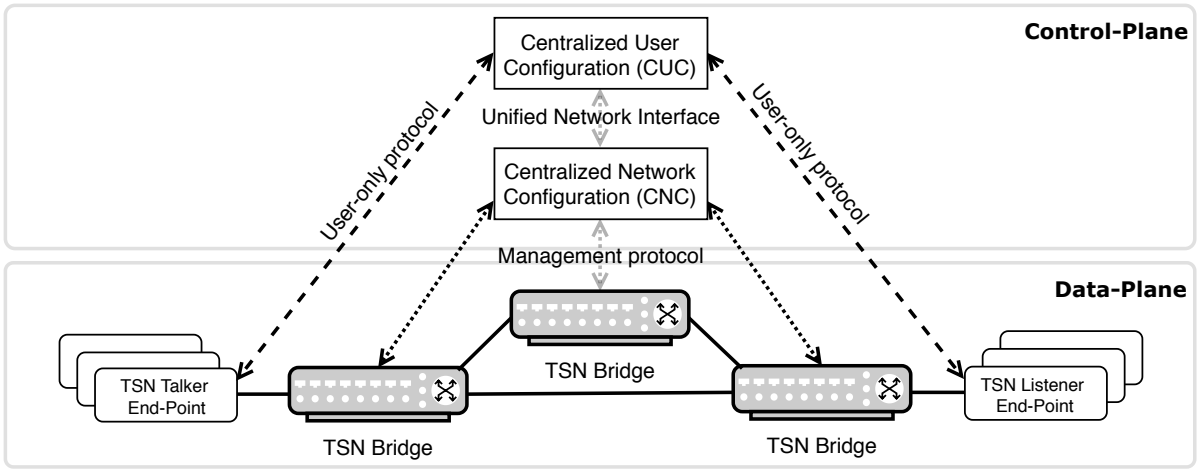


Figure 1: IEEE 802.1Qcc - Fully centralized model.

2.1 TSN Basics

TSN aims on deterministic communication in 802 networks. A bounded latency is achieved by the use of time-slots for network devices (IEEE 802.1Qbv - Enhancements for Scheduled Traffic [8]). Traffic is divided into traffic classes (TC) and assigned to time-slots with cyclical repetition. A configuration is specified in a Gate Control List (GCL). It defines the opening and closing of gates of queues based on the current time. An end-to-end connection in TSN is called a stream. It can for example be identified by the MAC address, IP address or the transport proto-col port. All devices on a TSN-stream path have to be configured properly to transfer frames of a TC. This requires a network-wide precise time-synchronization (IEEE 802.1AS-Rev - Timing and Synchronization for Time-Sensitive Applications [9]). This standard specifies the use of the Precision Time Protocol (PTP) in the context of TSN. Furthermore TSN provides a standard for reliability (IEEE 802.1CB - Frame Replication and Elimination for Reliability [10]), where frames are replicated to be transferred over multiple paths while the duplicated packet is eliminated later. Another standard provides frame preemption (IEEE 802.1Qbu - Frame Preemption [11]), where time-critical frames can suspend the transmission of a non-time-critical frame which will be resumed later.

2.2 Dynamic Stream Reservation

TSN also introduces on demand stream reservation for deterministic streams. An interface for stream requests, a module for schedule calculation for all network devices and the deployment of the configurations are addressed (IEEE 802.1Qcc - Stream Reservation Protocol (SRP) Enhancements and Performance Improvements).

The standard defines three different architectural models for the realization.

1) Fully distributed model: In a decentralized manner without any centralized configuration entities, applications can request their streams directly over the network by propagating the request along the topology using an UNI protocol. Each bridge on a path configures itself with the requirement information given in the request within their limited knowledge of the network.

2) Centralized network/distributed user model: Due to the computational complexity which raises with the amount of devices and streams, a centralized entity, called Centralized Network Configuration (CNC), is introduced. The CNC has global knowledge over all streams and devices in a network. Similar to the fully distributed model, stream requests are send directly over the UNI. The first bridge directs the request to the CNC which configures the bridges after finishing the computation and the generation of the bridges GCLs.

3) Fully centralized model: For more complex use cases, where the talkers and listeners have to be configured too, a Centralized User Configuration (CUC) is introduced as visualized in Figure 1. It discovers end stations and their capabilities, handles application requirements and configures TSN features in the end stations. The CUC forwards the stream information to the CNC using the UNI.

2.3 Software-Defined Networking

Software-Defined Networking (SDN) decouples the data-plane from the control-plane which are conventionally located in the same devices like in

switches and routers. The control-plane defines how frames/packets are forwarded in a devices specific forwarding table, called flow table. In an SDN-domain, all SDN-switches are connected to a logically centralized SDN-controller. Forwarding rules are specified in the application-plane where applications for different purposes decide how to route the traffic. An application can be a shortest path routing application or a firewall. Communicationbetween the SDN-switches and the SDN-controller takes place over the southbound interface. Here, the OpenFlow [6] protocol is dominant and usually supported by all SDN-switches.

3 RELATED WORK

An approach to combine TSN and SDN was first mentioned by Nayak et al. [12]. Their work, called Time-Sensitive Software-Defined Networking, focused on the calculation of schedules using Integer Linear Programming (ILP). Dürkop et al. presented an approach for the automatic configuration of real-time Ethernet (RTE) solutions [13]. Their approach was based on Dynamic Host Configuration Protocol (DHCP). Du and Herlich et al. also proposed the usage of SDN for the network management in RTE [14, 15]. Their proof-of-concept implementation is based on the Powerlink protocol [16]. This RTE implementation works with off-the-shelf switches. Changes on the data-plane are not required. The Powerlink protocol uses a special token to provide deterministic media access. The Powerlink protocol used by Du and Herlich et al. uses different concepts compared to IEEE 802.1 TSN. This paper proposes OpenFlow as a network management protocol for TSN-networks.

4 REQUIREMENTS

This Chapter describes requirements for a TSN-controller. Based on these, OpenFlow is evaluated. Additional features which OpenFlow can not provide are described.

The following requirements were identified for the TSN-controller which includes the CUC and CNC.

- **Topology detection:** The controller needs to be able to detect the topology of the bridges associated.
- **Host detection:** The controller needs to detect each talker and listener in its TSN-network.
- **Time-synchronization:** TSN-bridges, talkers, listeners and the controller need a common time-base (IEEE 802.1As-rev) for the use of IEEE 802.1Qbv.
- **Time-Aware Shaper:** The GCL of each TSN-bridge needs to be calculated and configured.
- **Traffic Classes:** The controller needs to assign Ethernet frames to a queue of the time-aware shaper.
- **Ingress/Egress Policing and Metering:** The controller requires a mechanism to assure that each TSN-stream adheres to the amount of bandwidth it requested.
- **UNI for Talkers/Listeners:** The controller needs to provide a user/network interface (UNI) which offers the ability to request TSN-streams.

Table 1 verifies, if the requirements for the TSN-controller can be fulfilled with the OpenFlow protocol.

OpenFlow does not provide management functionalities for time-aware shapers and also does not provide time-synchronization. For the time-synchronization an additional protocol like Precision Time Protocol (PTP) [17], which is a master/slave protocol, has to be used. The management of the time-aware shapers can either be implemented as a protocol extension of the OpenFlow protocol in the form of experimenter messages or by the use of existing configuration protocols like NETCONF [18]. The UNI for the stream request can be implemented as an extension of the controller. Besides the control-plane, data-plane devices need to support time-synchronization (e.g. PTP) and time-aware shapers.

5 ARCHITECTURE

This chapter presents an architecture for a TSN-controller based on the requirements presented in Chapter 4. The architecture is shown in Figure 2. On the top, it shows the TSN-controller while at the bottom, the functions for a compatible TSN-bridge are shown. Both will further be described.

5.1 TSN-Controller

First of all, the TSN-controller is separated by the CUC and the CNC. The CUC constists of an *Endpoint Request Handler* to provide an UNI which is compatible to IEEE 802.1Qcc [3]. It can be implemented as a REST API. Requests are forwarded from the CUC to the CNC. The *Path Control and Reservation* module has global knowledge about the network and finds paths through the network while re-

TSN-Controller Functions	OpenFlow
Topology detection	OpenFlow networks use LLDP to detect available links.
Host detection	In OpenFlow networks, the controller detects new hosts with the ARP protocol. Each ARP frame received by an OpenFlow-switch is copied and forwarded to the controller.
Time-synchronization	OpenFlow does not provide mechanisms for time-synchronization.
Time-Aware Shaper	OpenFlow provides credit-based shaping to reserve bandwidth for a specific traffic class. There are currently no time-based shapers in the OpenFlow specification.
Traffic Classes	OpenFlow provides an enqueue action which can be used to assign an Ethernet frame to a queue. These queues can be used to implement traffic classes.
Ingress/Egress Policing and Metering	OpenFlow provides ingress and egress metering.
UNI for Talkers/Listeners	OpenFlow does not provide an UNI.

Table 1: OpenFlow protocol feature evaluation for TSN network management

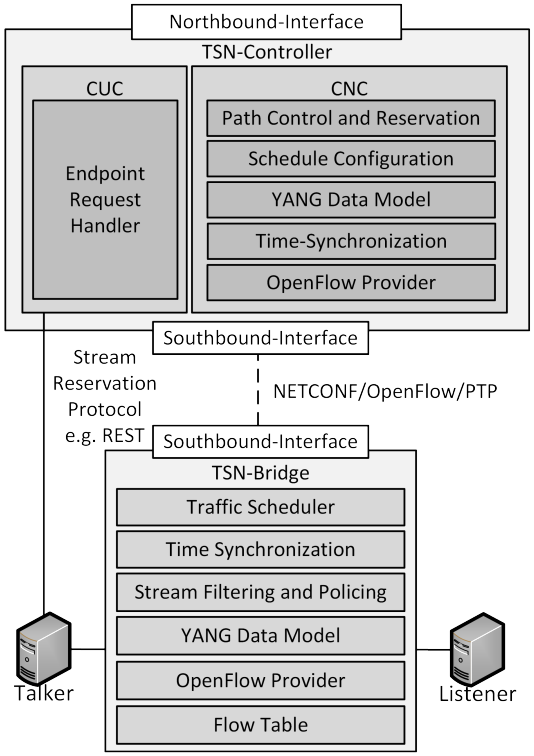


Figure 2: Architecture of the TSN-controller.

specting the TSN-bridges utilization. In the *Sched-ule Configuration* module, a configuration (GCL) for each device on a path is calculated. This process has a very high algorithmic complexity and a lot of research is taking place in this area [19, 20]. The configura-tion has to be represented in a format which can be applied by the TSN-bridges. Here, the *YANG Data Model* is used which is transferred using the *NET-CONF* protocol. For the *Time-Synchronization*, the TSN-controller needs to be part of a PTP-domain.

Logically, the master clock should be located in the controller. The *OpenFlow Provider* is used to configure the forwarding behaviour of the TSN-bridges.

5.2 TSN-Bridge

First of all, TSN-bridges need to support a *Traffic Scheduler* resp. IEEE 802.1Qbv. For proper functioning of the *Traffic Scheduler*, the TSN-bridge is timely synchronized. The *Stream Filtering and Policing* module takes care, that TSN-streams do not exceed their requested resources. The TSN-bridge needs to be compatible with a *YANG Data Model* to offer flexible reconfiguration. Over the *OpenFlow Provider*, the TSN-bridge is able to be configured with the OpenFlow protocol. Forwarding behaviour is located in the *Flow Table*.

6 IMPLEMENTATION AND EVALUATION

This chapter describes a proof-of-concept implementation based on the architecture presented in Chapter 5. Later the proof-of-concept will be evaluated.

6.1 Implementation

The proof-of-concept implementation is based on the open-source SDN-controller Ryu [21] written in Python. The data-plane consists of two TSN-bridges (Trustnode) from the company Innoroute which already support IEEE 802.1Qbv, PTP, Netconf and OpenFlow. The Openflow implementation on the bridges is based on Open vSwitch [22]. The talker uses an Intel i210 network card and a kernel extension

to support time stamped packet transmission for TSN-traffic [23]. Except the TSN-briges, all system are based on Ubuntu 18.04 using an Intel Core i7-6700 CPU and 16GB RAM. Every module from Figure 2 is implemented separately. The schedule calculation is simplified due to the complexity of this module. For the CUC interface, the existing REST API from the Ryu controller has been extended.

6.2 Evaluation

To evaluate the proof-of-concept implementation, a test-bed has been set-up as visualized in Figure 3.

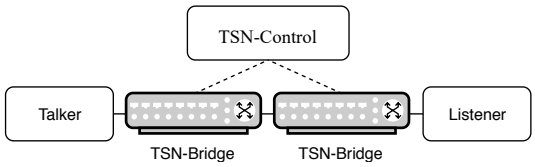


Figure 3: Architecture of the TSN test-bed.

Two aspects of the system have been evaluated. The function and performance of the time-aware shaper and the overall set-up time of a TSN-stream. To test the time-aware shaper, two streams have been requested using two different traffic classes. TC 1 has a time-slot length of 7 ms and TC 2 has a time-slot length of 3 ms resulting in a total cycle length of 10 ms. The talker generates frames at a rate of 100 μs (10 frames per ms) for both TC. Figure 4 shows the arrival of the packets on the listener site. The packets are distributed according to their specified time-slots. Once a time-slot starts, all buffered frames, which arrived outside of their time-slot, are sent.

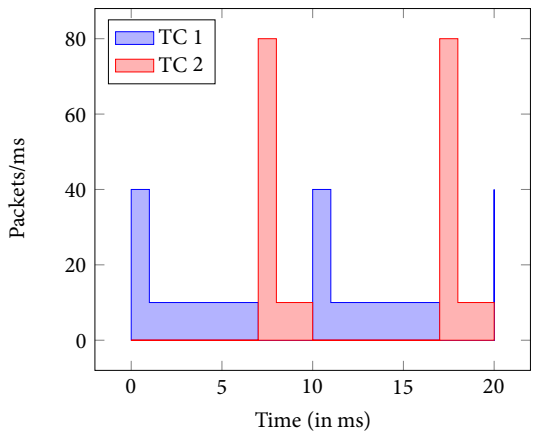


Figure 4: IEEE 802.1Qbv scheduled traffic - 10ms cycle time, 2 traffic classes (7ms and 3ms).

The second measured aspect was the set-up time of the UNI. Here, the talker requests 1 stream every second. Each stream is removed before the next one is created. The UNI protocol request response time, the time between request and response, is measured using Wireshark. For 1000 requests an average set-up time of 3.176 ms with a standard deviation of 1.031 ms has been measured.

The results show, that existing software which is originally developed for SDN can be easily extended to support TSN. It also shows, that TSN-streams can be requested within a few milliseconds. It has to be noted, that the calculation for the schedule is simplified in this proof-of-concept implementation.

7 CONCLUSION

This paper evaluated the use of OpenFlow for an implementation of a TSN-controller with respect to IEEE 802.1Qcc. Requirements for a TSN-controller have been identified, and verified if OpenFlow can fulfill these requirements. Potential extensions and companion protocols have been discussed and an architecture for a TSN-controller has been presented. Later, a proof-of-concept has been implemented with real hardware. The implementation has been evaluated while achieving an average set-up time of 3.176 ms.

OpenFlow itself is only able to partly fulfill the requirements identified. More protocols are needed for a full-featured TSN-controller. The proof-of-concept shows the feasibility of dynamic TSN-stream registrations.

In the future more TSN standards like frame preemption have to be investigated and added to the architecture presented in this paper. OpenFlow should also be considered for the realization of IEEE 802.1CB (Frame Replication and Elimination for Reliability). OpenFlow allows the duplication of frames and the forwarding on multiple output ports.

ACKNOWLEDGEMENTS

This work was partly funded by the Ministry for Science and Culture of Lower Saxony as a part of the research project SecuRIn (VWZN3224) and the Federal Ministry for Education and Research within the KMU-innovativ program as a part of MONAT (16KIS0782).

REFERENCES

[1] M. Wollschlaeger, T. Sauter and J. Jasperneite, "The future of industrial communication: Automation net-works in the era of the internet of things and industry 4.0," IEEE Industrial Electronics Magazine, vol. 11, no. 1, 2017, pp. 17-27.

[2] IEEE 802.1 Working Group, "Time-Sensitive Networking (TSN) Task Group", [Online]. Avaible: https://1.ieee802.org/tsn/.

[3] "IEEE Standard for Local and Metropolitan Area Networks – Bridges and Bridged Networks – Amendment 31: Stream Reservation Protocol (SRP) Enhancements and Performance Improvements," IEEE Std 802.1Qcc-2018 (Amendment to IEEE Std 802.1Q-2018 as amended by IEEE Std 802.1Qcp-2018), pp. 1-208, October 2018.

[4] D. Bruckner, R. Blair, M. Stanica, A. Ademaj, W. Skeffington, D. Kutscher, S. Schriegel, R. Wilmes, K. Wachswender, L. Leurs et al., "Opc ua tsn-a new solution for industrial communication," Whitepaper. Shaper Group, 2018.

[5] OPC Foundation. Part 14: Pubsub. [Online]. Available: https://opcfoundation.org/developer-tools/specifications-unified-architecture/part-14-pubsub

[6] Open Networking Foundation. Openflow switch specification version 1.5.1. [Online]. Avail-able: https://www.opennetworking.org/wp-content/uploads/2014/10/openflow-switch-v1.5.1.pdf

[7] ONF, "SDN architecture," Open Networking Foundation, Tech. Rep. Issue 1, TR-502, 2014. [Online]. Available: https://www.opennetworking.org/wp-content/uploads/2013/02/TR_SDN_ARCH_1.0_06062014.pdf

[8] "IEEE Standard for Local and metropolitan area networks – Bridges and Bridged Networks – Amendment 25: Enhancements for Scheduled Traffic," IEEE Std 802.1Qbv-2015, pp. 1-57, March 2016.

[9] "IEEE Draft Standard for Local and Metropolitan Area Networks – Timing and Synchronization for Time-Sensitive Applications," IEEE P802.1AS-Rev/D7.0, March 2018, pp. 1-496, August 2018.

[10] "IEEE Standard for Local and metropolitan area networks – Frame Replication and Elimination for Reliability," IEEE Std 802.1CB-2017, pp. 1-102, Oct 2017.

[11] "IEEE Standard for Local and metropolitan area networks – Bridges and Bridged Networks – Amendment 26: Frame Preemption," IEEE Std 802.1Qbu-2016 (Amendment to IEEE Std 802.1Q-2014), pp. 1-52, August 2016.

[12] N.G. Nayak, F. Dürr, and K. Rothermel, "Software-defined Environment for Reconfigurable Manufactur-ing Systems," in 2015 5th International Conference on the Internet of Things (IOT). IEEE, 2015, pp. 122-129.

[13] L. Dürkop, J. Jasperneite and A. Fay, "An Analysis of Real-Time Ethernets With Regard to Their Automatic Configuration," in WFCS, 2015, pp. 1-8.

[14] J.L. Du and M. Herlich, "Software-defined Networking for Real-time Ethernet," in ICINCO (2), 2016, pp. 584-589.

[15] M. Herlich, J.L. Du, F. Schörghofer and P. Dorfinger, "Proof-of-concept for a Software-defined Real-time Ethernet," in 2016 IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA). IEEE, 2016, pp. 1-4.

[16] W. Wallner and J. Baumgartner, "openpowerlink in linux userspace: Implementation and performance evaluation of the real-time ethernet protocol stack in linux userspace," in Proc. 13th Real-Time Linux Workshop (RTLWS).(Prague, Czech Republic, 2011, pp. 155-164.

[17] "IEEE Standard Profile for Use of IEEE 1588 Precision Time Protocol in Power System Applications," IEEE Std C37.238-2017 (Revision of IEEE Std C37.238-2011), pp. 1-42, June 2017.

[18] A. Bierman and M. Bjorklund, "Network Configuration Protocol (NETCONF) Access Control Model," Internet Requests for Comments, RFC Editor, RFC 6536, March 2012.

[19] S.S. Craciunas, R.S. Oliver, and W. Steiner, "Formal scheduling constraints for time-sensitive networks," arXiv preprint arXiv:1712.02246, 2017.

[20] W. Steiner, S. S. Craciunas, and R.S. Oliver, "Traffic planning for time-sensitive communication," IEEE Communications Standards Magazine, vol. 2, no. 2, pp. 42-47, 2018.

[21] Ryu SDN Framework Community. Ryu sdn framework. [Online]. Available: https://osrg.github.io/ryu/

[22] B. Pfaff, J. Pettit, T. Koponen, E. J. Jackson, A. Zhou, J. Rajahalme, J. Gross, A. Wang, J. Stringer, P. Shelar et al., "The design and implementation of open vswitch." in NSDI, vol. 15, 2015, pp. 117-130.

[23] M. Kumar, M. Boehm, J. Ohms, O. Shulha, and O. Gebauer, "Evaluation of the time-aware priority queueing discipline with regard to time-sensitive networking in particular ieee 802.1 qbv," in Proceedings of International Conference on Applied Innovation in IT, vol. 7, no. 1. Anhalt University of Applied Sciences, 2019, pp. 1-6.

Ultra Reliable Advanced Framework for Emergency and Mission Critical Data for 5G Services

Simon Bojadzievski[1], Marija Kalendar[2] and Tomislav Shuminoski[2]
[1]A1 Macedonia, A1 Telekom Austria Group, Ploshtad Presveta Bogorodica 1, 1000 Skopje, R.N. Macedonia
[2]Ss. Cyril and Methodius University, Faculty of Electrical Engineering and Information Technologies,Rugjer Boshkovik 18, PO Box 574, 1000 Skopje, R.N. Macedonia
simon.bojadzievski@a1.mk, {marijaka, tomish}@feit.ukim.edu.mk

Keywords: 5G, Emergency, IoT, Mission Critical Data, Reliability, Ultra-Reliable Low Latency.

Abstract: The paper provides a novel approach of using heterogeneous devices and future networks, with impact on ultra-reliable services for the next generation 5G networks and emergency mission critical data. The user-centric framework used here, together with the network interoperability, as well as the symbiosis with emergency communications systems, complements the IoT systems and the heterogeneous networks, enabling reliable transfer of time critical data communication. A new approach is establishing reliable communication of time-critical data, where the user is at the center and is able to use multiple available data networks to deliver the service. The architecture of the EMCD system model is based on the fundamental principals of 5G architecture, allowing IoT devices to communicate with IoT applications, hosted in a cloud datacentre. The simulation results and analysis show superior performance with a high level of ultra-reliable and low latency communications in a variety of network conditions and different network coverage. The packet duplication, the proposed emergency and mission-critical data algorithm, and multi-connectivity architectures are the basic principles that provide solution for high reliability and low latency.

1 INTRODUCTION

Digital transformation is present everywhere in our society, changing our way of communication and transforming into the digital age, where ubiquitous and reliable data connectivity is the foundation for such a transformation. Internet of Things (IoT) will connect all of objects and devices to the Internet and will exploit the full potential benefits of devices equipped with enough sensing, acting and processing capabilities. Domination of data communications between devices in the following years will many times overcome today's data transfers initiated by humans. Wireless networks are the main enabler of connected IoT, working within an environment of heterogeneous devices and networks with a variety of data types, used in various applications, such as health, networked vehicles, industrial IoT and media. Wireless sensors network are not anymore short-range small ad-hoc networks, but part of a wider ecosystem called IoT. IoT is a system that includes various types of sensing devices that communicate with smart devices, which continue to confidently and securely transfer data to the appropriate cloud platforms, where data for the respective applications are stored, archived and subsequently processed.

On the other hand, the emergency and mission critical data networks cover services, important for the security of society and citizens, and also encounter challenges for acceptance of smart phone applications based on their capability and interoperability between heterogeneous networks.

Despite the enormous development advances of technology and the standardization of new technologies, commercial use of digital applications that require highly reliable communications in case of emergencies and mission-critical events is still in the early stages. Moreover, according to International Telecommunication Union (ITU) the 5G is classified into ultra-reliable low latency communications (URLLC), enhanced mobile broadband (eMBB), and massive machine-type communications (mMTC).

Motivated by such a situation, this paper presents analysis and simulations of how highly trusted services can be used in everyday life, by

defining the paradigm Emergency and Mission-Critical Data (EMCD) communication and presen-ting a system model for communication in case of critical events. EMCD requires ultra-reliable data communication in case of serious life threatening events, as well as possible high property damage if the critical data is not delivered in due time.

At present, the telecommunication networks are completely IP based, where applications are using higher TCP/IP layers. Emergency services, natively using data packets, can use the EMCD system model horizontally integrated between different network providers through isolated IP peering, dedicated to EMCD packet communication. End user devices should support EMCD packet communication, integrating an EMCD agent in the firmware, creating a possibility for standardization and wider usage of the EMCD system model and supporting devices within a global international EMCD data network, enabled by all telecom providers.

Next-generation data networks and 5G, in addition to availability and reliability, should provide consistent service with certain network parameters (delays, jitter, and packet loss) [1-2] that will secure new services. In comparison, existing wireless data networks are not satisfying the parameters for ultra-reliable communication. Within 5G frameworks, solutions are being investigated that could enable the required parameters for Ultra-Reliable Low Latency Communication - URLLC as in [3]. Moreover, with current 3GPP standardization, in new versions of Release 15 and 16, new mechanisms are required to address the challenges of ultra-reliable communication [4].

The paper is organized as follows. Section II gives an overview of the most relevant research work in this field. Section III presents our system model, architecture and EMCD framework. Section IV provides simulation results. Finally, Section V concludes this paper.

2 RELATED WORKS

The major goal in current and future mobile and wireless networks and services is providing a high level of QoS support for any given service and minimal latency for real-time services. One heterogeneous network environment, where 5G takes central place, requires collaboration and interoperability of all entities for greater availability and reliability [5]. 5G networks will support the massive spread of intelligent IoT nodes to support wider acceptance of mission critical communication services [6]. The proposed approach does not refer to the unification of currently most widely used WLAN and mobile networks, but to all wireless networks used for communication in the IoT systems and platforms. Also, the proposed approach considers many challenges of 5G URLLC with IoT devices [7].

On the other hand, when we focus on the architecture of the EMCD framework presented here, many similar schemes for dual-mode mobile equipment are being proposed. For example an UMTS/WLAN interworking network has been proposed in [8] and [9], but without emphasizing QoS issues and without ultra-reliable low latency issues. Similarly, a dual-mode mobile node for UMTS/WLAN is presented in [10], including implemented handover logic modules. The dual-mode user equipment design includes a monito-ring and reporting unit to determine the status of the interfaces and an interface selection unit to activate or deactivate the interfaces (UMTS and WLAN) for mobile handoff. The results indicate a smoother and seamless handoff process. The shortcoming of this model is in focusing only on mobile HO processes and not implementing any adaptive QoS framework for improving the results of other QoS parameters (including URLLC). Furthermore, [11] presents adaptive QoS framework implemented in dual-mode UMTS/WLAN mobile terminals. According to the presented results, the proposed dual-stack UMTS/WLAN mobile equipment with an adaptive QoS module, performs fairly well in different network conditions, achieving better performance but only in comparison to the cases when only WLAN or only UMTS mobile equipment has been used. Moreover, URLLC or any emergency and mission critical data are not considered.

Despite all related works, one of the advantages for our proposed EMCD system model is that it relies on the basic 5G postulates for integration of different Radio Access Technology (RAT) networks and ultra-reliability [12-14] with reliability >99,999% [15] based on the packed duplication methodology and dual RAT interfaces (or multi-connectivity [14]) on IoT devices.

Moreover, accompanying mission critical with emergency data will lead to more efficient use of resources that for the most part of the time are unused. Use of public communication infrastructures to enable emergency services for citizens, but also for IoT devices, is necessary to provide next

generation emergency service. Having in mind that separate networks for emergency data transfer, commonly owned by the government, usually exist, the symbiosis with these networks, in the event of a failure of commercial networks, can lead to greater reliability and interoperability, as proposed in [16].

The theoretical background for reliability of the proposed EMCD system is based on the packet duplication concept and corresponds to the basic principles of communication theory within system engineering [17].

3 SYSTEM MODEL AND ALGORITHM

As mentioned before, the proposed ECMD system architecture is based on the URLLC and should support a variety of services that request challenging reliability (99.999%) and latency (1 ms).

In the new 5G RAN networks, comple-mentary to improving existing Physical (PHY) layer techniques, a packet duplication protocol is intro-duced. The EMCD architecture is primarily being focused on higher layer solutions based on Packet Duplication (PD) as a practical and low complexity technique for URLLC. The theoretic framework behind PD is investigated, and the recent enhan-cements in the 5G Dual Connectivity (DC) architec-ture for supporting PD are discussed, without excessively increasing the complexity in the RAN.

The fundamental principle underlying PD, involves generating multiple instances of a packet at higher layers and transmitting the packets simultaneously over different uncorrelated channels or transmission links [18]. At the receiver, the redundancy and diversity in the channel conditions is exploited, such that higher transmission reliability is achieved. While the reliability with PD is achieved using multiple redundant links, low latency is realized by eliminating the need for packet retransmission. With PD, duplicate packets are proactively transmitted simultaneously, thus eliminating the need to use time-diversity schemes such as HARQ to satisfy the URLLC requirements.

The user plane comprises Packet Data Convergence Protocol (PDCP), RLC, MAC, and physical (PHY) layers, all of which are collectively responsible for ensuring reliable over-the-air transmission of packets in both uplink (UL) and downlink (DL) directions [19]. The radio resource control (RRC) entity, the primary control plane (CP) function in the RAN, is responsible for configuring all protocol layers in the network and the IoT device.

The architecture of Dual Connectivity with packet duplication PD in 5G is intended to provide high throughput and high reliability by enabling the use of radio resources from two access nodes with distinct schedulers of the same or different RATs [20]. So, both the master node (MN) and secondary node (SN) are connected over the Xn interface, which supports data forwarding, flow control functions, and should provide interconnectivity with guaranteed bandwidth and latency for EMCD packets. As such, only semi-static coordination at the RRC level is supported in DC, taking in consideration small packet size for IoT data communication related to EMCD. On the other hand, both MN and SN have greater flexibility in independently scheduling resources for the IoT devices, as shown in Figure 1.

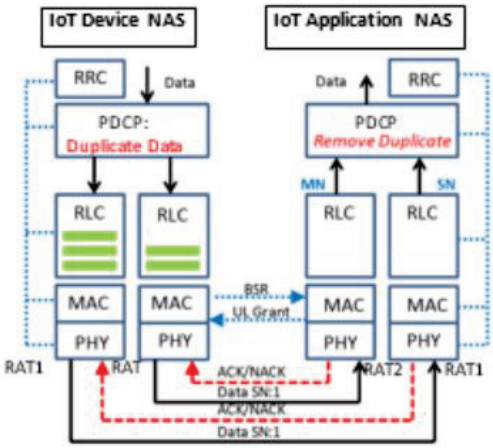


Figure 1: EMCD architecture and PD between IoT MN and SN.

In the case of 5G, both access nodes host the NR RAN protocol stack and are connected to the 5G Core Network (5GC) in a standalone NR-NR DC architecture [20] as shown in Fig. 2. In this case, the IoT MN and IoT SN are referred to as Master Next-Gen Node B (MgNB) and Secondary Next-Gen Node B (SgNB), respectively. The architecture of the referent EMCD system model (Figure 2) is based on the fundamental principals of 5G, C-RAN architecture and the architecture of the 5G Core Network, allowing IoT devices to communicate with an IoT application, hosted in a cloud datacenter. IoT devices communicate with the IoT application trough dual connectivity RAT establishing not only higher reliability, but also better handover in a case of mobility of wireless nodes with intention to reach lowest mobile interruption time.

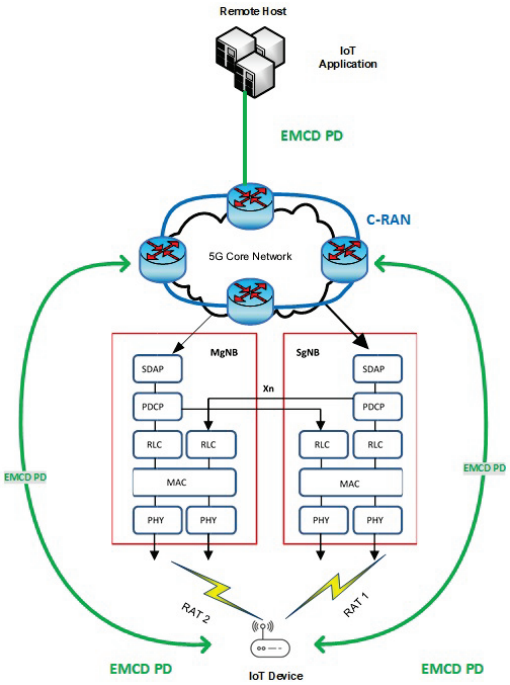


Figure 2: 5G IoT EMCD architecture.

Namely, a handover with 0 ms interruption is mandated, and extreme reliability will not tolerate any mobility failures. Consequently, softer handover concepts where the IoT device is multi-connected to a source, based on the dual connectivity principle is shown in [21].

The EMCD Algorithm for communication of the IoT device (MN) with the IoT application hosted in the edge of the RAT in a cloud (SN), is securing the reliable communication using redundant connecti-vity within the predefined WAN EMCD network.

The EMCD algorithm uses an EMCD user agent in the IoT device to send multiple copies of the same packet through independent RAN routes. The EMCD algorithm will be applied for the EMCD labelled IP packets in the IoT device and are related to emergency and mission critical situations that require fast and reliable communication. All other IP packets will not be replicated and will use a predefined network interface. Taking in considera-tion the fact that only a selected part of IP packets are replicated, the algorithm should secure a reliable mechanism, transparent in the application layer, for detection and management of duplicated IP packets.

Within the EMCD algorithm, the EMCD agent cannot control IP packets' route, accept the destination IP address and the choice of network interface to send IP packets through. The EMCD algorithm will transparently replicate IP packets, using different network interfaces connected to the EMCD WAN network segment, and will transmit them to the IP destination.

On the receiving side, the EMCD algorithm will receive the first IP packet and forward it to the upper layers, while the replicated packets will be silently discarded. The EMCD algorithm will be used in a secure and isolated EMCD WAN network environment that complies with the existing network protocols, and will not introduce additional security concerns. The EMCD WAN environment includes separate 5G network slices, separate PVC and VLAN's within the operator transport layers, and separated hidden SSID Wi-Fi segment on the end user CPE, where IoT devices will be connected. Since the EMCD algorithm will be used in heterogeneous public networks, it's obvious that the transparency of the network is one of the basic requirement, and the EMCD algorithm should be placed in the upper application layers. Additionally, part of the functionality can be moved on the transport layer, as in 5G, where the functions of replication are supported from the networks from same type.

Furthermore, since the EMCD algorithm is introduced on the application layer, all IoT devices should have an EMCD user agent incorporated, in order to support this algorithm, as well as an IoT application hosted in the cloud.

Once the EMCD session is established, the EMCD user agent is ready to transmit duplicated EMCD packets to the corresponding EMCD host. All packets related to EMCD are inspected, duplicated on the both network interfaces, and added an EMCD header containing a 32-bit sequence identification (SNID), representing a unique identifi-cation of an EMCD session.

4 SIMULATION RESULTS

Before presenting the simulation results and analysis, it is important to emphasize that the theoretical background for reliability of the EMCD system is based on packet duplication concept and corresponds to basic principles of communication theory within system engineering [17]. Reliability of the systems will increase with simultaneous use of multiple heterogenic uncorrelated RATs, where reliability of entire system can be presented as:

$$R = 1 - \prod_{i=1}^{N} (1 - R_i) \quad (1)$$

where R_i presents the subsystems' reliability and where N is the number of simultaneous fully uncorrelated RATs at a location. In the

heterogeneous wireless and mobile system with multiple sub-networks, ultra-high reliability can be achieved with sending duplicated packets trough all available networks.

For URLLC, the reliability achievable over link *i* is determined as:

$$R_i = P(l_i \leq LT)P(SNR_i > SNR_i^T)P(b_i > BT) \quad (2)$$

where:

- $P(l_i \leq LT)$ is the probability that the overall latency li (processing and propagation) over link *i* is less than the URLLC latency requirement *LT*.
- $P(SNR_i > SNR_i^T)$ is the probability that the signal-to-noise ratio (SNR) achievable on link *i* is greater than SNR_i^T, the *SNR* threshold for achieving a target block error ratio (BLER) value on link *i*.
- $P(b_i > BT)$ is the probability that b_i, the bandwidth allocation on link *i*, is greater than *BT*, the bandwidth required to transport an URLLC packet.

We will present simulations for end-to-end reliability depending on the link delays and reliability for different network coverage, based on the EMCD system model using multiple communication links. As described previously, in the EMCD model, usage of multiple links is in the same time connected with duplication of packets, while copies of packets are delivered at the same time through multiple links, according to the EMCD algorithm. Since mobile and IoT devices today have multiple wireless radio interfaces (4G, WLAN, NFC, Bluetooth etc.), the next generation of devices based on 5G are expected to support all backward radio technologies. Existence of multiple links connectivity, used with PD will improve our system towards the ultra-reliable EMCD services.

Packet Duplication with the EMCD algorithm and Multi-Connectivity architectures in the EMCD system model are the basic principles that provide solutions for high reliability and low latency. Based on these principles, losses within the radio networks due to fading and interference on individual links or possible network outage will be compensated with the copies of the packet travelling through the diverse infrastructure. Having in mind that packet duplication is used, for the EMCD data transfer the handover time is equal to zero.

Since, each geographical area has its own specificity in terms of terrain and RAT coverage, in order to create a simulation model that can realistically include these parameters, the geographic areas will be divided into three generic parts related to population: urban, suburban and rural areas. Since, the raw data we use, incorporate the information about availability per cell, we decided to divide the geographical areas according to population, thus forming groups of base stations and calculating the reliability of each of these areas for different technologies.

Usually, the areas that are densely populated, have a high number of wireless and mobile networks that cover all varieties of mobile and IoT devices (with incorporated EMCD). The mobile networks of all operators usually have full coverage of urban areas, using a large number of overlapping mobile cells and providing high-quality spatial and traffic signal coverage.

For further analysis of the regional impact by the EMCD framework, we will take advantage of the fact that the coverage of LTE networks is uneven across regions and network reliability in certain regions is lower. In addition to simulating our EMCD framework, the system uses historical real data as a reference value when comparing single current technologies. Furthermore, we will consider that the reliability of the second LTE network has different reliability in the aforementioned regions and we will see its impact over the overall reliability.

Table 1: Regional reliability of 2G, 3G, 4G and WLAN.

	2G (EDGE)	3G (HSPA)	4G (LTE)	WLAN
Urban	0.998	0.988	0.998	0.965
Sub-urban	0.988	0.986	0.983	0.931
Rural	0.92	0.91	0.912	0.882

Furthermore, Table 1 and 2 present regional reliability for urban, sub-urban and rural areas for different mobile and wireless technologies for the three simulation scenarios. As shown in Table 2, the confidentiality ratio of the two LTE networks, for the three scenarios, is given separately.

In the first scenario we will consider simulating the EMCD framework compared to single 2G, 3G and LTE reference models. The assumption is that the second mobile network has lower reliability than the first one, and it is different for each of the areas: 90%, 80% and 60% respectively. In the first scenario, we will consider simulating the EMCD framework compared to single 2G, 3G and LTE reference models. The assumption is that the second mobile network has lower reliability than the first one, and it is different for each of the areas: 90%, 80% and 60% respectively.

Figure 3 depicts the simulation results for the average reliability of different networks and of our

proposed EMCD framework in the urban, sub-urban and rural regions, when simulation scenario 1 is used.

Table 2: Regional reliability for proposed scenarios.

	LTE2/LTE1 Scenario1	LTE2/LTE1 Scenario2	LTE2/LTE1 Scenario3
Urban	0.9	0.9	0.95
Sub-urban	0.8	0.9	0.95
Rural	0.6	0.8	0.95

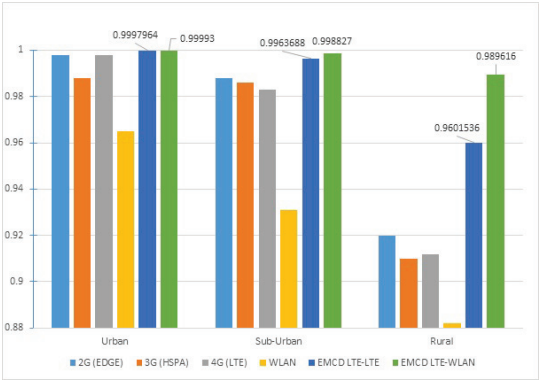


Figure 3: Regional reliability of different networks and EMCD framework for Scenario 1.

In Scenario 1, each technology has different values for the respective areas. As expected, the reliability for each of the technologies is particularly high in urban areas. Consequently, the simulation values for the EMCD framework in this area are very close to ultra-reliable requirements, because the reliability in urban areas is higher than in other regions. Next, comes the suburban area, and the lowest reliability is presented in the rural area.

Even though single systems (in each technology), don't have redundant links, they present comparable results to the EMCD framework, but for urban areas only. This is because the mobile system is redundant itself, i.e. within a single mobile system if the device is in the zone of coverage of two or more base stations and if one of them gets interrupted, the others within range of the device have predefined algorithms to seamlessly connect the device to the next available one. At the same time, to meet the high demand for data, in urban areas a significant number of base stations from several mobile operators and independent WLAN networks exist. The EMCD framework offers greater advantages in sub-urban and especially in rural areas compared to urban areas, with the greatest improvement over a single link being presented for the EMCD framework in rural areas.

Comparing the reliability of the EMCD model simulated with two LTE networks to the EMCD model using LTE and WLAN we can see that if we have one of the LTE networks with significantly lower reliability in certain regions, the $EMCD_{LTE\text{-}LTE}$ implicit reliability shows lower reliability than $EMCD_{LTE\text{-}WLAN}$.

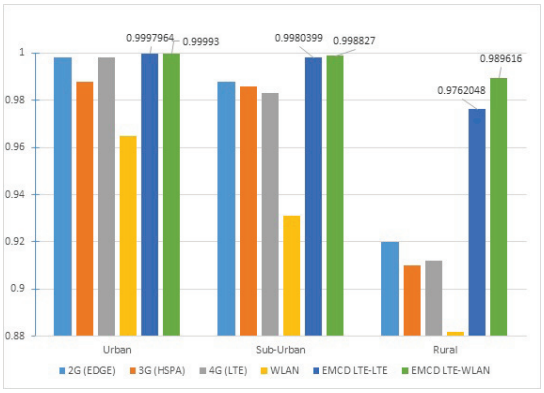


Figure 4: Regional reliability of different networks and EMCD framework for Scenario 2.

In addition, Figure 4 shows simulation results for the average reliability of different networks and of our proposed EMCD framework, for Scenario 2.

The scenario 2 presents the simulation of both types of EMCD frameworks compared to single 2G, 3G and LTE reference models. The assumption is that the second mobile network has lower reliability than the first one, and it is different for each of the areas: 90%, 90% and 80% respectively. Although the second LTE network in the second scenario has significantly better reliability in the rural areas, still the effective reliability of $EMCD_{LTE\text{-}LTE}$ is lower than that of $EMCD_{LTE\text{-}WLAN}$.

Scenario 3 presents simulation of the two types of EMCD models, compared to each other and with respect to single technologies. Again the assumption is that the second mobile network has lower reliability than the first one and is 95% for all regions. Again, Figure 5 presents the simulation results for the average reliability of the different networks and our proposed EMCD framework for this third scenario. In this Scenario 3, we consider that the second LTE network has reliability that is 5% lower than the reliability of the first LTE network. However, despite relatively minor difference in reliability of both mobile networks, we see that this has an impact on the $EMCD_{LTE\text{-}LTE}$ model.

Figure 5 shows that for these reliability values regarding the cellular networks, the reliability values for both EMCD models are almost identical in each of the regions, with insignificant differences.

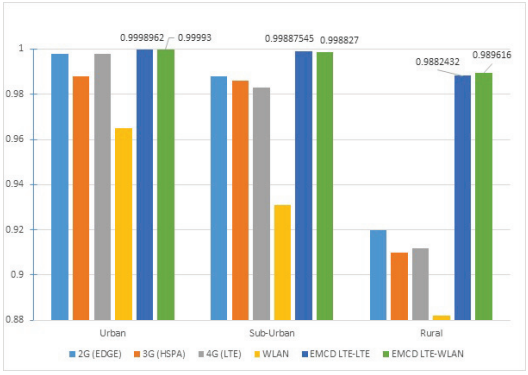


Figure 5: Regional reliability of different networks and EMCD framework for Scenario 3.

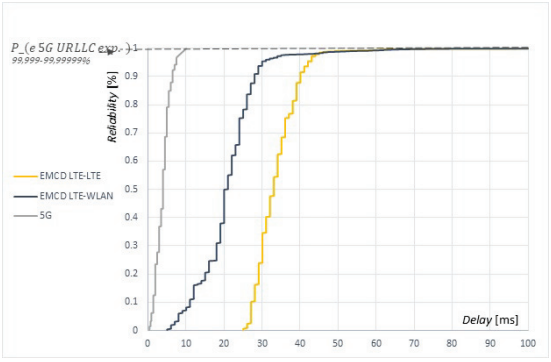


Figure 6: Average reliability of different EMCD models and 5G for URLLC vs delay.

Figure 6 shows a comparison of the EMCD model to the expected/desired theoretical values of 5G technology for URLLC. As it can be seen, simulated reliability values for the EMCD model are quite comparable to 5G technology, but with a significant difference in delays. Based on the results from the three scenarios, we conclude that differences in reliability between the networks of two different operators, plays a great influence on the resultant reliability for the EMCD system. Taking into consideration the regional reliability of each of the networks, the reliability of proposed EMCD framework may vary from one region to another, depending on the reliability of single networking radio technologies. The simulations of the EMCD model in the previous sections were generally aimed at simulated system reliability with respect to ultra-reliable applications. On the other hand, one of the interesting feature reuirement introduced by 5G technology is the URLLC support, which requires high reliability of the system and small latency at the same time.

Thus, taking this into account, Fig. 6 presents the simulation of the two EMCD models: $EMCD_{LTE-LTE}$ and $EMCD_{LTE-WLAN}$ compared to each other and with respect to the parameters that are expected to be supported by the 5G network, through their cumulative distribution functions for URLLC.

Based on the records for system measurements of the networks for average delay per technology within the same period, we can present reliability in correlation with the delay with cumulative distribution function (CDF) [22].

Comparing the two EMCD models, $EMCD_{LTE-LTE}$ and $EMCD_{LTE-WLAN}$ in terms of delay we can see that the simulation of the $EMCD_{LTE-WLAN}$ model presents better delay features.

Namely, this can be accounted to the use of a heterogeneous combination of networks, mobile LTE and WLAN networks. Moreover, as the model defines, the IoT devices send duplicate packets over the LTE and WLAN networks to the IoT application server. We can assume that WLAN's link delay is shorter (going through fixed broadband network). Thus it can be shown that packets will first arrive at the IoT application server via this link, and will use the link with better characteristics, amounting to a combined increased performance, with the resultant graph shown in Figure 6.

Ultimately, Figure 6 clearly presents the final conclusion that the EMCD model brings comparable reliability performance to 5G and can be proposed as a transitional model applicable for the presented network architectures and algorithms. Consequently the EMCD model enables utilizing current heterogeneous LTE and WLAN technologies, while presenting comparable performance to 5G envisioned URLLC standards.

3 CONCLUSIONS

This paper proposes a novel Emergency and Mission-Critical Data framework for mobile and wireless IoT devices in heterogeneous wireless environments, using ultra-reliable applications. The proposed EMCD model has been tested using several simulation scenarios, with the aim to obtain its statistical characteristics and to compare it with existing cases, when a single radio access technology is used by a single mobile terminal. According to the presented results, the proposed dual-stack EMCD framework performs fairly well in different network conditions and coverage, achieving better

performance in comparison to the cases when only one RAT is used and is comparable to 5G reliability performance. One of the major innovations, on which we focused in this paper, introduced by 5G technology as well, is the URLLC support, where despite the high reliability of the system it is required to support a short delay at the same time.

The results have shown performance gain by the EMCD module in the dual network scenario that can easily be generalized to a multi wireless and mobile networks scenario, including any 5G and Next Generation Radio Access Network, as well as IoT network access technologies. The EMCD framework brings comparable 5G reliability performance and can be recommended as a transitional model applicable to the proposed architecture and algorithms, especially for ultra-reliable low latency applications and multimedia services that require high reliability.

REFERENCES

[1] M. Maier, C. Mahfuzulhoq, B.P. Rimal, D. Pham Van,"The Tactile Internet: Vision, Recent Progress, and Open Challenges"; IEEE Communications Magazine 54(5), February 2016, doi: 10.1109/MCOM.2016.7470948.

[2] Aijaz, Adnan & Dawy, Zaher & Pappas, Nikolaos & Simsek, Meryem & Oteafy, Sharief & Holland, Oliver, "Toward a Tactile Internet Reference Architecture: Vision and Progress of the IEEE P1918.1 Standard", July 2018.

[3] M. R. Palattella, M. Dohler, A. Grieco Senior, G. Rizzo, J. Torsner, T. Engel, and L. Ladid, "Internet of Things in the 5G Era:Enablers, Architecture and Business Models" IEEE Journal on Selected Areas in Communications, Vol.: 34, Issue: 3, March 2016.

[4] 3GPP TS 23.282 V16.0.0 (2018-09), Functional architecture and information flows to support Mission Critical Data (MCData); Stage 2, (Release 16)

[5] O. Akrivopoulos, I. Chatzigiannakis, C. Tselios, and A. Antoniou, "On the Deployment of Healthcate Applications over Fog Computing Infrastructure", Proc. IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), vol. 2, pp. 288-295, June 2017.

[6] Zhang, Qi & Fitzek, Frank, "Mission Critical IoT Communication in 5G". FABULOUS 2015, Ohrid, Republic of Macedonia, 35-41. 10.1007/978-3-319-27072-2_5, September 2015.

[7] Siddiqi, Murtaza & Yu, Jaehyung & Joung, "5G Ultra-Reliable Low-Latency Communication Implementation Challenges and Operational Issues with IoT Devices", [Online]. Available: 8. 981. 10.3390/electronics8090981, September 2019.

[8] Y. Zhou, Y. Rong, H.A. Choi, J.H. Kim, J.K. Sohn, and H. I. Choi, "A Dual-Mode Mobile Station Modules for WLAN/UMTS Internetworking Systems," Proc. OPNETWORK 2007, pp. 27-31, Washington, DC, August 2007.

[9] N.Baldo, F.Maguolo, M. Miozzo, M.Rossi, M.Zorzi, "Ns2-MIRACLE: a Modular Framework for Multi-Technology and Cross-Layer Support in Network Simulator 2," NSTools '07, October 22, 2007.

[10] A.A. Al-Helali, A. Mahmoud, T. Al-Kharobi, and T. Sheltami, "Simulation Of a Novel Dual-Mode User Equipment Design For B3G Networks Using OPNET," Third International Conference on Modeling, Simulation and Applied Optimization, Sharjah, U.A.E, January 20-22, 2009.

[11] T. Shuminoski and T. Janevski, "Novel Adaptive QoS Framework for Integrated UMTS/WLAN Environment", Telfor Journal, Vol. 5, No. 1, 2013.

[12] G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5G be?" Selected Areas in Communications, IEEE Journal on, vol. 32, no. 6, pp. 1065–1082, 2014.

[13] J. F. Monserrat, G. Mange, V. Braun, H. Tullberg, G. Zimmermann, and O. Bulakci, "Metis research advances towards the 5G mobile and wireless¨ system definition," EURASIP Journal on Wireless Communications and Networking, vol. 2015, no. 1, pp. 1–16, 2015.

[14] P. Popovski et al., "Wireless Access in Ultra-Reliable Low-Latency Communication (URLLC)," in IEEE Transactions on Communications, vol. 67, no. 8, pp. 5783-5801, Aug. 2019. doi: 10.1109/TCOMM.2019.2914652

[15] 3GPP Tech. Rep. 38.913 v14.1.0, "Study on Scenarios and Requirements for Next Generation Access Technologies," Januare 2017.

[16] H.Djuphammar, N.Spångberg, C.Meyer and H.Basilier, "Ensuring critical communication with a secure national symbiotic network", Ericsson white paper GFMC-18:000199, May 2018.

[17] 3GPP Tech. Spec. 38.323, "NR; Packet Data ConvergenceProtocol (PDCP) Specification," v0.2.1, August 2017.

[18] Huawei, HiSilicon, "R2-1700172: Evaluation on Packet Duplication in Multi-Connectivity", 3GPP TSG RAN-WG2 NR Ad-hoc Meeting, Spokane, WA,USA, Jan. 2017

[19] 3GPP Tech. Spec. 36.300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description; Stage 2," v14.3.0, June 2017

[20] 3GPP Technical Specification 37.340, "NR; Multi-Connectivity;Overall Description; Stage-2," August 2017.

[21] I. Viering, H. Martikainen, A. Lobinger, and B. Wegmann, "Zero-Zero Mobility: Intra-Frequency Handovers with Zero Interruption and Zero Failures", IEEE Network March/April 2018.

[22] M. Rausand and A. Høyland, "System reliability theory: models, statistical methods, and applications", John Wiley & Sons, vol. 396, 2004.

Resource Sharing Challenge for Micro Operator Pattern in 5G SDN / NFV Network

Mariia Skulysh[1], Larysa Globa[1] and Eduard Siemens[2]
[1]*Institute of Telecommunication Systems, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Prosp. Peremohy 37, Kyiv, Ukraine*
[2]*Future Internet Lab Anhalt, Anhalt University of Applied Sciences, Bernburger Str. 57, Köthen, Germany*
mskulysh@gmail.com, lgloba@its.kpi.ua, eduard.siemens@hs-anhalt.de

Keywords: Communication Networks, SDN, 5G, Smart Migration System, Network Slicing, Micro Operator.

Abstract: To expand the capabilities of their networks, large operators turn to smaller operators for help, which allows serving more users. This is possible because software-defined network technologies and virtualization of network functions are used. However, the distribution of subscriber flows between the micro-operators networks problem arises. Micro-operator networks have limited technical resources. The network resources consumed by the services are unevenly in time distributed. There are situations when the network resources of the operator are not enough. At the same time, service consumers want to receive services at a given level of QoS. For dynamic control of the sufficiency of micro-operator resources, the article proposes the dynamic flow control method. The method algorithm include the stages: the flows and node resources use monitoring, the optimal node load calculation, prediction of exceeding the permissible load value, and automatic live migration. The modeling proposed algorithm results showed that there are no more overloads of the micro operator networks. The level of service delays decreased by 5%.

1 INTRODUCTION

A significant part of the new operators prefers to cooperate with small micro-telecommunication networks that cover the interior after the onset of the 5G era. Since these operators can provide various networks such as 3G, 4G, 5G and even Wi-Fi [1, 2]. Up-to-day technology use more and more Network Function [7] that is a functional unit within a network infrastructure that has clearly defined external interfaces and well-defined functional behaviour. In practice, a network function is today a network node or physical device.

Thanks to new SDN/NFV technologies the Internet provider can deploy its networks more flexibly and dynamically [3, 8]. The Internet provider can so provide localized services e.g. in public buildings. This opportunity not only changes the mechanism of network deployment but also fosters micro operator's creation [4-6]. The following factors contribute to the creation of a Mobile Virtual Network Operator:

- lack of spectrum resources;
- efficient usage of bandwidth;
- the scale of the mobile communications market;
- the ability of consumers to diversify services, which, in turn, lead to more diverse applications and services in the telecommunications market.

Mobile Virtual Network Operator can be applied to any wireless service provider that provide wireless services to consumers and do not have its own wireless network infrastructure. It provides completely new opportunities for development in a mature mobile communications market and stimulates the telecommunications market to move towards service-oriented competition.

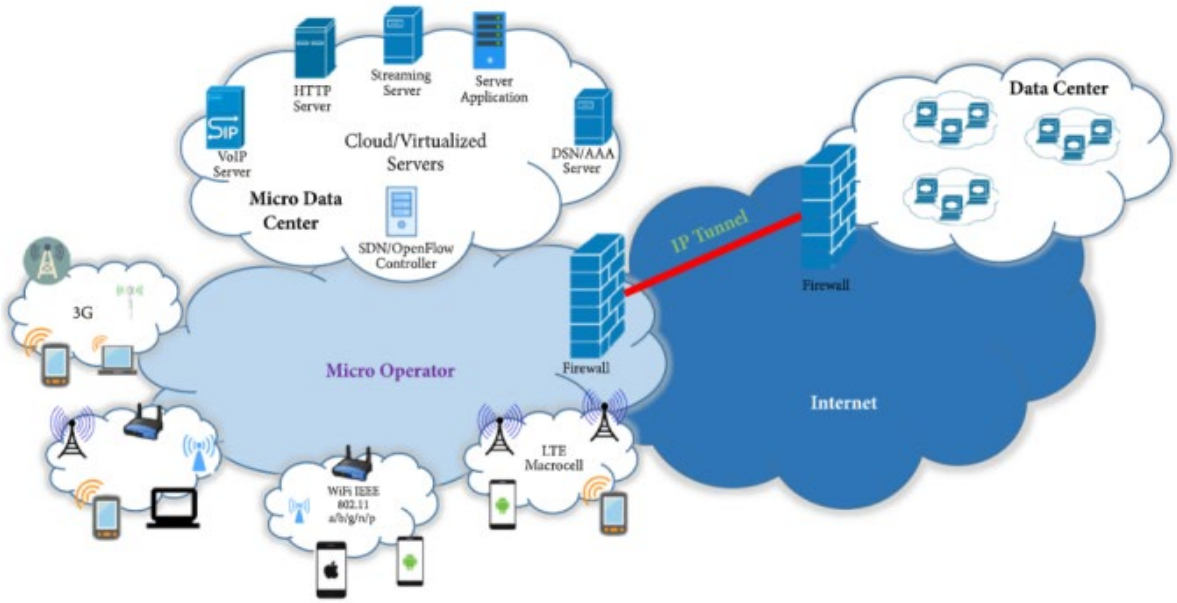


Figure 1: Micro operator network architecture.

Figure 1 presents the Micro Operator network architecture [8]. The micro-operator itself is small in scale and limited in telecommunication (network, hardware and others) and computation (servers, storages and so) resources to provide the necessary services to a certain number of users. The radius of access for mobile devices is limited by the scope of a particular local network (small cells) service which depends on the corresponding available network resources. Institutions such as hospitals, schools, large conferences, sports centres, shopping centres, hypermarkets and factories can use local network services to meet the needs of users in all types of applications.

Regionalization is the nature of the service, which allows it to provide under conditions with limited hardware infrastructure and resources various regional services in different regions allowing mobile users to access the services in other regions. In addition to reduce the consumption of bandwidth resources by providing neighboring network services this type of service localization can also transfer applications, data and computing services from nodes in the data center to the cloud to border nodes in a logical LANs which must be processed and implemented. The computing environment organized by the Fog computing technology reduces network latency and meets 5G requirements.

2 THE NETWORK RESOURCES ALLOCATION TASK

2.1 Reference and Related Work

The Internet of Things (IoT) is supposed to become the killer application of 5G networks and to foster new communications markets. The result will be the formation of different new application scenarios and more diversified network requirements. However, with regard to telecommunications, despite the fact that there is now a globally agreed IoT requirement, built with 5G characters such as speed transfer, capacity, coverage and security, there is still room for the 5G business model to improve. The 5G mobile broadband network focuses on small cells/base stations, enhances internal coverage, provides faster user maintenance and reduces network delays, creating a serious problem for telecom operators. In addition, the emergence of MVNO has brought new opportunities for development of mobile operators who have not yet received frequency licenses for mobile communications. MVNOs use the spectrum and network of mobile operators to provide individual mobile services, corporate virtual private networks for specific businesses or many other micro-markets where operators have not yet expanded their services to less-performing or regional emerging markets. There is a tendency for regionalization between small cellular/base stations, and regional services do

not handle the public network, the network of industrial applications of the network. The Micro Operator business is a new network business model that began to evolve [8]. The manual provides a mechanism for transferring access to a common spectrum of micro-operators and shared single-mode physical network infrastructure using virtual technology to fully utilize a valuable bandwidth resource.

2.2 Micro Operator Design Model in 5G Networks and the Network Selection Step

The interaction between SDN and NFV allows the introduction of a system design model for 5G network infrastructure[10-14]. Flexibility is ensured through a modular approach. Separate problems are solved in the appropriate modules, which are connected as needed. Therefore, we have the flexibility to manage the network in which the deployed networks of micro operators. The operator can switch to using technology cuts network virtualization and networks, as mentioned above, the flexibility to split a physical network into several independent networks and isolated User Interface under different scenarios.

Thus, the task is to organize the maintenance of independent slices within the existing information and telecommunication infrastructure. To ensure the maintenance of independent slices, the following features must be considered:

- it is required to provide each with sufficient amount of resources, both telecommunication and computing ones, for servicing virtualized network functions in order to provide service at a given quality level.
- it is necessary to take into account the nature of the load change in each slice for optimal allocation of resources between slices during the day.
- determine the conditions for the migration of slices in the telecommunications infrastructure, which will ensure the smooth operation of the system.

With regard to infrastructure design, this document uses technologies such as SDN and NFV base and combines network technology and tunnels to build the network infrastructure for microservices. Infrastructure allows users to connect multiple IVS using tunneling technology and running fast network connection to effectively strengthen the relationship networks.

The design model is shown in Figure 2 [8], where the network threading technology implements the logical section of ND networks through OpenVirtex. The connection between the base network and Micro Operator network are performed by a tunnel constructed using the SDN (as a border gateway (BG). The Internet Data Center and the data center are developed using the ETSI-defined NFV infrastructure to save equipment investments. The mission of the SDN controller is to use OpenFlow to construct the path between the BG at the edge of the base network and the BG at the edge of the Micro Operator network. When the SDN controller begins to coordinate and concatenate between the network passages, the Micro Operator network can build a tunnel connection and a basic backbone network.

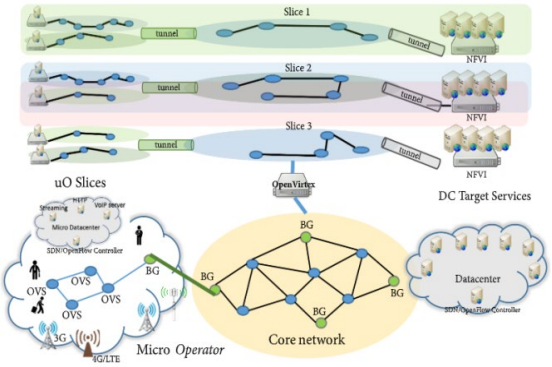


Figure 2: A Micro Operator (μO) design pattern with network slicing.

The Micro Operator can continue to complete the virtual network construction using OpenVirtex virtual winding technology, which allows users to access data in the nearest micro datacenter. Users can also, through tunnels, connect to a cloud datacenter on the Internet to access the service from a specific application network. The proposed architecture combines threading and network tunneling to implement a communication model for Micro Operator and further integrates the bandwidth management technology that applies to the bandwidth application of Micro Operator network. This will increase the utilization of network resources and the efficiency of traffic flow and will lead to a better QoE experience for network users.

In response to the demand of Micro Operator's network resource distribution that allows users to gain access of nearby network resources, the paper [8] proposes network selection mechanism for a Micro Operator and uses decision tree theory to serve as the reference in determining the SDN traffic flows path. The proposed method disadvantage is

that traffic will be distributed without taking into account the all network resources load..

The method shown in Figure 3 application will allow to predict the moments micro operator network overload and to migrate the subscriber sessions between micro-operators network tunnels in time, which will allow to provide conditionally infinite bandwidth micro-operators network.

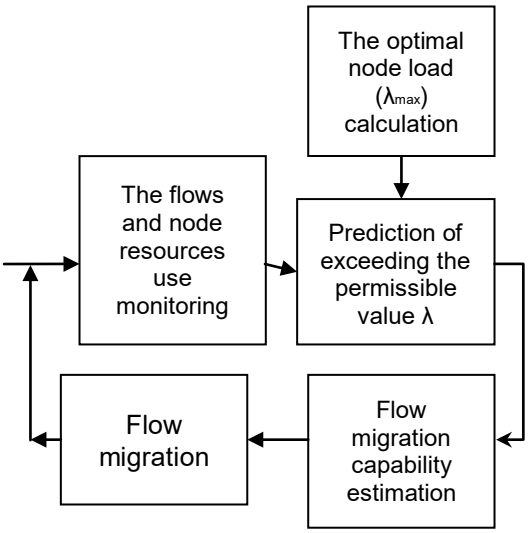


Figure 3: Dynamic Flow Control Algorithm.

"The flows and node resources use monitoring" block involves the accumulation the communication channels congestion information, and information about the resource use dynamics by each service.

For the addition of the possibility of exceeding the permissible value λ (the admissible input stream intensity), the method proposed in [15] is used. The basic method idea is to formulate requirements for the average input load on the basis of ergodic distribution for the possible states of the system, which will allow to make the most efficient use of the available physical resources of servicing the incoming application flow.

For prediction of exceeding the permissible value λ we propose to use the method [7] consists of two stages: the calculation of the prediction interval based on the operation servicing node statistics and directly periodic forecasting of the load and the control of the sufficiency resources.

If periodic forecasting of the load showed that it overload is expected and the available slice resources are not enough to provide services at a given level, then the migration mechanism starts.

2.3 Method of Automatic Live Migration

It is necessary to provide automatic load balancing of physical resources of telecommunications nodes while avoiding overloading one node and an inefficient use of another one. The mechanism of this balancing is called "smart migration" (Figure 4).

Formulation of the problem. The system shall provide:

- the migration process should be invisible to the services user;
- the migration process should be aimed at optimizing the telecommunication network state;
- when planning a migration, one shall assure that two channels (operative and backup) of one slice will not be located in the same physical telecommunications node, so it must support high availability;
- One of the important requirements for migration is the maximum time to complete it, as long migration time can negatively affect the state of the system;
- the system must provide protection against looping, that is, from the endless migration of the same slice;
- it shall provide protection against failures and work in a cluster mode, which is especially important for multiple migrations.

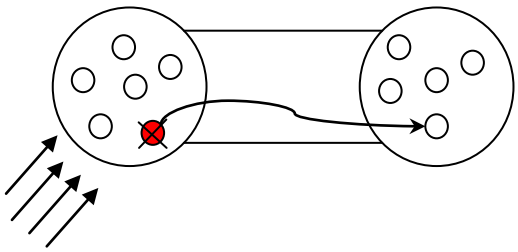


Figure 4: Flow migration between the neighboring micro-operator networks.

Migration System Architecture. There is information about the statistics of the telecommunications nodes and the load that the slice creates on the telecommunications node. This data is periodically read and transformed into metrics that are stored in a specific repository. Thus, by accessing this repository it is possible to obtain information on the dynamics of resource consumption on a separate telecommunications node or slice at any time. information is also available on the number of resources that physical servers have.

The decision on the need for migration as well as on what and where the control unit should

migrate. After selecting candidates for migration, they are placed in a distributed queue, which is processed by a special process. This process analyzes information on the number of items in the queue and, based on this information, selects the item to be migrated. The process of physical migration is synchronous without any discontinuities or returns, while there is a special mechanism that ensures that there are no errors and delays. The main task of the process is to ensure the selection of a candidate for migration and to move it to a certain telecommunications node in such a way as to optimize the state of the system. Thus, there is a multi-dimensional optimization problem. There are several algorithms for solving this problem.

Simple algorithm:

1) Looking for the most loaded node.
2) The server selects the most optimal node in terms of the amount of resources consumed.
3) Move the slice.

This algorithm is easy to use, however, with large number of communication nodes, the algorithm is not yet optimal. For such cases, you need to use a more complex algorithm. Its essence is as follows. Special rules are defined that must be met by the optimal solution. There are strict rules that can't be violated under any circumstances and soft rules that can be neglected in some cases. In addition, the types of problem solutions are determined:

- possible solutions - solutions that are achieved in violation of strict rules (bad decisions);
- feasible solutions - decisions that do not violate hard conditions, but do not fulfill a part of soft ones;
- optimal solutions - solutions that fulfill both types of conditions;
- optimal solutions are the best solutions calculated in the shortest time.

To solve a problem using a complex algorithm, it is necessary to pre-define soft and hard constraints.

Hard restrictions:

1) The amount of resources of the target node must be sufficient to move the slice. In addition, redundancy resources must be provided.
2) Slice cannot migrate to its own physical node.
3) On the same physical node should not be located streams of the same slice. This condition ensures that the minimum amount of data is lost in the event of a system failure.

Soft restrictions:

1) Migrate the most loaded slice streams.
2) The target physical node shall be the least loaded one.

The main disadvantage of this algorithm is the lack of tools to account for trends in the rate of resource consumption by various slices. A study was conducted on the effectiveness of accounting for statistical data in the process of selecting a stream for migration, as well as a node to which the migration will be carried out.

In order to correctly select the node to which the migration will be performed, it is necessary to assess the trend of changes in the resource usage dynamics of the selected server while taking into account the load that the migrating slice stream will create.

To determine the moment of migration, it is necessary with a specified time interval to evaluate the current statistics of resource utilization, to build a statistical trend on the number of serviced requests. Based on the trend, an assessment is made of the likelihood that the maintenance of containers located on the node under study will exceed the allowable amount of resources, then the migration process will start. The method of assessing the adequacy of resources is presented in [10].

Thus, based on the current load statistics generated by the sum of the flows of the individual slices; estimates of the upper limit of the capacity of the telecommunications node will be decided on the need for migration.

3 THE SIMULATION RESULTS

The simulation was conducted in the Matlab environment. The initial model data the were mobile telecom operator statistical data from the resources and services monitoring system per day.

Table 1: The comparative analysis results.

	The services distribution between micro operators is fixed	Dynamic Flow Control Algorithm
Resource overload	10%	0%
Number of service migrations	-	10
Service delays	8%	3%

Network resources were conditionally divided into separate micro operator network slices. The micro operator network resources were fixed. Services usage statistics by network subscribers was analyzed as follows:

1) Without balancing. The resource usage indicators of the micro operator's network by assigned subscribers group over the norm were estimated.
2) The number of subscribers assigned to the micro-operator changed according to the Dynamic Flow Control Algorithm.

The comparative analysis results are shown in Table 1. Thus, in order to avoid the micro operator network overloading, it is necessary to use the method of dynamic flow control, which includes the mechanisms of live flows migration and provides conditionally infinite service resources.

4 CONCLUSION

This document uses SDN and NFV technologies as the basis and combines network streaming and tunneling technologies to create a network infrastructure model for MSO using a smart migration mechanism. This model allows users of different MOs to connect using tunneling technology, and then implement a fast network connection to effectively improve network interaction, while balancing the load between all nodes of a given network. To meet the needs of the regional micro-operator service, this article proposes a DTBFR mechanism that uses decision tree theory as the basis for making SDN-based traffic decisions. As a distribution and control of the load on the nodes, we use the method of slices working together in the existing telecommunication foreign infrastructure, which ensures the automatic distribution of telecommunication and computing resources of the system depending on the load and allows solving the problem of peak loads and idle resources. This method of automatic load balancing of telecommunication nodes ("smart migration") does not allow overloading one node and downtime of another node. The functions used by the regional micro-operator service can effectively reduce the load on the datacenter on the Internet and accelerate the development of the regional computer service in the future 5G network. And the "smart migration" method will allow rational use of network resources.

REFERENCES

[1] A. Raschellà, F. Bouhafs, G.C. Deepak, and M. Mackay, "QoS aware radio access technology selection framework in heterogeneous networks using SDN," Journal of Communications and Networks, vol. 19, no. 6, 2017, pp. 577-586.

[2] M. Matinmikko, M. Latva-aho, P. Ahokangas, S. Yrjölä, and T. Koivumäki, "Micro operators to boost local service delivery in 5G," Wireless Personal Communications, vol. 95, no. 1, 2017, pp. 69-82.

[3] J. S. Walia, H. Hammainen, and M. Matinmikko, "5G Micro-operators for the future campus: A techno-economic study," in Proceedings of the 2017 Internet of Things - Business Models, Users, and Networks, Copenhagen, Denmark, pp. 1-8, November 2017.

[4] M. Matinmikko-Blue and M. Latva-aho, "Micro operators accelerating 5G deployment," in Proceedings of the IEEE International Conference on Industrial and Information Systems (ICIIS), Peradeniya, Sri Lanka, pp. 1-5, December 2017.

[5] P. Ahokangas, S. Moqaddamerad, and M. Matinmikko, "Future micro operators business models in 5G," The Business and Management Review, vol. 7, no. 5, 2016, pp. 143-149.

[6] P. Mach and Z. Becvar, "Mobile Edge Computing: A Survey on Architecture and Computation Offloading," IEEE Communications Surveys & Tutorials, vol. 19, no. 3, 2017, pp. 1628-1656.

[7] L. Globa, M. Skulysh, O. Romanov and M. Nesterenko, "Quality Control for Mobile Communication Management Services in Hybrid Environment". International Conference on Information and Telecommunication Technologies and Radio Electronics, Springer, Cham, pp. 76-100, November 2018.

[8] M.A. Skulysh, O.I. Romanov, L.S. Globa and I.I. Husyeva, "Managing the Process of Servicing Hybrid Telecommunications Services. Quality Control and Interaction Procedure of Service Subsystems". In International Multi-Conference on Advanced Computer Systems. Springer, Cham, pp. 244-256, September 2018.

[9] Chia-Wei Tseng, Yu-Kai Huang, Fan-Hsun Tseng, Yao-Tsung Yang, Chien-Chang Liu and Li-Der Chou, "Micro Operator Design Pattern in 5G SDN/NFV Network," Wireless Communications and Mobile Computing, vol. 2018, Article ID 3471610, 14 p., 2018.

[10] O. Semenova, A. Semenov and O. Voitsekhovska, "Neuro-Fuzzy Controller for Handover Operation in 5G Heterogeneous Networks". IEEE 3rd International Conference on Advanced Information and Communications Technologies (AICT), pp. 382-386, July 2019.

[11] M. Skulysh, "The method of resources involvement scheduling based on the long-term statistics ensuring quality and performance parameters". International Conference on Information and Telecommunication Technologies and Radio Electronics (UkrMiCo), pp. 1-4, September 2017.

[12] M. Skulysh and O. Romanov, "The structure of a mobile provider network with network functions virtualization". 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET). IEEE, 2018, pp. 1032-1034.

[13] L. Globa, V. Kurdecha, I. Ishchenko, A. Zakharchuk and N. Kunieva, "The intellectual IoT-system for monitoring the base station quality of service". IEEE International Black

Sea Conference on Communications and Networking (BlackSeaCom), pp. 1-5, June 2018.

[14] O.I. Romanov, Y.S. Hordashnyk and T.T. Dong, “Method for calculating the energy loss of a light signal in a telecommunication Li-Fi system”. International conference on information and telecommunication technologies and radio electronics, (UkrMiCo), pp. 1-7, September 2017.

[15] M. Skulysh, F. Shilov and A. Safaryan, “Investigation of the method of computing resources optimal choice for billing systems effectiveness." Control, Navigation And Communication Systems. Academic Journal 3.49, pp. 147-152, 2018, doi: https://doi.org/10.26906/SUNZ.2018.3.147.

[16] L.S. Globa, I. Ishchenko, V. Kurdecha, A. Zakharchuk and O. Zvonarov, “An approach to the Internet of Things system with nomadic units developing”. IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), pp. 1-5, June 2016.

Improved Approach to Quality Control of Telecommunication Service Providers

Larysa Globa, Maryna Popova and Nataliia Yushko
Institute of Telecommunication Systems, National Technical University of Ukraine "Igor Sikorsky Kyiv Politechnic Institute", Peremohy av., Kyiv, Ukraine
l.globa@its.kpi.ua, pma1701@gmail.com, natalia.yushko@outlook.com

Keywords: Ontology, Workflow, Data Structure, Microservices, Orchestrator, Computer-Aided Design, Quality of Service, Telecommunication.

Abstract: The modern development of telecommunications and telecommunication providers requires an increasing level of service provision. It is explained by the fact that the formation of a market for network services has increased attention to issues of quality control, both by regulators and by providers themselves. To ensure compliance with the specified level of quality of service provision, telecommunication providers develop algorithms and solutions to control the quality of service provision, based on different criteria, by themselves. However, these solutions are not universal for different types (quantitative, qualitative, etc.) of service quality indicators. This article proposes an improved approach to quality control of telecommunication service providers. The implementation of the proposed approach is performed using the ontological model of service quality indicators given by the provider and the dynamically changing workflow, which provides versatility and computer-aided quality of service control. The proposed approach allows to make the process of quality control of service delivery transparent and reduce the involvement of expert analysts in this process.

1 INTRODUCTION

In the modern world, there is growing number of providers which provides services of various types: from Internet access and telephony to content providers, over the top services etc. This growth is due to the fact that digital services are gaining more relevance and demand, also the competition in the environment is increasing. This is a reason why it is necessary to calculate the quality of service provision, to standardize it and to comply with the rules dictating the competitive environment.

The purpose of the standard is to determine a unified approach for providers and other stakeholders to evaluate the quality of communications services based on user opinions and to calculate some technical indicators. Based on this standard, business entities operating in the field of telecommunication services can develop their own (internal) regulatory algorithms for controlling the quality of telecommunication services, taking into account different parameters of the service delivery system. The results of the control of the quality of telecommunication services are used in the analysis of the effectiveness of the functioning of the quality management system, certification of services, making management decisions on improving the quality of telecommunication services.

Considering that there are many factors that affect the quality of service which are interconnected in different ways, it becomes necessary to computer-aided control process.

There are two types of parameters related to the service quality control: parameters (indicators) of the quality of network operation (Network Performance, NP) and indicators quality of service (QoS). NP is determined by the performance of individual network elements or the performance of the entire network as a whole. QoS parameters characterize the quality of services provided from a user perspective and may not always be expressed in technical terms.

Considering the above factors, numerous studies are being conducted in the world, aimed at the tools development and optimization of the quality of service control processes.

Therefore, in the context of the topic of analysis of quality of service, it's an important question how to realize the process automation and encapsulation

of complex algorithms from the end-user in order to provide a more accurate control of the parameters of the quality of service and increase the efficiency of the provider company by simplifying the work with this process.

This paper is organized as follows. After the Introduction, section 2 contains the state of the art of the approach to control the quality of services provided by the telecommunications provider. Section 3 is devoted to the formalized description of computer-aided business process design based on ontology meta-models. Also, the third section provides the solution to the problem and further research perspectives. Section 4 – conclusion – includes the summary and outlook on future work.

2 STATE OF THE ART AND BACKGROUND

Service quality research has been conducted on a number of occasions, using different methods and algorithms.

For example, taking QoS control into account, such groups of non-technical parameters as preliminary service information (integrity of preliminary information, price transparency, availability, time to look for information), provider-user contract (integrity of contract information, compliance with the contract, simple extension of the contract), provision of services (fulfilment of the term of service, speed of service, completeness of the contract, punctuality of delivery of devices [1]), replacement of service (service replacement time, punctuality, simplicity of replacement), technical support (complaints management, complaints response, complaints management effectiveness), etc. [2].

It is also worth noting that often to describe the quality of service, the technical indicators of the telecommunication system, such as download speed, upload speed, voice quality, network coverage, reliability, uptime, downtime, etc.[3], are explored.

Each of the parameters is described by a set of criteria, which in turn impose limits on the allowed parameter values and thus normalize the quality of service delivery. This means that in order to control such indicators, it is necessary to involve a telecommunications expert who possesses a certain set of knowledge and skills that is not always beneficial for organizations.

It is proposed computer-aided process of quality of telecommunication services control by forming and executing a workflow that is generated from the ontology of the research area. This will make the control process more encapsulated and therefore easier to use by the end-user without additional programming costs to save material and time resources.

3 PROBLEM DEFINITION

Based on previous research [4-6], the purpose of the work is to develop a solution for executing complex computational scenarios through coordinated interaction of web services (microservices) on the basis of service-oriented (microservice) architecture using an ontological knowledge base.

The main task is to modify the centralized form of service interaction ("orchestration") in such a way as to exclude low-level details of the description of interaction scenarios while maintaining the rules of service interaction in the knowledge base [7,8]. By orchestration, we mean centralized coordination of components of a distributed software system in order to organize a coordinated interaction to achieve the desired effect – the implementation of a given workflow to implement the appropriate process. We will consider a separate process in the context of a particular area, in this case – in the context of controlling the quality of service provided by a telecommunications provider. This will preserve the approach of some encapsulation since the end-user (for example, expert analyst, management of the provider organization) will be excluded from direct control process according to certain algorithms. Also, this solution can be used as a universal solution, because when you replace the main algorithms in the workflow, you don't need to replace other indicators.

An ontological data model is proposed as a domain data model that can be represented as a tree structure.

In general, an ontology means a system of concepts of some domain, which is represented as a set of entities connected by different relationships. Ontologies are used for the formal specification of concepts and relationships that characterize a particular area of knowledge. The advantage of ontologies as a way of presenting knowledge is their formal structure, which simplifies their computer processing [9,10].

In the general case, the ontology domain is formally represented by an ordered three:

$$O = \{X(w, s, q), R(w, s, q), F\}, \quad (1)$$

where X, R, F – the finite sets appropriately: X – the set of registries (X_w – workflow registry, X_s – services registry, X_q – query registry,), R – the set of relationship between registries, F – the set of the interpretation functions X and/or R.

The ontological structure allows working with heterogeneous, unstructured data.

In order to create an ontological structure, it is necessary to analyze the domain, identify the basic structural bases and describe them.

For example, four major registries have been identified to build a solution for the quality of service control by telecommunications providers:

- Services registry;
- Microservices registry;
- Workflow registry;
- Query registry.

Services registry is an ontological structure that contains a list of all services provided by a telecommunications provider.

The registry lists the main services of the provider: they are divided into classes (Internet, TV, Over-The-Top Content (OTT)), the level of service (top lever offers, second-level offerings), the service characteristics, type of users, valid values, etc. are specified. All service data can be viewed in a structured manner using an object ontology.

Microservices registry is an ontological structure that is a tree of services that will be used to control complex technical parameters, which in turn are part of an overall assessment of service quality.

To describe each microservice, a mathematical model, represented as (2), was used:

$$M_i = \{Ip_{i..j}, Op_{i..j}, A_i, T_i, P_{i..j}\}, \quad (2)$$

where $Ip_{i..j}$ – input data which requires for the microservices; $Op_{i..j}$ – output data; A_i – algorithm of microservice; T_i – the type of microservice output data (quantitative, logical); $P_{i..j}$ – processes which should provision before current microservice.

Following this description for each microservice, it is necessary to take into account all the indicators that will be required when performing the algorithm of controlling a certain quality indicator.

In the developing solution, all microservices will play the role of a "black box", which means that the expert creates such a microservice once and uses it in further calculations without going into detail.

Workflow registry is an ontological structure containing a set of computational scripts to execute a workflow with the described parameters.

Workflow consists of an orchestrated and repetitive structure, which is provided by the systematic organization of resources into processes that transform materials, provide services or process information. This can be depicted as a sequence of operations, the work of a person or group, the organization of the work of staff, or one or more simple or complex mechanisms. From a more abstract or higher level, a workflow can be considered as a kind or representation of a sequence of execution, taking into account the content of the stages of real work [11]. The described flow may refer to a document, service, or product that is transmitted from one step to the next. Workflow can be seen as one of the main components that must be combined with other parts of the organization such as information technology, teams, projects and hierarchies [12]. In this case, the workflow will consist of microservices and describe their sequential or parallel execution.

The mathematical model presented in (3, 4, 5) is used to describe each workflow.

$$W_i = \{Ip_w, Op_w, D_w, Dg_w, Pr_w\}, \quad (3)$$

where Ip_w – input workflow data; Op_w – output workflow data; D_w – diagrams description; Dg_w – BPMN diagram which connected with specific workflow and its representation; Pr_w – priority of workflow provisioning.

$$Ip_w = \bigcup_{M_n}^{M_k} Ip_{i..j}, \quad (4)$$

where Ip_w – input workflow's data, which represented as the intersection of all input data $Ip_{i..j}$ by all microservices $M_n...M_k$.

$$Op_w = \bigcup_{M_n}^{M_k} Op_{i..j}, \quad (5)$$

where Op_w – output data for workflow which represented as the intersection of all output parameters $Op_{i..j}$ by all microservices $M_n...M_k$.

As indicated in (3), an important component of the workflow registry description is the BPMN diagram, which is a representation of the workflow execution process. Business Process Model and Notation (BPMN) is a notation system for modelling business processes [13]. BPMN is a standard for business process modelling, providing graphical notation for defining a business process in the form of a Business Process Diagram (BPD). Such a diagram is based on a business process representation in the form of a flowchart that is semantically similar to an activity diagram [14].

BPMN aims to support the modelling and management of processes or microservices. At the same time, a unified business process model should be clear to all users (stakeholders). Nevertheless, the notation makes it possible to define complex semantics of processes [15, 16].

Workflow registry provides a description of each workflow step in the form of a BPMN diagram showing the sequence of microservices execution and the dependency between them, which is to some extent the orchestration of services.

Query registry is an ontological query structure, presented in the form of query descriptions and their corresponding workflow. The mathematical model of the query registry is presented in (6).

$$Q_i = \sigma_c(W_i), \quad (6)$$

where Q_i – query, which should be entered by the user; σ_c – the function of selection appropriate workflow by the condition C; W_i – appropriate workflow.

Each query is matched to the corresponding workflow and causes its execution.

So, in sum, all four registries are semantically interconnected. Because the query registry is a set of queries that can be entered by the user and invokes the corresponding workflow from the workflow registry. In turn, the workflow starts orchestrating the microservices in the order that is reflected in the workflow using the service registry. And running microservices can query the service registry for more information about a particular service and its data. Figure 1 shows a flowchart for the process described.

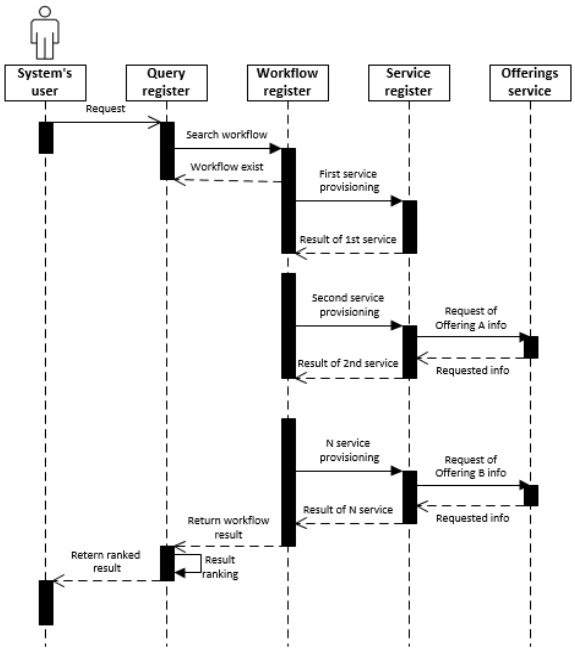


Figure 1: The sequence diagram of the solution's registries.

As a result of the algorithm, if the user prompts a query to calculate the quality of service, the system, having received the result, is forced to rank it by certain criteria. Therefore, the following criteria were developed to evaluate the result (Table 1).

Table 1: Criteria for ranking the quality of service delivery.

Mark	Description
-3	The quality of service is quite poor. Most of the technical indicators are much lower, user reviews are negative. The deviation of the real indicators from the norm exceeds on average 20%.
-2	The quality of service is not satisfactory. Technical indicators are lower than normal, average deviation within 10-15%. There are complaints from users. The information provided to users is in a difficult place.
-1	The quality of service is not satisfactory. Technical indicators are lower, by an average of 5-10%. There are complaints from users. Very little information is provided to users.
0	The quality of service is neither unsatisfactory nor satisfactory. There was no quality assessment or no input to get started.
1	The quality of service is low but satisfactory. The deviation from the norm of some technical indicators is in the range of 0-5%. There are almost no complaints from users. All the information that the user needs is publicly available and several queries are required to use it.
2	The quality of service is medium and satisfactory. The deviation from the norm of technical indicators is quite small, in the range of 0-3%. There are no complaints from users about the quality of services. All the information that the user needs is publicly available and several queries are required to use it.
3	The quality of service is high and satisfactory. Deviation of technical indicators from the norm is almost not observed. All parameters are within the range of acceptable values. There are no user complaints. All information the user needs is publicly available.

Thus, an approach to determine the quality of service by telecommunication providers is proposed. The essence of it is to formalize the 4 main entities (service registry; microservice registry; workflow

registry; query registry) involved in the process of their control, as well as the universal mechanism of formation workflow when determining the quality of service, taking into account the specificities of the contract of a particular user.

The proposed approach has the following advantages over the known ones:

- Versatility of use in any subject area, by downloading the relevant data in the registries.
- Computer-aided process of quality of service control, which contains a large number of heterogeneous parameters, which, in turn, are calculated by complex algorithms.
- Encapsulation of the solution, enabling the use of the presented solution by employees of any level, without the involvement of an expert analyst.
- Ability to computer-aided workflow modification when performing quality of service control procedures in accordance with an ontological structure that describes service requirements.

CONCLUSIONS

Development of software components for executing complex computational scenarios through coordinated interaction of web services (microservices) on the basis of service-oriented (microservice) architecture with the use of ontological knowledge base will allow computer-aided process of quality of services control by telecommunication providers.

This approach is universal and takes into account the peculiarities of the domain due to the fact that each procedure for the control of a specific indicator is implemented in the form of microservices, and workflow is dynamically formed from a set of microservices that correspond to those involved in the process of quality control indicators described in the ontological model.

The proposed approach to the computer-aided business processes design and their components (microservices, communications, and interaction rules) based on computer-aided generation of a set of services that are components of workflows, as well as computer-aided formation the sequence of their execution through the use of ontology – a meta-model of workflow, domain, logical rules, services that establish relationships between functional services.

Using the described approach to computer-aided workflow construction allows to choice of a functional input processing service, and this is a very important factor in real-time systems because depending on the data, the most efficient method for processing in the shortest period of time can be selected.

Using the domain ontology as a registry of functional services will help computer-aided business process building and identify a functional service from many other services that most closely matches the conditions of use that are determined by the incoming data stream.

Further research will be devoted to a more detailed consideration of the computer-aided workflows design from microservice sets and, in part, the program code computer-aided generation for the workflows' execution.

REFERENCES

[1] V.P. Lupanin and T.A Kulikova, "The system of quality service-oriented consumers", Technology of information society, 2016, pp. 78-84.

[2] International Telecommunication Union. SERIES E: overall network operation, telephone service, service operation and human factors / International Telecommunication Union., 2012, p. 52, Rec. ITU-T E.803 (12/2011).

[3] ETSI EG 202 843 V1.1.2 (201107) User Group; Quality of ICT Services; Definitions and Methods for Assessing the QoS parameters of the Customer Relationship Stages other than utilization.

[4] L.S. Globa, O.S. Shtogrina and M.Y. Ternovoy, "Integration of ontology-based databases and knowledge bases". Collection of scientific works of NTUU "KPI", no. 1, 2011.

[5] L. Globa, M. Kovalskyi and O. Strvzhak, "Increasing web services discovery relevancy in the multi-ontological environment", Springer, Cham, 2015, pp. 335-344.

[6] A. Koval, L. Globa and R. Novogrudska, "The approach to web services composition", Springer international publication AG, no. 534, 2017, pp. 293-304.

[7] W. Tan, Y. Fan, A. Ghoneim, M.A. Hossain and S. Dustdar, "From the Service-Oriented Architecture to the Web API economy", IEEE Internet Computing, vol. 20, no. 4, 2016, pp. 64-68.

[8] C. Peltz, "Web services orchestration and choreography", Computer, vol. 36, 2003, no. 10, pp. 46-52.

[9] L. Koo, N. Trokanas, A. Panteli, E. Kalemi, N. Shah et. al., "Integration of CAPE Models and Data for the Domain of Biorefining: InterCAPEmodel Ontology Design", Computer-Aided Chemical Engineering, Elsevier, vol. 40, 2017, pp. 2341-2346, doi: 10.1016/B978-0-444-63965-3.50392-5.

[10] D. Oberle, "How ontologies benefit enterprise applications", Semantic Web, vol. 5, 2014, pp. 473-491, doi: 10.3233/SW-130114.

[11] A. Bitkowska, "Business Process Management Centre of Excellence as a Source of Knowledge",

Business. Management and Education. no. 16, 2018, pp. 121-132, doi: 10.3846/bme.2018.2190.

[12] B. Ludäscher, I. Altintas, S. Bowers, J. Cummings et al., "Scientific Process Automation and Workflow Management", 2009, doi: 10.1201/9781420069815-c13.

[13] M. Rosing, S. White, F. Cummins and H. Man, "Business Process Model And Notation (BPMN)". The Complete Business Process Handbook, 2015, pp. 429-453, doi: 10.1016/B978-0-12-799959-3.00021-5.

[14] A. Stephen, "Process Modeling Notations and Workflow Patterns", in Wayback Machine, IBM Corporation, 2006.

[15] B. Silver, BPMN Method and Style, 2nd Edition. 2011, Cody-Cassidy Press, ISBN 0982368119.

[16] F. Curbera, et al. Business process execution language for web services, 2002, [Online]. Available: http://www106.ibm.com/developerworks/webservices/library/wsbpel/.

Mutual Influence of Opposite TCP Flows in a Congested Network

Nikolai Mareev, Dmytro Syzov, Dmytry Kachan, Kirill Karpov, Maksim Iushchenko and Eduard Siemens

Future Internet Lab Anhalt, Anhalt University of Applied Sciences, Bernburger Str. 57, Köthen, Germany
{*nikolai.mareev, dmytro.syzov, dmitry.kachan, kirill.karpov, maksim.iushchenko, eduard.siemens*}*@hs-anhalt.de*

Keywords: TCP, IP, Highspeed, Congestion Control Algorithm, Internet, Performance, Bidirectional, Two-Way, 10G Networks, BBR, CUBIC, Coexistence.

Abstract: With the rapid growth of the Internet community, some of the simple and familiar tasks related to the field of data transfer are becoming increasingly complex. A modern worldwide network can offer high-speed channels and many opportunities for IT companies that provide high load through the Internet. This creates a bunch of new problems for software solutions and algorithms in the field of high-speed digital c ommunications. This article observes one of these problems: the mutual influence between two mutually opposite single-threaded TCP flows with the various congestion control algorithms. In this paper, some of the most efficient congestion control algorithms were tested on a real network using channel emulation equipment. The test results presented in the article show that two-way TCP data transfer with modern congestion control algorithms can lead to a significant performance drop.

1 INTRODUCTION

Transmission Control Protocol (TCP) provides a set of functions for automatically controlling sender parameters during data transfer in TCP/IP networks. One of these functions is the congestion control algorithm, that addresses three features:

- Prevent network devices from overloading.
- Achieve high bottleneck bandwidth utilization.
- Share the network resources with other flows.

The network congestion is a situation when a network node receives more data than it can handle or forward. Network congestion results in an overloaded transmission buffer on network devices, additional network delay, and packet drops. Congestion control algorithms (CCA) can be divided into groups according to the main indicator of congestion - the data transfer parameter, which corresponds to network congestion. Key congestion indicators are network delay, packet loss, and available bandwidth. Delay-based congestion control algorithms (VENO [1], VE-GAS [2]) are designed to proactively detect network congestion - before packet loss occurs. Common issues of such algorithms are unfair resource sharing and low bottleneck bandwidth utilization. Loss-based and loss-delay-based algorithms (CUBIC [3], YEAH [4]) treat packet loss as network congestion. Achieving high bottleneck bandwidth utilization is another important challenge for congestion control algorithm. Different types of CCAs use different data rate control schemes and require different depths of the bottleneck queue buffers to fully utilize the bottleneck bandwidth. The third challenge for congestion control algorithms is resource sharing. Network resources, such as bottleneck bandwidth or port queue depth, are limited. Sharing network resources require additional methods in the algorithm and rely on the congestion indicators dynamics. BBR [5] is a congestion-based congestion control algorithm developed by Google past few years. This algorithm uses the bottleneck bandwidth estimation as the primary indicator and the round trip time as the secondary indicator of congestion. BBR can achieve relatively high data transfer performance in cases where packet loss can occur on a non-congested link.

The main purpose of this article is to present a study of the mutual influence of two mutually opposite TCP data streams in a congested network. Particular attention was paid to eliminating hardware, cross-traffic, and other possible impacts on the results. Work has been performed in Future Internet Lab Anhalt [6].

The rest of this document is organized as follows: The second Section provides a brief overview of TCP coexistence issues. Section 3 describes the experimental setup and properties of the experiment. Test results and evaluation are presented in Section 4. Sec-

tion 5 contains a discussion of the results provided, and Section 6 contains a conclusion.

2 TCP COEXISTENCE

The simultaneous coexistence of different TCP data streams in the same channel requires special behavior of the congestion control algorithms for the fair sharing of network resources. Essentially, during data transfer, the congestion control algorithm probes the bandwidth by changing the data transfer rate and measures the parameters of the connection i.e. congestion indication. In the case of loss-based congestion control algorithms, packet losses considered as a sign of congestion. This leads to a certain behavior during data transfer: the amount of data increases until the bottleneck of the port buffer is overloaded and some amount of data packets are dropped. Such algorithms relatively fairly share network resources among themselves and can provide high data transfer performance. The modern trend is to increase the depth of the queue buffers over the network. In cases with fat network buffers, loss-based congestion control algorithms have a strong negative effect on network delay [7]. However, most TCP connections are controlled by congestion control algorithms based on loss or loss-delay congestion indication.

Delay-based congestion control algorithms use changes in the network delay as an indication of network congestion. This allows keeping the load level of the bottleneck queue buffers at some lower level than loss-based algorithms do. Such algorithms have less aggressive behavior compared to loss- or loss-delay based algorithms, it leads to unfair sharing of the network resources. However, there are several different strategies for achieving fairness between loss- and delay-based congestion control algorithms [8].

A relatively new solution, the BBR congestion control algorithm, uses probing cycles to estimate available bandwidth, network delay, and channel state. BBR tends to keep low bottleneck queue buffer load level and achieve high bandwidth utilization. Another important feature of BBR is packet losses tolerance and high performance in lossy networks. This strategy allows in most cases to nearly fairly share network resources during coexistence with loss-based TCP flows. However, BBR is still under development and has several performance issues [9, 10, 11].

Congestion control algorithms use the dynamics of congestion indicators to mutually influence each other during coexistence and change the data transfer rate for the main purpose of sharing network resources. In case of one-way congestion, the dynamic behavior of congestion indicators is expected in network latency and available bandwidth. In case of two-way network congestion, main congestion indicators may have unexpected behavior due to the influence of the two-way data stream, and lead to performance issues.

3 EXPERIMENTAL SETUP

Testbed network is presented in Figure 1. Core elements in the network are Netropy 10G and Netropy 10G2 - WAN emulators from Apposite Technologies [12]. These devices allow to emulate various network conditions by setting the properties of the channel (see Table 1) and saving per second statistics of the forwarded data stream, such as data transfer rate, queue buffer load level, packet loss, etc. All data flow statistics in this work are collected by Netropy devices.

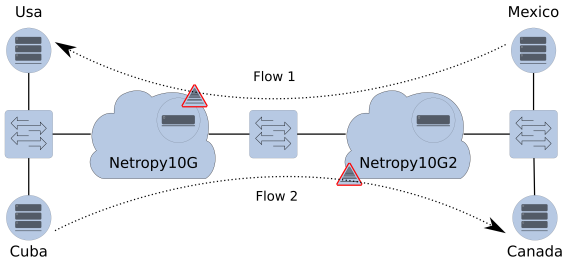


Figure 1: Experimental network.

Table 1: Netropy WAN emulators description.

Label	Netropy10G	Netropy10G2
Max. Agg. Throughput	20Gbps	40Gbps
Max. Packet Rate	29Mpps	59.5Mpps
Bandwidth	from 100 bps to 10 Gbps	
Queuing	RED or tail drop queue management; priority or round robin queuing;	
Queue depth	up to 100MB	
Latency	0 ms – 10000 ms or greater in each direction in 0.01 ms increments; constant, uniform, exponential, normal distributions with or without reordering; accumulate and burst delay;	
Packet loss	random, burst, periodic, BER, Gilbert-Elliott, or recorded packet loss; data corruption; network outage	

The second important element in the testbed is a network switch - Extreme Networks Summit x650-24x [13]. It has 24 10GBASE-X SFP+ inter-

faces, 488 Gbps maximum aggregated bandwidth and 363 Mpps maximum packet throughput. It is an edge-level network switch with tiny shared queue port buffer. The last elements on the scheme are servers (named as follows: Usa, Mexico, Cuba and Canada) with common specifications:

- 64GB DDR4 of RAM.
- Intel Corporation 82599ES 10-Gigabit SFI/SFP+ NIC.
- Linux 5.3.0-24-generic x86_64 Kernel.
- Intel(R) Xeon(R) CPU E5-2643 v4 3.40GHz CPU.

Provided tests require the exclusion of a possiblenegative impact from the OS and hardware on the process of transferring data. Each test case includes the following features.

- To exclude OS-level resource sharing / competition / queuing, a separate pair of servers were used for each TCP data stream.
- Bottleneck queues in both directions were configured separately on different WAN emulators in order to eliminate possible specific queue management problems in cases of two-way congestion.
- The emulated bottleneck bandwidth in all tests was configured at a level that is significantly lower than the maximum bandwidth of network devices in the testbed.
- The maximum data transfer rate was significantly lower than the maximum aggregated throughput of the tested devices.
- The emulated network delay was configured on 20ms to exclude possible overreact issues on the TCP congestion control side (TCP congestion control can show unexpected behavior in cases of LAN network delay)
- Bottleneck buffer queue depth has been set as 2.5 MB (tail drop queuing algorithm) according to the rule-of-thumb recommendations mentioned in [14, 15].

All tests have been performed with iperf3 ver. 3.6 TCP traffic generation utility [16].

4 EXPERIMENTAL RESULTS

To observe the behavior of data transfer of both streams separately and in coexistence TCP flows were started with a time interval of 50 seconds between each other. The interaction period of oncoming traffic is 150 seconds and shows the mutual influence of TCP data streams in case of two-way network congestion.

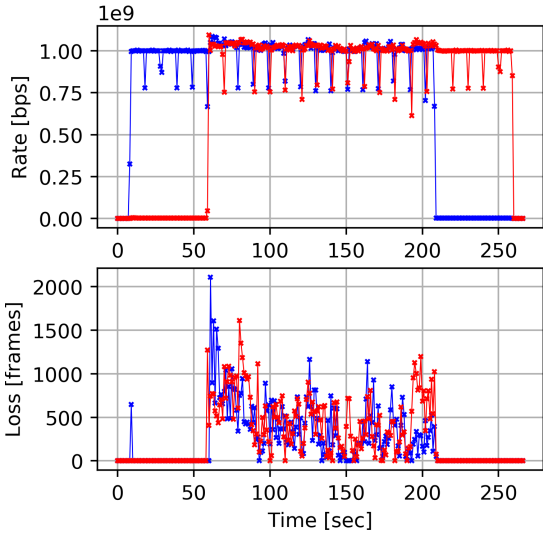


Figure 2: Two TCP BBR mutually reverse data flows.

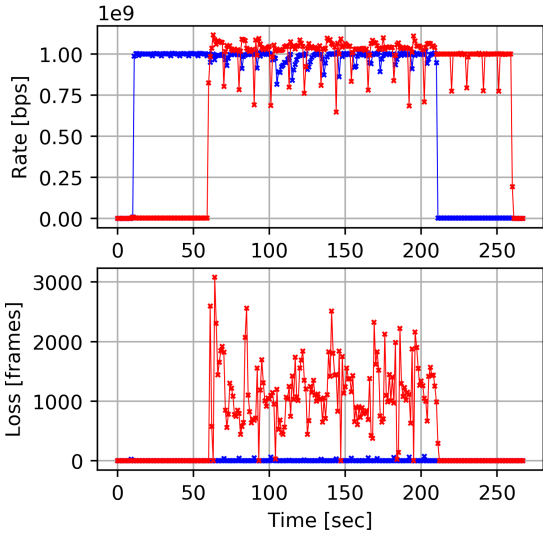


Figure 3: TCP CUBIC (blue) and TCP BBR (red) mutually reverse data flows.

On the Figure 2, an example of the mutual influence of two counter TCP BBR data flows is presented. TCP BBR requires relatively low bottleneck queue buffer during data transmission and it perfectly fits in the given test environment. Bottleneck bandwidth is fully utilized and no packet losses detected until the second TCP BBR flow appears in the link. The interaction of two data flows on a this link leads to overloaded bottleneck queue buffers and massive packet drops in both directions. It breaks the resource sharing ability of an algorithm and excludes any additional loss- or loss-delay based congestion control TCP flow in this link. However, the bottleneck bandwidth is utilized fully during the coexistence period.

The mutual influence of TCP counter BBR and

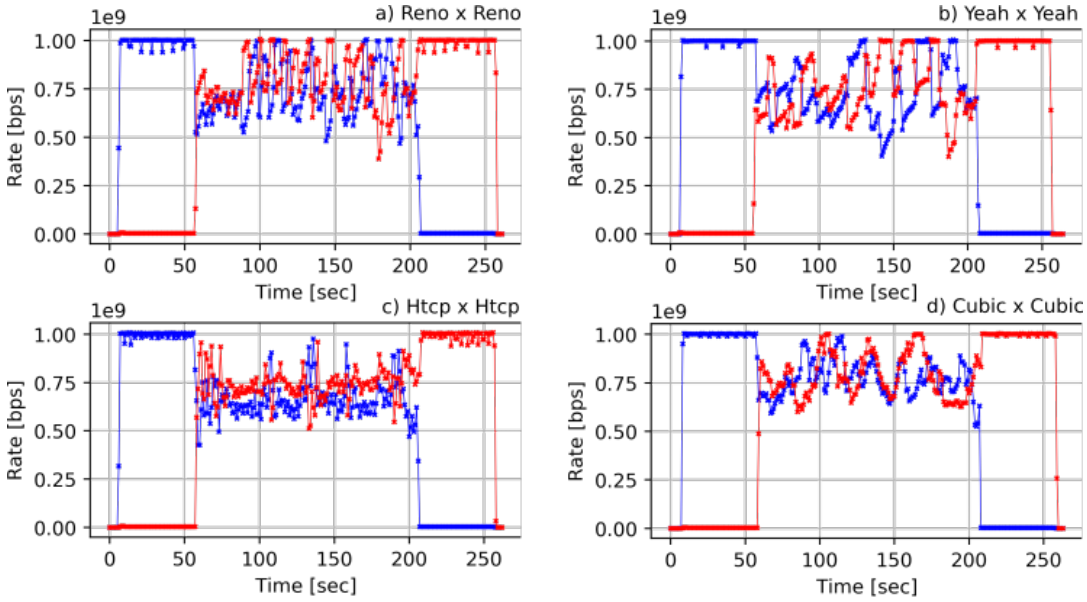


Figure 4: Different loss-based TCP congestion control mutually reverse data flows.

TCP CUBIC traffic is shown in Figure 3. The TCP BBR stream loses twice as many packets as the previous case. The TCP flow in the opposite direction to the BBR is controlled by the TCP CUBIC congestion control and shows a slight decrease in data rate during coexistence.

Highly efficient congestion control algorithms are observed in the [17] by Lukaseder T. et al., these CCAs was decided to test in the proposed case. Figure 4 shows the inter-protocol mutual influence of TCP streams with various loss and loss-delay congestion control algorithms: TCP RENO, TCP YEAH, HIGH-SPEED TCP, and TCP CUBIC. Each tested congestion control algorithm fits in the channel and utilizes full available bandwidth until another data flow started. Compared to TCP BBR, loss-based algorithms show a much higher influence on each other during coexistence. Performance degradation during this type of coexistence can be described by reduction of bottleneck bandwidth utilization by up to 25 %.

5 DISCUSSION

Delay-based congestion control algorithms have well-known issues of resource sharing during coexistence [18] and were not included in the article. The main goal of the article is to observe the behavior of the most popular congestion control algorithms. The issue of the performance drop during the two-way network congestion is the influence on the congestion indicators in both directions. In such a case the round-trip-time delay (RTT) measured by first flow would be influenced by queuing delay load in the opposite direction caused by the second data flow. Loss-based congestion control algorithms treat changes in the network delay and packet losses as the signals to release the bandwidth, like in one-way coexistence. This behavior leads to a drop in the bottleneck bandwidth utilization. Another influence is caused by packets in the feedback channel of the flows. A lot of service packets from the downstream flow are including in the data packets of the upstream data flow disturbing a bottleneck queue and provide an additional network delay and packet losses.

A possible solution for this issue could be the usage of one-way network delay (OWD) as the congestion indication instead of a round-trip-time delay. This would exclude the influence of a feedback channel on the congestion indication. It would also exclude additional network delay jitter in the feedback channel and, probably, increase the data transmission performance. Nevertheless, clock drift is a serious problem, and such a strategy requires additional algorithms for proper operation. Low priority TCP congestion control algorithms like TCP LP [19] or TCP LEDBAT [20] also shows performance drop in case of bidirectional network congestion. It is confusing because these algorithms use one-way delay instead of RTT for the congestion indication. Probably the implementation of these algorithms in the Linux kernel is actually using RTT instead of OWD.

6 CONCLUSION

In this article a mutual influence of counter TCP data flows in the case of bidirectional network congestion was observed. loss- and delay-based congestion control algorithm demonstrates significant performance degradation during such a test case, up to 25% data rate drop. TCP BBR, a congestion based congestion control algorithm demonstrates still high bottleneck bandwidth utilization, however, two-way network congestion leads to massive packet losses and impossibility of share the bandwidth with other loss-based flows. Future work including OWD-based congestion indication implementation in RMDT [21] protocol or/and research of the congestion indication in the TCP low priority congestion control solutions in the Linux kernel.

ACKNOWLEDGEMENTS

This work has been funded by Volkswagen Foundation for trilateral partnership between scholars and scientists from Ukraine, Russia and Germany within the project CloudBDT: Algorithms and Methods for Big Data Transport in Cloud Environments.

REFERENCES

[1] Ch. Peng Fu and S. Liew, "TCP veno: TCP enhancement for transmission over wireless access networks," IEEE Journal on Selected Areas in Communications, vol. 21, no. 2, pp. 216–228, February 2003.

[2] L.S. Brakmo and L.L. Peterson, "TCP vegas: end to end congestion avoidance on a global internet," IEEE Journal on Selected Areas in Communications, vol. 13, no. 8, pp. 1465-1480, October 1995.

[3] S. Ha, I. Rhee and L. Xu, "CUBIC: a new TCP-friendly high-speed TCP variant," ACM SIGOPS Op-erating Systems Review, vol. 42, no. 5, pp. 64-74, July 2008.

[4] A. Baiocchi, A. P. Castellani, and F. Vacirca, "YeAH-TCP: Yet another highspeed TCP," In proc. The fifth PFLDNET workshop, February 2007.

[5] N. Cardwell, Y. Cheng, C. S. Gunn, S. H. Yeganeh, and Van Jacobson, "BBR: congestion-based congestion control," vol. 60, no. 2, pp. 58-66, January 2017.

[6] Future Internet Lab Anhalt. [Online]. Available: https://fila-lab.de/

[7] J. Gettys, "Bufferbloat: Dark Buffers in the Internet," IEEE Internet Computing, vol. 15, no. 3, pp. 96-96, May 2011.

[8] M. Hock, R. Bless and M. Zitterbart, "Toward coexistence of different congestion control mechanisms," IEEE 41st Conference on Local Computer Net-works (LCN), pp. 567-570, November 2016.

[9] K. Miyazawa, K. Sasaki, N. Oda and S. Yamaguchi, "Cycle and divergence of performance on TCP BBR," IEEE 7th International Conference on Cloud Networking (CloudNet), October 2018.

[10] N. Mareev, D. Kachan, K. Karpov, D. Syzov and Siemens, "Efficiency of BQL Congestion Control under High Bandwidth - Delay Product Network Conditions," Proc. of the 7th International Conference on Applied Innovations in IT, (ICAIIT), pp. 19–22, March 2019.

[11] K. Sasaki, M. Hanai, K. Miyazawa, A. Kobayashi, N. Oda and S. Yamaguchi, "TCP Fairness Among Modern TCP Congestion Control Algorithms Including TCP BBR," IEEE 7th International Conference on Cloud Networking (CloudNet), October 2018.

[12] Leaders in network emulation and testing. [Online]. Available: https://www.apposite-tech.com/

[13] End-to-end cloud driven networking solutions. [Online]. Available: https://www.extremenetworks.com/

[14] A. Dhamdhere, Hao Jiang and C. Dovrolis, "Buffer sizing for congested internet links," Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 2, 2005, pp. 1072-1083.

[15] C. S. Curtis Villamizar, "High performance TCP in ANSNET," ACM Computer Communications Review, pp. 45-60, September 1994.

[16] iPerf - the TCP, UDP and SCTP network bandwidth measurement tool. [Online]. Available: https://iperf.fr/

[17] T. Lukaseder, L. Bradatsch, B. Erb, R. W. Heijden and F. Kargl, "A Comparison of TCP Congestion Control Algorithms in 10G Networks," IEEE 41st Conference on Local Computer Networks (LCN), pp. 706-714, November 2016.

[18] R. Al-Saadi, G. Armitage, J. But and P. Branch, "A survey of delay-based and hybrid TCP congestion control algorithms," IEEE Communications Surveys & Tutorials, vol. 21, no. 4, pp. 3609-3638, 2019.

[19] A. Kuzmanovic and E. Knightly, "TCP-LP: a dis-tributed algorithm for low priority data transfer," IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 3, pp. 1691-1701, 2003.

[20] D. Rossi, C. Testa, S. Valenti and L. Muscariello, "LEDBAT: The new BitTorrent congestion control protocol," in Proc. of 19th International Conference on Computer Communications and Networks (IC-CCN 2010), August 2010.

[21] D. Syzov, D. Kachan and E. Siemens, "High-speed UDP data transmission with multithreading and automatic resource allocation," Proc. of the 4th International Conference on Applied Innovations in IT,(ICAIIT), pp. 51-55, March 2016.

The Possibilities for Deployment Eco-Friendly Indoor Wireless Networks Based on LiFi Technology

Oleksandr Romanov, Thi Tho Dong and Mikola Nesterenko
Institute of Telecommunication Systems, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Prosp. Peremohy 37, Kyiv, Ukraine
a_i_romanov@ukr.net, dongthitho1993@gmail.com, nikolaiy.nesterenko@gmail.com

Keywords: Optical Wireless Communication (OWC), Light Fidelity (LiFi), LEDs, PoE, Possibility to Deploy Indoor Wireless Networks, Optical Received Power, Light-of-Sight (LoS), SNR.

Abstract: Recently, traffic demand for wireless data has increased the need for extending the spectrum to transmit numerous signals. However, the bandwidth of radio waves has been scrunching in the last few years with tremendous development of technology such as 4G, 5G, Internet of Thing (IoT) and so on. Therefore, in a few recent years, LiFi technology which is considered a complement solution for radio frequency technology has attracted a lot of attention in scientific community. LiFi, which is a part of optical wireless communication (OWC), employs visible light from the green and eco-friendly light emitting diode (LED) to forward signals. In addition, LiFi is best-known for high-speed, bi-directional connection, save energy, security network, and safe for human. Because of the potential advantages that LiFi benefits in, we investigated to analyze the possibility to deploy indoor wireless networks based on LiFi technology with the existing infrastructure of LEDs and PoE cables. Moreover, in this study, we proposed the wireless network model in a typical office with nine LED luminaires positioned in the center of the room ceiling that support to intensify the illumination and communication in an entire room. Additionally, the received power of optical signals and signal-to-noise ratio (SNR) level in the proposed model were calculated and simulated with the MATLAB program. The study of these parameters advocates deploying the network system more effectively.

1 INTRODUCTION

According to an analysis in [1] that has pointed out the exponential increase of wireless data and significant growth of approximately 80 billion IoT devices which connect to the wireless network by 2020 and beyond. Due to the explosion of the amount of wireless data to be transmitted, the radio waves spectrum is quickly reaching its limit, as well as breeding some problems that are gaining more cause electromagnetic interference with high-frequency, not enough spectrum capacity to send out the enormous amount of data, and difficulty in security data [2]. Because of these limitations in wireless technologies which uses radio waves (RF) to carry signals, a number of researchers have drawn attention to other parts of the electromagnetic spectrum in order to find out the new approaches to deal with this issue. Therefore, LiFi, which had proposed since 2011 by professor Harald Haas, is considered as one of the robust trusted technologies, and widely known as a solution to complement wireless technology, especially WiFi.

In particular, the visible light that has a range of spectrum which is larger more than radio frequency approximately 10000 times, and unregulated that means LiFi no need to have a license to operate [3,4, 15-20]. Moreover, LiFi network system offers the dual functions of light used for lighting and communication purposes. Furthermore, LiFi could be compatible with PoE cables used for backbone networks, owing to simplify the network system and save energy [5-6]. Thanks to this, LiFi technology may have huge opportunities to develop and become more widespread in the future. Additionally, it should be noted that there have not been many pieces of research in evaluating the possibility of combining the benefits of both LEDs and PoE technology to use an indoor wireless network. For this reason, in this paper, LiFi's characteristics were studied, the possible prospect of deployment indoor wireless networks based on LiFi technology was explored, and we desire to investigate as much as

possible in this issue, especially in typical office network systems. Simultaneously, parameters of physical elements in the network system which impact on process transmitting data were analyzed and simulated in MATLAB program to choose the suitable type of LEDs luminaires for setting up the position of LED luminaires array efficiently.

The rest of the study is constructed as follows: Section II assesses the possibility of deploying indoor wireless networks based on LiFi technology. Section III described the light of sight (LOS) propagation model in LiFi system. In section IV, the model of the wireless network in a typical office was proposed, and the received optical power and SNRlevel were calculated and simulated in MATLAB program. In section V, simulation results, which impact on the connectivity quality and the selecting different kind of LED luminaires for setting up a wireless network, were discussed. Finally, Section VI sums out the study and orientates future studies.

2 PROSPECTS OF DEPLOYING INDOOR WIRELESS NETWORKS BASED ON LIFI TECHNOLOGY

2.1 Structure of LiFi Network

Basically, LiFi network consists of two main components that are LED transmitter and LiFi dongle. Both of them have a built-in infrared uplink sensor. LiFi access point (AP) can be connected directly to the base network using PoE cables that contain data and power as one. In this case, LiFi technology has following merits:

- Environmentally safe wireless connection.
- Can be used in places that do not allow the use of WiFi (such as in aircraft, hospitals, etc).
- High bandwidth.
- High efficiency, energy saving, no required license, energy saving, and can be compatible with infrastructure based on lighting system using green and eco-friendly LEDs and PoE technology.
- Safe for humans (especially children and pregnant women).

The structure of the LiFi network was shown in Figure 1, in which the benefits listed above can be realized. First of all, LEDs are semiconductor devices, therefore, the impulse of light emitted by

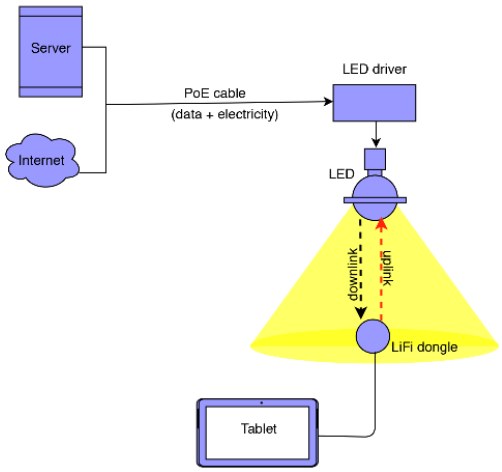


Figure 1: The structure of LiFi network system.

LEDs can be changed expeditiously. Recently, the time of switching LEDs from one state to another can reach nanoseconds. It allows transferring information in a wide variety of speeds. At the same time fluctuations in the light are not captured by the human eye. And in LiFi dongle, the signal is detected and light intensity changes are interpreted as data. In addition, it can integrate a built-in infrared detector for the uplink.

Then, the work process of the LiFi network system is as follows. Firstly, data from the Internet or Server of department store, office, hospital,…are sent to the LED driver which controls the process of converting electrical signals into optical ones using PoE cables. At the present time, almost LiFi luminaries are integrated LED drivers inside. Secondly, the modulated signals from LEDs are transmitted to the photodetector (PD) integrated into LiFi dongle. Finally, LiFi dongle which is plugged into end-user equipment like computers, laptops or smartphones to receive optical signals from LEDs and converts data from photon to electronic forms. The LiFi dongle integrated an infrared LED that modulates the transmission data from users to return to the LEDs and to the network.

2.2 Possibility for the Deployment of the Indoor Wireless Network System Base on Existing Infrastructure

Due to rising electricity prices, LEDs are now being implemented everywhere to decrease the cost. And there is a number of leading companies which deal with lighting business note that the use of LED lamps provides them with more than 50% of their income. In particular, Acuity Brands (67%),

OSRAM (65%), Philips (61%), Hubbell (55%), and Zumtobel (73%). According to Strategies Unlimited at the end of 2016 LEDs accounted for 11% of the total number of installed lighting devices, and by 2022 this value would reach 20% as shown in Figure 2 [7,16].

In addition, the search for energy-efficient solutions attracted attention to the PoE technology which allows the transmission of electrical energy and data using the same cable. For example, there are already high-rise buildings in Amsterdam which have over 6500 PoE connections to LEDs, that help to reduce installation cost by 25% and system deployment time by more than 50%. Likewise, by 2024, the PoE market will be expected to reach $105.2 million, an increase of 13% on CAGR data as illustrated in Figure 3. So the growth rate of the LEDs and PoE market is a great prospect for the development wireless network based on LiFi technology, as it is ready for the basic wireless network infrastructure.

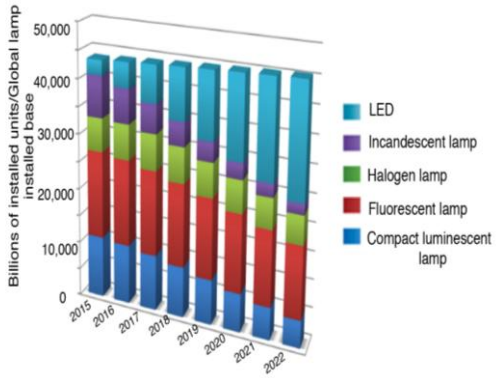


Figure 2: The trend growth some kind of lamps in the market from 2015 to 2022.

Currently, there are some pioneer companies (e.g. Oledcomm, Purelifi, Vlncomm, and others) that offer LiFi devices used to build optical wireless optical networks. On top of that, the world's first optical LiFi elements designed for mobile integration embedded into a standard HP laptop to facilitate high-speed LiFi with Gbps connectivity has published by pureLiFi company at the Mobile World Congress on February in 2019. The Gigabit LiFi system can transmit data up to 1 Gbit/s for downlink, and 377 Mbps for uplink [8]. However, this can be seen the LiFi components are still relatively high-priced. The cost is a range from $1000 to $2000 for a set of the transmitter. In addition, the LiFi network requires more transmitters than Wi-Fi due to standard illumination. According to experts, it is necessary to reduce the price of equipment up to $100 per piece, so that they were competitive compared to WiFi devices. On the other hand, it is necessary to scale down the size of the LiFi devices so that smartphones and laptops can accommodate signal receivers inside.

Additionally, in 2018 the IEEE 802.11 Working Group on LiFi Communications worked with manufacturers, operators, and end-users to develop a new standard [9]. The goal was to ensure that the new standard was completed in May 2021. However, it is possible to use the early version of the standard at this time. Therefore, the developing of the new LiFi equipment is in full swing.

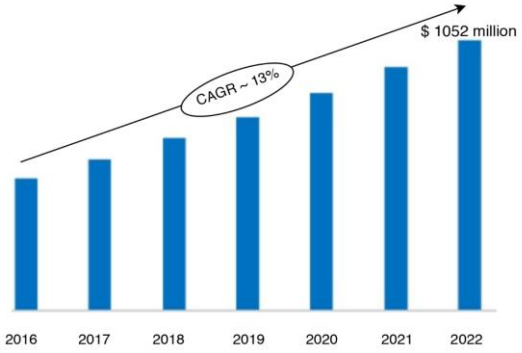


Figure 3: The trend of PoE market growth from 2016 to 2022.

3 LIGHT OF SIGHT PROPAGATION MODEL

In the indoor places, light reflects from the ceiling, walls or mirror surfaces but does not penetrate obstacles, while in external environment light is dissipated and absorbed in atmospheric conditions. There are several ways in which optical paths can be physically configured. They are usually grouped into two main system configurations: LOS (directed light) and non-LOS (undirected light). The LOS connection to the LiFi network offers lots of benefits including faster data transfer speeds over Gb/s, security, and low power consumption [10,11]. Therefore, in this study, we just dominated analyzing basic parameters which are relevant for the deployment network in the LOS path as shown in Figure 4.

The attenuation coefficient H_{LOS_i} of the optical signals from the LEDs to the receiver located at a distance d_i and the angle φ_i in LOS path can be simplified as [12]:

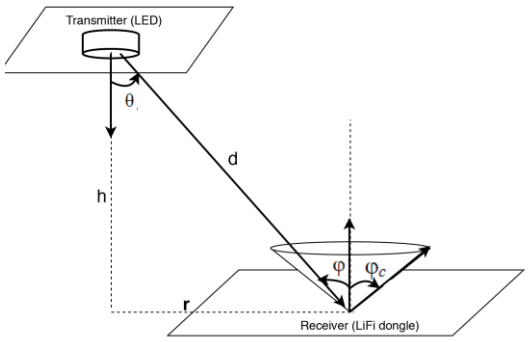


Figure 4: The LOS propagation model in optical channel.

$$H_{LOS\_i} = \begin{cases} \frac{m+1).A_{PD}.cos^m(\theta_i).T_s.g(\varphi_i).cos\,\varphi_i}{2\pi {d_i}^2} \\ 0, \varphi_i > \varphi_c \end{cases} \quad (1)$$

where m is the Lambert order number representing the direction of the beam of the source, which is given as [12]:

$$m = \frac{-ln2}{ln(cos(\phi_{1/2}))}, \quad (2)$$

where $\phi_{1/2}$ is the angle at which the intensity of the source is half compared to the intensity when the source is viewed directly on the axis; A_{PD} is the physical area of the photodetector; θ_i is the incidence angle; φ_i is the angle at which the source is considered; T_s is the gain of optical filter; $g(\varphi_c)$ is the gain of optical concentrator, which is expressed as [12]:

$$g(\varphi_c) = \begin{cases} \frac{n^2}{sin^2(\varphi_i)}, \varphi_i \le \varphi_c \\ 0\,, \ \varphi_i > \varphi_c \end{cases}, \quad (3)$$

where $0 < \varphi_c < 90^\circ$ is the maximum angle at which the light that falls on the optics can be successfully sent to the detector.

For simplicity, we assumed that $\theta_i = \varphi_i$ and $cos(\theta_i) = \frac{h}{d_i}$.

Then H_{LOS_i} can be obtained as:

$$H_{LOS\_i} = \frac{m+1).A_E.T_s.h^{m+1}}{2\pi(d_i)^{m+3}}, \quad (4)$$

where A_E which is the light collection zone and is usually expressed as the area of the detector multiplied by the optical gain of any optics, can be written as [12]:

$$A_E = A_{PD} * g(\varphi_c). \quad (5)$$

4 CALCULATING THE RECEIVED OPTICAL POWER AND SNR LEVEL

4.1 Select quantity of LEDs

In LiFi technology, light from LEDs is used effectively with dual functions: lighting and data transmission. Therefore, when setting up the LiFi optical wireless network, we must consider the allowable illuminance of the LED to provide minimal illuminance for the network location. Especially for the indoor network, e.g. schools, hospitals, offices…

In this work, optical wireless network system LiFi in typical office is modeled. Thus, the illuminance requirement range is between 350 lux and 500 lux level satisfied the ISO standard [13]. General, the illuminance can be determined by the formula:

$$L_X = \frac{P_T .N_{LED} .\emptyset}{S}, \quad (6)$$

where L_X is the illumination; P_T is the power of LED; N_{LED} is the number of LEDs; $\emptyset$ is the energy efficiency of the system; and S = a × b, is the office size.

To guarantee the minimum illuminance for the network place, the number of LEDs should be accounted for. Consequently, from (6) the number of LEDs IN LED can be calculated as follow:

$$\frac{L_{Xmin} .S}{P_T.\emptyset} < N_{LED} < \frac{L_{Xmax} .S}{P_T.\emptyset}. \quad (7)$$

In this simulation, we selected each LED with a transmitted power P_T of each LED is equal to 35 W, an office size is 6m × 6 m, and $\emptyset$ is 50 lumen W. Hence, the quantity of LEDs is expressed as:

$$\frac{350.36}{35.50} < N_{LED} < \frac{500.36}{35.50}. \quad (8)$$

Equation (8) is equivalent to:

$$7.2 < N_{LED} < 10.3. \quad (9)$$

With the results from (9), we picked out 9 LEDs for our simulation.

4.2 Proposed indoor wireless network system configuration in the typical office

To calculate the parameters of the system, we used the following scenario. The fixed components of the

Li-Fi network consist of LEDs located around the center ceiling of the office as depicted in Figure 5. The distance between adjacent LED lamps is 2 m. The coordinates of each LED are given by (X_T, Y_T). The mobile component of the Li-Fi network consists of terminals moving at the office at very low speeds. The coordinates of a receiver is (X_R, Y_R). Therefore, the distance d_i can be calculated by:

$$d_i = \sqrt{(X_R - X_T)^2 + (Y_R - Y_T)^2 + h^2}. \quad (10)$$

And 100 × 100 points are selected in the room. These sample positions are uniformly allocated on the plane where the receiver is positioned. Other parameters of the LiFi system, presented here, are shown in Table I.

In this simulation, the selected LEDs have the same parameters, so if the receiving device is located under directly the LED at these positions (1,1), (1,2), (1,3), (2,1), (2,2), (2,3), (3,1), (3,2), (3,3), their received power is the same. The signal receiving signal is given by the formula :

$$P_R = P_T . H_{LOS\_i} = P_T . \frac{(m+1).A_E.T_s.h^{m+1}}{2\pi(d_i)^{m+3}}. \quad (11)$$

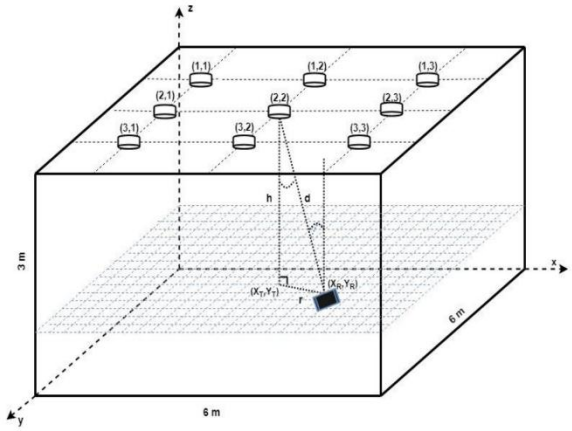


Figure 5: Proposed indoor wireless network system model with 9 LED luminaires.

4.3 Simulating in MATLAB program

Normally, LEDs have various radiation models that correspond to the value of $\varphi_{1/2}$. In this study, we simulated three scenarios with three different LED models selected, i.e. a planar lens, $\phi_{1/2} = 60^{\circ}$; a semispherical lens, $\phi_{1/2} = 40^{\circ}$; a parabolic lens,

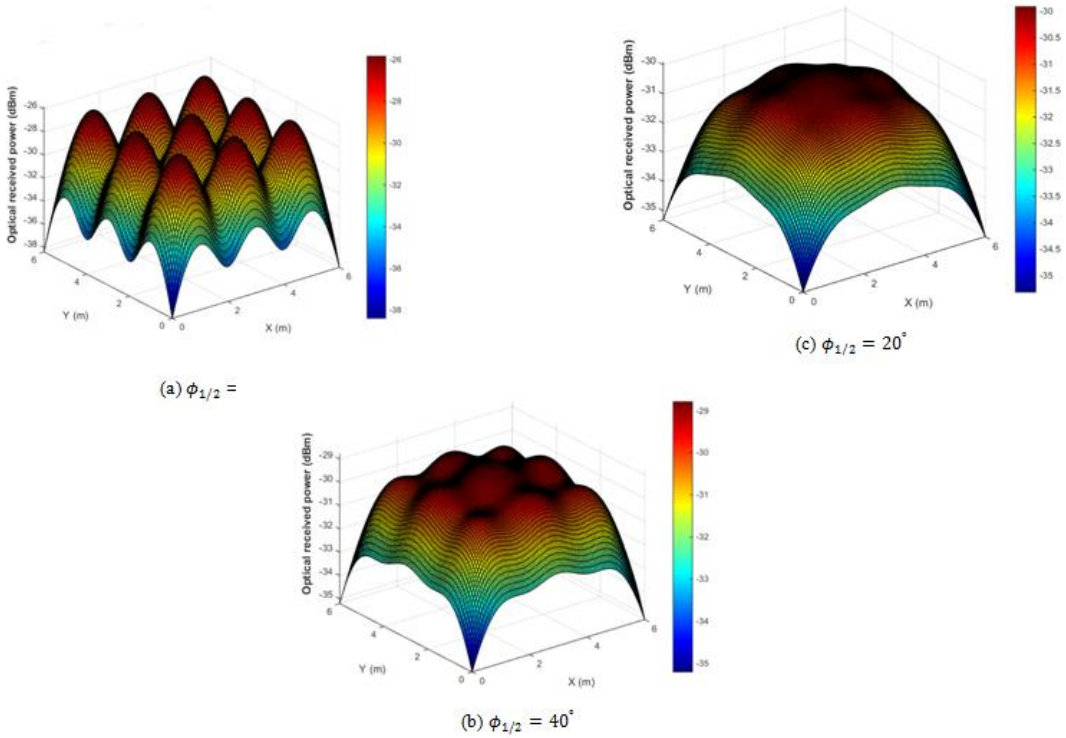


Figure 6: The optical received power distribution with diffirent LED model, (a) $\phi_{1/2} = 20^{\circ}$, (b) $\phi_{1/2} = 40^{\circ}$ and (c) $\phi_{1/2} = 60^{\circ}$.

$\phi_{1/2} = 20^0$. Note that the vertical stick on the right side of the picture indicates the relationship between P_R (or *SNR)* and color: blue represents the smallest value of P_R, and yellow is the highest P_R value.

Table 1: Simulated Parameters

Parameter	Value
Office size	6m×6m×3m
Vertical distance between transmitter and receiver plane, h	2 m
The half-intensity angle, $\phi_{1/2}$	20°, 40°, 60°
Transmitted optical power for an LED, PT	35 W
Number of LEDs, N_{LED}	9 LEDs
Field of view, φ_c	60°
Optical filter gain, TS	2
Refractive index of the lens in the photodetector, n	1,5
Physical area of the photodetector, APD	25.10-6, m2

4.4 Simulated results

4.4.1 Optical received power

4.4.1.1 Scenario I: $\boldsymbol{\phi_{1/2} = 20^0}$

The maximum received optical power, P_{R_max}, is -25.8360 dBm, when the receiver is located directly below the LED, and the minimum P_{R_min} is -38.3878 dBm, when the receiver is located in the corner of the room. Thus, the difference from peak to deflection is 12.5518 dBm.

4.4.1.2 Scenario II: $\boldsymbol{\phi_{1/2} = 400}$

P_{R_max} = -28.7871dBm
P_{R_min} = -35.2049 dBm
$\Delta P_R = P_{R\_max} - P_{R\_min}$ = 6.4178 dBm

4.4.1.3 Scenario III: $\boldsymbol{\phi_{1/2} = 600}$

P_{R_max} = -29.9097 dBm
P_{R_min} = -35.3150 dBm
$\Delta P_R = P_{R\_max} - P_{R\_min}$ = 5.4053 dBm

4.4.2 SNR level

4.4.2.1 Scenario I: $\boldsymbol{\phi_{1/2} = 200}$

The same as the received optical power, the maximum $SNR_{\_max}$ is 55.4376 dB when the receiver is positioned straight under the LED, while the minimum $SNR_{\_min}$ is 42.8738 dB when the receiver is positioned in the edge of the room. Thus, the variation SNR is 12.5638 dB.

4.4.2.2 Scenario III: $\boldsymbol{\phi_{1/2} = 400}$

SNR_{max} = 52.4859 dB
SNR_{min} = 46.0634 dB
$\Delta SNR = SNR_{max} - SNR_{min}$ = 6.4225 dB
SNR_{max} = 51.3629 dB
SNR_{min} = 45.9531 dB
$\Delta SNR = SNR_{max} - SNR_{min}$ = 5.4098 dB

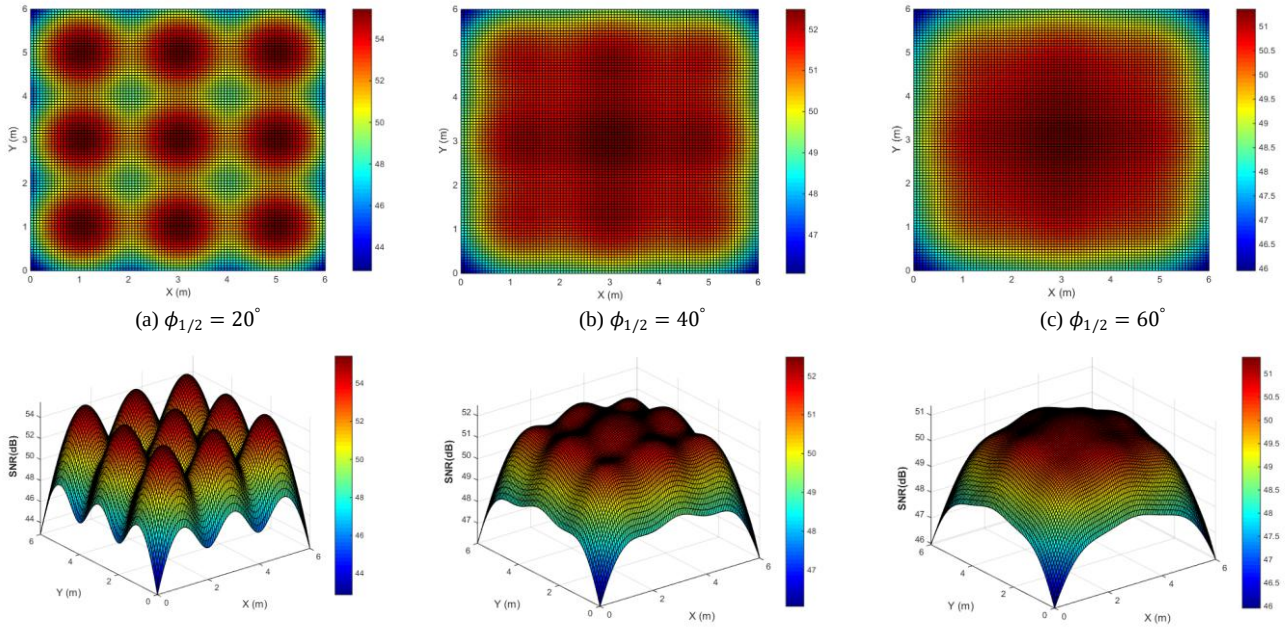
5 DISCUSSING SIMULATION RESULTS

As mention in the earlier studies [14], to stabilize the connection between transmitters and receivers the optical received power at the receivers requires higher than the receiver sensitivity (about - 36 dBm). As shown in Figure 4, the received power value is ranging from about -35 to -30 dBm in most of the places in a proposed model. Therefore, these results match the result mention before. Consequently, in these scenarios, all lighting configuration can get full connectivity.

Furthermore, observing Figure 6 and Figure 7, it can be seen that SNR is proportional to P_R. When the value of P_R increases, SNR simultaneously raises. In addition, the optical signal power is strongest at the area under directly each LED and it becomes more weaker moving towards the corners.

In the first case when $\phi_{1/2} = 20^{\circ}$, the value of optical received power and the SNR level is the highest level at small areas (ΔSNR=12.5638 dBm) and appear blind spots on received plane including four corners and overlap areas. It means receiver maybe cannot get the light or receive a small amount of light from the LEDs in these areas.

In the second scenario when $\phi_{1/2} = 40^{\circ}$, the optical received power is distributed more uniformly on the receiving plane due to the difference between the highest and the lowest value is not too large about 6.4178 dBm, and the SNR level is at the medium level.



(a) $\phi_{1/2} = 20^{\circ}$ (b) $\phi_{1/2} = 40^{\circ}$ (c) $\phi_{1/2} = 60^{\circ}$

(d) $\phi_{1/2} = 20^{\circ}$ (e) $\phi_{1/2} = 40^{\circ}$ (f) $\phi_{1/2} = 60^{\circ}$

Figure 5: 2D plot for SNR distribution with diffirent LED model, (a) $\phi_{1/2} = 20^{\circ}$, (b) $\phi_{1/2} = 40^{\circ}$, and (c) $\phi_{1/2} = 60^{\circ}$. 3D plot for SNR distribution with diffirent LED model, (d) $\phi_{1/2} = 20^{\circ}$, (e) $\phi_{1/2} = 40^{\circ}$, and (f) $\phi_{1/2} = 60^{\circ}$.

5.1 Scenario III: $\phi_{1/2} = 600$

In the third case when $\phi_{1/2} = 60^{\circ}$, in the overlap areas, the value of the SNR level is declined gradually in spite of the fact that the receivers get more optical power from different LEDs. This is because, in the overlap area, the total received power is the sum of the desired signal power and noise power. And if the noise power increases that leads to the reducing of the channel quality and rising the BER level.

6 CONCLUSIONS

In this work, the possibility of deploying indoor wireless networks based on LiFi technology with the available infrastructure of LEDs and PoE cables is investigated. Additionally, we proposed the LiFi network model in the typical office with a size of 6m × 6 m × 3m and 9 LED luminaires. Furthermore, we were calculating the optical received power and SNR level in order to evaluate the quaility of the channel. Moreover, our simulation results have shown the relation between the optical received power and SNR level, i.e. when PR at the highest level, SNR also at the highest level. In order to reach the best connectivity, the half-intensity angle, in other words, it is the type of LEDs should be considered for setting the LiFi network system. Overall, with LEDs has the half-intensity angle that is equal to 40 or 60 will provide more uniform distribution of light more than the angle is 20. In conclusion, the result of this study provides wider support for researchers to investigate LiFi technology, and it could be used for companies to design effectively any practical LiFi network systems. Further studies with more focus on LiFi should be done to research the applicability of LiFi technology in hospitals and on airplanes.

REFERENCES

[1] A Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, pp. 2017-2022.

[2] M. Ayyash, H. Elgala, A. Khreishah, V. Jungnickel, Th. Little, S. Shao, M. Rahaim, D. Schulz, J. Hilt, and R. Freund, “Coexistence of WiFi and LiFi Toward 5G: Concepts, Opportunities, and Challenges”, IEEE Communications Magazine, February 2016.

[3] H. Haas, "LiFi is a paradigm-shifting 5G technology", Reviews in Physics,Volume 3, pp. 26-31, November 2018.
[4] O.I. Romanov, Y.S. Hordashnyk and T.T. Dong, "Method for calculating the energy loss of a light signal in a telecommunication Li-Fi system", International Conference on Information and Telecommunication Technologies and Radio Electronics, IEEE Xplore Digital Library, doi: 10.1109/ UkrMiCo.2017.8095404), pp. 1-7, September 2017.
[5] S. Dimitrov, German Aerospace Center (DLR), Harald Has, "Principles of LED Light Communications Towards Networked Li-Fi", Cambridge University Press 2015.
[6] Harald Haas, Liang Yin, Yunlu Wang, and Cheng Chen, "What is LiFi?", Journal of lightwave technology, 2015, pp. 8724-8733.
[7] P. Morgan Pattison, Monica Hansen, Norman Bardsley, Lisa Pattison, "2017 DOE SSL R&D Plan", Technical Report, September 2017.
[8] S. Scace, "PureLiFi shows laptop powered by Gigabit LiFi speeds at Mobile World Congress", February 2019.
[9] IEEE P802.11 - Light Communication (LC), Task Group (TG).
[10] G. Cossu, et al., "3.4 Gbit/s visible optical wireless transmission based on RGB LED", ©2012 Optical Society of America,vol. 20, no. 26, December 2012.
[11] D. Tsonev, S. Videv, H. Haas, "Towards a 100 Gb/s visible light wireless access network", Opt. Express vol. 23 (2), 2015, pp. 1627-1637.
[12] Taylor Francis Group, "Optical Wireless Communications System and Channel Modelling with MATLAB®",CRC Press 2013, 13: 978-1-4398-5235-4.
[13] Institute of Electrical and Electronics Engineers (IEEE), Standard for Local Area Networks, 802.15.7 IEEE, 2011.
[14] A. Burton, H. Le Minh, Z. Ghasemlooy and S. Rajbhandari, "A Study of LED Lumination Uniformity with Mobility for Visible Light Communications", in Proceedings of the International Workshop on Optical Wireless Communications (IWOW), 2012.
[15] O.I. Romanov, M.M. Nesterenko, L.A. Veres, Y.S. Hordashnyk, "IMS: Model and calculation method of telecommunication network's capacity", International Conference on Information and Telecommunication Technologies and Radio Electronics, IEEE Xplore Digital Library, doi: 10.1109/UkrMiCo.2017.8095412), pp. 1-4, September 2017.
[16] M.A. Skulysh, O.I. Romanov, L.S. Globa and I.I. Husyeva, "Managing the process of servicing hybrid telecommunications services". Quality Control and Interaction Procedure of Service Subsystems. In International Multi-Conference on Advanced Computer Systems, Springer, Cham, pp. 244-256, September 2018.
[17] M. Skulysh and O. Romanov. "The structure of a mobile provider network with network functions virtualization". 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET). IEEE, 2018. pp. 1032-1034. doi: 10.1109/TCSET.2018.8336370.
[18] L. Globa, M. Skulysh, O. Romanov and M. Nesterenko, "Quality Control for Mobile Communication Management Services in Hybrid Environment". In The International Conference on Information and Telecommunication Technologies and Radio Electronics, Springer, Cham, pp. 76-100, November 2018.
[19] M. Skulysh, "The method of resources involvement scheduling based on the long-term statistics ensuring quality and performance parameters". International Conference on Information and Telecommunication Technologies and Radio Electronics (UkrMiCo). IEEE, 2017. pp. 1-4.
[20] L. Globa, M. Skulysh, A. Zastavenko, "The method of resources allocation for processing requests in online charging system". In: The Experience of Designing and Application of CAD Systems in Microelectronics. IEEE, 2015, pp. 211-213.

The Steepest Descent Method Using the Empirical Mode Gradient Decomposition

Vasiliy Esaulov and Roman Sinetsky
Platov South-Russian State Technical University, Prosvescheniya Str. 132, Novocherkassk, Russia
rmsin@srspu.ru, esaul_va@mail.ru

Keywords: Steepest Descent Method, Gradient Descent Method, Empirical Mode Decomposition, Optimization Problem.

Abstract: The aim of the article is to study the possibility of improving gradient optimization methods. The leading approach to the chosen concept is based on the possibility of a featured description of the gradient that sets the direction of the search for a solution. A modification of the method of steepest descent of global optimization based on the Hilbert-Huang transform is proposed. The proposed solution is based on the decomposition of the gradient of the objective function into empirical modes. The main results of the work are iterative optimization methods, in which, in addition to the gradient, its empirical modes are also taken into account. New estimates of the descent step are obtained, which could not be deduced in the classical formulation of the steepest descent method. Their correctness is due to the fact that in the absence of the possibility of gradient decomposition, they are reduced to existing estimates for the steepest descent method. The theoretical significance of the results lies in the possibility of expanding the existing gradient methods by a previously not used gradient description method. The practical significance is that the proposed recommendations can help accelerate the convergence of gradient methods and improve the accuracy of their results. Using the Python language, computational experiments were carried out, as a result of which the adequacy of the proposed method and its robustness were confirmed.

1 INTRODUCTION

The optimization problem is a significant mathematical model in a wide class of disciplines. Its methods are applied in areas such as computer-aided design, machine learning, mathematical modeling, and others. As one of the main statements of the optimization problem, we will further consider the problem of finding the minimum of a function. Let the task of finding the minimum

$$F(X) \to \min,\ X \in R^n, \qquad (1)$$

where $F(X)$ – objective function; X – objective function parameters.

The formula for the coordinate descent process for (1) in the case of applying the gradient has the form

$$X_{k+1} = X_k - \lambda_k \nabla F(X_k),\ k = 0,1,2...., \qquad (2)$$

where $\nabla F(X_k)$ – objective function gradient; X_k, X_{k+1} – objective function parameters at k and k+1 iteration respectively; λ_k – step value, $\lambda_k \geq 0$.

The essence of the steepest descent method is the selection of such λ_k, where, with a known X_k, the condition is satisfied.

$$F(X_k - \lambda_k \nabla F(X_k)) \to \min,\ \lambda_k \geq 0. \qquad (3)$$

Let us consider the possibility of modifying the steepest descent method based on the representation of the gradient of the objective function in some basis.

Indeed, we can consider gradient $\nabla F(X_k)$ as a discrete one-dimensional signal having a length equal to the dimension of the search space. This makes it possible to apply the methods of signal processing theory to it.

Having a spatial decomposition of the gradient of the objective function in some basis, one can both improve the convergence of gradient methods and get the opportunity to synthesize their modifications with fundamentally new properties.

2 EMPIRICAL MODE DECOMPOSITION OF GRADIENT

Let us consider the possibility of applying the empirical mode decomposition method to the gradient of the objective function [2, 3]. The principle of decomposition into empirical modes developed relatively recently. Its main specialization is the analysis of non-stationary processes. It is quite well established in a broad range of problems [6, 7].

One of the significant advantages of the empirical mode decomposition (EMD) method is that it does not require a choice of basis. Unlike Fourier or wavelet analysis, a mathematical apparatus is less developed for it. However, this fact does not reduce interest in studying the effectiveness of its application for practical problems.

Let us consider some general features of the empirical mode method. The basic functions used in the decomposition are extracted directly from the original signal. This, in turn, allows us to take into account its individual structural features.

The qualitative basis of the apparatus of empirical modes is to use the multiple addition of white noise to the signal. Next, the average value of the distinguished components is calculated by the classical method of decomposition as the result.

As a result of decomposition, the signal is presented in the time-frequency domain, which allows revealing hidden modulations and energy concentration regions. Since the decomposition is based on the data of a specific local time domain of the signals, it is also applicable to non-stationary signals. Using EMD, it is possible to determine the instantaneous frequency as a function of time, which allows you to get a clear idea of the internal structure of the signal [2–5].

An *empirical mode* (or *intrinsic mode function*, IMF) is such a function that has the following properties [7]:

1) The number of function extrema (maxima and minima) and the number of zero intersections should not differ by more than one.
2) At any point, the average value of the envelopes defined by local maxima and local minima should be zero.

IMF is an oscillatory function, but instead of a constant amplitude and frequency, as in a simple harmonic, IMF can have a variable amplitude and frequency, as functions of an independent variable (time, coordinate, etc.).

The first property guarantees that the local maxima of the function are always positive, the local minima are respectively negative, and between them, there always are intersections of the zero line.

The second property ensures that the instantaneous frequencies of the function will not have undesirable fluctuations resulting from the asymmetric waveform.

Any function and any arbitrary signal that initially contains an arbitrary sequence of local extrema (minimum 2) can be divided into the IMFs family and the residual trend. If the data are devoid of extrema, but contain inflection points (“hidden” extrema of superimposing mode functions and steep trends), then signal differentiation can be used to “open” extrema [8, 9].

Suppose that there is an arbitrary signal $x(t)$. The essence of the EMD method consists in sequentially calculating the functions of the empirical modes $c_j(t)$ and the residues $r_j(t) = r_{j-1}(t) - c_j(t)$, where $j = 1, 2, ...,n$ at $r_0 = x(t)$. The decomposition result will be the representation of the signal as a sum of mode functions (IMFs) and the final residual [6–9]:

$$x(t) = \sum_{j=1}^{n} c_j(t) + r_n(t), \qquad (4)$$

where n is the number of IMFs that is established during the calculations.

3 EMD ALGORITHM

The block diagram of the EMD algorithm is presented in Figure 1 [4–6].

The EMD algorithm consists of the following operations:

1) For any data $x(t)$, all local extrema are identified.
2) Based on the extrema, the upper, $u(t)$, and lower, $l(t)$, envelopes are formed (in this case, cubic spline interpolation can be used).
3) Envelope mean value is calculated as
4) $m(t) = [u(t) + l(t)] / 2$.
5) The difference between the original signal and the average value is considered as IMF
6) $h(t) = x(t) - m(t)$.
7) The current $h(t)$ value is evaluated for IMF compliance.
8) If $h(t)$ does not satisfy the definition of IMF, go to steps 1-5. Otherwise, the IMF is accepted as component $c(t)$.
9) The residual function $r(t) = x(t) - c(t)$ is determined. Steps 1 to 6 are repeated for $r(t)$.
10) The operation ends when $r(t)$ contains no more than one extremum.

So, using the EMD method, let decomposition of $\nabla F(X_k)$ constructed on basis of modes $\{H_i(X_k)\}$, i=1,…,m in such a way that

$$\nabla F(X_k) = \sum_{j=1}^{m} H_j(X_k) . \quad (5)$$

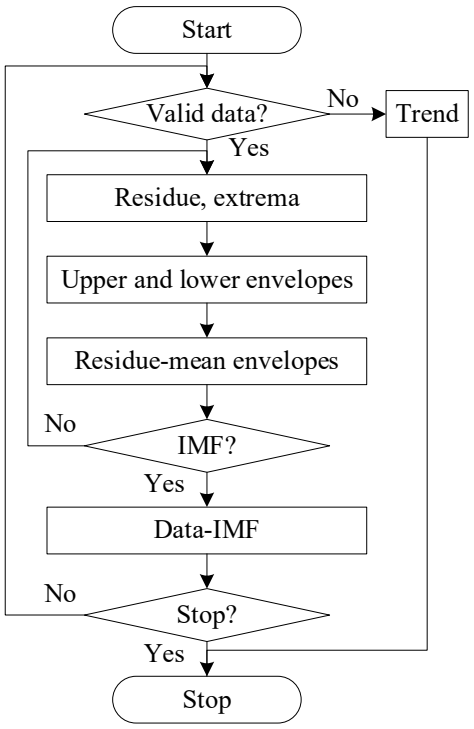


Figure 1: The block diagram of the EMD algorithm.

Set a descent for each of the mods

$$X_{k+1}^{j} = X_k - \lambda_k H_j(X_k), \; j = 1,2,\text{K},m . \quad (6)$$

The obtained set of points $\{X^i_{k+1}\}$, i=1,…,m can be considered as a combination of some alternative results obtained by approximating the gradient of the objective function while maintaining its structural properties. At a qualitative level, this makes the process of finding the optimum nonlocal.

The presence of many possible alternatives to the solution can accelerate the convergence of the search process. On the other hand, it can be expected that it will be more stable in situations where the initial approximation is given far enough from the optimal solution.

4 DESCENT STEP SELECTION

From statements (5), (6), two variants of modification of the search rule for step λ_k were formulated in the paper. In the first case, we can require the following condition

$$\sum_{j=1}^{m} F(X_{k+1}^{j}) = \sum_{j=1}^{m} F(X_k - \lambda_k H_j(X_k)) \to \min . \quad (7)$$

Thus, the above condition consists in finding a step that minimizes the sum of the function values calculated at points obtained by descent along all components of the EMD of gradient $\nabla F(X_k)$. The decrease in the value of the function on average over the totality of values on the set $\{X^i_{k+1}\}$, i=1,…,m allows us to talk about the global nature of optimization.

We can estimate the optimal descent step in (7). Expand the left side in a Maclaurin series

$$F(X_{k+1}^{j}) \approx F(X_k) - \lambda_k \left(\nabla F(X_k), H_j(X_k)\right) + \\ + \lambda_k^2 H_j^T(X_k) G(X_k) H_j(X_k) , \quad (8)$$

where $G(X_k)$ – Hessian matrix of objective function;

Then for (7) we obtain the following approximation

$$f(\lambda_k) = \sum_{j=1}^{m} F(X_{k+1}^{j}) \approx$$

$$\approx \sum_{j=1}^{m} \left(F(X_k) - \lambda_k \left(\nabla F(X_k), H_j(X_k)\right) + \right.$$

$$\left. + \lambda_k^2 H_j^T(X_k) G(X_k) H_j(X_k) \right)$$

The necessary optimality condition in this case has the form

$$\frac{df(\lambda_k)}{d\lambda_k} = 0 . \quad (9)$$

From (9) we obtain an estimation for the step

$$\lambda_k = \frac{\sum_{j=1}^{m} \left(\nabla F(X_k), H_j(X_k)\right)}{\sum_{j=1}^{m} H_j^T(X_k) G(X_k) H_j(X_k)} = \\ = \frac{\left(\nabla F(X_k), \nabla F(X_k)\right)}{\sum_{j=1}^{m} H_j^T(X_k) G(X_k) H_j(X_k)} \quad (10)$$

Expression (9) allows us to estimate the descent step taking into account several empirical modes, which are the levels of gradient decomposition in the space of empirical modes.

The solution to problem (7) can be used in two ways. The first of them is that the obtained λ_k can be used to go over to the next approximation in (2). Another way is to iterate over alternatives from the set $\{X^i_{k+1}\}$, i=1,…,m. The point X^i_{k+1} at which the smallest value is reached can be considered as the next approximation to which process (7) can be reapplied.

Another search option for λ_k is to minimize the expression

$$\inf_{X \in \{X_{k+1}^i\}_{i=1}^m} F(X) \to \min . \quad (11)$$

It can be seen from the above that the use of EMD allows us to obtain many alternative solutions to problem (1). Moreover, it can be reformulated in terms of alternative expressions (7), (10).

5 EXPERIMENTS

As a technique that allows us to evaluate the algorithms proposed in (7), (9), we used the solution of test problems of multidimensional optimization problems reduced to statement (1). The optimization results were compared with those obtained using the standard steepest descent algorithm. The maximum number of iterations was set as 50000, the convergence error was set as 10^{-11}. As the first test function, a quadratic function of the form

$$F(X) = \sum_{i=1}^{n} i \cdot x_i^2 \quad (12)$$

was used.

The test value of the function is $F^* = 0$. The initial approximation has the form x_0 = (-2, 2, -2, 2, -2, 2, -2, 2, -2, 2,-2, 2,). The simulation results are shown in table 1.

As the second test function, the Rastrigin function was used [1]. The test value of the function is $F^* = 0$. The initial approximation has the form x_0 = (-5, 5, -5, 5, -5, 5, -5, 5, -5, 5,-5, 5,). The simulation results are shown in table 2.

As the third test function, the Rosenbrock function [1] was used. The test value of the function is $F^* = 0$. The initial approximation has the form x_0 = (-2, 2, -2, 2, -2, 2, -2, 2, -2, 2,-2, 2,). The simulation results are shown in table 1.

Table 1: Experiments results for function (12).

Method	Function value	Number of iterations
Method (2), (3)	0	108
Method (7) with search	0	103
Method (7), (2)	–	–
Method (11) with search	0	107

Table 2: Experiments results for Rastrigin function.

Method	Function value	Number of iterations
Method (2), (3)	0	14
Method (7) with search	0	12
Method (7), (2)	0	15
Method (11) with search	0	10

Table 3: Experiments results for Rosenbrock function.

Method	Function value	Number of iterations
Method (2), (3)	$4.485 \cdot 10^{-17}$	30905
Method (7) with search	$1.415 \cdot 10^{-16}$	30884
Method (7), (2)	$4.968 \cdot 10^{-17}$	10052
Method (11) with search	$9.4 \cdot 10^{-4}$	1924

6 CONCLUSION

From the presented results it is seen that the gradient decomposition in the case of applying the EMD method gives adequate optimization results. At the same time, when trying to combine it with the traditional steepest descent method, a situation of solution divergence may arise. On the other hand, the application of methods (7), (11) can lead to a decrease in the number of iterations in comparison with the traditional method of steepest descent. Thus, the possibilities of a refined search for the descent step that exist in (7), (11), as well as the choice of an approximation obtained from many alternative options, are the strengths of the method proposed in the work.

REFERENCES

[1] A.A. Zhiglyavskii, A.G. Zhilinskas, “Metody poiska global'nogo optimuma” [Methods of search of global optimum]. Moscow, Nauka Publ., 1991, p. 247.

[2] N.E. Huang, Z. Shen, S.R. Long, M.C. Wu, H.H. Shih, Q. Zheng, N.-C. Yen, C.C. Tung and H.H Liu, “The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis”, Proc. R. Soc. London Ser. A., vol. 454, 1998, p. 903.

[3] N.E. Huang, Z. Shen and S.R. Long, “A new view of nonlinear water waves”, the Hilbert spectrum, Annu. Rev. Fluid Mech, vol. 31, 1999, p. 417.

[4] N.E. Huang, “The Hilbert–Huang transform and its applications”, S.S.P. Shen. Singapore, World Scientific, 2005.

[5] K.T. Coughlin and K.K. Tung, “11-year solar cycle in the stratosphere extracted by the empirical mode decomposition method”, Adv. Space Res., vol. 34, 2004, p. 39.

[6] E.P.S. Neto, M.A. Custaud, C.J. Cejka, P. Abry, J. Frutoso, C. Gharib and P. Flandrin, “Assessment of cardiovascular autonomic control by the empirical mode decomposition”, Method. Inform. Med., vol. 43, 2004, p. 60.

[7] Z. Wu and N.E. Huang, “A study of the characteristics of white noise using the empirical mode decomposition method”, Proc. R. Soc. London, Ser. A., vol. 460, 2004, p. 1597.

[8] P.O. Pavlovichev and A.L. Priorov, “The empirical mode decomposition for reduction in speech signals”,

DSPA: Digital Signal Processing Issues, vol. 6, no. 2, 2016, pp. 398-403.

[9] A.K. Alimuradov, Y.S. Kvitka, A.P. Zaretskiy and A.P. Kuleshov, “Noiseproof processing of speech signals based on the complementary ensemble emperical mode decomposition”, Proceedings of Moscow Institute of Physics and Technology, vol. 3 (31), 2016, pp. 43-53.

Prediction of Air Pollution Concentration Using Weather Data and Regression Models

Aleksandar Trenchevski, Marija Kalendar, Hristijan Gjoreski and Danijela Efnusheva
Ss. Cyril and Methodius University, Faculty of Electrical Engineering and Information Technologies,
Rugjer Boshkovik 18, PO Box 574, 1000 Skopje, R.N. Macedonia
atrenchevski@gmail.com, {marijaka, hristijang, danijela}@feit.ukim.edu.mk

Keywords: Air Pollution, Feature Selection, Machine Learning, Prediction, Regression Models.

Abstract: Air pollution is becoming a global environmental problem, in both developed and developing countries. It has greatly impacted the health and lives of millions of people, thus increasing mortality rates and pollution induced diseases reports. This paper proposes machine learning methods for predicting the rates of possibly increased air pollution in several areas, by processing the gathered data from multiple weather and air quality meter stations. The data has been gathered over a period of several years including air quality and pollution data and weather data including temperature, humidity and wind characteristics. The development process included feature extraction, feature selection for removing redundancy, and finally training multiple regression models and hyperparameter optimization. Pollutants and air quality index (AQI) were used as target variables, and appropriate regression models were trained. The performed experiments show that XGBoost is the most accurate, achieving MAE of 8.9 for Center, 8.9 for Karpos and 7.3 for Kumanovo municipality for the PM10 pollutant. The improvements over the baseline, Dummy regressor are significant, reducing the MAE for 12 on average.

1 INTRODUCTION

The continuing increase in computing power and the development of many machine learning methods, currently opens up the possibility for massive data processing. One quite interesting and very particular research area of interest among scientists around the world encompasses climate changes and atmospheric impact by human behavior. Consequently, a quite vast amount of data describing atmospheric characteristics is being collected over a number of previous years. The existing data and the novel ways of data processing using machine learning algorithms, enable the scientists to gather, process and connect the data, and subsequently produce a novel view, relations and deductions. These new methods make it possible to detect the interconnections within the data, to present these results effectively, as well as to make some predictions based on the previous data occurrences.

Due to increasing demand for energy, population growth, economic development, urbanization and transportation, the problem of air pollution becomes the focus of modern society, primarily because of its adverse effects on human health, the environment and the climate system. It can be noted that the concentration of harmful substances in the atmosphere is constantly increasing, but this problem has only recently been approached with greater care. The main pollutants in the air are carbon oxides (carbon monoxide and carbon dioxide), nitrogen and sulphur oxides, particulate matter (PM2.5 and PM10), ammonia, some toxic metals, volatile organic compounds, etc.

Realizing that mere air monitoring means just scraping the surface of this enormous problem, the next step which is a challenge for scientists is the possible prediction of increased air pollution rates at particular time periods. This information could possibly aid the human population for health preservation, as well as governmental organizations responsible for controlling traffic and industrial capacities that have been identified as main polluters and source of the toxic materials present in the air.

Taking into account the seriousness of the researched area, this paper focuses on predicting the

hourly air pollution for multiple locations across the country for the year 2018 using vast data merged in a dataset created from data gathered in the previous four years (2015 - 2018). Multiple regression models were used for processing the data in order to benchmark and pinpoint the most precise model that could be further developed and improved for the designated cause.

The rest of the paper is organized as follows: Section 2 presents related work in similar areas of research. Section 3 layouts the data gathering and preparation process. The used methodology has been described in Section 4. Section 5 presents the experimental results, and finally Section 6 concludes the paper.

2 RELATED WORKS

As a global environmental problem, air pollution has become a highly researched topic. Most researchers focus on monitoring and predicting the air quality index (AQI). As presented in [11] the prediction of air pollutants is extremely important for early warning and control of environmental pollution. Thus, developing models and forecasting the AQI (PM2.5 concentration) is very important to enable prevention and control of air pollution [12].

As a result, air pollution has been deeply researched all around the world and a vast variety of predictive models have been proposed. Researchers in Brazil [13] used a multilayer neural network for predicting hourly concentration of PM2.5 in Santiago, and identified the small dataset as reason for the poor predictions. Other researchers in China, [17], were evaluating hybrid regression models, EMD–SVR hybrid and EMD–IMF hybrid, achieving at most 80% accuracy, using only past AQI data to predict present AQI, not taking into consideration other correlation between different pollutants. Researchers in Italy [14] used feed-forward neural networks to predict ozone and PM10 in Milan. The predictions showed a satisfactory reliability, but the model still has the tendency toward overfitting. Another work, [15], uses recursive neural network model to forecast PM10 concentration for the next few days. The model showed 95% accuracy in predictions, but simultaneously yielding 30% false positives, which shows the limitations of neural networks models.

Another predictive model, the supplementary leaky integrator echo state network (SLI-ESN) is presented in [3]. This model aims to accurately predict the PM2.5 time series and, thus, implements different techniques to incorporate the historical information from the data and to consider the redundancy and correlation between multivariable time series. In order to achieve this, a minimum redundancy maximum relevance (mRMR) feature selection method is being introduced to reduce redundant and irrelevant information. The proposed model has been verified by experimenting with Beijing PM2.5 time series prediction. The experiments in [3] present the validity of the SLI-ESN model, showing high prediction accuracy in medium- and long-term projects, good generalization performance and good application prospects. Nevertheless, long-term predictions in [3] are not satisfactory and need to be improved.

The study presented in [4] focuses on using two regression algorithms, SVR and RFR, to build prediction models for the AQI in Beijing and the nitrogen oxides (NO_X) concentration in an Italian city based on publicly available datasets. The root-mean-square error (RMSE), correlation coefficient (r), and coefficient of determination (R2) were used to evaluate the performance of the regression models. Both models present good experimental results, but the complexity of the SVR model increased drastically with the increase of samples.

Focusing on our vicinity, [18] presents some results regarding developing multiple regression models for predicting the pollution mostly in suburban areas in Skopje. The models used vast number of features, subsequently reduced to the most important ones. Three approaches for building a model have been considered: single regression approach, ensemble approach and TPOT. Results showed that the ensemble-based methods (XGBoost) present quite good characteristics. Another conclusion is that PM10 appeared to be generally less predictable than the PM2.5 particles. Nevertheless, the obtained results from the used datasets were moderate, due to the incompleteness of the data, with major gaps of missing data.

3 DATA PREPARATION

Firstly, datasets publicly available at the Git repository of AirCare application [19], [20], were used in this research. The archived data of four years (2015 – 2018) gathered from air pollution measuring stations across the country of Republic of N. Macedonia has been used. Next, weather data gathered from the open API provided by DarkSky project [21], for cities and municipalities across the world was used, and in this case, data regarding

Republic of N. Macedonia. Both datasets have been combined into multiple reports, presenting merged information from pollutants as well as atmospheric details of the local weather.

A total of 10 weather/pollutant stations across the country were included in the research, and all of them had different measurement inconsistencies and huge time gaps that had to be conditioned when merging the data.

All of the available information about pollutants from the datasets were included in each of the consecutive steps: feature generation, feature selection, training and prediction phases, since they all contribute to the pollution rates and overall air quality.

Data conditioning included filtering out redundant and unnecessary data in the combined reports, since several variables had very few valid data values, depending on the station that monitored those values. Finally, the best decision was to eliminate such values. For other variables, filling out missing values had to be undertaken, using interpolation methods. This was justified for variables having missing values in very short time intervals, thus filling such missing information would not change the realistic values drastically and would not have great influence on the training phase. Some of the variables, like the air quality index (AQI), can be calculated as the maximum index of all pollutants. The last step included removal of potential outliers, finally completing the dataset to be ready for the next phase.

4 METHODOLOGY

4.1 Feature Selection

Feature selection is the process of selecting a subset of relevant features to use in the model construction. Appropriate feature selection enables accuracy improvement, overfitting risk reduction, speed up in training, improved data visualization, and increases the possibility for model understanding.

Time series data, affecting air pollution, contains rich, but also irrelevant and redundant information. This information reduces the accuracy of the predictions and efficiency of the model. Many feature selection algorithms exist, and they are distinguished by the evaluation metric into three main categories: filters, wrappers and embedded methods.

For this research, the backward elimination method from the wrapper category has been used, due to its precision for selecting relevant features based on the given machine learning model. Nevertheless, for a large number of features, the time complexity rises.

Backward Stepwise (Backward Elimination) Regression is a stepwise regression approach that begins with a full (saturated) model and at each step gradually eliminates variables from the regression model to find a reduced model that best explains the data. The stepwise approach is useful because it reduces the number of predictors, reducing the multi co-linearity problem and is one of the ways to resolve the overfitting.

The subset of features is generated separately for each station, target variable and machine learning model accordingly, in order to achieve maximum efficiency of the algorithm and better testing results.

4.2 Regression Learning

The selected subset of features is used as an input in the training of six regression models for predicting pollution values. The data for the year 2018 has been chosen to be used as the test dataset for each model.

The execution process starts from the first model and collects all the prediction values, errors from predicting, as well as the selected features for each iteration. There is need for some manual feature generation in order to add features derived from the timestamp and the previous value of the target variable, as well as categorical features which were needed to be hard-coded because the regression algorithm cannot process string object features.

Selecting proper parameters for tuning the efficiency of the model is calculated using randomized grid search due to the time complexity of the grid search algorithm for a large number of parameters.

4.2.1 Decision Tree Regression

The first regression model is the decision tree, which builds regression models in the form of a tree structure. It breaks down the dataset into smaller and smaller subsets, while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches, each representing values for the attribute tested. Leaf node represents a decision on the numerical target. The topmost decision node in a tree which

corresponds to the best predictor is called a root node. Decision trees can handle both categorical and numerical data. This model would prove to be one of the most accurate models, ranking mostly at second or third place.

4.2.2 Dummy Regression

The dummy regression model is a baseline model, since it calculates the predictions by following a set of simple rules. It can be set to predict a fixed value calculated as the mean, median and quantile of the training set or as a constant given by the user. It was used as a baseline model to compare the rest of the models.

4.2.3 Light GBM Regression

The third, light GBM regression model, is a fast, distributed, high-performance gradient boosting framework, based on the decision tree algorithm, used for ranking, classification, regression and many other machine learning tasks. Its most significant characteristic is splitting the tree leaf-wise, and not depth-wise or level-wise, as other algorithms do. This enables better, faster and more accurate reduction decisions. The downside of leaf-wise splitting is increase in complexity and possible overfitting, overcome by specifying max-depth parameter where the splitting ends. Another feature of Light GBM, leading to faster training and higher efficiency is the histogram-based algorithm that buckets continuous feature values into discrete bins, thus also resulting in lower memory usage. Light GBM is suitable for use with large datasets, where it presents significant reduction in training time as compared to XGBoost.

4.2.4 Linear Regression

Linear regression is a machine learning algorithm based on supervised learning. The regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between the variables and the forecasting values. The relationship between dependent and independent variables is one of the main differentiators between regression models, as is the number of independent variables being used.

This regression technique finds out a linear relationship between x (independent variable, input) and y (dependent variable, output). Training the linear regression model means trying to find out coefficients for the linear function that best describe the input variables.

While building a linear model, the main goal is to minimize the error made by the algorithm while making predictions, which is done by choosing a function to help measure the error also called a cost function. The cost function, that help measure the error of the linear regression is the Root Mean Squared Error (RMSE) between the predicted y value and the true y value.

4.2.5 Random Forest Regression

Random Forest is a flexible, easy to use machine learning algorithm that produces great results most of the time with minimum time spent on hyper-parameter tuning. It has gained popularity due to its simplicity and the fact that it can be used for a great amount of regression tasks.

Random Forest is an ensemble machine learning technique capable of performing regression tasks using multiple decision trees and a statistical technique called bagging.

This algorithm builds multiple decision trees and merges their predictions together to get a more accurate and stable prediction rather than relying on individual decision trees.

The advantages of this model include: reduction in overfitting, its easy to measure the relative importance of each feature on the prediction, and it has an in-built validation mechanism named Out-of-bag.

However, the Random Forest model's disadvantages include: more complex and computationally expensive, slow and ineffective for real-time predictions, cannot extrapolate at all to data that is outside the range that the algorithm has seen.

Thus, Random Forest is a technique of many simple ideas combined together to yield an extremely accurate model.

4.2.6 Support Vector Regression

Support Vector regression (SVR) is characterized by the use of kernels, sparse solution and VC control of the margin and the number of support vectors. Although less popular than Support Vector Machine (SVM), SVR has been proven to be an effective tool in real-value function estimation. As a supervised-learning approach, SVR trains using a symmetrical loss function, which equally penalizes high and low misestimates.

One of the main advantages of SVR is that its computational complexity does not depend on the dimensionality of the input space. Additionally, it

has excellent generalization capability, with high prediction accuracy.

4.2.7 XGBoost regression

XGBoost is an optimized distributed gradient boosting model designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solves many data science problems in a fast and accurate way. Gradient Boosting Machines fit into a category of machine learning called Ensemble Learning, which is a branch of machine learning methods that train and predict with many models at once to produce a single superior output.

Ensemble learning is broken up into three primary subsets:

- Bagging: Bootstrap Aggregation or Bagging presents two features defining its training and prediction. For training, it uses a Bootstrap procedure to separate the training data into different random subsamples, which different iterations of the model use to train on. For prediction, a bagging regression takes an average of all models to produce output.
- Stacking: A Stacking model is a "meta-model" which uses the outputs from a collection of many, different models as input features. The idea is that this can reduce overfitting and improve accuracy.
- Boosting: The core definition of boosting is a method that converts weak learners to strong learners and is typically applied to trees. More explicitly, a boosting algorithm adds iterations of the model sequentially, adjusting the weights of the weak learners along the way. This reduces bias from the model and typically improves accuracy.
  Bagging along with boosting are two of the most popular ensemble techniques which aim to deal with high variance and high bias.

In conclusion, the XGBoost algorithm is optimized for modern data science problems and tools, it is highly scalable/parallelizable, quick to execute and typically outperforms other algorithms.

5 EXPERIMENTAL RESULTS

This section shows the comparison of the 7 regression algorithms. The dataset was split into 2 parts, 75% of the data for training (including the first three years, 2015 – 2017) and 25% for testing - i.e., year 2018. A cross-validation algorithm known as Randomized Search was used to determine the maximum potential of each algorithm.

The evaluation metrics used for ranking each of the results was Mean Absolute Error (MAE) as one of the most often used metrics with regression models. In MAE the error is calculated as an average of absolute differences between the target values and the predictions. MAE is a linear score, meaning that all individual differences are weighted equally in the average.

Figure 1 shows the comparison of MAE for each of the 7 algorithms for the PM10 pollutant in Centar municipality. The figure shows that the XGBoost is the most accurate algorithm for predicting the PM10 pollutant with a MAE value of 8.9. Light GBM is the second-best algorithm with a MAE of 9.1 and Random Forest with 10.4. The XGBoost performance is significantly better compared to the baseline - Dummy regressor.

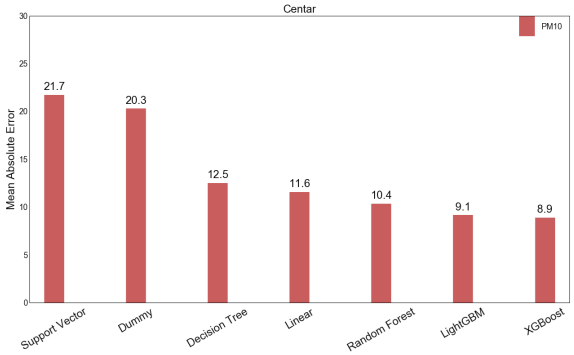


Figure 1: Mean Absolute Error plots for each algorithm predicting the PM10 pollutant in the center of Skopje.

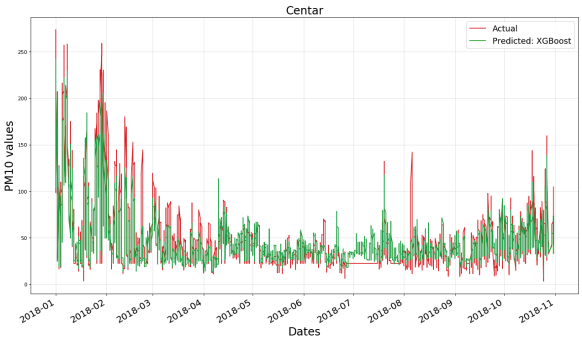


Figure 2: Actual and predicted values of the PM10 pollutant for each hour in 2018 in the center of Skopje.

Figure 2 shows the measured and the predicted values for the PM10 of the XGBoost algorithm. The values are shown for the period of 11 months in 2018. It can be noted that the predictions nicely follow the actual measurements.

Figure 3 shows the comparison of MAE for each of the 7 algorithms for the PM10 pollutant in Karpos municipality. The figure shows that XGBoost and LightGBM are the best performing, with a MAE value of 8.9 and 9.3 respectively, which is again a solid performance improvement by around 60% compared to the baseline - Dummy regressor.

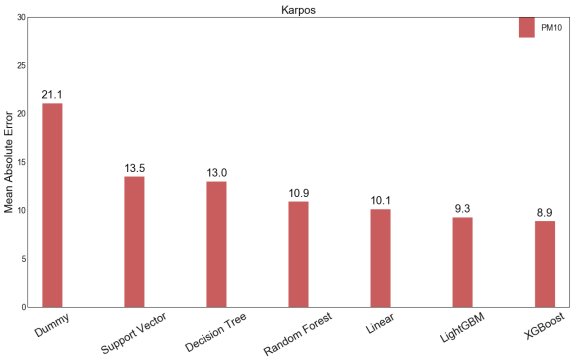


Figure 3: Mean Absolute Error for each algorithm predicting PM10 pollutant in Karpos municipality - Skopje.

Figure 4 represents the comparison between the actual pollution data and the predictions from the XGBoost algorithm for Karpos. It shows that in general the predictions follow the actual measurements, and that there are underestimations in the predictions for the 10th and 11th month.

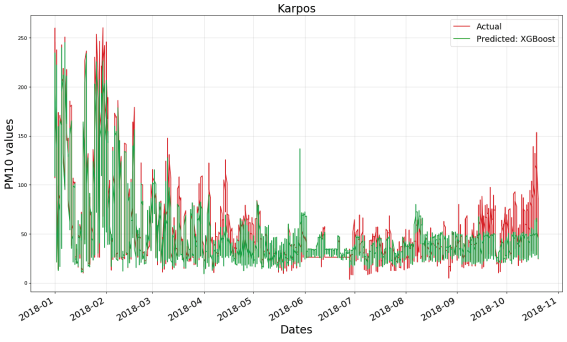


Figure 4: Actual and predicted values of PM10 pollutant for each hour in 2018 in the municipality of Karpos - Skopje.

Figure 5 shows the comparison of MAE for each of the 7 algorithms for the PM10 pollutant in Kumanovo municipality. Again, XGBoost and LightGBM are the best performing, with a MAE value of 7.3 and 8.0, respectively. This again results in a 60% performance improvement over the baseline algorithm.

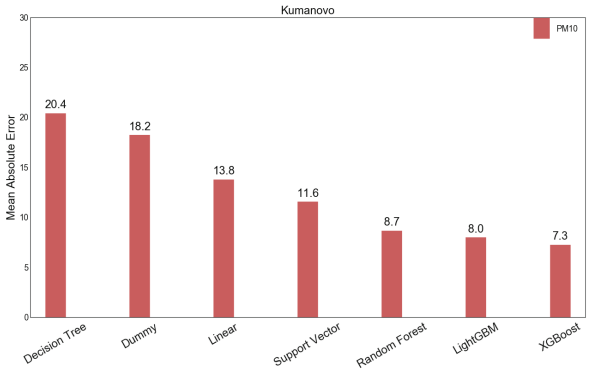


Figure 5: Mean Absolute Error for each algorithm predicting the PM10 pollutant in Kumanovo.

Figure 6 represents the comparison between the actual pollution data and the predictions from the XGBoost algorithm for Kumanovo. It shows that in general the predictions follow the actual measurements, and that there are underestimations in the predictions for the 5th month and the summer period.

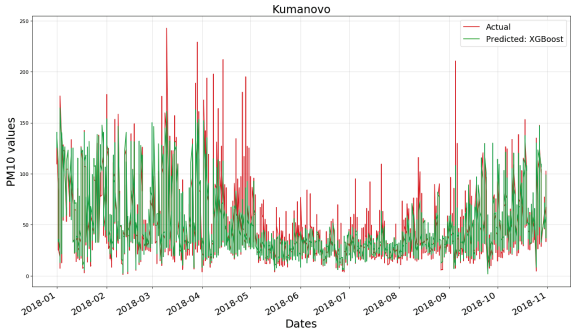


Figure 6: Actual and predicted values of the PM10 pollutant for each hour in 2018 in Kumanovo.

To summarize, the results show that XGBoost is the best performing algorithm, achieving MAE of 8.9 for Center, 8.9 for Karpos and 7.3 for Kumanovo. The improvements over the baseline, Dummy regressor are significant, reducing the MAE for 12 on average.

6 CONCLUSIONS

The paper presented a machine learning approach to predicting air pollution concentration, in particular PM10 concentration. The method uses the weather information and the previous pollution as an input, in order to calculate features and predict the PM10 concentration.

The first, and quite important step, is preparing and filtering the dataset so it can be ready for training and testing. This process eliminates unnecessary features, removes outliers that corrupt data and removes any inconsistencies with the target variables in order to preserve data integrity. The next step includes manual generation of useful features from already existing features in the dataset, using popular feature selection algorithms to improve overall dataset accuracy, as well as, using different cross-validation algorithms to achieve best results and obtain useful hyper parameters for each regression model. Finally, choosing evaluation metrics for dealing with prediction results from multiple regression models is necessary.

The overall results presented better performance than the baseline algorithm (Dummy Regression) by 60% and deliver a low mean absolute error which confirms the necessity of each mentioned step.

The incomplete data in the datasets played a major role in making the whole process harder to develop due to its inconsistency, a great deal of outliers, missing values that needed to be filled by interpolation or removed entirely. From around 30000 – 40000 rows of data it had to be cut down to around 15000 – 20000, which significantly lowers the accuracy of the model when training with half of the entire data.

REFERENCES

[1] Awad M., Khanna R., Support Vector Regression. In: Efficient Learning Machines. Apress, Berkeley, CA, 2015.

[2] Tuysuzoglu, G.; Birant, D.; Pala, A. Majority Voting Based Multi-Task Clustering of Air Quality Monitoring Network in Turkey. Appl. Sci., vol. 9, 2019, p. 1610.

[3] Xu, X.; Ren, W. Prediction of Air Pollution Concentration Based on mRMR and Echo State Network, Appl. Sci., vol. 9, 2019, p.1811.

[4] H. Liu, Q. Li, D. Yu, Yu Gu, Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms, Appl. Sci., vol. 9, 2019, p. 4069; doi:10.3390/app9194069.

[5] Backward Stepwise Regression. [Online] Available: http://www.analystsoft.com/en/products/statplus/content/help/analysis_regression_backward_stepwise_elimination_regression_model.html (28.12.2019).

[6] Decision Trees in Python with Scikit-Learn. [Online] Available: https://stackabuse.com/decision-trees-in-python-with-scikit-learn/ (28.12.2019).

[7] Linear Regression using Python. [Online] Available: https://medium.com/analytics-vidhya/linear-regression-using-python-ce21aa90ade6? (28.12.2019).

[8] Random Forest Regression model explained in depth. [Online] Available: https://gdcoder.com/random-forest-regressor-explained-in-depth/ (28.12.2019).

[9] Support Vector Regression Or SVR. [Online] Available: https://medium.com/coinmonks/support-vector-regression-or-svr-8eb3acf6d0ff (28.12.2019).

[10] A Step by Step Regression Tree Example. [Online] Available: https://sefiks.com/2018/08/28/a-step-by-step-regression-decision-tree-example/ (28.12.2019).

[11] S. Cai, Y. Wang, B. Zhao, S. Wang, X. Chang and J. Hao, "The impact of the "air pollution prevention and control action plan" on PM2.5 concentrations in Jing-Jin-Ji region during 2012-2020. Sci. Total Environ. 2017, 580, pp.197–209.

[12] L. Li, J.H. Zhang, W.Y. Qiu, J. Wang and Y. Fang, An Ensemble Spatiotemporal Model for Predicting PM2.5Concentrations. Int. J. Environ. Res. Public Health, vol. 14, 2017, p. 549.

[13] P. Pérez, A. Trier and J. Reyes, Prediction of PM2.5 concentrations several hours in advance using neural networks in Santiago, Chile. Atmos. Environ. 2000, 34, pp.1189-1196.

[14] G. Corani, Air quality prediction in Milan: Feed-forward neural networks, pruned neural networks and lazy learning. Ecol. Model. 2005, 185, 513–529.

[15] F. Biancofiore, M. Busilacchio, M. Verdecchia, B. Tomassetti, E. Aruo, S. Bianco, S. Di Tommaso, C. Colangeli, G. Rosatelli and P. Di Carlo, Recursive neural network model for analysis and forecast of PM10 and PM2.5. Atmos. Pollut. Res. 2017, 8, pp.652-659.

[16] G.W. Fuller, D.C. Carslaw and H.W. Lodge, "An empirical approach for the prediction of daily mean PM10 concentrations". Atmos. Environ. 2002, 36, pp.1431-1441.

[17] S. Zhu, X. Lian, H. Liu, J. Hu, Y. Wang, and J. Che, "Daily air quality index forecasting with hybrid models", A case in China. Environ. Pollut. 2017, 231, pp.1232-1244.

[18] P. Ilijevski, Gj. Smilevski, Predicting Air Pollution in Skopje, Project work for the course Data Warehouses and Data Processing.

[19] AirCare, Air Quality Visualized. [Online] Available: https://getaircare.com/.

[20] Pollution measurement data dumps, AirCare, December 2019. [Online] Available: https://github.com/jovanovski/MojVozduhExports.

[21] Dark Sky API, Weather Data on the Web., December 2019. [Online] Available: https://darksky.net/dev.

The Modelling Methodology of the New Product Release on the Open Market Based on the Production Systems and Rival Products Interaction Dynamics

Leonid Mylnikov, Dmitrii Vershinin and Rustam Faizrahmanov
Perm National Research Polytechnic University, Komsomolsky avenue 29, Perm, Russia
leonid.mylnikov@pstu.ru, nsenz@yandex.ru, fayzrakhmanov@gmail.com

Keywords: Agent-Based Modeling, Market Models, Diffusion, Volume-Schedule Planning, Decision Making Support, Production System Management, Innovation Project, Production System.

Abstract: The relevance of the problem is related to the necessity of improving production systems economical efficiency in constantly changing market conditions and their rivals' behaviour. In the article, it is suggested to use predictions that include diffusion component and production system models constructed on their basis which allows us to synchronize external and internal in relation to the system under consideration processes. To take into account the possible price and quality changes, the model based on markets equilibrium where there is only non-symmetrical information is suggested. The models' formalization received allows us to set a lot-scheduling planning problem that includes internal and external environmental changes. Eventually, solving this problem, we will get both lot-scheduling production plans and warehouse requirements for the products considered taking into account possible market and production systems constraints. The practical significance of the paper is related to the possibility to verify and clarify business planning data in production management tasks which makes it possible to increase the objectivity of decisions by increasing the production processes formalization rate, incorporation the diffusion factor when releasing new products and the possibility to take into consideration various rival behaviour strategies.

1 INTRODUCTION

The increasing number of market goods released leads to increased competition among production systems and goods. The competition is developing in the management and decision-making systems fields. The evolution of production systems management methods at present is an incorporation of many approaches and methods of such areas as system analysis, cybernetics, mathematical economics.

Nowadays, these practical methods by J. Von Neuman [1] and L. Kantorovich [2] of system analysis and their variations in fuzzy and interval formulations [3], [4]; production systems' efficiency management methods based on the allocation of limited key indicators and multi-criteria assessments [5], [6]; innovation and investment management methods [7], [8]; portfolio management methods [9], [10]; system dynamics methods [11]; theory of active systems [12], hierarchical systems management methods and the approaches based on the hierarchical systems management methods [13], different types of diffusion [14], [15], [16]; features of markets behaviour and equilibrium theory of markets [17], [18], [19] have become widespread.

Thus, we can conclude that the problem of economic efficiency is a complex task affecting many areas of knowledge which were developing parallel over time [20]. Despite the obvious statement, the combination of many approaches into a single theory, model or algorithm has not been done yet due to the complex nature of the problem. Neither existing models nor different methods can solve this problem properly since they make assumptions related to interaction simplification that is associated with the knowledge of decision-makers. Besides, decision-makers do not take into account the dynamic behaviour of the competing production systems due to the generalized market analysis [21], [22].

Agent-based modelling can be used to test the competing production systems' behaviour based on their models [23], [24], [25]. Systems based on agent

modelling are frequently used as a process research tool which depends on many dynamically changing parameters.

Several approaches have been developed for building agent-based models (inclusive agent interaction). Such approaches allow to tackle agent communication and synchronization problem [26], [27].

Thus, a mathematical and algorithmic base has been formed which allows us to build interaction models as shown in Figure 1.

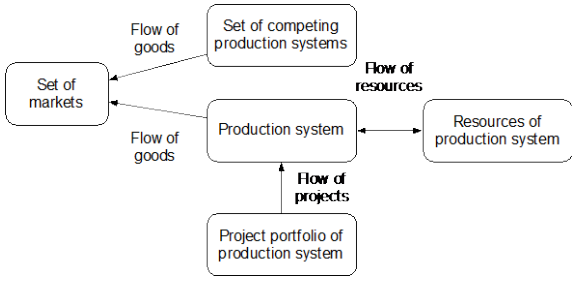


Figure 1: The scheme of production system interaction through their products on the market.

Constructing different models of the production systems this way makes it possible to take into account the behaviour of individual competing systems by building models for them or simulate situations when the other systems will perform similar actions (i.e. situations hindering the development of a managed production system).

2 MODELLING

The model development is associated with the implementation of a group of models that exchange data. The minimum required set of models is:

- The production system model that uses demand predictions for optimal planning.
- A warehouse model for accounting the volume of funds withdrawn from circulation, additional storage costs and a mechanism to meet customers' demand based on the products stockpiling when production capacity and variety of the products is limited.

As a result of the model evaluation, we can determine not only the volume-schedule plan for every product type but also products' price ranges and quality requirements. If we meet these requirements the production system will remain effective.

2.1 Production System and WareHouse Models

Let us formulate the production system management model as a model of volumetric scheduling [28]:

$$\sum_n (C_n(t) - (Q_m(t) + I_n(t))) \times X_n(t) \to max$$
$$X_n^H(t) \le X_n(t) \le X_n^B(t)$$
$$(C_n(t) - (Q_m(t) + I_n(t))) \ge 0$$
$$D_{nm}(t) \ge Z_{nm}$$
$$S_{pc} \ge S_n(t) \ge 0$$

where X_n - the product output volume n, $C_n(t)$ – the revenue of the n product, Q_m – the product net cost, I_n – packaging costs, X_n^H – the breakeven point, X_n^B – the market's volume constrains (for every product), n – the product type, m - the variety of its components, t – the time, S_n – the warehouse's empty places, S_{pc} – the warehouse's maximum capacity, D_{nm} - the matrix of available products, which has the form:

Variety of products	1	…	m
Quantity of products	D_{n1}	…	D_{nm}.

Z_{nm} - the matrix of the demanded products, which has the form:

Variety of products	1	…	m
Quantity of products	Z_{n1}	…	Z_{nm}.

In this problem, the values X_n^B will be determined based on the demand forecasting data with a coefficient corresponding to the market share claimed by the production system. The value $C_n(t)$ will be corrected on the base of market model.

In order to predict the X_n^B values, a wide variety of forecasting methods based on statistical data can be used [29]. However, if there is no statistic data available, sales volume data can be used instead.

In the modelling of the real market conditions, the maximum market capacity cannot be reached by any product immediately. The data on sales growth is in good accordance with the data obtained when the phenomenon of physical diffusion is used as an analogy. The diffusion affects the velocity of all the products diffusions on the market during the first stages of any project. In order to solve the management planning problem and estimate the products diffusion degree, the physical parallel (cloud particles' spreading from the point) was used to define the necessary volume forecast shrinking rate will be used:

$$G(x,i) = \frac{1}{\sqrt{4 \cdot \pi \cdot D \cdot i}} \cdot e^{-\frac{x^2}{4 \cdot D \cdot i}},$$

where D – the diffusion coefficient, empirically determined, either based on the products statistical data or expertly, based on the rival's reputation (its

value during the modelling process can vary due to new releases, price changes, the manufacturers' authority growth and so on according to the Walras equilibrium model [30]), i – the calculation step (to take into account the similar products' output time shift, it is crucial to reduce this value by the value of the products delay in the calculation steps), x – the market capacity.

To determine the effect on the magnitude of the forecasted value, it is important to define the value that will be subtracted from the forecast data excluding diffusion (the value of competitors' diffusion) according to the formula:

$$n(x,t) = \sum_k n_0(x_k) \cdot \Delta x_k \cdot G(x - x_k, t),$$

where $\Delta x_i = x_{i+1} - x_i$, $n_0(x_i)$ – the initial diffusion value (for equivalent products), released by the i production system, k – the number of the production system producing equivalents. The use of this component is justified since the release of competitors' products requires recalculation at each step, as the volume of the planned output and demand for the manufactured products will be changing.

As a result of the discussion above, we obtain the task that using modelling tools allows us to research the processes of new products release. The variables that affect this process are i – the product release point, n_0 – the similar products' diffusion rate by our release moment, and the values D – the product diffusion rate on the market (this value will vary depending on the market situation).

The warehouse model in this problem has the simplest form and takes into account only the cost of storing items and their availability. In this form, it can be integrated into the optimal production planning task as described above (adjusted criterial function and restrictions). Thus, the costs associated with overproduction, excess inventory and so on will be considered as well.

Describing every production system with this model and using only different sets of products, methods and data for demand forecasting, we can evaluate the effect of different production systems on each other in dynamics.

2.2 Market Model

The market model in our task is crucial for determining the fluctuation in the product prices (P) and demand. Для этого необходимо установление взаимосвязи между ними. To do so, we need to set connections between them. To get a dynamic behaviour for this dependence, we will use the parameters of quality (Q) and brand (B) (which can be measured in relative units from 0 to 1, where 1 is the maximum confidence, 0 is the minimum).

Then we can write the following models to describe the dependence of the parameters:

$$Q(t) = C_1 \cdot Q(t-1) + C_2 \cdot P(t).$$

By the price of the products in this formula we mean the competitors' products prices which will be predicted by one of the regression methods.

The brand significance assessment can be performed based on the prices, quality, advertisement rate of the product of the previous step (R):

$$B(t) = C_1 Y(t-1) + C_2 Q(t-1) + C_3 R(t-1).$$

Using these relationships, you can evaluate the redistribution of market shares. Analyzing two production systems, the competitor's market share will be calculated using this formula (the parameter will be in the range $0 \le M \le 1$):

$$M = \frac{B_r * P_r * Q_r}{B_o * P_o * Q_o} \cdot 0.5,$$

Where index r means that these are the values of the rivals' products, and index o – variables that describe the products of the production system.

The proposed interrelation allows performing calculations for researching possible changes in the strategy of the considered production system and the impact on the competitor's behavior strategy changes.

3 MODELLING AND PRODUCTION SYSTEM MARKET EFFICIENCY TASK SOLUTION

For testing and verifying purposes, we consider three situations that may occur in a market environment:

- manufacturers rivalry (the situation where two production systems produce the same product and the results of their labor compete for a place on the market);
- the rivalry of two products (the situation when two production systems offer similar goods fighting for the same audience on the same market);
- production system portfolio rivalry.

To give a dynamic character to the experiments, we will be changing the values of prices, volumes, quality, advertisement rate within the confidence interval.

For the first experiment, we used the Amazon online store sales data of the group of products. To set up the model, Feltcraft Doll Emily and Feltcraft Doll Molly sales' and prices' data were used.

As a result of the first experiment, we obtain the values given in Figure 2. The graphs show that the

sales and production volumes of every system grow and each of them gets its market share.

Doing the second experiment, for the initial setup of the model, we used sales data for Feltcraft Doll Emily and Toy Tidy Dolly Girl Design and obtained the results shown in Figure 3. This experiment also shows that as a result of competition observed in the first stages, production systems come to the equilibrium state. However, it is important to note that even in the search of steady state, production activity is effective. Thus, the model used allows the system to operate under the influence of factors that were not taken into account in the formulation of the task.

The research of the third situation is more curious since in this case, the values of the parameters characterizing the size of the brand are not stable (see Figure 4). For simulation, we selected the following products for the first and second production systems.: Mr Robot Soft Toy, Toy Tidy Spaceboy, Toy Tidy Dolly Girl Design.

As in the previous experiments, we can observe the diffusion of products in the market with a stabilization process after taking a certain market share.

Unlike the two previous experiments, the graphs show the influence of quality and brand factors on sales and profits. The production volume diagrams show that both on the market and inside the production system there is a competition for production capacity. Based on the graphs of production volumes in warehouses, we can be assumed that the model performs well in the rivalry market environment, and, therefore, sets market shares. Moreover, according to the price changes, we can observe patterns of the Walras spiral model [31], which is a factor that confirms the adequacy of the description.

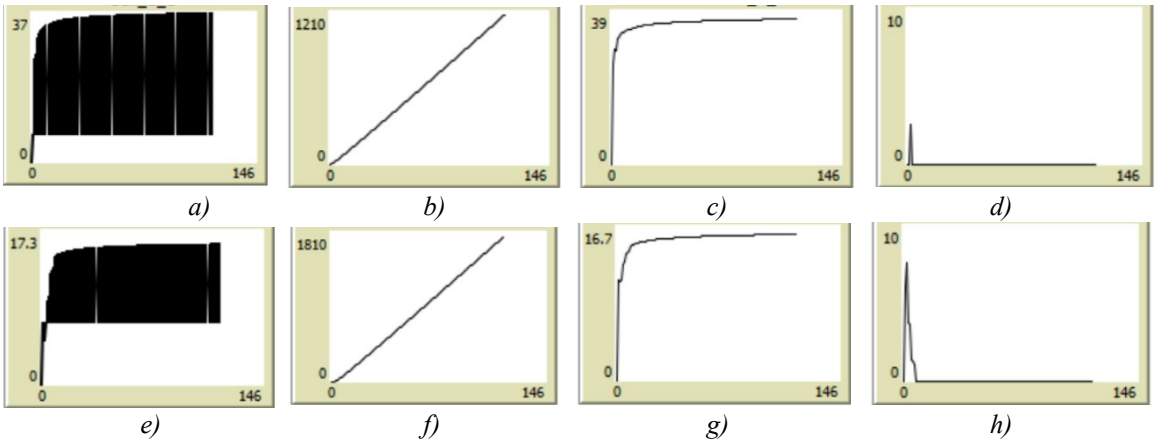


Figure 2: The first experiment modelling results (the first row describes the first production system, the second one – the second production system): a) and e) – the sales volume, b) and f) – the net profit, c) and g) – the production volume, d) and h) – the warehouse stock.

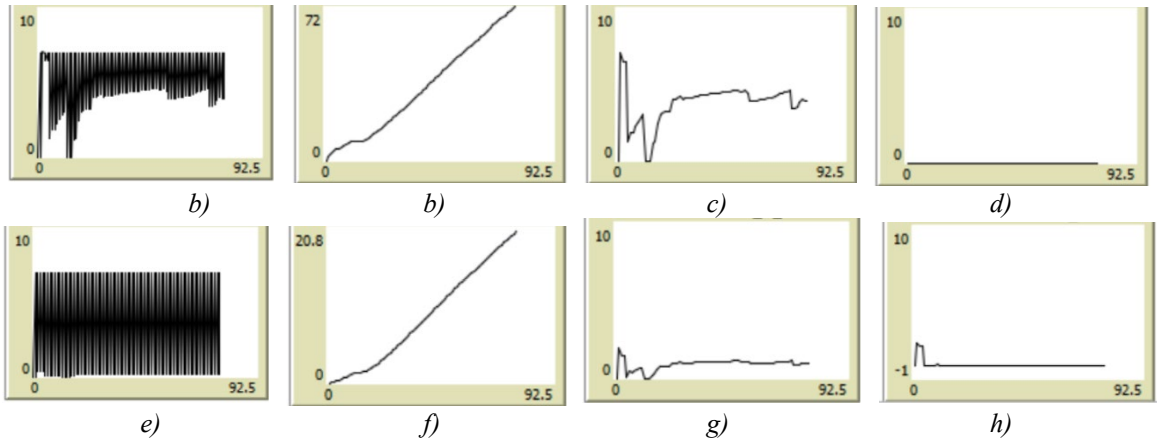
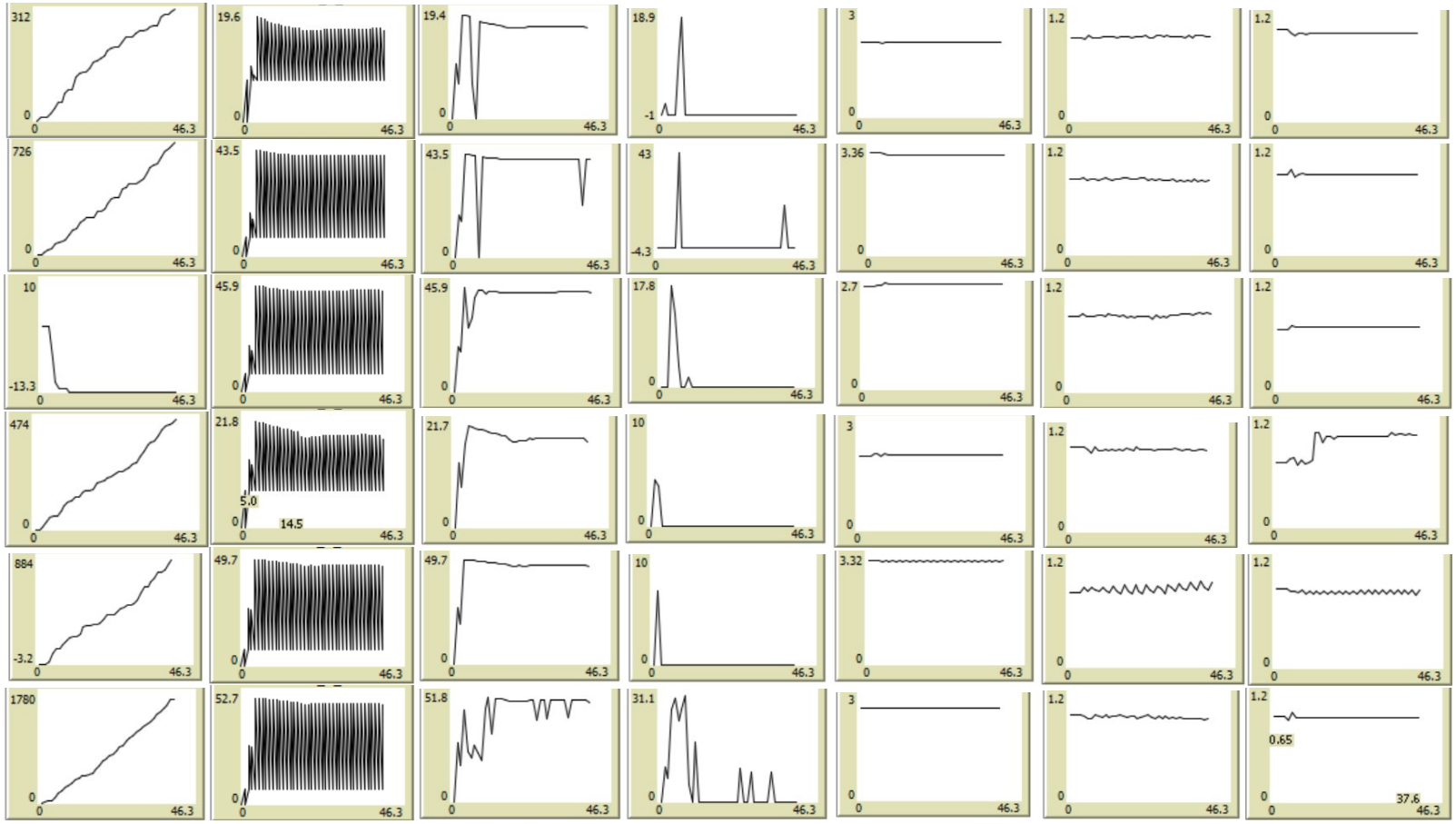


Figure 3: The second experiment modelling results (the first row describes the first production system, the second one – the second production system): a) and e) – the sales volume, b) and f) – the net profit, c) and g) – the production volume, d) and h) – the warehouse stock.



a) b) c) d) e) f) g)

Figure 4: The third experiment modelling results (1st – 3rd rows – three products of the first production system, 4th – 6th rows – three products of the second production system): a) sales volume group, b) instant sales group, c) production volume group, d) products in stock, e) price fluctuations, f) quality fluctuations, g) brand fluctuations.

Since the indicators of quality and pricing were considered as interrelated, we observe their interconnection on the graphs. The brand value in our model is used, apart from other things, as a compensation factor, aggregating non-economic indicators such as product advertisement rate, its recognition, fluctuation in its quality and price over time.

4 DISCUSSION

The article suggests a way how to simulate market competition based on a market model. The results of the simulation obtained correspond to well-known models like Walras model, a cobweb-like model, a market model with asymmetric information which indicates that the market situation will come to a steady-state after the disturbing effects.

However, unlike the others, our model can use different formalization methods and parameters describing production systems.

Another distinguishing feature of our model is the principle of using and adjusting forecast data based on the forecasting of our competitor's data and adjusting, if necessary, our input data for the following periods. As a result, the obtained data can be explained not only by theoretical knowledge of market behaviour but also by the statistical data of the retrospective periods (see. Figure 5).

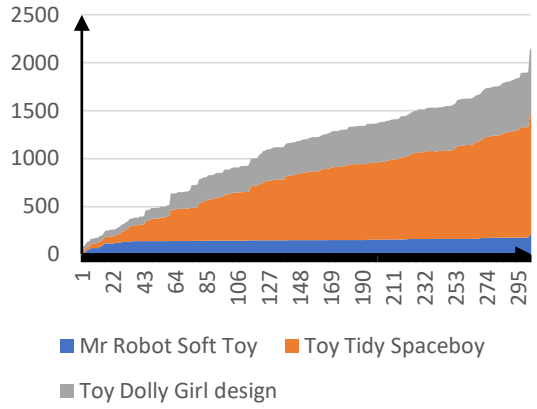


Figure 5: Real sales volume.

5 CONCLUSIONS

As a result of the discussions and calculations presented in the article, we can conclude that it is not enough to use only one approach to evaluate the situation in order to form a cost-effective project portfolio, it is also necessary to build a combination of methods and approaches which will be able to interact with each other through parameter values.

At the same time, system management does not come down to finding the optimal solution but to searching for the system state which will make it stable and effective in a constantly changing environment.

Moreover, it should be noted that the release of any new product is gradual. To describe this phenomenon, we use forecast correction based on the physical resemblance with the cloud of particles diffusion, which allows avoiding the accumulation of unclaimed products at the warehouses and heavy circulating assets freeze.

The results obtained correspond to the statistical data and theoretical knowledge in the innovation project management field. Thus, the structuring of models, the predictive approach, and the diffusion phenomenon, which we used to research the products release behaviour, should be considered as an essential part in modelling and analyzing the processes associated with the implementation of projects [16].

REFERENCES

[1] J. von Neumann and O. Morgenstern, "Theory of games and economic behavior", 60th anniv. Princeton University Press.

[2] Л.В. Канторович, "Математические методы организации планирования производства". Л.: Издание Ленинградского государственного университета, 1939.

[3] K.-I. Voigt, "Industrielles Management: Industriebetriebslehre aus prozessorientierter", Sicht. Berlin: Springer, 2008.

[4] А.А. Первозванский, "Математические модели в управлении производством". М.: Изд-во "Наука", 1975.

[5] D. Walczak and A. Rutkowska, "Project rankings for participatory budget based on the fuzzy TOPSIS method", Eur. J. Oper. Res., no. 260(2) , 2017, pp. 706-714.

[6] M.K. Sayadi, M. Heydari and K. Shahanaghi, "Extension of VIKOR method for decision making problem with interval numbers", vol. 33, no. 5, pp. 2257-2262.

[7] S. Mezhov and L. Mylnikov, "Specifics of project management on industrial innovation", Proceedings of International Conference on Applied Innovation in IT, vol. 6, iss. 1, 2018.

[8] П.Л. Виленский, В.Н. Лившиц и Р.А. Смоляк, "Оценка эффективности инвестиционных проектов". Теория и практика. М.: Дело, 2002.

[9] D. Jonas, A. Kock and H.G. Gemuenden, "Predicting Project Portfolio Success by Measuring Management Quality", IEEE Trans. Eng. Manag., vol. 60, iss. 2, 2013, pp. 215-226.

[10] H. M. Markowitz, "Portfolio selection: efficient diversification of investments", New Haven, Conn.: Yale Univ. Press, 1970.

[11] J.W. Forrester, "Industrial dynamics", J. Oper. Res. Soc., iss. 48(10) , 1997, pp. 1037-1041.

[12] V.N. Burkov and A.Y. Lerner, "Fairplay in control of active systems", Differential games and related topics ed., Amsterdam, London: North-Holland Publishing Company. H. W. Kuhn and G. P. Szego, 1971, pp. 164-168.

[13] M.D. Mesarovic, D. Macko and Y. Takahara, "Theory of hierarchical multilevel, systems". Cleveland, Ohio: Systems Research Center. Case Western Reserve University, 1970.

[14] E.M. Rogers, "Diffusion of innovations". New York, NY: Free Press, 2003.

[15] J. Mejía, R. Britto and O. Buitrago, "A forecast model for diffusion of innovations based on molecular diffusion", Ciência e Técnica Vitivinícola, vol. 30, 2015, pp. 41-54.

[16] F. Gault and E. von Hippel, "The Prevalence of User Innovation and free Innovation Transfers", Implic. Stat. Indic. Innov. Policy. MIT Sloan Sch. Manag. Pap. 4722-09, 2009.

[17] M. Rothschild and J. Stiglitz, "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information", vol. 90, iss. 4, 1976, pp. 629-649.

[18] L. Walras, "Elements of pure economics or The theory of social wealth", London, 1890.

[19] C.D. Aliprantis, B. Cornet and R. Tourky, "Economic Equilibrium: Optimality and Price Decentralization", Positivity, vol. 6, iss. 3, pp. 205-241, 2002.

[20] R.A. Faizrakhmanov and L.A. Mylnikov, "The foundations of modeling management processes for innovation projects in production-economics systems", vol. 50, iss. 3, pp. 84-90.

[21] K.J. Sandner, "Impacts of Rivalry on Types of Compensation – Competition vs. Co-operation between Multiple Agents under Technological Interdependencies", Zeitschrift für Betriebswirtschaft, vol. 79, iss. 4, pp. 427-471, April 2009.

[22] П. М. Симонов, "Экономико-математическое моделирование". Пермь: Ред.-изд. отд. Пермского гор. ун-та, 2010.

[23] M. Wooldridge and N.R. Jennings, "Intelligent agents: theory and practice", Knowl. Eng. Rev., vol. 10, iss. 2, 1995, p. 115.

[24] B. Linder, W. Hoek and J.-J. C. Meyer, "Formalising motivational attitudes of agents", Berlin, Heidelberg: Springer Berlin Heidelberg, 1995.

[25] J.-H. Lee and C.-O. Kim, "Multi-agent systems applications in manufacturing systems and supply chain management: a review paper", Int. J. Prod. Res., vol. 46, iss. 1, 2008, pp. 233-265.

[26] D. Pawlaszczyk, "Skalierbare Agentenbasierte Simulation – Verteilte Simulation agenten-basierter Modelle", KI - Künstliche Intelligenz, vol. 24, iss. 2, pp. 161-163, July. 2010.

[27] D. Nicol and R. Fujimoto, "Parallel simulation today", Ann. Oper. Res., vol. 53, iss. 1, pp. 249-285, December 1994.

[28] L. Mylnikov, D. Vershinin and D. Fatkhullin, "The use of optimal management tasks for verification and adjustment of new product release planning in discrete production systems", Proceedings of International Conference on Applied Innovation in IT, vol. 6, iss. 1, 2018.

[29] A.V. Seledkova, L.A. Mylnikov and K. Bernd, "Forecasting characteristics of time series to support managerial decision making process in production-And-economic systems", Proceedings of 2017 20th IEEE International Conference on Soft Computing and Measurements, SCM 2017, 2017.

[30] L. Walras and L. Walras's "Elements of Theoretical Economics". Cambridge University Press, 2014.

[31] J. Kol and P. de Wolff, "Tinbergen's work: Change and continuity", Economist (Leiden), 1993.

Concept Map for Clinical Recommendations Data and Knowledge Structuring

Giyzel Shakhmametova[1], Nafisa Yusupova[1], Rustem Zulkarneev[2] and Yevgeniy Khudoba[1]
[1]*Computer Science & Robotics Department, Ufa State Aviation Technical University, K.Marx Str. 12, Ufa, Russia*
[2]*Faculty of General Medicine, Bashkir State Medical University, Teatralnaya Str. 2a, Ufa, Russia*
shakhgouzel@mail.ru, yussupova@ugatu.ac.ru, zurustem@mail.ru, eugchud@gmail.com

Keywords: Structuring Data and Knowledge, Unstructured Text, Clinical Recommendations, Concept Map, Production Rules.

Abstract: The article deals with the problem of structuring medical texts of clinical recommendations, which are unstructured texts. A review of existing solutions in the field of analysis of unstructured texts of both non-specialized and medical nature was carried out, shortcomings of existing developments were identified, the need for a new software solution for structuring clinical recommendations was revealed, which, in turn, is demanded in clinical decision support systems. The method of structuring data and knowledge of clinical recommendations is described, as well as the general structure of the solution, as along with the process of forming a map of concepts, including graphematic, morphological, syntactic and semantic analysis of text. In conclusion, the results of implementation in the form of concept map fragments are presented, on the basis of which further product rules are formed, which are suitable for use in knowledge bases. The method is universal and can be applied to any clinical recommendations texts.

1 INTRODUCTION

Over the past decades, the volume of stored, processed and transmitted information has increased many times in almost all areas of human activity (i.e., research, economics, and business). This has led to a significant increase in researchers' interest in data and knowledge processing techniques and algorithms.

By degree of organization, data can be conditionally divided into two categories:

- Structured data, examples of which are databases, information system logs, sensor and sensor data.
- Unstructured data, such as text data, images and videos.

According to experts [1], about 80-90% of all information used in organizations is presented in unstructured form. There is therefore a need to reduce the labour, time, financial and other resources required to process such information. The most effective way to achieve this is to bring such information into a structured form (data structuring).

Interest in methods of extracting and classifying data in unstructured texts as tools of knowledge generation has long emerged (mid-20th century [2]), but only in the last two decades have the technologies necessary for such research been developed [2]. A significant increase in interest in this field of research was caused by the advent of data processing technologies such as Text Mining and Natural Language Processing.

One of the most important areas of human activity in which it is possible to apply technologies for structuring data presented in text form is medicine [3]. In particular, the analysis of clinical documentation, i.e., medical records, survey results, operational intervention logs, etc., is of great practical importance [4] also in the context of improving health care services [5].

Among the least studied tasks in this field to date is the task of analyzing the texts of clinical recommendations. Clinical recommendations are specialized documents developed to support decision-making by a practitioner to provide appropriate medical care in a particular clinical situation. In fact, this document is the guide of the specialist in patient management, diagnosis and treatment. Clinical recommendations contain unstructured data and knowledge that guide a person skilled in the art of prescribing treatment,

examinations, and other decisions that affect the outcome of a's patient's disease [5]. In its original form, these data and knowledge are unsuitable for automated processing, and therefore clinical recommendations in medical practice are analyzed manually. If the data and knowledge of the clinical recommendations are adjusted to a structured form, it is possible to apply them in clinical decision support systems (CDSS) for the diagnosis and selection of the's patient's treatment trajectory [6].

In this article, the clinical recommendations texts structuring method is considered and examples of results obtained from texts of clinical recommendations for bronchopulmonary diseases treatment are shown. The method is universal and can be applied to other clinical recommendations texts.

2 RELATED WORKS

In the field of structuring text data for both research and application purposes, a large number of software solutions has been developed.

2.1 Solutions for Analyzing General Texts

- SAS Text Miner, an integrated component of the SAS system designed to analyze text data, provides a large set of linguistic and analytical modeling tools designed specifically to discover and extract knowledge from text information collections [7].
- GATE (General Architecture for Text Engineering) is an open source natural language processing system that uses Java component sets [8].
- STATISTICA Text Miner is an optional extension of the STATISTICA Data Miner designed to extract knowledge from unstructured texts [9].
- Natural Language Toolkit (NLTK) is a package of libraries and programs for symbolic and statistical processing of natural language written in Python programming language, containing graphical representations and sample data [10].

2.2 Solutions for Analysis of Medical Texts

The software solutions discussed above are oriented towards processing of texts of a general nature, such as news reports, for example. At the same time, it should be noted that the style of clinical texts is very different from the style of texts from other subject areas, so that their analysis requires considerable improvement of existing methods and tools for the analysis of natural language texts. Therefore, the analysis of clinical texts was identified as a separate area of research.

Research in this area has led to the development of a number of applications and platforms specializing in integrated computer language analysis of medical texts, some of which are already being used in clinics to improve the quality of medical services. Let us take a closer look at some of the most popular ones.

- UMLS (Unified Medical Language System) is a tool for the development of computer systems for the analysis of biomedical information and other types of information in the field of health care. Developed in 1986 at the National Library of Medicine (NLM) [11].
- MedLEE (Medical Language Extraction and Coding System) - a system for extracting, structuring and encoding clinical information contained in various types of medical reports (e.g., X-ray, mammography and echocardiological studies) [12].
- cTAKES (Clinical Text Analysis and Knowledge Extraction System) is an open source natural language processing system that extracts clinical information from unstructured electronic medical card texts [13].

All of the above systems have a significant disadvantage within the framework of our task: none of them have built-in support of the Russian language.

2.3 Solutions for Automatically Building Ontologies from Text Documents

A possible means of solving the problem for presenting data and knowledge of clinical recommendations is ontology, a description of the subject area presented in the form of a conceptual diagram. We looked at the most famous means of automatically generating ontologies based on text files:

- Text-To-Onto is a software solution developed by the University of Karlsruhe researchers that automatically builds ontologies based on natural language texts by identifying key

concepts in them and discovering links between them [14].

- DOG4DAG (Dresden Oncology Generator for Directed Acoustic Graphs) is a tool for automatic generation of ontologies based on natural language texts. It is presented in the form of a plugin for Protégé 4.1 and OBOEdit 2.1. This plugin allows you to use PubMed articles, web pages, or PDF documents as input. The generation of ontology in DOG4DAG is carried out by building a hierarchical model of classes connected by relationships of the form "is subclass of" [15].

As a result of the analysis on means of automatic ontologies generation (such as, for example, ASIA, Syndikate, WebKB, etc.), it was found that support for the vast majority of them is currently discontinued, many of them are unavailable, as well as none of them support Russian text processing.
In this regard, in order to solve the problem of structuring data and knowledge of clinical recommendations, it is necessary to develop an algorithm for extracting data and knowledge from Russian-language texts of medical topics.

3 PROBLEM DEFINITION

An analysis of the current state of research showed a lack of ready-made solutions in this area. Therefore, it is necessary to develop a method and algorithm for structuring data and knowledge of Russian-language texts of clinical recommendations and bringing them into a form suitable for further processing in clinical decision-making support systems, as well as their software implementation.

The software solution being developed shall have the following characteristics:

- The software decision should accept texts of clinical recommendations in Russian.
- Text processing should result in a set of rules suitable for use in clinical decision support systems (Figure1).

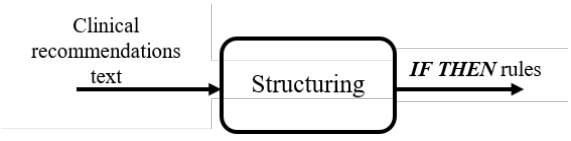


Figure 1: Task setting.

4 CLINICAL RECOMMENDATIONS DATA AND KNOWLEDGE STRUCTURING METHOD DEVELOPMENT

4.1. General Structure of a Method

In order to structure the texts on clinical recommendations, the authors proposed and developed the “concept map” object and identified the following main stages of text processing:

- Definition of keywords (performed automatically by means of specialized algorithms or manually, 2 modes).
- Mapping concepts based on keywords.
- Highlight the concepts of rules in the map and represent them in a form suitable for use in the HACT.

A diagram illustrating the above steps is shown in Figure 2.

4.2 Concept Map

The concept map is essentially a form of semantic networks and is an oriented graph whose vertices record the concepts of the subject area, and in the edges, the relations between them. Relations between concepts can be taxonomic (i.e., forming a hierarchy of concepts) and of other kinds.

A key feature of the concepts proposed by the authors of maps is the ability to display rules for links of objects based on conditional constructs. They are displayed in the diagram using dashed lines; the signature of communications for a conditional part of the rule is followed by a prefix "at", and the conclusions of the rule, by the prefix "3". An example of a concept map fragment to be developed is shown in Figure 3.

4.3 Algorithm Structure

The stage of automatic concept mapping is essential and most time-consuming, because it involves a large number of natural language processing algorithms.

The concept map development algorithm is shown in Figure 4.

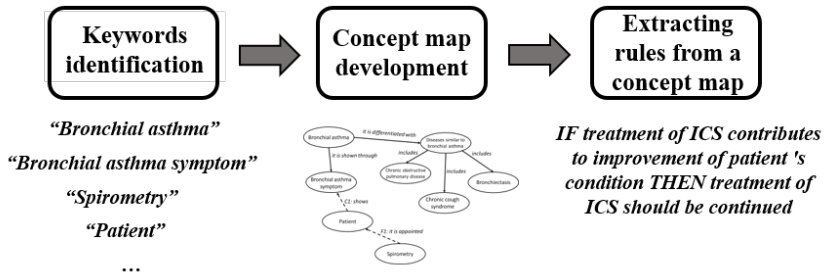


Figure 2: Stages of the process for structuring data and knowledge of clinical recommendations.

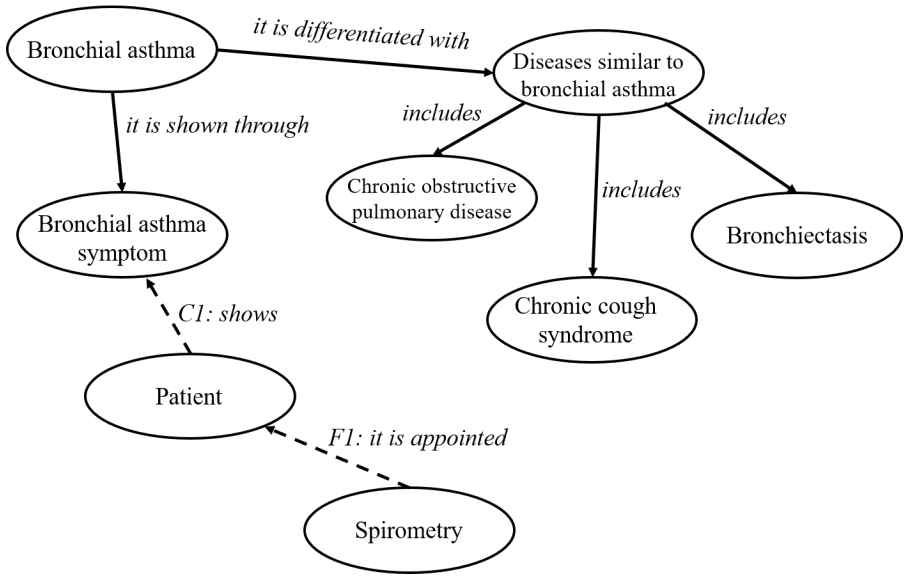


Figure 3: Concept map fragment.

In the process of constructing a concept map through this software service, the following main steps can be distinguished:

- Graphical analysis. This stage is preparatory; during this process, the text is preprocessed as required for the next steps. The tasks of this stage are to divide the source text into paragraphs, sentences, words, as well as to highlight specific words and phrases in the text (for example, proper names).
- Morphological analysis. The purpose of this step is to construct a morphological interpretation of the words of the input text. In the process of text processing, word forms are extracted from the text and subsequently normalized (lemmatized), i.e., reduction to the initial morphological form, for example, for nouns it will be a nominative case, singular, for verbs an infinitive, etc.
- Parsing. This step determines the link structure of word forms in sentences. The result of the analysis is usually presented as a so-called syntax tree (a graph of a tree structure whose nodes display word forms and whose branches are links). One of the most popular methods of conducting such analysis is the use of link grammar.
- Semantic analysis. This stage consists of highlighting semantic relations and forming semantic representation of texts. Typically, during processing, key entities (word forms or phrases) are highlighted in the text, weighted, and the strength of the links between them is counted. The result of text processing at this stage is also a graph in whose vertices concepts are placed, and in nodes - links between them. Among the methods used in practice are graph methods, varieties of

Markov random fields method, and methods of context-dependent analysis.

Building a concept map. A final step in which the graph obtained in the previous step is used to map concepts suitable for further processing.

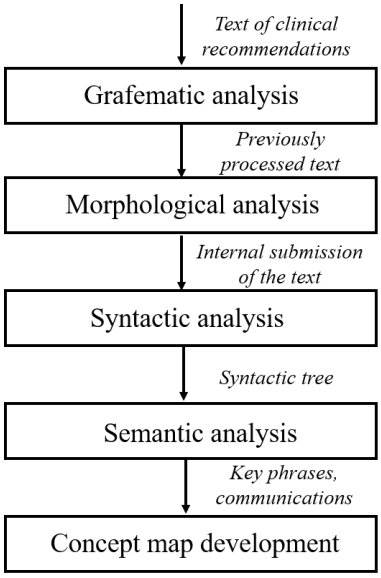


Figure 4: Concept map development algorithm.

5 REALIZATION RESULTS

The result of processing clinical recommendation text using software developed based on the methods and algorithms discussed above is a concept map, which is a mapping of key concepts found in clinical recommendation texts and the relationships between them.

Examples of the resulting concept map for clinical recommendations on the treatment of chronic obstructive pulmonary disease are shown in Figure 5.

This concept map shows the key concepts that appear in the text of clinical recommendations, as well as the relationships between them. It is necessary to separately note the presence of two conditional links represented on the map by dashed lines with prefixes (C for the conditional part of the rule and F for the final part, respectively).

Based on these conditional relationships, two rules can be generated in a format suitable for use in decision support systems:

- IF treatment of ICS (Inhaled Corticosteroids) contributes to improvement of 's patient's condition then treatment of ICS should be continued.
- IF treatment with ICS did not cause a significant change in pulmonary function THEN COPD (chronic obstructive pulmonary disease) is a probable diagnosis.

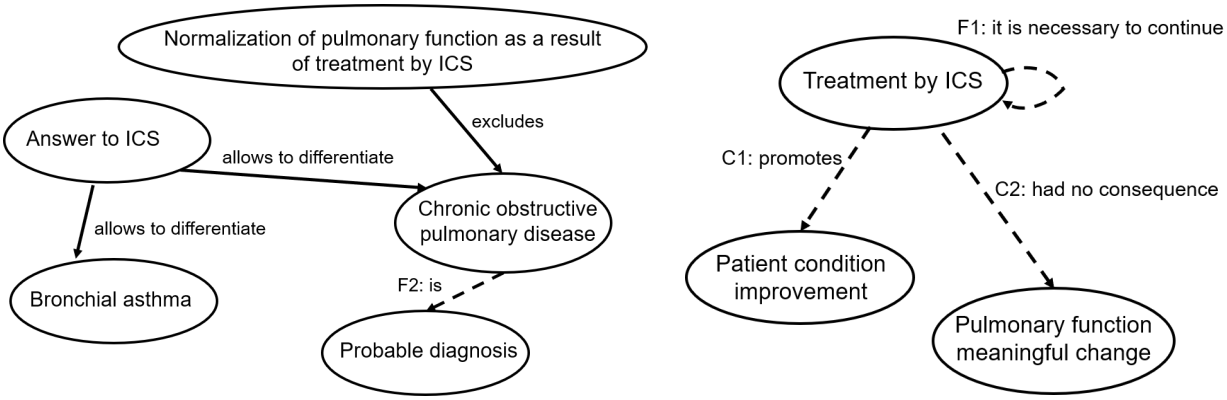


Figure 5: Fragments of the realization results.

6 CONCLUSIONS

An analysis of existing solutions for the task of automatic extraction of data and knowledge from the texts of clinical recommendations has shown that none of the currently available software is suitable for the task in question. In this regard, it was concluded that a new method of structuring data and knowledge of clinical recommendations should be developed and implemented as a software solution.

The proposed method distinguishes the use of a new means for presenting data and knowledge in a structured form called a "concept map"; i.e., it is possible to represent the relationships between the concepts containing the conditional and the final part. Since one of the key stages of the developed method for automatic extraction of data and knowledge from clinical recommendations is generation of product rules from obtained maps of concepts, it is possible to further apply these rules in knowledge bases of clinical decision support systems.

ACKNOWLEDGMENTS

The reported study was funded by RFBR according to the research projects № 19-07-00780, 19-07-00709.

REFERENCES

[1] R. Grishman, 'Twenty-five years of information extraction". Natural Language Engineering, vol. 25(6), 2019, pp. 677-692, doi: 10.1017/ S1351324919000512.

[2] S. Grimes, "A Brief History of Text Analytics". Eye Network. Retrieved, June 2016.

[3] B. Presannan, N. Ramasubramanian and A.S. Vijayan, "Disease risk prediction from clinical texts", 2020, doi:10.1007/978-981-32-9515-5_30.

[4] F. Dhombres, J. Charlet and Section Editors for the IMIA Yearbook Section on Knowledge Representation and Management. Formal medical knowledge representation supports deep learning algorithms, bioinformatics pipelines, genomics data analysis, and big data processes. Yearbook of Medical Informatics, vol. 28 (1), 2019, pp. 152-155, doi: 10.1055/s-0039-1677933.

[5] B. Séroussi, L.F. Soualmia and J.H. Holmes, Transforming data into knowledge: How to improve the efficiency of clinical care? Yearbook of Medical Informatics, vol. 26(1), 2017, pp. 4-6, doi:10.1055/s-0038-1637768.

[6] C. Combi and G. Pozzi. "Clinical information systems and artificial intelligence: Recent research trends". Yearbook of Medical Informatics, vol. 28(1), 2019, pp. 83-94, doi:10.1055/s-0039-16779156.

[7] Text Mining Software, SAS Text Miner | SAS. Renewal date: 18.03.2019. [Online]. Available: https://www.sas.com/en_us/ software/text-miner.html.

[8] General Architecture for Text Engineering. Renewal date: 18.03.2019. [Online]. Available: https://gate.ac.uk.

[9] STATISTICA Text Miner. Renewal date: 18.11.2019. [Online]. Available: http://statsoft.ru/products/STATISTICA_Data_Miner /STATISTICA_Text_Miner.

[10] Natural Language Toolkit NLTK 3.4 documentation. Renewal date: 18.03.2019. [Online]. Available: https://www.nltk.org.

[11] Unified Medical Language System (UMLS). Renewal date: 18.11.2019. [Online]. Available: https://www.nlm.nih.gov/ research/umls.

[12] MedLEE | MedLingMap. Renewal date: 18.11.2019. [Online]. Available: http://www.medlingmap.org/ taxonomy/term/80.

[13] Apache cTAKES - clinical Text Analysis Knowledge Extraction System. Renewal date: 18.11.2019. [Online]. Available: http://ctakes.apache.org.

[14] A. Maedche and S. Staab, "The TEXT-TO-ONTO Ontology Learning Environment" (PDF). ResearchGate, July 2000.

[15] Dresden Ontology Generator for Directed Acyclic Graphs (DOG4DAG). Renewal date: 30.10.2019. [Online]. Available: http://www.biotec.tu-dresden.de/ research/schroeder/dog4dag.

Hidden Authentication of the User Based on Neural Network Analysis of the Dynamic Profile

Anastasiya Sivova, Alexey Vulfin, Konstantin Mironov and Anastasiya Kirillova
Department of Computer Engineering and Information Security Ufa State Aviation Technical University, K. Marks Str. 12, Ufa, Russia
sivova.ae@net.ugatu.su, vulfin.alexey@gmail.com, mironovconst@gmail.com, kirillova.andm@gmail.com

Keywords: Keyboard Handwriting, Hidden Authentication, Neural Network Classifier.

Abstract: The problem of continuous hidden user authentication based on the analysis of keyboard handwriting is considered. The main purpose of the analysis is to continuously verify the identity of the subject during his work on the keyboard. The aim of the work is to increase the efficiency of hidden user authentication algorithms based on a neural network analysis of a dynamic profile, formed by keyboard handwriting. The idea of user authentication using keyboard handwriting is based on measuring the time of keystrokes and the intervals between keystrokes, followed by comparing the resulting data set with the stored dynamic user profile. Studies have shown that analyzing the average value of the time each key is pressed is inefficient. It is proposed to analyze the holding time of a combination of several keys and the time between their presses. An approach in which not the times of pressing individual keys, but the parameters of pressing the most common letter combinations are analyzed, will increase the accuracy of recognition of dynamic images. An algorithm and software implementation for Russian keyboard layout have been developed, experiments conducted on field data allow us to conclude that the proposed method is effectively used to authenticate the user using keyboard handwriting.

1 INTRODUCTION

User authentication based on tracking his or her behavioral features is a relatively novel technique. Behavioral features are acquired, when the user is working with various manipulators: a computer mouse, keyboard, etc. The term “user information handwriting” (UIH) proposed means the style of the work with such manipulators [1]. UIH tracking allow us to define a unique pattern, which may be used as a mean of authentication or assessing the user’s state, level of computer literacy etc.

The aim of our work is to increase the efficiency of hidden user authentication based on a neural network analysis of a keyboard handwriting. The sub-tasks of our work are the following:

1) Development of an algorithm for analyzing the user's keyboard handwriting with use of neural network.
2) Development of a system for hidden authentication with use of above-mentioned algorithm.
3) System evaluation based on accumulated data.

Margins, column widths, line spacing, and type styles are built in. Some components, such as multileveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

2 BIOMETRIC AUTHENTICATION BASED ON KEYBOARD HANDWRITING

Keyboard handwriting is a dynamic biometric pattern including speed of typing, use of the main and additional parts of the keyboard, specific keystrokes and specific techniques and methods of working with the keyboard [2]. With the improvement of keyboard skills, individualism of the keyboard handwriting also grows [3]. The listed individual features are a part of the dynamic user profile and may be used to authenticate the user. The methods of continuous hidden keyboard monitoring

make it possible to detect the substitution of a legitimate operator and block the system from intruder intrusion. The probability of false recognition, when using 5-letter word, is about 10^{-33} [4].

Dynamic authentication systems are capable to make biometric patterns hidden. The attacker is not able to use the previously prepared dummy (which is possible, when using static patterns such as fingerprint). The main disadvantage of dynamic biometric systems is that their functioning is affected by the psychophysiological state of a person [5, 6, 7]. He or she may be worried or calm, tired or alert, healthy or sick, etc.

Analyzing a user's dynamic profile allow the system to verify continuously the identity of the subject and to control that this particular subject is working on the computer. The principle of continuous authentication consists measuring the durations of keystrokes and intervals between them. These data are compared to available pattern of the user. According to a large-scale study of this approach, conducted by The National Institute of Standards and Technology, USA [8] the probability of correct recognition for users with established keyboard skills is 98%, which is enough for practical applications.

Keyboard handwriting authentication system should include three modules [9, 10, 11]:

1) Keylogger, which tracks keystrokes.

2) Module for generating reference templates for the handwriting based on the data from the keylogger. This template is generated, while the user is working on the computer.

3) Authentication Decision Module, which analyzes the characteristics of the current user and compares them with a reference sample.

Let:

$n = 1\text{K}\ 1024$ is the number of all possible combinations of two keys in Russian alphabet (“а”, “б”, ..., “я”);

T_i – the experimental time interval between keystrokes of the n-th combination;

T_i^R – the reference time interval between keystrokes of the n-th combination.

Then the feature vector of the k-th user, generated from the average values of keystrokes of the i-th key, is determined as (1):

$$T_i^{R.avg} = \frac{1}{m}\sum_{j=1}^{1} T_{ij}^{\ R}\ , \qquad (1)$$

where m – the number of keystrokes of the i-th key. The vector may be expanded with the variance or mean square deviation for the i-th key.

The state-of-the-art algorithm for obtaining the vector of dynamic characteristics of the user consist of the following steps:

1) Generation of the reference feature vector for all K users: $V_k = \left\{ T_i^{R.avg} \mid i = \overline{1,n} \right\},\ k = \overline{1,K}.$
2) Formation of a feature vector based on user signature: $V_k = \left\{ T_i^{avg} \mid i = \overline{1,n} \right\}.$
3) Search for the most similar vector in the database (DB).

The disadvantage of the state-of-the-art is that it analyzes the average values of the retention time of each individual key and the time after it is pressed. If we consider the location of the keys on the keyboard, we can conclude that the time between keystrokes of adjacent keys will be significantly lower than the time of keystrokes located more remotely from each other. Therefore, the distribution of the collected parameters would not be normal. So, using the obtained values as a vector of biometric features is inefficient.

We propose to analyze holding time and the timeout between presses within the most widespread combinations of several keys. This may increase the accuracy of pattern recognition, since the user's actions in this case are automatic and the parameters would be normally distributed.

We propose to analyze the typing time of N-graphs: sequences of several keys pressed in a row. Analysis of digraphs, i.e. sequence of two keys pressed in a row, allow determination of three indicators: the holding time of two keys and the time between them. To classify the obtained values, it may not be enough, therefore, it is proposed to use N-graphs of higher dimension.

3 ALGORITHM FOR ANALYSIS OF KEYBOARD HANDWRITING

3.1 Biometric Features of the Keyboard Handwriting

The approach for identifying a subject based on continuous hidden authentication of computer system user in the process of working at a computer is proposed. As identification characteristics, the features of the user’s work with the keyboard are used – his keyboard handwriting, which is characterized by the key holding times (KHT) and

the times between keystrokes (TBK). These characteristics can be measured by a standard keyboard.

Retention time also depends on overlays. Overlapping keystrokes occurs when one key has not yet been released and the other is already being pressed. There is a tendency to increase the number of overlays with increasing dialing speed. The vast majority of overlays occur when the keys of adjacent letters in a word are pressed with different fingers. However, with very fast sliding dialing, overlays are also possible. Out of the total amount of text entered by the user during the working day, it is proposed to process not individual keystrokes, but the so-called N-graphs – trigraphs and tetraphs – sequences of three or four consecutive keys.

The controlled input parameters are the reference values of Key Holding Times (KHT) t_1, t_2, t_3,K ,t_n the for each key in the reference, as well as the time intervals between pressing adjacent keys (times between keystrokes, TBK) t_{12}, t_{23}, t_{34},K $t_{(n-1)n}$, i.e. exclusively time parameters, which may be measured by a standard keyboard.

KHT also depends on overlays. Overlapping keystrokes occurs when one key has not yet been released and the other is already being pressed. There is a tendency to increase the number of overlays with increasing dialing speed. Most overlays occur when the keys of adjacent letters in a word are pressed with different fingers. However, with very fast sliding dialing, overlays are also possible. In case of overlapping, the parameter $t_{(n-1)n}$ becomes negative. The controlled parameters t_n and $t_{(n-1)n}$ significantly depend on how many fingers are used during typing, as well as on user-specific combinations of typing movements.

An artificial temporary function, which represent the entire process of typing a phrase and include all the necessary information about the user's keyboard handwriting, is shown on Figure 1.

Let the user enter a phrase containing n characters over a period T from the keyboard. When this phrase is entered, $r = n + m$ keyboard events will occur: n key holdings and $n = m - 1$ pauses between holds. The temporary function at the time t_i will take the value $q(t_i)$, where q is scan-code i.e. key identifier on the keyboard. Overlapping is interpreted as a sum of two scan-codes of the pressed keys.

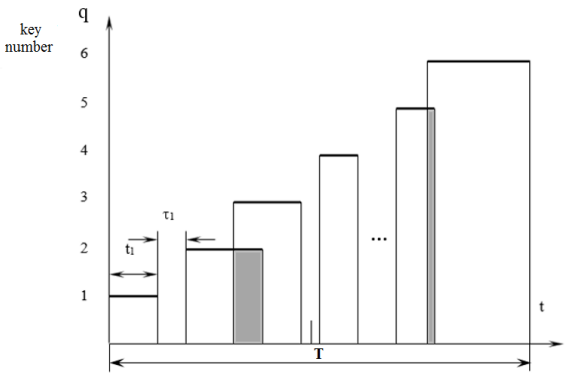


Figure 1: Time diagram of phrase typing.

The process of entering control phrase with $r = 11, m = 6, n = 5$, is illustrated by a time diagram (Figure 1). Temporary layout of the process is individual for each user and acts as a standard for keyboard handwriting.

As a feature vector of biometric (keyboard) features of V, we use the values of the function $q(\Delta t)$, where Δ t is determined as (2).

$$\Delta t = \frac{t_1 - t_0}{n} \tag{2}$$

The typing of certain N-graph for various users differs. Therefore, it is necessary to bring the feature vector to a single length. For this purpose, vector normalization is applied. Thus, the length of the vector of biometric features for each of the users will be equal to n – the number of samples equal to 32.

Development of a reference sample for one user require a series of L samples, which constitute a representative sampling for the s-class $\Psi^{(s)} = \{V_i\}, i = \overline{1, L}$.

In general, the system can include multiple users $K = \{k_1, k_2, \text{K}, k_M\}$. Each user is represented by its reference pattern and associated with a certain class from the set $s = \{s_1, s_2, \text{K}, s_M\}$. Thus, an unambiguous mapping from the set of users $\{K\}$ to the set of classes $\{s\}$ is developed. Development of the reference samples for M legitimate users require M training samplings respectively $\Psi^{(s_1)}, \Psi^{(s_2)}, \text{K}, \Psi^{(s_M)}$.

When the system is in the authentication mode, unknown user (x) presents a sample of keyboard handwriting as a vector of biometric parameters $V^{(s_x)} = \{v_j\}$, $j = \overline{1, N}$. The system should form a description of the unknown x-class on the basis of

the vector $V^{(s_x)}$, compare it with the standards of all users registered in the system $\{k_1, k_2, \text{K}, k_M\}$ and make the authentication decision based on the results.

In this formulation, the task is classifying the vector $V^{(s_x)}$ into $M+1$ exclusive classes: M classes from the set $s = \{s_1, s_2, \text{K}, s_M\}$, and $(M+1)$-th class reserved for all other users, united by the concept of "aliens". If there is a procedure for preliminary authorization of users, the task is simplified and reduces to the classification of the vector $V^{(s_x)}$ into two classes: s_o – "own", that is, belonging to any class from $\{s\}$, and s_a – "alien", that is, not belonging to any class from $\{s\}$.

3.2 Selection of Informative Values of Biometric Features

Only the most frequent N-graphs are processed by the algorithm. A frequency dictionary of trigraphs and tetragraphs is generated in order to choose the most frequent of them. These N-graphs are selected for analysis. In the authentication mode users are authenticated based on the analysis of these N-graphs.

3.3 Model of User Authentication

Authentication decision is made based on the difference between actual data and reference pattern. The input information consists of the values of KHT and TBK.

The minimum number of neurons in the hidden layer, which provide the solution to the interpolation task, is determined by the expression (3) [12]:

$$n_2 = \text{int}\left[\frac{m(R-1)}{n+m+1}\right], \qquad (3)$$

where n_2 – the number of neurons in the hidden layer;
n – the number of neurons in the input layer;
m – the number of neurons in the output layer;
R – the dimension of the training sampling. Operation int() means rounding up to an integer.

Substituting the values in the (3), we get:

$$n_2 = \text{int}\left[\frac{3(30-1)}{3+5+1}\right] = \text{int}(9.7) = 10$$

The classical neural network for biometrical authentication is shown in Figure 2.

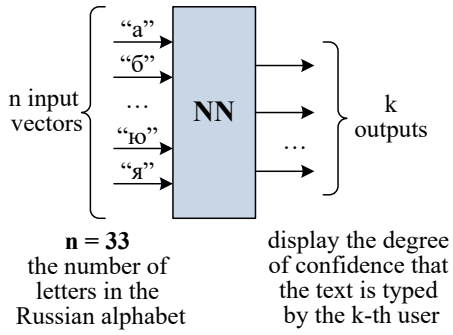


Figure 2: Classical neural network for biometrical authentication.

Average KHT are analyzed in this approach, which is inefficient. In the proposed system, the KHT and TBK of several consecutively pressed keys act as a vector of biometric features. In addition, we propose to use the modular structure of a neural network, in which each network make a decision for only one of the selected N-graphs. The modular approach allows us to divide the authentication task into subtasks, solve them individually with separate neural networks, and then combine the results.

Large neural network can suffer from interference, as new data can dramatically change existing connections. The modular approach makes it easy to scale the network, because adding or removing modules for a specific N-graphs is possible without retraining the entire network.

Depending on which feature vector is fed to the input of the neural network, it is proposed to use two approaches for the modular structure of the network.

First approach. A vector of user biometric features normalized to 32 samples is supplied to the input of the first neural network (let us call it "network of the first type"). The first network is responsible for recognizing the input N-graph. It activates the second network ("network of the second type"), which was trained on this image. The second network determines from which user the biometric feature vector was received, containing the number of representations of the recognizable N-graph. The output is the values characterizing the degree of confidence that the text was typed by each of k users. The final decision is made by the decision unit, analyzing the data received from the neural networks. Thus, the decision to authenticate the user will be made based on data received from several neural networks at once. The structural diagram of the described approach is shown in Figure 3.

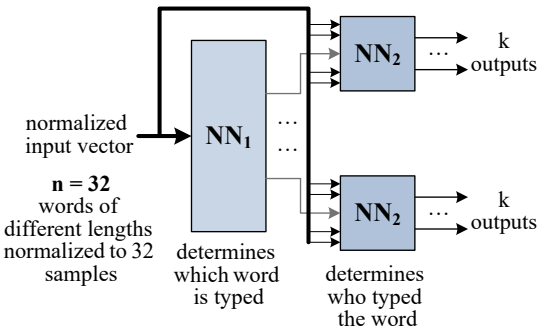


Figure 3: Diagram of the first approach.

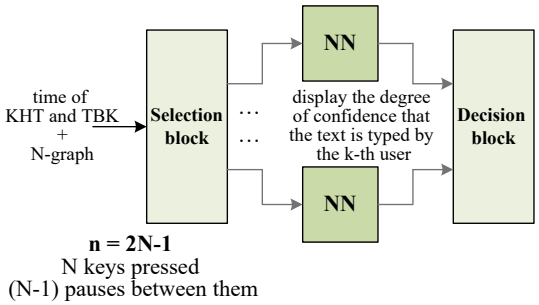


Figure 4: Structural diagram of the second approach.

Second approach. If each network of the second type is trained on only one N-graph, only the values of the KHT and TBK can be fed into the network input, without considering, which sequence of characters was typed.

Thus, it is possible to throughout the first network and use another classifier instead of it. The input of the classifier consist of KHT, TBK and the identifier of the N-graph. Based on these data, the classifier the neural network of the second type. The final decision is also made by the decision unit as in the first approach.

The structure of the second approach is shown in Figure 4.

Input vector is much smaller in the second approach than in the first one and consists of only 5 or 7 signs for trigraphs and tetraphs, respectively, compared to 32 in the first approach. This will allow the neural network to learn faster and with smaller samplings; moreover, the probability of getting into local minima with this approach is reduced.

4 SYSTEM FOR HIDDEN AUTHENTICATION

4.1 Algorithm of Hidden Authentication

The hidden authentication system consists of three modules [13, 14, 15]:

- Module for collecting the data;
- Module for data preprocessing and generating feature vector;
- Module for learning and recognition using neural network classifier.

The block diagram of the authentication system is shown in Figure 5.

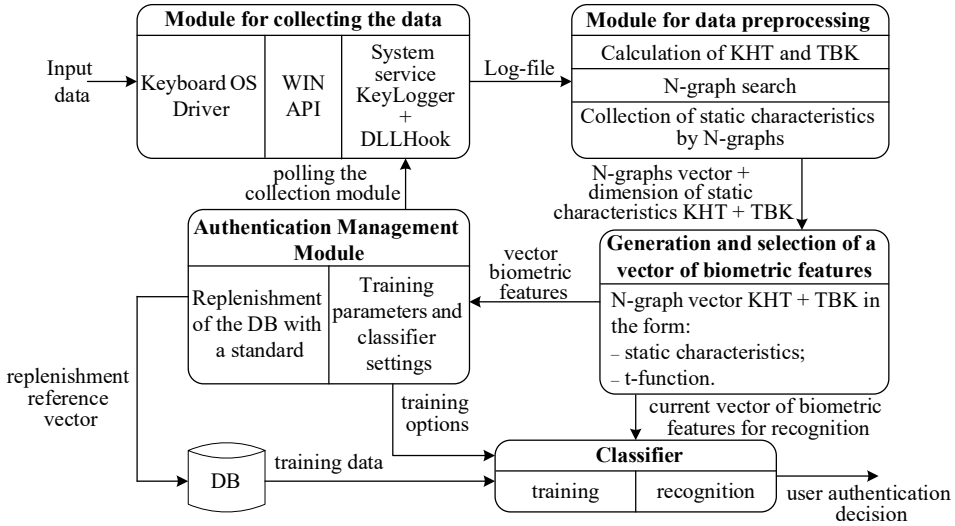


Figure 5: Structural diagram of the second approach.

Let us consider each of the modules more precisely.

Biometric authentication is based on the creation of reference representations of identifiable users. A reference is created when the system is in data collection mode. Registration of keyboard handwriting is carried out by the KeyLogger software module (keylogger).

The algorithm of the pre-processing and feature generation module is depicted as a flowchart in Figure 6. This module analyzes the data obtained using the keylogger. The logbook is analyzed line by line and a list of N-graphs is compiled, including the characteristics of KHT and TBK. The obtained values are used as an input vector of biometric features for the neural network.

The second and third modules are logically combined and executed sequentially. The resulting set of examples is divided into learning sampling and test sampling for cross-validation. Then the procedure of supervised training of the neural network is applied.

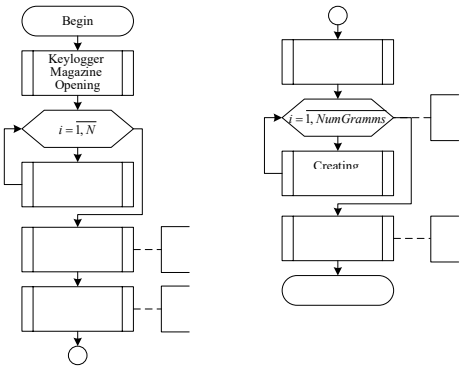


Figure 6: Algorithm for pre-processing and generating feature vector.

The general algorithm for obtaining feature vector with the subsequent provision of the obtained training sampling is presented as a flowchart in Figure 7.

4.2 Functioning of the System for Hidden Authentication

The system consists of several modules, which carry out their work invisibly for the user.

KeyLogger write a specialized log-file, which include timing of key pressing and two versions of key codes: scan code and virtual code used by the operating system to identify keys. Thus, each line in the log include the following data: scan code, key status, timing datum and virtual key code.

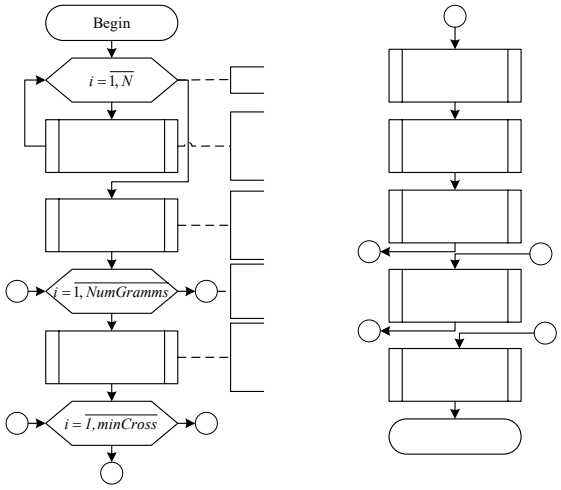


Figure 7: Algorithm for training and validation of the classifier.

The next module is used for creating biometric feature vector, which will be given to the input of the neural network. The program extracts the data from the keylogger log. They are parsed line by line, and all values are entered into an array of sample structures. For each keystroke, a search is made for the moment it is pressed, given that the first keystroke begins at time $t = 0$. The time of key releasing is added to the sample array, and the line that previously contained this parameter is deleted.

The next stage of the program is building an array of N-graphs. The values of the virtual key codes of the sample array are analyzed for this purpose. Starting with the first element in step 1, the values of N consecutive keys are entered into a new word array. Only N-graphs, which are found in the text more than 15 times and typed by all users, are selected All other N-graphs are deleted. For the subsequent analysis, only the values of the most frequently encountered N-graphs are left with the data on pressing and releasing each key that make up the N-graph itself.

Since the time of typing a phrase is different for all users, normalization of the vector and the time chart by the number of samples n equal to 32 is carried out. As a result, a normalized vector of biometric features is constructed.

4.3 Test Results for the Prototype of Hidden Authentication System

Jarque-Bera test [16] and Lilliefors test [17] allow us to validate the hypothesis that the variables analyzed in the classical approach do not obey the law of the normal probability distribution. The experiment

showed that analysis of the average timing for single key pressing is inefficient. The easiest graphical way to check the nature of the data distribution is to plot a histogram. If it has a bell-shaped symmetrical appearance, we can conclude that the analyzed variable has an approximately normal distribution. For example, Figure 8 shows the histograms of the distribution of the retention time for the keys “a” and “б”.

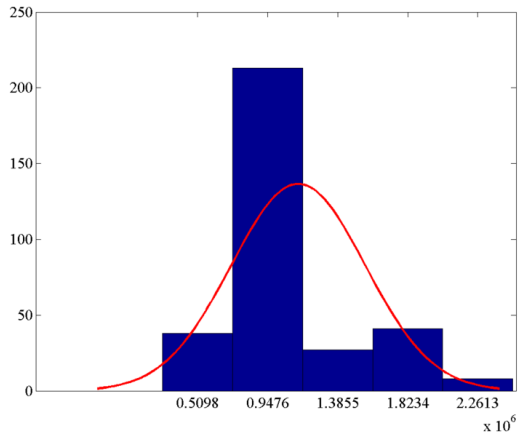


Figure 8: Histograms of the distribution of the retention time for the keys “a”.

Another graphical way to check the nature of the distribution is to build the so-called quantile plots (Q-Q plots, Quantile-Quantile plots). The quantile graphs for the distribution of the retention time of the “a” keys are shown in Figure 9.

The obtained graphs confirm the hypothesis that the average values for the retention time of individual keys do not obey the normal distribution are obtained.

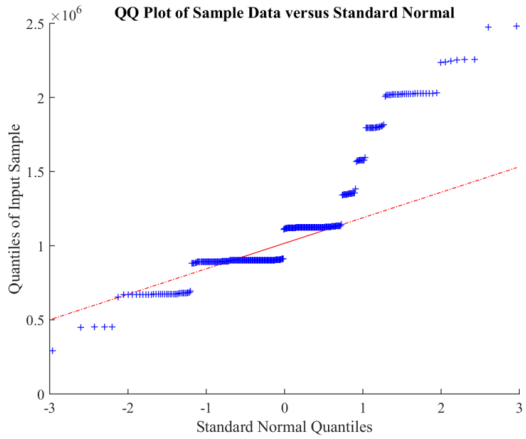


Figure 9: Graphs of the distribution quantiles for the retention time of the keys “a”.

Figure 10 show the distribution of the average time that the “a” key was pressed, depending on the phrase in which the letter was used.

In addition, the time between pressing two adjacent keys, depending on the typed combination, is also different, as shown in Figure 11.

As can be seen from Figures 10, 11, the average time of keystrokes in different combinations is different, as well as the time between holding the keys. Therefore, it was proposed to use KHT and TBK of the most key combinations.

During the experiment, User 1 used the “ст” combination 75 times in his work behind the keyboard. The ordered values of the key holding time “с” in the specified combination are shown in Figure 12.

In the same key combination, the ordered time values between the keystrokes “с” and “т” are shown in Figure 13.

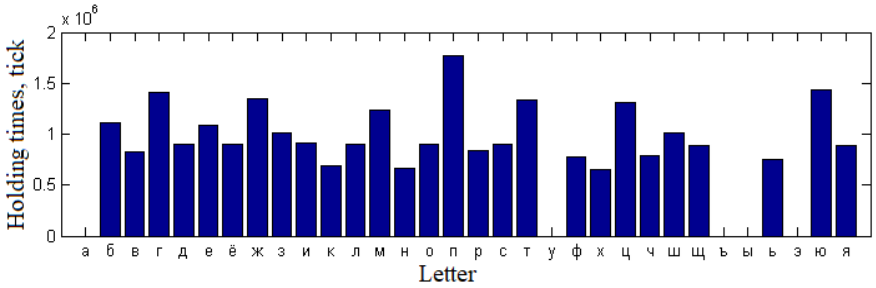


Figure 10: Average holding time of the “a” key for pairs “аа” ... “ая”.

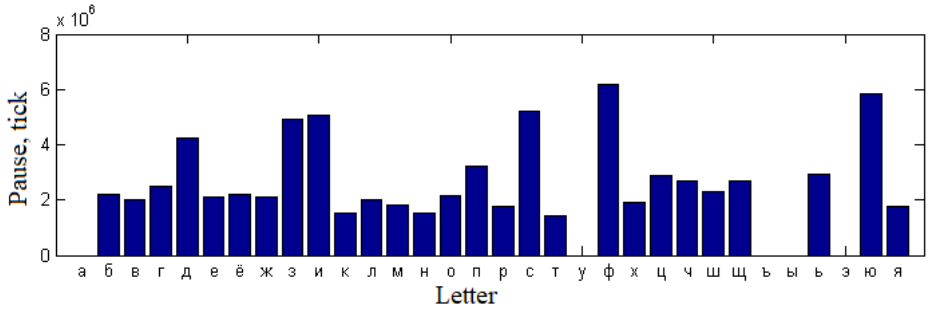


Figure 11: Average time between keystrokes for pairs “аа” ... “ая”.

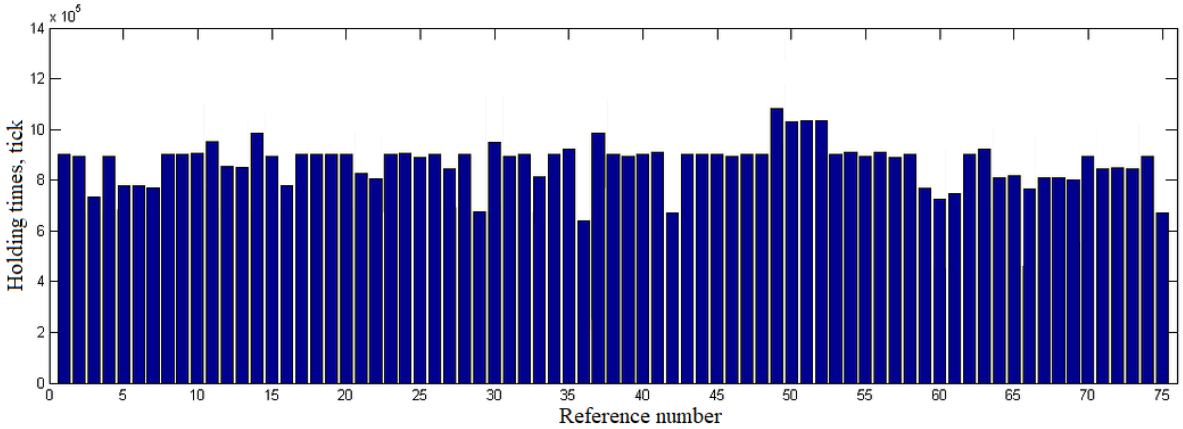


Figure 12: Ordered values of the key holding time “с” in the combination “ст”.

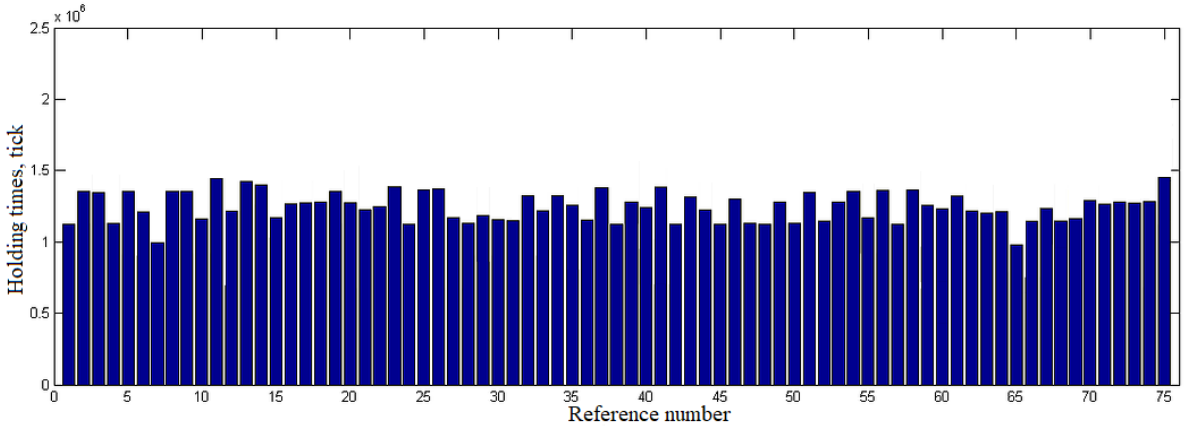


Figure 13: Ordered values of time between keystrokes “с” and “т” in the combination “ст”.

As can be seen from Figures 12, 13, the user types the same key combination in a similar way, therefore, the data on the typing time of corresponding N-graph can be used for training and testing a neural network.

Three users took part in the experiment. Trigraphs selected by frequency of occurrence, as well as a general list, are presented in Table 1.

The whole set of obtained vectors was divided into 10 subsets for 10-validation. The results are listed in Table 2.

Table 1: Selected trigraphs for various users.

User 1		User 2		User 3		Total information
N-graph	qty	N-graph	qty	N-graph	qty	
али	15	ани	17	або	16	ени
ана	16	ени	20	ани	28	льн
ель	23	ите	15	еле	23	ния
ени	24	льн	15	ени	19	нны
ите	15	мен	15	ефо	19	про
льн	24	ния	17	леф	19	
нал	15	нны	17	льн	19	
ния	18	при	15	ния	24	
нны	17	про	19	нно	22	ени
нов	16	чен	17	нны	24	льн
ные	17			ног	15	ния
ных	21			ной	29	нны
ова	17			ные	16	про
ого	16			ова	23	
ост	28			ого	22	
пол	20			онн	19	
при	17			оро	16	
про	19			ост	30	
				пол	18	
				про	24	
				ред	17	
				ров	17	
				ств	17	
				фон	19	

Table 2: Trigraph recognition percentage.

Trigraph		ени	льн	ния	нны	про	**Total**
Quantity		63	58	59	58	62	
Correct recog-nition (%)	Method 1	96.34	98.48	100	99.81	100	98.926
	Method 2	99.12	100	99.62	98.48	98.6	99.164

The table shows the values of the correct user recognition for each selected trigraph when using two methods. The results obtained using both methods are averaged and entered in the final column of the table. The results of the test in the first two passes during training on the N-graph "ния" are the following (Table 3):

Table 3: Inaccuracy matrix.

18	0	0
0	17	0
0	0	24
0	**0**	**0**
Sensitivity = 1 Specificity = 1 Correctness = 100%		

16	0	1
0	16	0
0	0	21
0	**0**	**0**
Sensitivity = 1 Specificity = 0.94 Correctness = 98.15%		

5 CONCLUSIONS

The following results were obtained:

- algorithm for transforming keyboard handwriting log into feature vector has been developed;
- algorithm for analyzing the user's keyboard handwriting based on neural network classifiers has been developed;
- modular structure of the neural network has been developed that correctly recognizes users in 99.164 % of cases;
- prototype system of hidden user authentication was developed.

Proposed system allows one to:

- authenticate the user according to the typed text (i.e. answer the following question: is it really that particular employee or someone else?);
- detect the substitution of the user in cases where an employee without access rights is trying to get the access through the computer of the qualified colleague;
- find the author of a specific text – which of the users in the company entered text on this PC in a suspicious period of time;
- identify the user in an atypical state and the specific period of time during which the user remained in this state;
- prevent the attempt of unauthorized access to the system in cases where the attacker managed to circumvent all previous lines of protection.

ACKNOWLEDGMENTS

The reported study was funded by RFBR according to the research project No. 20-08-00668 "Development and research of the methodology, models and methods of complex analysis and cybersecurity risk management of process control systems of industrial facilities using cognitive modeling technology and data mining".

REFERENCES

[1] M.N. Eshwarappa and M.V. Latte, "Multimodal biometric person authentication using speech, signature and handwriting features," International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence. 2011, pp. 77-86.

[2] T.V. Zhashkova, O.M. Sharunova and E.Sh. Isyanova, “Neural network identification of a person’s personality type by keyboard handwriting,” International Student Scientific Herald, 2015, no. 3.

[3] M. Cortopassi and E. Endejan, “Method and apparatus for using pressure information for improved computer controlled handwriting recognition, data entry and user authentication,” U.S. Patent, no. 6,707,942, 24 March 2004.

[4] S.M. Didenko, “Development and research of a computer model for the dynamics of the user-mouse system”, Tyumen, 2007.

[5] ГОСТ Р 54412-2011 – ISO/IEC/TR 24741:2007 “Information technology. Biometrics. Biometrics tutorial,” Standartinform, 2012.

[6] ГОСТ Р ISO/IEC 19794-2008 “Automatic identification. Biometric identification. Formats for the exchange of biometric data,” Standartinform, 2009.

[7] ГОСТ Р ISO/IEC 1978-4-2014 “Information technology. Biometrics. Biometric software interface,” Standartinform, 2016.

[8] A.V. Skubitsky, “Analysis of the applicability of the method of reconstructing dynamic systems in biometric identification systems by keyboard handwriting,” Informacionnye tehnologii, vol. 6, no. 1, 2008.

[9] R. Sharipov, M. Tumbinskaya and A. Abzalov, “Analysis of Users' Keyboard Handwriting based on Gaussian Reference Signals,” 2019 International Russian Automation Conference (RusAutoCon). IEEE, 2019, pp. 1-5.

[10] O. Vysotska and A. Davydenko, “Keystroke Pattern Authentication of Computer Systems Users as One of the Steps of Multifactor Authentication,” International Conference on Computer Science, Engineering and Education Applications. Springer, Cham, 2019, pp. 356-368.

[11] R. Chen, S. Kutten and E. Biham, “User authentication system and methods,” U.S. Patent no. 9,680,644, 13 June 2017.

[12] V.I. Vasiliev and B.G. Ilyasov, “Intelligent management systems. Theory and practice,” tutorial. M: M.: Radiotehnika, 2009, 392 p.

[13] Z.H.U. Yunzhou, and X. Jiang, “System and method for user authentication with exposed and hidden keys,” U.S. Patent no. 8,132,020, 6 March 2012.

[14] A. Schwartz and G.A. Woodward, “Composition and method for hidden identification,” U.S. Patent no. 4,767,205, 30 August 1988.

[15] N. Harun, W.L. Woo and S.S. Dlay, “Performance of keystroke biometrics authentication system using artificial neural network (ANN) and distance classifier method,” International Conference on Computer and Communication Engineering (ICCCE'10), IEEE, 2010, pp. 1-6.

[16] T. Thadewald and H. Büning, “Jarque–Bera test and its competitors for testing normality–a power comparison,” Journal of Applied Statistics, vol. 34, no. 1, 2007, pp. 87-105.

[17] N.M. Razali and Y.B. Wah, “Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests,” Journal of statistical modeling and analytics, vol. 2, no. 1, 2011, pp. 21-33.

Robotic System Position Control Algorithm Based on Target Object Recognition

Pavel Slivnitsin, Andrey Bachurin and Leonid Mylnikov
Electrical Engeneering Faculty, Perm National Research Polytechnic University, Komsomolsky ave. 29, Perm, Russia
slivnitsin.pavel@gmail.com, bachurin@ellips.ru, leonid.mylnikov@pstu.ru

Keywords: Pattern Recognition, Image Recognition, Object Detection, Neural Network, Robotic System, Identification, Self-Positioning.

Abstract: Creation of robotic systems capable of manipulating objects of the real world is an actual problem allowing both raising labor efficiency and reducing traumatism risk for a person. In the paper, the task of large-node replacement of the outdoor lighting luminaires with the use of robotic system is considered. For this, the accompanying tasks connected with the object (luminaire) identification and positioning of robot gripper concerning object have been solved. The resulting algorithms allow us to solve tasks in conditions of varying visibility, different backgrounds, overlapping objects, if necessary, positioning on certain parts of the target object. They can be used to identify of any elongated shape objects after appropriate training of the neural network. These algorithms allow us to position the robot arm in such a way that it can take the necessary object with the help of the gripper. The practical significance of the solved problem is connected with the possibility of robotics systems practical application in the human environment and the creation of anthropomorphic robots.

1 INTRODUCTION

The development of automation tools and the algorithmic base so far has allowed to automate the majority of monotonous and similar operations, and to replace people on such operations and to increase labor productivity. Modern tasks facing automated and robotic systems [1] are related to the issues of independent collection and processing of information about the surrounding environment and its use for decision-making. This is largely due to the fact that the use of robotic and intelligent devices has left from the factory floor to the outside, where the location of the objects is not fixed. In addition, there are many modifications of objects that perform similar functions and various spatial combinations of objects.

The development of production systems within the framework of the Industry 4.0 concept will lead to the creation of virtual and intelligent automatic production that will have to operate not only according to predetermined algorithms, but also in emergency situations.

One such task is the outdoor lighting maintenance[2], namely the replacement of defective lighting equipment (luminaires) using a robotic system. This task is a good example of identification and self-positioning relative to a given object (a luminaire and a connector), by a robotic system for the replacement of outdoor lighting luminaires using large-node replacement technology [3]. To implement this technology it is possible to use the special connector [4] (Figure 1). Implementation of this technology is necessary for the device to be able to remove the defective luminaire and install a new one. When using this method in a large lighting network, the operation of replacing outdoor lighting luminaires becomes the same for each lighting installation and can be automatically performed using a robotic system (Figure 2).

To identify objects, the industry now uses technology based on the determination of their coordinates (device coordinates) based on binding to global coordinates, positioning based on predefined maps, technology based on the use of radio tags to identify objects, etc. Such technologies show good results in production, where all changes in space configuration are clearly tracked. When working in the environment, such technologies become extremely expensive due to the impossibility to equip all devices with tags and keep the

environmental maps up to date. In such cases, image recognition technologies based on video and photo processing are used. Solutions for the recognition of some well-detectable objects (e.g. car license plates) have already been widely used. However, for self-positioning of technical devices, it is necessary to distinguish more complex objects under different conditions. The object recognition tasks are currently being solved as classification tasks, for which neural networks and machine learning methods have been used. Among neural networks, modifications of recurrent and convolutional neural networks (CNN) such as LSTM are widely used. Probabilistic neural networks (PNN) are used to identify objects with a variable shape and structure. SVM, decision trees, and Dalal-Triggs [5] and Viola-Jones methods [6] are used for individual images are the most frequently used among machine learning methods [7]. In some cases, methods of scale-invariant features transform (SIFT and SURF) are used. Preprocessing of images and models of multilayer perceptrons (MLP) using neural networks can be used to increase accuracy.

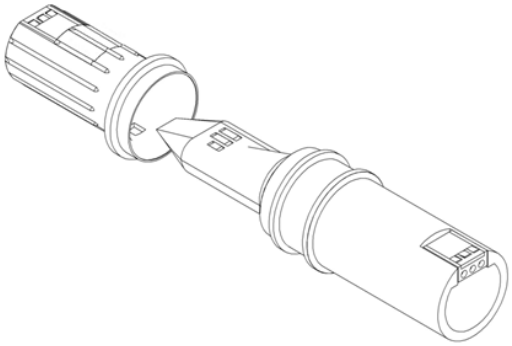


Figure 1: The connector scheme.

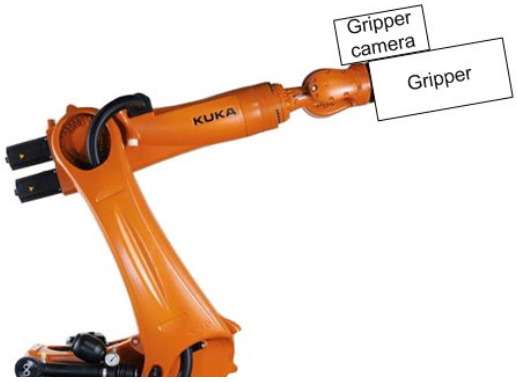


Figure 2: The structure of the robotic system.

The choice of a method to solve an applied task is associated with the speed of the identification and the accuracy of objects identification (Table 1), including the cases when there may be variability of shape and position, when the objects overlap, and when there are requirements to the speed of work and learning.

Table 1: Comparison of object recognition methods.

Method	Accuracy
SURF [8]	63%-90%
SIFT [8]	71%-91%, for some tasks 90%-99%
SVM [9]	77,69%-91,54%
Dalala-Triggs [5]	>90%
CNN [10]	90% - 97%

As we can see in the table, all modern methods for a certain class of tasks show good results. Therefore, the choice of the method will be based on the object peculiarities, the learning process and use of the object.

In the considered task, robotic system is an autonomous low performance device which operates in real time, that leads to high requirements to the speed of objects identification. Thus, training of model can occur independently, on a separate server or cluster, and the set of training sample is constantly replenished by the new images collected in the course of autonomous operation of the robotic system.

In such conditions, it is effective to use convolutional neural networks, which require the use of deep learning methods for their learning. The trained model can work effectively on low-power computing resources and gives high accuracy without using additional algorithms to increase accuracy [11].

2 METHODOLOGY

In outdoor lighting maintenance using a robotic system, the robot is delivered to the place of replacement using a lifting platform car. The car stops near the necessary support, and the robot identifies the luminaire. Then, it approaches the luminaire and the connector at a distance sufficient to perform the replacement procedure according to the algorithm shown in Figure 3.

The implementation of the above algorithm implies performing actions related to the object identification and robot positioning (rotation, shift and approaching operations).

Step 1: Target object = OUTDOOR LIGHTING LUMINAIRE.
Step 2: Photo / video shooting of the surrounding area.
Step 3: Target object identification.
Step 4: Perform a test move and determine the distance to the target object.
Step 5: Determine the Δ_1 step value for approach to the target object.
Step 6: Having approached the target object (step Δ_1), re-photograph the area and determine, the position relative to the target object, based on the analysis of two images.
Step 7: If the distance to the object is greater than Δ_2 then return to step 5, otherwise if the target object = OUTDOOR LIGHTING LUMINAIRE then the target object = LUMINAIRE CONNECTOR and go to step 5.
Step 8: Execution of the luminaire replacement algorithm.

Figure 3. The algorithm of the robotic system for the replacement of outdoor lighting luminaires.

2.1 Target Object Position Identification in the Image

To solve the task of object recognition, we will build a neural network on the basis of convolution neural network *Inception–v2* ([12] and [13]). The peculiarity of the network is in its structure, that uses complex structures representing modules as neurons (typical structures of neural networks): filtering and signal normalization modules, convolutional networks with the number of layers and branches from 2 to 5 and the number of neurons in each branch from 1 to 5, typical multilayer neural networks in which all neurons of the previous layer are connected to all neurons of the next layer with the number of layers from 1 to 5 and the number of neurons in each layer from 1 to 5. The use of such modules makes it possible to construct from them networks in which each layer is organized as a separate neural network with normalization of input and output signals, and the number of modules in each layer increases. To identify the luminaires, a network with 20 layers was used, 10 of which are Inception modules described in [13]. This structure in the learning process introduces specialization for constituent modules that begin to identify various special elements of an identifiable object. The structure of the selected neural network can be represented as shown in Table 2.

Table 2: Neural network architecture [13].

Layer type	Patch size/stride or remarks	Input size
Conv	3 × 3 / 2	299 × 299 × 3
Conv	3 × 3 / 1	149 × 149 × 32
Conv padded	3 × 3 / 1	147 × 147 × 32
Pool	3 × 3 / 2	147 × 147 × 64
Conv	3 × 3 / 1	73 × 73 × 64
Conv	3 × 3 / 2	71 × 71 × 80
Conv	3 × 3 / 1	35 × 35 × 192
3 × Inception module	As in figure 5 in [13]	35 × 35 × 228
5 × Inception module	As in figure 6 in [13]	17 × 17 × 768
2 × Inception module	As in figure 5 in [13]	8 × 8 × 1280
Pool	7 × 8	8 × 8 × 2048
Linear	Logits	1 × 1 × 2048
Softmax	Classifier	1 × 1 × 1000

The structure of the neural network affects the training process. Therefore, after selecting the structure of the neural network, it is necessary to optimize it. One of the main reasons why this is necessary is that when training the network, it is possible to move incorrect information from the end of the network to all the weights inside. In this case, if to change the parameters of the input layer is a simple task, then to get access to the parameters of the layers behind the first one is not a simple task. It is possible to write formulae for updating the weights within the network. However, due to the fact that each neuron depends on the other with which it has a connection, various new problems may appear, such as: 1) sticking at local minima; 2) infrequent updating of rare features, which negatively affects the capabilities of the generalizing network rule, but on the contrary, great emphasis on rare features leads to the network retraining.

When implementing a neural network, neurons with the ReLU activation function should be used to identify objects. Calculation of the sigmoid and hyperbolic tangent requires more resource-intensive operations, such as exponentiation, while the ReLU can be implemented using a simple threshold conversion of the activation matrix at zero. In addition, the ReLU is not saturated.

The use of ReLU increases the learning speed significantly. That is caused by linear character and absence of saturation of this function.

Special attention should be paid to the choice of optimizer. The speed of the learning mechanism and the stability of the result will depend on this

(comparison of the work of several optimizers is shown in Figure 4).

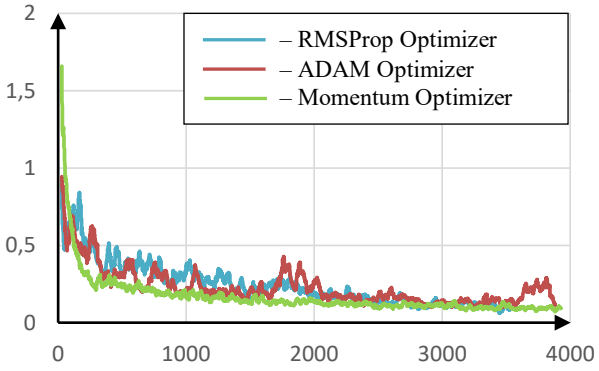


Figure 4: Graphs of changes in the value of the loss function when using various optimizers.

2.2 Rotation Operation

It is not enough to identify the object in the image, it is also necessary to define the robot position relative to the target object. First of all, the task of gripper rotation is solved. The solution of this task is to determine the angle of rotation for the correct orientation relative to the luminaire (target object) (Figure 5).

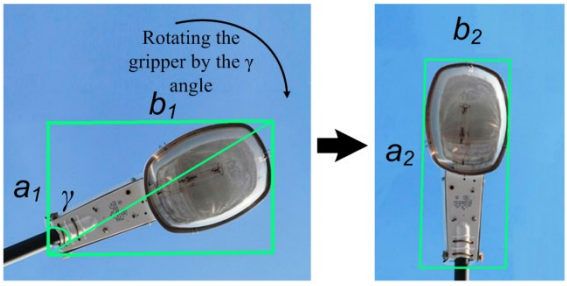


Figure 5: Determination of the rotation angle.

The solution of this task is performed using formulae (1) and (2).

$$\frac{b_1}{a_1} \neq \frac{b_2}{a_2}; \frac{b_2}{a_2} = \frac{B}{A}; \quad (1)$$

$$\gamma = \arctan\left(\frac{b_2}{a_2}\right); \quad (2)$$

where a_1, b_1, a_2, b_2 are the frame dimensions of a detected object (luminaire) in the image in pixels; A, B are the actual luminaire size characteristics in meters; γ is the manipulator gripper rotation angle.

2.3 Shift Operation

After the rotation operation and before approaching to the object, it is necessary center the object relative to the center of the image (Figure 6). This is necessary so that there is no gripper shift relative to the object when approaching it.

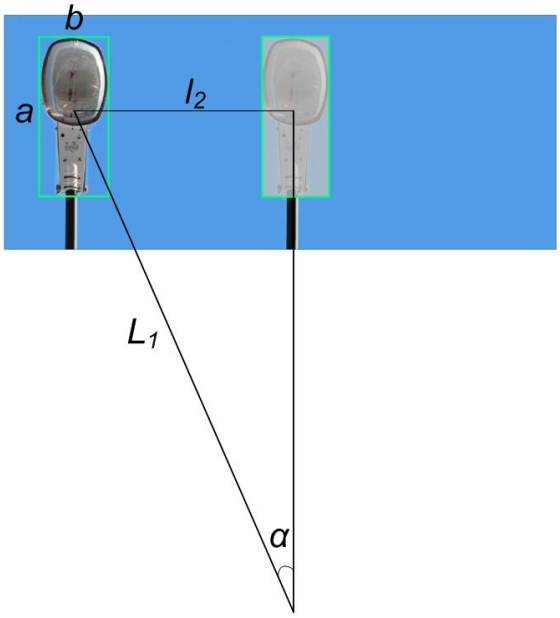


Figure 6: Determination of the angle of robot gripper.

To do this, it is necessary to determine the angle through which the robot should be rotated; this angle is determined by the Formula 3.

$$\alpha = \arcsin\left(\frac{L_2}{L_1}\right) = \arcsin\left(\frac{l_2 A}{L_1 a}\right); \quad (3)$$

where α is the angle through which the robot must be rotated; l_2 is the distance from the object to the center of the image in pixels; a, b are the frame dimensions of a detected object (luminaire) in the image in pixels; A, B are the actual luminaire size characteristics in meters; L_2 is the distance from the object to the center of the image converted into meters; L_1 is a measured distance from the gripper camera to the object in meters (approaching operation).

2.4 Approaching Operation

The next operation to be performed is to define the value of the next step to approach to the object. For this, it is necessary to define the distance to the object (Figure 6).

Considering that the object can be photographed at each step of the algorithm (Figure 3), it is possible to compare the last two images, namely, to compare the geometric dimensions of the detected object and to estimate the distance to the object. For this, it is necessary to solve the system of equations (4) (Figure 7).

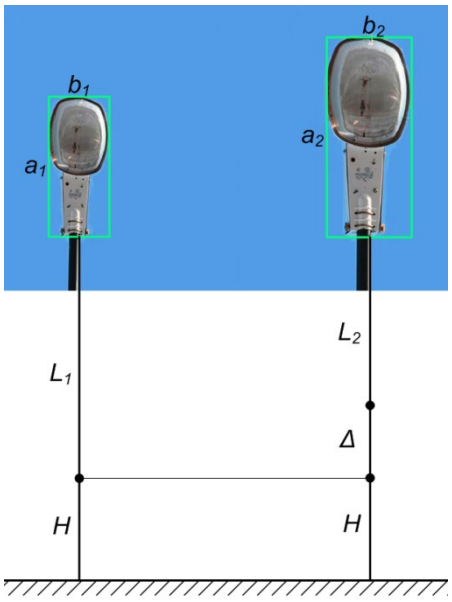


Figure 7: Determining the distance to the object.

$$\begin{cases} \frac{L_1}{L_2} = \frac{L_2 + \Delta}{L_2} = k\frac{a_2}{a_1} \\ \frac{L_1}{L_2} = \frac{L_2 + \Delta}{L_2} = k\frac{b_2}{b_1} \end{cases}; \quad (4)$$

where H is the camera height in meters; Δ is the distance travelled towards the object in meters; a_1, b_1, a_2, b_2 are the frame dimensions of the detected object (luminaire) in the image in pixels; k is the pixel-to-meter conversion factor (due to the use of two images, the values obtained from the system (4) are independent of the k value); L_1, L_2 are the required distances from the camera to objects in meters.

3 IMPLEMENTATION

Implementation of an algorithm to identify outdoor lighting luminaires requires neural network training.

To train a neural network, it is necessary to form a training sample of images with luminaires labelled on them. To label luminaires in the image, a rectangular area is selected, which coordinates are recorded in an XML file (Figure 8):

```
<annotation>
    <folder>train</folder>

 <filename>IMG_20191101_164622.jpg</fil
ename>

 <path>C:\tensorflow1.1\models\research
\object_detection\images\train\IMG_2019
1101_164622.jpg</path>
    <source>
        <database>Unknown</database>
    </source>
    <size>
        <width>195</width>
        <height>260</height>
        <depth>3</depth>
    </size>
    <segmented>0</segmented>
    <object>
        <name>luminaire</name>
        <pose>Unspecified</pose>
        <truncated>0</truncated>
        <difficult>0</difficult>
        <bndbox>
            <xmin>30</xmin>
            <ymin>44</ymin>
            <xmax>161</xmax>
            <ymax>144</ymax>
        </bndbox>
    </object>
</annotation>
```

To create labels, we use some tools (for e.g. *LabelImg*).

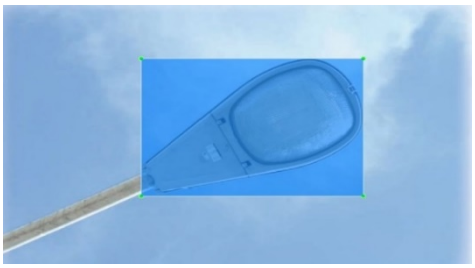


Figure 8: An example of labelling an object in an image.

After all images have been labelled, the coordinates of all selected areas are recorded to a ".csv" file, which contains the image names and coordinates of the selected areas in them.

For correct training of the neural network, it is necessary to use images of different types of luminaires from different angles, in different

conditions in which we can observe them (examples in Figure 9).

The initial dataset for training of the neural network consisted of 250 images.

To test the object identification in the image, the convolutional neural network, the *TensorFlow* library to work with neural networks and its add-on for object detection in images (*Object Detection API*) were used.

Having prepared a set of images for training, select the previously prepared model [14]. *TensorFlow Object Detection API* provides a number of pre-trained models. The use of a pre-trained model significantly reduces training time. The model *"faster_rcnn_inception_v2_coco"* is selected for testing.

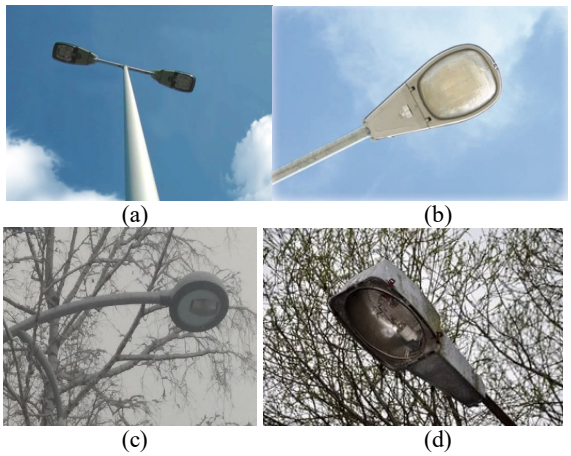


Figure 9: Examples of images from the training set.

As a result of the neural network training, rotation method implementation and approach-to-target method application, the algorithm shown in Figure 3 is implemented as a sequence of steps, an example of which is shown in Figure 10.

4 DISCUSSION

The task considered in this paper is an element of a complex solution consisting of two operations: the removing operation and installing operation of an outdoor lighting luminaire. Each of these operations consists of many subtasks, some of which can be invariant. For example, if we consider the luminaire removing operation, it can be decomposed into a number of algorithmic subtasks, as shown in Figure 11.

The solution of some tasks can be invariant. For example, the task of approaching to the target object in practice, using the proposed solution, can be performed in different ways. They will depend on the initial position, conditions affecting the quality of the images (for example, weather conditions and overlapping), and this does not take into account obstacles that may be in the way of the robot (for example, tree branches). Identification and grasping of objects by robotic systems are repeated many times. It means that the accumulated experience (photos and motion trajectories leading to successful results and failures) can be used for the training of the object detection model and improvement of the trajectory choice model based on machine learning methods or using intelligent algorithms.

Other tasks, which need to be solved for the luminaire removing operation, also have the property of invariance and the possibility of the algorithm modernization.

In this regard, at the moment, one can observe a variety of ways to solve these tasks, that combine attempts to create universal algorithms, the use of which will allow them to be used without relating to a particular object. One of the methods described in the literature when using neural networks for task recognition is the search for universal structures of such networks and use of Siamese networks [15].

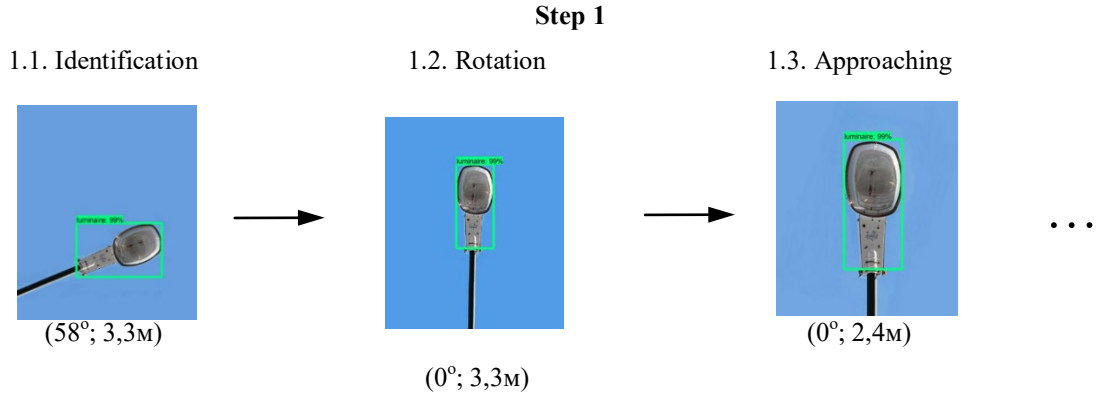


Figure 10: Identification and determination of the distance to the luminaire.

The subtasks shown in Figure 11 are not unique. They constantly arise in the control and creation of robotic systems. Therefore, today there is a great variety of task solutions to find an object and grasp it, approach the manipulator to the target object and the gripping point. In literature, there are algorithms using stereo vision [16]; identification of objects based on the choice of primitives that they consist of [17]; search for objects in the image and their classification using neural networks [18]; performing operations related to inserting one object into another (insertion of wires in the terminals, in our case installation of a new luminaires) [19]; tasks related to the movement of the gripper with an object that are considered as tasks of bending around obstacles on the basis of predetermined trajectories [20].

The presence of a large number of algorithms indicates the relevance of solving these algorithmic subtasks. The solution of tasks is associated with compromises between quality and complexity of the hardware, complexity of the algorithms and the speed of their work. Today, there are solutions that demonstrate good results in laboratory conditions. However, they do not always work well in real conditions due to the factors of the environment in which they operate.

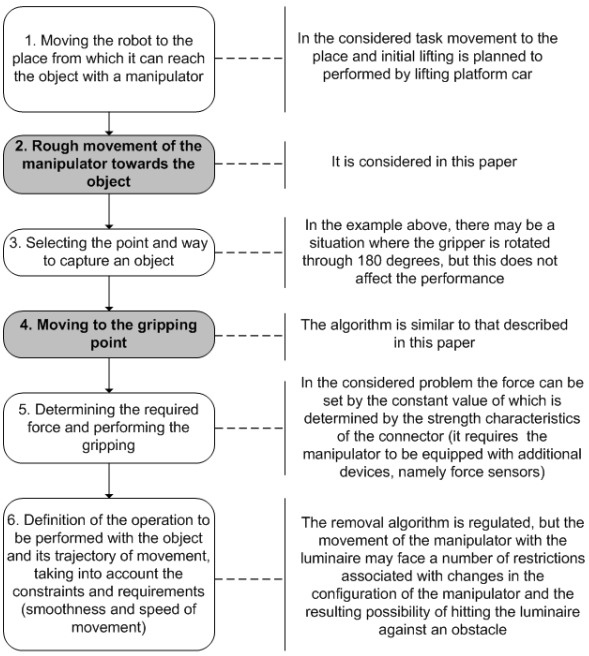


Figure 11: The scheme of dividing the luminaire removing task into algorithmic subtasks (the tasks considered in the paper are highlighted in gray).

5 CONCLUSIONS

The key task that was solved in the study described in the paper is the task of identifying objects. The use of neural networks is an effective tool that does not depend on a particular subject area.

The resulting model, trained on a dataset of 250 images, identifies the outdoor lighting luminaires in the images with high accuracy, but for some objects, much more images may be required to improve accuracy.

The approach considered in the paper relates to the subject area only through the object, the recognition of which takes place. In this regard, it can be considered universal and used to solve other applied tasks associated with the fact that an object must be taken and certain actions must be performed with it. Performing such manipulations is especially important when implementing humanoid robots capable of performing actions related to household and industrial activities. [21]. The creation of such robots is one of the priorities of robotics and is required to implement a complete Turing test [22].

REFERENCES

[1] A. Dudarev, “The Problem Sensitization Robotic Complex Drilling and Milling of Sandwich Shells of Polymer Composites,” Proc. 4Th Int. Conf. Appl. Innov., vol. 4, no. March, 2016, pp. 15-19.

[2] N. Pavlov, A. Bachurin, and E. Siemens, “Analysis of Outdoor Lighting Control Systems and Devices for the Creation of Outdoor Lighting Automatic Control System Using the Traffic Flow Value,” Proc. Int. Conf. Appl. Innov. IT, no. March, 2017, pp. 95-100.

[3] P. Slivnitsin and A. Bachurin, “A modern way of outdoor lighting maintenance,” International Conference on Innovation Energy, 2019, J. Phys.: Conf. Ser. 1415 012010.

[4] А.А. Бачурин и П.А. Сливницин, “Соединительное устройство для монтажа и подключения светильника наружного освещения,” 2695631, 2019.

[5] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” Proc. - 2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition, CVPR 2005, vol. I, no. 16, 2005, pp. 886-893.

[6] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," in Proceedings IEEE Conf. on Computer Vision and Pattern Recognition, pp. 511-518, 2001.

[7] L. Mylnikov, B. Krause, M. Kütz, K. Bade and I. Shmidt, Intelligent data analysis in the management of production systems (approaches and methods), Shaker Verlag {GmbH}.

[8] V. Renò et al., "A SIFT-based software system for the photo-identification of the Risso's dolphin," Ecol. Inform., vol. 50, pp. 95-101, January 2019.

[9] N. Kumar et al., "Leafsnap: A computer vision system for automatic plant species identification," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7573 LNCS, no. PART 2, 2012, pp. 502-516.

[10] H. Zhou, C. Yan and H. Huang, "Tree species identification based on convolutional neural networks," in Proceedings - 2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics, IHMSC 2016, vol. 2, 2016, pp. 103-106.

[11] F. Al-Azzo, A. M. Taqi and M. Milanova, "Human related-health actions detection using Android Camera based on TensorFlow Object Detection API," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 10, 2018, pp. 9-23.

[12] C. Szegedy et al., "Going deeper with convolutions," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 07-12-June, pp. 1-9, June 2015.

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-Decem, pp. 2818-2826, December 2016.

[14] "Tensorflow detection model zoo." [Online]. Available: https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/detection_model_zoo.md.

[15] S. Ravichandiran, Hands-On Meta Learning with Python: Meta learning using one-shot learning, MAML, Reptile, and Meta-SGD with TensorFlow. Packt Publishing Ltd.

[16] F. Chen, M. Selvaggio and D.G. Caldwell, "Dexterous Grasping by Manipulability Selection for Mobile Manipulator with Visual Guidance," IEEE Trans. Ind. Informatics, vol. 15, no. 2, 2019, pp. 1202-1210.

[17] H. Dong, E. Asadi, G. Sun, D. K. Prasad and I.M. Chen, "Real-Time Robotic Manipulation of Cylindrical Objects in Dynamic Scenarios Through Elliptic Shape Primitives," IEEE Trans. Robot., vol. 35, no. 1, 2019, pp. 95-113.

[18] C.H. Corbato, M. Bharatheesha, J. Van Egmond, J. Ju and M. Wisse, "Integrating different levels of automation: Lessons from winning the amazon robotics challenge 2016," IEEE Trans. Ind. Informatics, vol. 14, no. 11, 2018, pp. 4916-4926.

[19] D. De Gregorio, R. Zanella, G. Palli, S. Pirozzi and C. Melchiorri, "Integration of robotic vision and tactile sensing for wire-terminal insertion tasks," IEEE Trans. Autom. Sci. Eng., vol. 16, no. 2, 2019, pp. 585-598.

[20] Z. Cao, N. Gu, J. Jiao, S. Nahavandi, C. Zhou and M. Tan, "A Novel Geometric Transportation Approach for Multiple Mobile Manipulators in Unknown Environments," IEEE Syst. J., vol. 12, no. 2, 2018, pp. 1447-1455.

[21] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," Science (80), vol. 364, no. 6446, p. eaat8414, June 2019.

[22] A.M. Turing, "Computing machinery and intelligence," Mind, vol. LIX, no. 236, 1950, pp. 433-460.

Investigation of Brillouin Reflectometry Method Application for Mechanical Stresses Diagnostics in Optical Fiber

Igor Bogachkov[1], Nikolay Gorlov[2], Tatiana Monastyrskaya[2] and Evgenia Kitova[3]
[1]*Department of Communications and Information Security, Mira Str. 11, Omsk State Technical University, Omsk, Russia*
[2]*Department of Communication Lines, Department of Sociology, Siberian State University of Telecommunications and Computer Science, Kirova Str. 86 , Novosibirsk, Russia*
[3]*Department of Foreign Languages, Novosibirsk State Technical University, Prospekt K. Marksa 20, Novosibirsk, Russia*
bogachkov@mail.ru, gorlovnik@yandex.ru, t.monastyrskaya@mail.ru, kitovaet@mail.ru

Keywords: Optic Fiber, Communication Line, Tension, Brillouin Reflectometry Method, Degradation, Scattering.

Abstract: The paper strength characteristics and data based on the theory of optic fiber strength. The possibility of application of Brillouin Reflectometry Method for mechanical stresses control in optical fiber is studied. A special attention is given to the analysis of displacement of spectral components in Brillouin scattering of light depending on tension value or fiber elongation. The proposed method makes it possible do detect potentially dangerous sections in optical cable during all production stages and to improve production technologies, used in the cable manufacturing process as well as in the process of building and technical operational support of optical fiber communication lines. Obviously, it is the only optical method which makes it possible to measure the absolute tension value in the optic fiber. In addition, by use of Brillouin scattering spectral allocation along the fiber it is possible to determine the allocation of tension along the fiber line. Determination of the tension value in cable optic fiber allows predicting the reliability of the line as well as timely detecting the risk sections in the communication line.

1 INTRODUCTION

To ensure the long-term operation of the physical channel in optical access networks, it is necessary to guarantee the absence of mechanical tension in the optical fiber (OF). Even a slight tension of an optical fiber can lead to a significant decrease in its service life; therefore, to evaluate the reliability, information is required about the tension of an optic fiber in a cable.

Three groups of linear deformations of optic fiber can be distinguished: safe – up to 0.3%, dangerous from 0.3% to 0.6% and unacceptable – more than 0.6%. Therefore, it is necessary to create a measuring basis to control the deformations of the optical fiber in order to ensure the reliable operation of the fiber-optic communication lines. The most urgent tasks are to study the deformation fields (mechanical stresses) and temperature [1]. The principles of optical fiber deformations measurement are based on a variety of physical effects. However, the greatest practical interest can be given to methods based on the use of Mandelshtam – Brillouin scattering (MBS – acoustic phonons scattering). The Brillouin reflectometry method has two main advantages.

Firstly, it is practically the only optical method that allows to measure the absolute tension of the fiber. To do this, it is necessary to measure only the frequency of the maximum signal in the MBS spectrum and there is no need to undertake additional stretching of the fiber. In other well-known optical methods, the magnitude of the fiber elongation that occurs under additional tension in the fiber is measured, which makes these methods unsuitable for determining the tension of the fiber laid in the transmission line.

Secondly, MBS leads to the formation of a back wave in the fiber. Therefore, by probing the fiber with short pulses and scanning the frequency of these pulses, it is possible to find the distribution along the fiber of the MBS spectrum and, accordingly, the frequency of the maximum signal in this spectrum. And since this frequency is proportional to the tension in the fiber, this shows how this tension along the fiber is propagated [2]. It is known that glass starts breaking from the surface. The degree of its destruction is determined by shell defects. At the same time, the majority of measuring

instruments for fiber-optic communication lines are based on measurements of the parameters of optical radiation propagating through a fiber at a wavelength of an optical signal. More than 90% of the optical radiation power of this signal is concentrated in the core of the fiber [3]. Obviously, only the state of the core is controlled. To identify the defects of the shell and, accordingly, to assess the state of the optic fibers, instruments are used based on measurements of the Stimulated MBS parameters (SBS). It is known that SBS is scattering by acoustic phonons and its parameters such as capacity and frequency depend on the internal mechanical stresses in the fiber, and, consequently, on the degree of destruction of the shell [4 – 6].

2 METHODS

2.1 Measurements in Deformations

Various monitoring techniques are currently applied to control fiber condition. These techniques could be classified into several groups:

- Single wavelength Optical Time Domain Reflectometer (OTDR), including upstream OTDR measurements, active bypass, semi-passive bypass, reference reflector, switchable reflective element;
- Tunable OTDR, including wavelength routing and reference reflector;
- Brillouin OTDR (BOTDR);
- Embedded OTDR;
- Optical frequency domain reflectometer (OFDR);
- Optical coding;
- Self-injection locked reflective semiconductor optical amplifier (SL-RSOA);
- Reflective signal.

A detailed description of the monitoring techniques mentioned is given in the work [7]. The most efficient techniques to measure and evaluate deformations in OF are Brillouin reflectometry methods [4 – 6].

In recent few years, the Brillouin reflectometer has proven itself to be a very informative tool capable of detecting sections with loaded fiber in the laid communication lines during their operation. Analysis of the data obtained by using the Brillouin reflectometer allows operating organizations to determine with great accuracy the location of the cable line section with highly loaded fibers, estimate the level of their stresses and evaluate the reliability of the cable line. It is quite natural to use the device at the factory – manufacturer of optical cables during the tests, both in the process of developing the design of new products, and when conducting typical and periodic tests.

Regardless of the type and configuration, the cable is gradually stretched, controlling its stretching and tension. Essential differences in test methods are considered only when choosing the method for recording fiber elongation. The first method consists in measuring the attenuation gain of a fiber in a cable, and it is assumed that the fiber tension is accompanied by such an increase. This method of determining the beginning of the tension of the fiber is very inaccurate and depends on the design of the cable and the device used to stretch it. For example, when testing a cable with a central tube, it is not clear why a damping increase should occur, if you do not take into account edge effects. For a longitudinally stretched fiber that does not touch the cable walls, there are no direct mechanisms for generating attenuation gain. The second method, the phase shift method, or the method for detecting the propagation time of light pulses in a stretchable fiber, practically has no drawbacks. At the moment when the fiber begins to stretch in a stretchable cable, the optical path length for light pulses begins to grow, their propagation time increases, and the device registers it. But in a real situation, as the dimensions of the setup are limited, the device detects the accumulated effect along the length of both in the stretched fiber and in the transient region, where the tension smoothly increases. Moreover, in case when several fibers of the cable under test are welded into the cable, there is an additional uncertainty associated with the possible spread of excess lengths of different fibers. It is clear that the level of fiber elongation detected by this method is average with an unpredictable level of error.

2.2 Brillouin Reflectometry Method

The Brillouin reflectometer has at least one indisputable advantage – it makes it possible to measure the distribution of fiber tension level along the length. This removes all the uncertainties mentioned above. The result of measurements carried out by the Brillouin reflectometer is a well-localized distribution of tension, which makes it possible to isolate and take into account the edge effects and the variation of the tension in different fibers in case of their welding into a stub. MBS is the scattering of optical radiation by condensed media (solids and liquids) as a result of its interaction with own elastic oscillations of these

media. It is accompanied by a change in the set of frequencies (wavelengths) characterizing the radiation – its spectral composition. In the spectrum of the backward wave in the fiber, in addition to the unbiased component due to Rayleigh light scattering, there are also spectral components caused by Brillouin (MBS) and Raman light scattering (RS) on Figure 1 [6, 8].

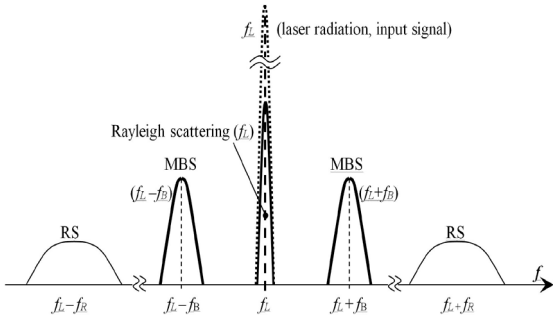


Figure 1: Spectrum of light scatterings in the fiber ($f_B \sim 10 \ldots 11$ GHz, $f_R \sim 13$ THz).

The occurrence of these components can be explained as follows. In the case of Rayleigh scattering, the light is scattered on the fluctuations of the refractive index frozen in the fiber, and therefore the frequency of the scattered light does not change.

In MBS and Raman scattering, the frequency of scattered light changes, since scattering occurs on time-varying fluctuations of the refractive index (caused, respectively, by thermal fluctuations of the medium density and intramolecular vibrations) [6, 9].

When the power of light pulses is less than 25 ... 30 dBm, Rayleigh scattering makes the main contribution to the power of the reverse wave. For comparison, the spontaneous MBS (SPBS) coefficient $\alpha_B \cong 0.03/\lambda^4$ is approximately 14 dB less than the Rayleigh scattering coefficient $\alpha_0 \cong 0.75/\lambda^4$, where λ is the radiation wavelength in microns [1]. When the threshold value is reached (~ 5 dBm with continuous pumping), the dependence of the reflected power on the pumping power becomes nonlinear. At the threshold increase, the result from SBS scattering becomes comparable with Rayleigh scattering. With an increase in the pump power several times, almost all the power is reflected from the fiber [2 – 4].

Weaker spectral components due to SPBS can be distinguished using an optical filter, since they are different in frequency.

2.3 Determination of the Fiber Tension

The threshold power can be increased by reducing the effective length of the interaction of the light wave with the acoustic wave. For a single pulse, the effective length is equal to half the pulse length [8]:

$$L_{eff} = L_p = \frac{c \cdot \tau}{2 \cdot n}, \quad (1)$$

where τ is the pulse duration;

c is the speed of light in vacuum;

n is the group index of refraction of the fiber ($n = 1.5$).

For a typical value of τ = 1 μs, we obtain $L_{eff} = L_p$ = 0.1 km, which is approximately two orders less than the magnitude of the effective interaction length (L_{eff} = 20 km) for a narrow-band radiation source.

The spectral components due to MBS have such an important property for practical applications as their frequency which is shifted by an amount proportional to the tension (relative elongation ε) of the fiber. For a standard single-mode fiber (SMF – G.652), the measured value of the coefficient is [8 – 10]:

$$K = (f_B - f_{B0})/\varepsilon, \quad (2)$$

where f_B is the frequency shift when registering the fiber tension (Brillouin frequency shift);

f_{B0} is frequency shift in the absence of fiber tension (strain);

ε is the relative elongation of optic fiber.

Thermal fluctuations in the medium density can be considered as a combination of elastic waves propagating in a medium in all possible directions and possessing all sorts of frequencies. Each flat sound wave is similar to a diffraction grating, since in places with the increased density, the refractive index of the medium is greater than in discharged locations.

For a light wave of length λ, there will always be a grating with a suitable period providing the maximum reflection of light in the opposite direction. The length of the corresponding sound wave is determined by the Bragg – Wulf condition: $\Lambda = \lambda/2$. In the fiber, MBS is observed only backward (the frequency shift between the pump and the wave scattered in the forward direction is zero). The wave reflected from such a moving diffraction grating, due to the Doppler effect, will be shifted in frequency by the value [9]:

$$f_B = \frac{2 \cdot n \cdot v_A}{\lambda}, \quad (3)$$

where $v_A \approx 5.7$ km/s is the speed of the acoustic wave in the fiber core [6].

At the wavelength λ = 1550 nm, the frequency shift (f_B) is 10.8 … 10.9 GHz [6, 11].

The fiber tension affects the sound speed v_A and the refractive index n. In turn, the formula for the speed of sound in the OF has the form:

$$v_A = 1.05\sqrt{\frac{E_Y}{\rho}}, \quad (4)$$

where E_Y – Young's modulus, when stretched.

E_Y is equal to 72 GPa = $7.2 \cdot 10^{10}$ N/m^2 for quartz glass; ρ = 2201 kg/m^3 is the density of quartz glass [6, 9].

The most contribution to the change in the frequency of the scattered light results from a change in the Young's modulus [9 – 11].

3 RESULTS AND DISCUSSION OF THE EXPERIMENTAL RESEARCH

BOTDR-graphs of the "bad" (potentially dangerous) fiber place located in the laid optical cable (OC), some section of which was under the influence of a strong displacement force is shown on Figure 2 [12].

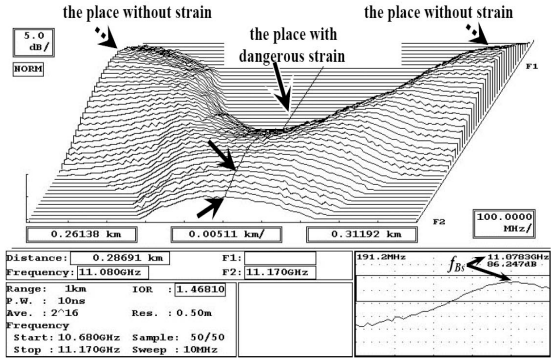


Figure 2: MBS graphs in "bad" OF section.

The significant shift of the maximum of MBS-spectrum (from 10.84 GHz (f_{B0}) up to the frequency 11.08 GHz (f_{Bs})) is observed in the place with dangerous strain [12].

Figure 3 shows the BOTDR multi-traces corresponding to MBS graphs presented in Fig. 2. It can be seen from the multi-traces that the strain in the place of mechanical action applied to the OC increased by more than 0.45 %, which is dangerous for the OF [12].

The analysis of BOTDR-traces for OF in OC under test showed that the sections of the route were localized which required a further investigation of factors that caused dangerous mechanical loads on the OC. After eliminating these factors and restoring the "bad" OF sections, the fiber optical communication line returned to its normal operation.

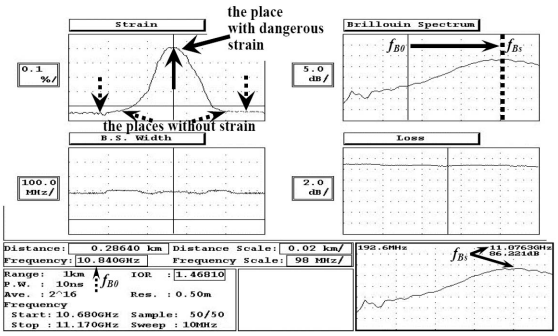


Figure 3: Multi-trace graphs in "bad" OF section.

In addition to fiber diagnostics in optic communication cables, the proposed Brillouin reflectometry method could be applied in large variety of fields, which involve mechanical deformation measurements and temperature measurements. For example, in avionics and auto electronics under conditions of low (up to –70°C) and high (up to +150°C) temperatures in electromagnetic interferences, as well as in power industry to test power electric cable under pressure, bending and stretching conditions. In oil and gas industry the proposed technique could be used to control temperature in a well up to 4 km deep with stretching load up to 520 kg. The technique could be applied in waste nuclear fuel storage.

4 CONCLUSIONS

It is recommended to include BOTDR in the control system to monitor the characteristics of the OF to detect locations of optical cable with increased mechanical stress and temperature changes. This will allow identifying potentially hazardous areas in the cable at different stages of manufacturing and improving the technologies used in the production process. In the process of optic fiber and cable production, it is easy to access both ends of the OF, which makes it possible to use phase methods or BOTDA. During the construction and operation of the cable, access is possible only to one end of the OF, and this makes it possible to use only reflectometric methods (BOTDR). Measurement of the tension value in the optical fiber of the cable will make it possible to predict the operating time before failure and timely identify areas of risk [6].

Information-measuring systems using in their operation MBS effect are very reliable in extreme operation conditions which result in their great demand in the market.

REFERENCES

[1] T. Horiguchi, T. Kurashima and M. Tateda, "Nondestructive measurement of optical-fiber tensile strain distribution based on Brillouin spectroscopy", Trans. IEICE Japan, vol. J73-B-1, no. 2, 1990, pp. 144-152.

[2] D. Iida, N. Honda, H. Izumita and F. Ito, "Design of identification fibers with individually assigned Brillouin frequency shifts for monitoring passive optical networks", J. Lightwave Technology, vol. 25, no. 5, pp. 1290-1297, May 2007.

[3] D. Iida, N. Honda, H. Izumita and F. Ito, "Detection sensitivity of Brillouin sensors located near Fresnel reflection," in Proc. OFC2007, 2007, vol. OMQ2.

[4] A. Kobyakov, M. Sauer and D. Chowdhury, "Stimulated Brillouin scattering in optical fibers", Advances in Optics and Photonics, vol. 2 (1), 2010, pp. 1-59.

[5] J. Fang, M. Sun, D. Che, M. Myers, F. Bao, C. Prohasky and W. Shieh, "Complex Brillouin optical time-domain analysis", J. Lightwave Technology, vol. 36, no. 10, 2018, pp. 1840-1850.

[6] I.V. Bogachkov, A.I. Trukhina and N.I. Gorlov, "Research of the features of Mandelstam – Brillouin backscattering in optical fibers of various types, International Siberian Conference on Control and Communications (SIBCON–2019), Tomsk, 2019, pp. 1-7.

[7] M.A. Esmail and H.A. Fathallah, "Physical Layer Monitoring Techniques for TDM-Passive Optical Networks: A Survey", IEEE Communications Surveys & Tutorials, vol. 15, no. 2, 2013, pp. 943-958.

[8] G.P. Agraval, "Nonlinear Fiber Optics", Elsevier, 2007.

[9] J. Smith, A. Brown, M. DeMerchant and X. Bao, "Pulsewidth dependence of the Brillouin loss spectrum," Opt. Comm., vol. 168, 1999, pp. 393-398.

[10] A. Yariv, Photonics: Optical Electronics in Modern Communications, ed. Oxford, U.K.: Oxford Univ. Press, 2005.

[11] X. Bao and L. Chen, "Recent Progress in Brillouin Scattering Based Fiber Sensors", Sensors, 2011, vol. 11, pp. 4152-4187.

[12] I.V. Bogachkov, "The detection of pre-crash sections of the optical fibers using the Brillouin reflectometry method", Journal of Physics: Conference Series, vol. 1210, 2019, pp. 1-11.

Development of the Oil Well Electrotechnical Complex Model in LabVIEW: Application Work Package

Anton Petrochenkov, Alexander Romodin, Sergey Mishurinskikh and Pavel Speshilov
Perm National Research Polytechnic University, Komsomolsky avenue 29, Perm, Russia
{pab, romodin, msv, spm}@ msa.pstu.ru

Keywords: Electrical Submersible Pump, Electrotechnical Complex, Oil Producing, Modelling.

Abstract: When planning the technological modes of an oil and gas producing enterprises, an important task is the energy consumption assessment of the electrical complex equipment. Most of the electrical and mechanical equipment operates in non-nominal modes, respectively, this equipment has non-nominal operating parameters. In this paper, it is proposed a method for building the electrotechnical complex models of an oil well in LabVIEW environment using the principles of object-oriented programming. The methodology is tested on the example of two wells. The results evaluation shows that the proposed approach gives results which is permitted for rapid assessment.

1 INTRODUCTION

Considering application work package is aimed to development of the scientific-methodical complex for modeling systems and power system optimization, support of power complex operational reliability; development of off-optimum situation training systems for personnel training and staff certification.

It is an universal software, allowing to check in complex the possibility of application of technical decisions connected with the circuit formation at the stages of variants' comparison during the process of design and within the changing conditions of exploitation of industrial enterprises' electric supply systems by means of calculation methods. The program is intended for investigation of the enterprise's electric supply system and as a result, for increase of its operating reliability and also for running on a schedule of the electrical equipment's scheduled-preventive works.

The program-technical complex is based on the unified analytical environments and is a component of the technological process automatic control system and communicates with the higher hierarchy levels of the automated control and electric power calculation system and anti-damage automatics.

Deliverables of this package are:

- Training and Simulation Software.
- User's Manual.
- HelpDesk.

2 STRUCTURE OF ELECTROTECHNICAL COMPLEX

Let's consider the structure of electrotechnical complex regarding to oil prodicing enterprises. The oil production process is a complex technological process, which interact various technological subsystems. In a mechanized method of oil production using electric drive centrifugal pumps, an electromechanical energy converter is a submersible induction motor (SIM). In addition, the oil well electrotechnical complex (ETC) includes: cable line (CL), transformer (T) and control station (CS) [1].

Modeling of power consumption in branched power supply systems (PSS) is performed, as a rule, by matrix calculation methods. Usually L- and T-shaped equivalent circuits are used. But sometimes more complex multi-circuit equivalent circuits of PSS elements can be used [2-4]. With large dimensions of the object, this may require large computing power. Another problem in modeling the electrical modes of an oil producing enterprises (OPE) is that the technological process parameters are not fully taken into account [5, 6]. In turn, taking into account the equipment features and the technological process may require significant complication of computational algorithms.

Thus, the development elements models of the ETC of an oil producing enterprise and models of

their interaction taking into account technological parameters is an urgent task.

According to well-known approaches, the structure of OPE ETC can be represented as a three-level system [6, 7]:

- Level of transformer substation (TS). The power source is an external power grid (PG). Usually, the power supply voltage at this level is 35-110 kV.
- Level of complete transformer substation (CTS). The power source is the TS, and the distribution of electricity through the CTS occurs. Usually, the power supply voltage is 6-20 kV.
- Electric centrifugal pumps (ESP). The power source is the CTS, and the ETC of ESP installation is directly supplied. Usually, the power supply voltage at this level is 0.4-1 kV.

In the process of oil production, subsystems of various physical nature interact with each other. In our case, these are electrical, mechanical and hydraulic subsystems. The oil reservoir and the ESP interact as the hydraulic and mechanical subsystems. ESP and SIM interact as the mechanical and electrical subsystems. The scheme of the ESP elements interaction is shown in Figure 1.

The algorithm for the ETC elements parameters calculation taking into account the oil reservoir parameters and technological parameters, as well as an algorithm for calculation of the electric mode parameters is shown below.

2.1 ETC Elements Parameters Calculation

2.1.1 Electric Centrifugal Pump

The power required to drive the pump is calculated by the formula:

$$P_{\text{ESP}} = \frac{\rho_{\text{l}} g \left(H_{\text{dyn}} + \frac{P_{\text{wh}}}{\rho_{\text{l}} g} \right) \cdot Q_{\text{ESP}}}{\eta_{\text{ESP}}} \cdot 10^{-3}, \qquad (1)$$

where ρ_{fl} – liquid density, kg/m^3; H_{dyn} – dynamic liquid level, m; P_{wh} – wellhead pressure, Pa; Q_{ESP} – ESP pumping rate, m^3/day; η_{ESP} – ESP efficiency.

It should be noted that the ESP efficiency is determined by its nameplate characteristic, taking into account the operating point and frequency regulation.

2.1.2 Submersible Induction Motor

Based on the power needed to drive the pump and the nameplate motor power, the motor load factor is determined by the formula:

$$k_{\text{LOAD}} = \frac{P_{\text{ESP}}}{P_{\text{SIM.NP}}} \cdot 100\%, \qquad (2)$$

where P_{SIMNP} – nameplate motor power, kW.

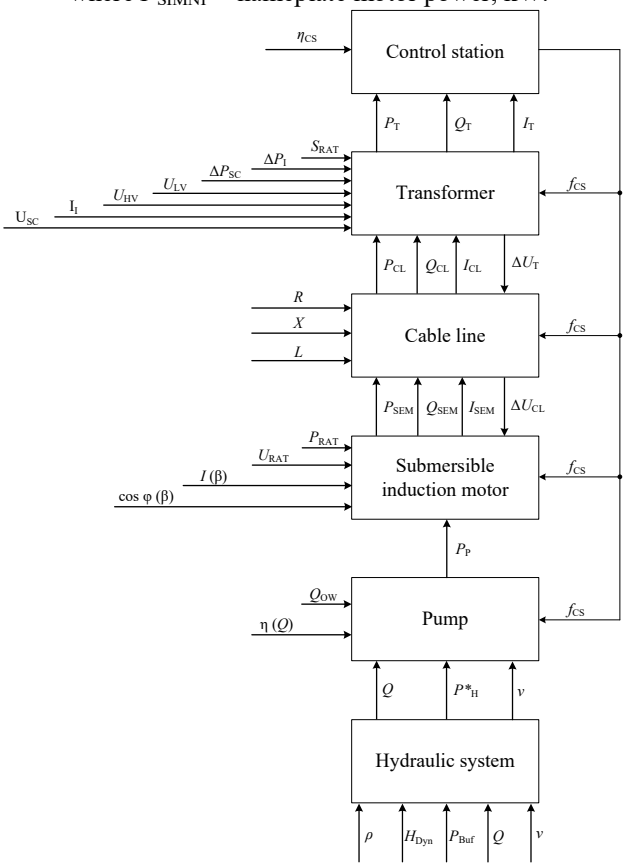


Figure 1: Scheme of the ESP elements interaction.

Then, based on the motor load characteristics, which are presented in the equipment catalog, the motor power factor and efficiency at the current load are determined. After that, the active and reactive power consumed by the motor are determined according to the formulas:

$$P_{\text{SIM}} = P_{\text{ESP}} + P_{\text{ESP}} \cdot \left(1 - \eta_{\text{SIM}}\right) \qquad (3)$$

$$Q_{\text{SIM}} = P_{\text{SIM}} \cdot tg(\arccos\varphi) \qquad (4)$$

2.1.3 Cable Line

$$r_{\text{CL}} = r_0 \cdot l_{\text{CL}}, \qquad (5)$$

$$x_{CL} = x_0 \cdot l_{CL} \cdot \frac{f}{50}, \quad (6)$$

where r_0, x_0 – specific CL active and reactive resistances, Ohm; f – voltage frequency, Hz.

CL power losses are calculated by the formula:

$$\Delta \dot{S}_{CL} = \left|\dot{I}_{CL}\right|^2 \cdot \left(R_{CL} + iX_{CL}\right), \quad (7)$$

where $\left|\dot{I}_{CL}\right|$ – CL current modulus.

2.1.4 Double-Wound Transformer

The equivalent circuit parameters of a double-wound transformer are calculated by the formulas:

$$r_T = \frac{\Delta P_{SC} \cdot U_{HV}^2}{S_{T,NP}^2}, \quad (8)$$

$$x_T = \sqrt{\left(\frac{U_{SC} U_{HV}^2}{100 S_{T,NP}}\right)^2 - r_T^2}, \quad (9)$$

where U_{HV} – transformer high voltage wound voltage, kV; $S_{T,NP}$ – transformer nameplate power, kVA; ΔP_{SC} – transformer short circuit losses, W; U_{SC} – transformer short circuit voltage, %.

Transformer active and reactive power losses are calculated by the formulas:

$$\Delta P_T = \Delta P_I \cdot \left(\frac{f}{50}\right)^{1,3} + \Delta P_{SC} \cdot \beta^2, \quad (10)$$

$$\Delta Q_T = \left(\frac{I_I}{100} + \beta^2 \cdot \frac{U_{SC}}{100}\right) \cdot S_{T,NP}, \quad (11)$$

where ΔP_I – transformer idle losses, W; I_I – transformer idle current, %; β – transformer loading, p.u.

2.1.5 Control Station

Control station power losses are calculated by the formula:

$$\Delta P_{CS} = P_{CS} \cdot \left(1 - \eta_{CS}\right), \quad (12)$$

where P_{CS} – control station output power, kW; η_{CS} – control station efficiency, p.u [1, 8].

2.2 Algorithm of the Electric Mode Parameters Calculation

The calculation of the PSS steady mode parameters is performed by the iteration method. The method algorithm implements the following basic relations for determining the mode parameters [9]:

- currents in branches of load:

$$I_{ij}^{(p)} = \frac{\left|\dot{S}_i\right|}{\sqrt{3} \cdot U_i^{(p-1)}}, \quad (13)$$

where p – iteration number; $\dot{S}$ – total power, kVA; U – voltage, kV.

- branches power:

$$\dot{S}_{ij}^{(p)} = \sum_{j \in n^*} \dot{S}_j^{(p)} + \Delta \dot{S}_{ij}^{(p)}, \quad (14)$$

where $\Delta \dot{S}_{ij}$ – branch power losses, kVA; $i = \overline{1,n}$; n^* – the set of nodes incident to node j, except node i.

- branches currents:

$$\dot{I}_{ij}^{(p)} = \frac{1}{N_{ij}} \cdot \sum_{k \in m^*} \dot{I}_{jk}^{(p)}, \quad (15)$$

where $i = \overline{1,n}$; N_{ij} – ij branch transformation ratio; m^* – the set of nodes incident to node j, except node i, however, this set must include at least 2 nodes, excluding node i.

- branches voltage losses:

$$\Delta U_{ij}^{(p)} = \frac{P_{ij}^{(p)} \cdot R_{ij} + Q_{ij}^{(p)} \cdot X_{ij}}{U_i}, \quad (16)$$

where P_{ij} – active power in branch; Q_{ij} – reactive power in branch.

- node voltage:

$$U_j^{(p)} = \frac{1}{N_{ij}} \cdot \left(U_i^{(p)} - \Delta U_{ij}^{(p)}\right), \quad (17)$$

- condition for convergence of the iterative process:

$$\left|U_j^{(p)} - U_j^{(p-1)}\right| \le \varepsilon. \quad (18)$$

The algorithm works as follows: it is assumed that initial voltage are equal supply source voltages and equipment nameplate voltages in load nodes; the reverse direction (the first stage) is the load currents are calculated from the end of the network (load) to the beginning of the network (supply source); the forward direction (second stage) consists in determination of branches voltage losses nodes voltages from the beginning of the network to end of the network, as well as in controlling convergence and iterative process.

3 MODELLING

The proposed algorithm is implemented in the LabVIEW development environment. Figure 2 shows fragment of the virtual instrument with implementation of the proposed approach.

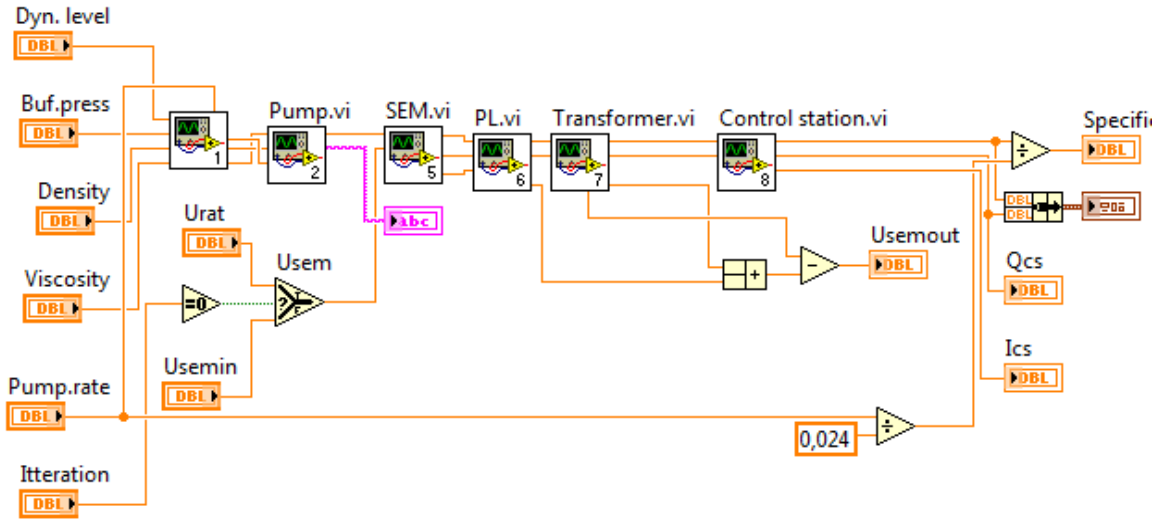


Figure 2: ESP Virtual Instrument.

4 RESULTS

For the test example calculation, the CTS-2310 of the Sukharev's field of the "LUKOIL-PERM" Ltd. is chosen. It feeds 2 wells.

The used initial data for the calculation corresponds to the technological mode on June 18, 2019 and are presented in Table 1. The parameters of the electrical equipment are presented in Table 2.

Table 1: Technological process parameters.

Well position	Pumping rate, m^3/day	Liquid density, kg/m^3	Dynamic level, m	Wellhead pressure, MPa
115	60,9	891	626	2,9
318	72,3	846	908	1,78

Table 2: Electrical equipment parameters.

Well posi-tion	Frequ-ency, Hz	SIM nameplate power, kW	Cable line	Pump nameplate flow, m^3/day
115	43	40	KPBP 3x16 L=1941 m	50
318	48	45	KPBP 3x16 L=2028 m	60

The absolute relative calculation error was calculated as the ratio modulus of the difference between the calculated and measurement values to the measurement value [10]:

$$\delta = 100 \cdot \left| \frac{x - X}{X} \right|, \qquad (19)$$

where x – calculated value, X – measurement value.

Table 3 presents a comparison of the modelling results with the data from the SIM control station.

Table 3: SIM modelling results.

Well posi-tion	Parameter	Measur-ment value	Calcu-lated value	Error δ, %
115	Current, A	19,2	19,5	1,56
	Power factor, p.u.	0,65	0,69	6,15
	Loading, %	56,1	58,6	4,46
318	Current, A	18,9	17,7	6,35
	Power factor, p.u.	0,64	0,69	7,81
	Loading, %	54,0	51,3	5,00

Table 4 presents a comparison of the modelling results with the results of instrumental measurements on CTP low voltage buses.

Table 4: ESP modelling results.

Well posi-tion	Parameter	Measurment value	Calculated value	Error δ, %
115	Active power, kW	28,6	27,5	3,85
	Reactive power, kvar	-	34,2	-
318	Active power, kW	36,3	34,0	6,34
	Reactive power, kvar	-	38,9	-

5 CONCLUSION

Based on the simulation results, it was revealed that the proposed models allow us to accurately determine the current parameters of the SIM. The error in determining the SIM current is not more than 6.35%, when determining the power factor is not more than 7.81%, and when determining the loading is not more than 5.00%.

The results of modeling the electricity consumption of an ESP (on low voltage buses of a complete transformer substation) show that the maximum error in active power is 6.34%. For reactive power, error estimation was not carried out due to the fact that during measurements these values were not measured.

These errors can be caused by the fact that the used technological parameters are averaged values over several days. This is due to the collecting and accounting features for these parameters in the enterprise under study.

The proposed approach allows a comprehensive assessment of electricity consumption by installing an electric centrifugal pump, taking into account the influence of technological factors, parameters of the electrical network, as well as electrical and mechanical equipment. A feature of the model is that it can take into account the mutual influence of wells through the parameters of the electric mode. Evaluation of the results shows that the proposed approach gives results that are acceptable for rapid assessment.

In the future, it is planned to accurate the models of electrical equipment and take into account its features, as well as the features of its functioning in non-nominal operating modes.

The methods and engineering strategies described above were tried on the set of mineral resource enterprises (LUKOIL, GAZPROM etc). The project is also aimed at supporting of a new Master's program "Conceptual design and engineering to improve energy efficiency" for preparing of engineers, scientists and administrative specialists in power industry, network companies, and related industries [1].

REFERENCES

[1] A.B. Petrochenkov, A.V. Romodin, S.V. Mishurin-skikh, V.V. Seleznev and V.A. Shamaev, "Experience in Developing a Physical Model of Submersible Electrical Equipment for Simulator Systems: Research and Training Tasks on the Agenda of a Key Employer". PTES, Saint Petersburg, Russia, 2018, pp. 114-117, doi: 10.1109/ PTES.2018.8604169, wos:000458979100033.

[2] B.V. Kavalerov, A.B. Petrochenkov, K. Odin and V.A. Tarasov, "Modeling of the Interaction of Structural Elements". Russian Electrical Engineering, 2013, vol. 84, no. 1, pp. 9-13, doi: 10.3103/S1068371213010033.

[3] N. Zhou, F. Ye, Q. Wang, X. Lou and Y. Zhang, "Short-Circuit Calculation in Distribution Networks with Distributed Induction Generators. Energies", vol. 9 (4) , 2016 , pp. 1-21, doi:10.3390/en9040277.

[4] V.A. Venikov, "Theory of Similarity and Simulation with Applications to Problems in Electrical Power Engineering". Macdonald Techincal & Scientific, London, 1969.

[5] A. Abur, "Power System State Estimation. Theory and Implementation", ISBN: 0-8247-5570-7.

[6] F. Milano, "Power System Modelling and Scripting", doi: 10.1007/978-3-642-13669-6.

[7] A.V. Lyakhomskiy, L.A. Plashchansky, S.N. Reshetnyak and M.Y. Reshetnyak, "High-voltage unit for automated monitoring of electrical energy quality in underground networks of coal mines". Mining Informational and Analytical Bulletin, vol. 7, 2019, pp. 207-213. doi:10.25018/0236-1493-2019-07-0-207-213.

[8] M. Bagajewicz, "A review of techniques for instrumentation design and upgrade in process plants", Canadian Journal of Chemical Engineering, vol. 80 (1), 2002, pp. 3-16.

[9] K. Szendy, Korszerű Hálózatszámítási Módszerek, Budapest:Akad. Kiado, 1967.

[10] R. John, "An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements", Taylor: Books, doi: 9780935702750.

Data Processing and Analysis of Glucose Concentration According to the Immittance Meter

Sergey Shipilov[1], Andrey Klokov[1], Ekaterina Yurchenko[1], Kseniya Zavyalova[1] and Alexey Yurchenko[2]
[1]*Tomsk State University, Department of Radiophysics, Lenin Ave 36, Tomsk, Russia*
[2]*Tomsk Politehnic University, Research School of High-Energy Physics, Lenin Ave 30, Tomsk, Russia*
s.shipilov@gmail.com, *niipp@inbox.ru*

Keywords: Glucose Measurements, Data Analysis, Immittance Measurement, Data Processing.

Abstract: This article proposes a method for measuring glucose concentration as a possible promising way to create a non-invasive glucometer. The dependence of the complex resistance of a solution of sodium chloride with different concentrations of glucose is investigated. In fist the glucose concentrations were calculated in saline to carry out these measurements. Further we prepared solutions (test samples) with a content of 3, 5, and 10 mmol per liter. We also created a board layout for measurements, that allowed us to vary solutions with different glucose concentrations. Using an immittance meter, we measured the complex resistance as a function of frequency in the range from 25 Hz to 1 MHz. At a medium frequency, we observed a resonance. We also found that at a frequency of 2 kHz there is a clear dependence of the complex resistance on the increase in glucose in solution. The obtained measurement data is of great interest for non-contact monitoring of glucose concentration in biological fluids.

1 INTRODUCTION

Monitoring glucose levels is an important indicator of normal human activities, especially for people with diabetes. The incidence of diabetes is increasing every year. Diabetes mellitus is a chronic disease in which blood sugar levels are elevated. This happens as a result of the fact that the pancreas either does not produce insulin, or the synthesized insulin cannot work effectively. The number of people with diabetes over the past 35 years has increased 4 times. Now more than 400 million people in the world have diabetes, and the prevalence of the disease continues to grow [1-5]. The World Health Organization estimates that diabetes will be the seventh leading cause of death by 2030. Possible methods of treatment include regulation of blood glucose levels by dietary methods, oral medication or insulin administration, all presented methods have an adverse effect on daily life [6-10].

Currently, it is recommended for patients with diabetes to regularly check their blood glucose with a glucometer. This practice can help closely monitor your blood glucose. In this way, patients with diabetes and their doctors can get a clear picture of blood glucose levels to optimize therapy. This is an indicator for adjusting the dose of insulin among patients with diabetes who need daily injections of insulin.

Typically, all blood glucose meters are invasive and require a finger puncture to take a blood sample. However, finger pricking to determine blood glucose levels for diabetics who check their blood sugar daily several times a day is difficult [11-14]. Puncture of a finger causes pain and subsequently leads to tissue damage. It also increases the risk of infection.

2 SUGGESTED MEASUREMENT METHOD

It is proposed to investigate the dependence of the complex resistance of a sodium chloride solution with different glucose concentrations.

2.1 Method for Calculating Concentrations in Solution

Molar concentration is the amount of solute (number of moles) per unit volume of solution. The molar

concentration in the SI system is measured in mol / m³, but in practice it is much more often expressed in mol / l or mmol / l. The test solution consists of saline (NaCl 0.9%) and glucose solution (C6H12O6 40%).

The level of glucose in the blood drops to 3.5 mmol per liter with hypoglycemia. This concentration of glucose in the blood leads to various disorders in the body. Normal blood glucose levels are on average between 3.5 and 5.5 mmol per liter. The level of glucose in the blood increases from 6.5 or more with hyperglycemia, it's negatively affects the body. We selected solutions according to all 3 states based on these data. You may notice that these data are too generalized. However, we believe that these extremes are enough for the first tests. At the first stage, it is important to understand the applicability of this approach - to understand sensitivity.

Solution calculation steps:

1) Calculate the molar mass of glucose. The chemical formula for glucose is C6H12O6. Molar mass of carbon - C = 12.011 Da. The molar mass of hydrogen is H = 1.008 Da. Molar mass of oxygen - O = 15.999 Da. Molar mass C6H12O6 = C * 6 + H * 12 + O * 6 = 72.066 + 12.096 + 95.994 = 180.156 Da. (weight 1 mol in grams).
2) 1 milliliter of glucose solution contains 0.4 grams of glucose therefore 0.00222 moles of C6H12O6.
3) Based on this, to obtain a solution with a content of 3 mmol per liter of C6H12O6, it is necessary to add 1.36 ml of a solution of 40% glucose in a 200 ml capacity with saline solution.
4) For a solution with a content of 5 mmol per liter of C6H12O6, it is necessary to add 2.25 ml of a solution of 40% glucose in a 200 ml capacity with saline.
5) For a solution with a content of 10 mmol per liter of C6H12O6, it is necessary to add 4.5 ml of a solution of 40% glucose in a 200 ml capacity with saline.

2.2 Preparing solutions (test samples)

Test samples were made on basis of the concentration calculation method presented above. For to appropriate concentrations (test samples) we mixed solution of 0.9% sodium chloride and 40% solution of glucose in the necessary proportions (Figure 1).

Figure 1: NaCl solution 0.9% and glucose solution 40%.

The procedure for the manufacture of solutions:

1) Removed the aluminium plug from the tank with NaCl solution.
2) Disconnect the top of the ampoule with 40% glucose solution.
3) Using a sterile syringe, the required amount of a 40% glucose solution was set for each sample and added to a container with a solution of 0.9% sodium chloride.
4) The contents of the sample containers were mixed and signed for measurements.

2.3 Making Layout for Measurements

The layout was created for measurements. The measuring model is a printed circuit board with a fixed latex tube on it. The prepared solutions of different glucose concentrations were supplied through this nylon tube. There are 4 contacts located at the same distance from each other on the printed circuit (Figure 2).

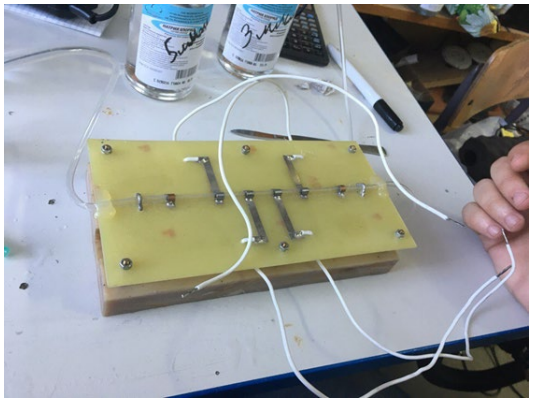


Figure 2: Measurement Layout.

The procedure for making the layout for measurements:

1) Production of printed circuit board in the copper sulfate.

2) The nylon tube was taken from a medical dropper.
3) The cylindrical contacts are soldered to the PCB.
4) 4 wires are soldered to the board for ease of use.

3 CARRYING OUT THE EXPERIMENT

The immittance meter E7-20 was used for measurements. E7-20 Immittance Meter (RLC) is a precision instrument of accuracy class 0.1, with a wide range of operating frequencies of 25 Hz - 1 MHz and a high measurement speed (up to 25 measurements per second) [5].

We researched the complex resistance module using the Immittance meter E7-20. The measurement layout and the E7-20 immittance meter were connected using a connecting device (CD) (Figure 3).

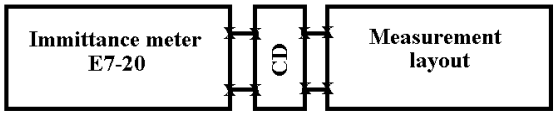


Figure 3: Scheme of the measurements.

To start work, it is necessary to carry out zero correction in open circuit mode and then in short circuit mode. Connected to the COM port of a personal computer and launched the program for measurements. We connected our model for measurements to the connecting device CD. Further we placed a test sample (solutions) in the model for measurements.

The measurements were carried out as follows. In the program, the frequencies for measurements were selected: 25, 50, 60, 100, 120, 200, 500, 1000, 5000, 10000, 20,000, 50,000, 100,000, 200,000, 500,000, 1,000,000 Hz. We used all frequencies at which E7 - 20 can measure to find the dependence of the complex resistance at frequencies on the glucose content in the samples.

At each frequency, measurements were performed 100 times for each sample. This number of measurements were made to establish the true average value. Since we know that the standard error of the mean is estimated as the sample standard deviation divided by the square root of the sample size:

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}},$$

where s - corrected sample standard deviation:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

and n - size of the sample. It follows that the more measurements are taken, the easier it will be to identify the true average value:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

After taking measurements, we started processing and analysing the measurement results.

4 RESULTS AND DISCUSSIONS

For data processing, we used the Microsoft Excel program, as it meets our requirements and is easy to use. As an example of the processed data, we took 3 frequencies: 25 Hz, 2 kHz and 50 kHz, that is, at low medium and high frequencies. We plotted probability density function for all data on the measurement frequencies (Figure 4).

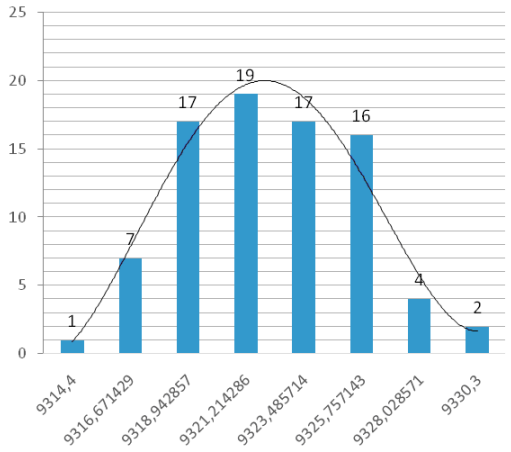


Figure 4: Histogram and density function for solution without glucose.

Density function was a normal distribution for all dataset. The normal distribution was parametrized in terms of the mean and the variance:

$$f(x;\bar{x},\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\bar{x})^2}{2\sigma^2}}.$$

At a frequency of 25 Hz, histograms of the distribution of the complex resistance module were obtained. From the results it was found that at a frequency of 25 Hz in a solution without glucose

there is a high reading of the complex resistance modulus and a slight change when measuring solutions with glucose.

We examined the behaviour of the complex resistance module at a frequency of 2 kHz. The results showed that at a frequency of 2 kHz, with an increase in glucose concentration, the complex resistance modulus increases.

At a frequency of 50 kHz, the distribution of the complex resistance module was also obtained. From the results it became clear that at a frequency of 50 kHz, the complex resistance modulus does not explicitly depend on the frequency.

It can be seen in our measurements that the scatter of values is present, but this will not affect the average value of the repetition rate. To establish the law of distribution and a more accurate true value in the future, we will conduct more measurements.

To visualize the values of the readings of the complex resistance modulus at all frequencies, we plotted the dependence of the average value of the samples on the measurement frequency (Figure 5).

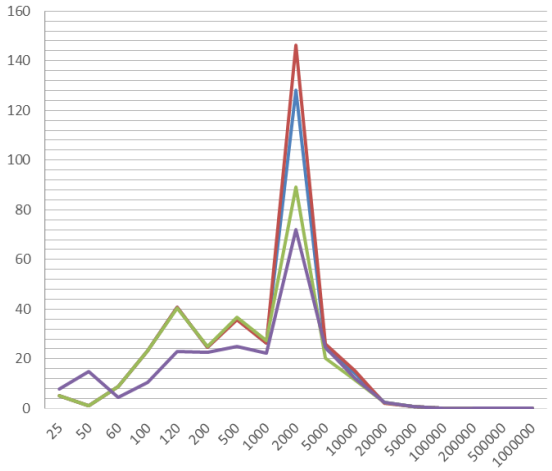


Figure 5: The graph of the true average module of complex resistance Z (vertical, MOhm) on measurement frequency (horizontal, Hz): lilac line - 0 mmol / liter; green line - 3 mmol / liter; blue line - 5 mmol / liter; the red line is 10 mmol / liter.

This graph shows that at a frequency of 2000 Hz we observe an unexplained resonance. This resonance will subsequently be investigated by us. It is also seen that at low frequencies we have a large scatter of values and the influence of glucose concentration on the frequency dependence of the complex resistance modulus is not observed, but at high scatter of values it is insignificant.

5 CONCLUSIONS

In this paper, we have shown the relevance of a non-invasive glucometer in the daily activities of a person with diabetes and the need to create it. We presented a method for measuring the level of glucose in the blood. We developed a model for measurements and made measurements and their processing.

The measurement results showed that the creation of such a glucometer can be implemented in practice. It can be seen from the above measurement results that it is not advisable to measure the complex resistance modulus Z at low and high frequencies. There are observe resonance at the middle frequency. Also, at a frequency of 2 kHz, the dependence of the complex resistance modulus Z on the increase in glucose concentration in the solution is observed.

In the future, it is planned to study the results in more detail. It is necessary to increase the number of measurements in order to reveal a clear dependence of the measured parameters on glucose. In addition, accuracy is important in such measurements (measurement error should not exceed 10-15 percent). Therefore, it is necessary to study the errors that can affect the measurement results, the influence of skin layers and other physiological data of a person. However, we believe that the results obtained now are important and have the prospect of further development.

ACKNOWLEDGMENTS

This study was supported by the Russian Science Foundation № 18-75-10101.

REFERENCES

[1] The official website of the World Health Organization media center. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs312/en/ Date of access: 09.10.2018.

[2] Official site MedicineNet, Definition of a puncture of a finger. [Online]. Available: http://www.medicinenet.com/script/main/art.asp?articlekey=39540 Date of access: 09/10/2018.

[3] S. Gozani, “Nerve damage in diabetes”, Bulletin of the Harvard Institute of Neurology. D. Mahoney, The Brain Magazine, no. 5, 1996, pp. 44-56.

[4] Diabetes mellitus standards for diabetes care, diabetes treatment, 39 supplement. 1, 2016, pp. 4-42.

[5] V.M. Lozovsky, A.G. Varakomsky and V.V. Bahur "Immitance meter E7-20", Operation manual, OJSC "MNIPI", 2006, pp.4-13.

[6] World Health Organization (WHO). Diabetes. [Online]. Available: http://www.who.int/news-room/fact-sheets/detail/diabetes (accessed on 3 October 2018).

[7] Healthline. "The Effects of Low Blood Sugar on Your Body". [Online]. Available: https://www.healthline.com/health/low-blood-sugar-effects-on-body#6 (accessed on 2 October 2018).

[8] L.C. Clark and C. Lyons, "Electrode systems for continuous monitoring in cardiovascular surgery". Ann. N. Y. Acad. Sci., no 102, 1962, pp. 29-45.

[9] C.-F. So; K.-S. Choi; T.K.S. Wong and J.W.Y. "Chung, Recent advances in noninvasive glucose monitoring". Med. Dev. (Auckl.), no 5, 2012, pp. 45-52.

[10] Y. Uwadaira and A. Ikehata, "Noninvasive Blood Glucose Measurement". In Nutritional and Therapeutic Interventions for Diabetes and Metabolic Syndrome, 2nd ed.; Bagchi, D., Nair, S., Eds.; Academic Press: San Diego, CA, USA, 2018; pp. 489-504.

[11] N.H. Cho, J.E. Shaw, S. Karuranga, Y. Huang, J.D. da Rocha Fernandes, A.W. Ohlrogge and B. Malanda, "IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045". Diabetes Res. Clin. Pract. no. 138, 2018, pp. 271- 281.

[12] J. Lin, T.J. Thompson, Y.J. Cheng, X. Zhuo, P. Zhang, E. Gregg and D.B. Rolka, "Projection of the future diabetes burden in the United States through 2060". Popul. Health Metr., no. 16, 2018, pp. 1-9.

[13] A. Tura, S. Sbrignadello, D. Cianciavicchia, G. Pacini and P. Ravazzani, "A Low Frequency Electromagnetic Sensor for Indirect Measurement of Glucose Concentration: In Vitro Experiments in Different Conductive Solutions". Sensors, no. 10, 2010, pp. 5346-5358.

[14] C. Chen, X.-L. Zhao, Z.-H. Li, Z.-G. Zhu, S.-H. Qian and A.J. Flewitt, "Current and Emerging Technology for Continuous Glucose Monitoring". Sensors, no. 17, 2017, p. 182.

[15] T. Lin, A. Gal, Y. Mayzel, K. Horman and K. Bahartan, "Non-invasive Glucose Monitoring: A Review of Challenges and Recent Advances", Curr. Trends Biomed. Eng. Biosci, no. 6, 2017, pp. 1-8.

[16] B.J. Van Enter, E. von Hauff, "Challenges and perspectives in continuous glucose monitoring". Chem. Commun, no. 54, 2018, pp. 5032-5045.

[17] O.S. Khalil, "Spectroscopic and Clinical Aspects of Noninvasive Glucose Measurements". Clin. Chem. no. 45, 1999, p. 165.