Vladimir M. Vishnevskiy
Dmitry V. Kozyrev (Eds.)

Communications in Computer and Information Science 919

Distributed Computer and Communication Networks

21st International Conference, DCCN 2018
Moscow, Russia, September 17–21, 2018
Proceedings

Springer

Communications in Computer and Information Science 919

Commenced Publication in 2007
Founding and Former Series Editors:
Phoebe Chen, Alfredo Cuzzocrea, Xiaoyong Du, Orhun Kara, Ting Liu, Dominik Ślęzak, and Xiaokang Yang

Editorial Board

More information about this series at http://www.springer.com/series/7899

Vladimir M. Vishnevskiy · Dmitry V. Kozyrev (Eds.)

Distributed Computer and Communication Networks

21st International Conference, DCCN 2018
Moscow, Russia, September 17–21, 2018
Proceedings

Springer

Editors
Vladimir M. Vishnevskiy
V.A. Trapeznikov Institute of Control Sciences
Russian Academy of Sciences
Moscow
Russia

Dmitry V. Kozyrev
V.A. Trapeznikov Institute of Control Sciences
Russian Academy of Sciences
Moscow
Russia

and

RUDN University
Moscow
Russia

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-319-99446-8 ISBN 978-3-319-99447-5 (eBook)
https://doi.org/10.1007/978-3-319-99447-5

Library of Congress Control Number: 2018951637

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This volume contains a collection of revised selected full-text papers presented at the 21st International Conference on Distributed Computer and Communication Networks (DCCN 2018), held in Moscow, Russia, September 17–21, 2018.

The conference is a continuation of traditional international conferences of the DCCN series, which took place in Bulgaria (Sofia, 1995, 2005, 2006, 2008, 2009, 2014), Israel (Tel Aviv, 1996, 1997, 1999, 2001), and Russia (Moscow, 1998, 2000, 2003, 2007, 2010, 2011, 2013, 2015, 2016, 2017) in the past 21 years. The main idea of the conference is to provide a platform and forum for researchers and developers from academia and industry from various countries working in the area of theory and applications of distributed computer and communication networks, mathematical modeling, methods of control, and optimization of distributed systems, by offering them a unique opportunity to share their views as well as to discuss the developments and pursue collaborations in this area. The content of this volume is related to the following subjects:

1. Computer and communications networking architecture optimization
2. Control in computer and telecommunication systems
3. Performance analysis and QoS/QoE evaluation in wireless networks
4. Analytical modeling and simulation of next-generation communications systems
5. Wireless 4G/5G networks, centimeter- and millimeter-wave radio technologies
6. RFID technologies and their application in intelligent transportation networks
7. Internet of Things, wearables, and applications of distributed information systems
8. Distributed and cloud computing systems, big data analytics
9. Probabilistic and statistical models in information systems
10. Queuing theory and reliability theory applications
11. High-altitude unmanned telecommunications platforms

The DCCN 2018 conference gathered 156 submissions from authors from 17 different countries. From these, 132 high-quality papers written in English were accepted and presented during the conference, 50 of which were recommended by session chairs and selected by the Program Committee for the Springer proceedings.

All the papers selected for the proceedings are given in the form presented by the authors. These papers are of interest to everyone working in the field of computer and communication networks.

We thank all the authors for their interest in DCCN, the members of the Program Committee for their contributions, and the reviewers for their peer-reviewing efforts.

September 2018

Vladimir Vishnevskiy
Konstantin Samouylov

Organization

DCCN 2018 was jointly organized by the Russian Academy of Sciences (RAS), the V.A. Trapeznikov Institute of Control Sciences of RAS (ICS RAS), the People's Friendship University of Russia (RUDN), the National Research Tomsk State University, and the Institute of Information and Communication Technologies of the Bulgarian Academy of Sciences (IICT BAS).

International Program Committee

V. M. Vishnevskiy (Chair)	ICS RAS, Russia
K. E. Samouylov (Co-chair)	RUDN University, Russia
S. Andreev	Tampere University of Technology, Finland
A. M. Andronov	Transport and Telecommunication Institute, Latvia
L. I. Abrosimov	Moscow Power Engineering Institute, Russia
Mo Adda	University of Portsmouth, UK
T. I. Aliev	ITMO University, Russia
A. S. Bugaev	Moscow Institute of Physics and Technology, Russia
T. Czachorski	Institute of informatics of Polish Academy of Sciences, Poland
A. N. Dudin	Belarusian State University, Belarus
V. V. Devyatkov	Bauman Moscow State Technical University, Russia
D. Deng	National Changhua University of Education, Taiwan
A. V. Dvorkovich	Moscow Institute of Physics and Technology, Russia
V. P. Dvorkovich	Moscow Institute of Physics and Technology, Russia
V. A. Efimushkin	Central Science Research Telecommunication Institute (ZNIIS), Russia
D. V. Efrosinin	Johannes Kepler University, Austria
M. A. Fedotkin	State University of Nizhni Novgorod, Russia
Yu. V. Gaidamaka	RUDN University, Russia
E. Gelenbe	Imperial College London, UK
A. Gelman	IEEE Communications Society, USA
P. Gaj	Silesian University of Technology, Poland
D. Grace	York University, UK
Yu. A. Gromakov	Moscow Aviation Institute, Russia
V. C. Joshua	CMS College, India
N. Kolev	University of São Paulo, Brazil
J. Kolodziej	Cracow University of Technology, Poland
G. Kotsis	Johannes Kepler University Linz, Austria
U. Krieger	University of Bamberg, Germany
A. Krishnamoorthy	Cochin University of Science and Technology, India

A. E. Koucheryavy	Bonch-Bruevich Saint-Petersburg State University of Telecommunications, Russia
Ye. A. Koucheryavy	Tampere University of Technology, Finland
N. A. Kuznetsov	Moscow Institute of Physics and Technology, Russia
L. Lakatos	Budapest University, Hungary
E. Levner	Holon Institute of Technology, Israel
S. D. Margenov	Institute of Information and Communication Technologies of Bulgarian Academy of Sciences, Bulgaria
O. Martikainen	Aalto University, Finland
A. Melikov	Institute of Cybernetics of the Azerbaijan National Academy of Sciences, Azerbaijan
G. K. Mishkoy	Academy of Sciences of Moldova, Moldavia
E. V. Morozov	Institute of Applied Mathematical Research of the Karelian Research Centre RAS, Russia
V. A. Naumov	Service Innovation Research Institute (PIKE), Finland
A. A. Nazarov	Tomsk State University, Russia
I. V. Nikiforov	Université de Technologie de Troyes, France
P. Nikitin	University of Washington, USA
S. A. Nikitov	Institute of Radio Engineering and Electronics of RAS, Russia
D. A. Novikov	ICS RAS, Russia
M. Pagano	Pisa University, Italy
V. V. Rykov	Gubkin Russian State University of Oil and Gas, Russia
L. A. Sevastianov	RUDN University, Russia
M. A. Sneps-Sneppe	Ventspils University College, Latvia
P. Stanchev	Kettering University, USA
S. N. Stepanov	Moscow Technical University of Communication and Informatics, Russia
S. P. Suschenko	Tomsk State University, Russia
H. Tijms	Vrije Universiteit Amsterdam, The Netherlands
S. N. Vasiliev	ICS RAS, Russia
E. Yakubov	Holon Institute of Technology, Israel
Y. P. Zaychenko	Kiev Polytechnic Institute, Ukraine

Organizing Committee

V. M. Vishnevskiy (Chair)	ICS RAS, Russia
K. E. Samouylov (Vice Chair)	RUDN University, Russia
S. P. Moiseeva	Tomsk State University, Russia
T. Atanasova	IIICT BAS, Bulgaria
D. V. Kozyrev	RUDN University and ICS RAS, Russia
A. A. Larionov	ICS RAS, Russia
R. E. Ivanov	ICS RAS, Russia

O. V. Semenova	ICS RAS, Russia
I. A. Gudkova	RUDN University, Russia
E. R. Zaripova	RUDN University, Russia
E. V. Markova	RUDN University, Russia
S. N. Kupriyakhina	ICS RAS, Russia

Organizers and Partners

Organizers

Russian Academy of Sciences, RUDN University,
V.A. Trapeznikov Institute of Control Sciences of RAS,
National Research Tomsk State University,
Institute of Information and Communication Technologies of Bulgarian Academy of Sciences,
Research and Development Company Information and Networking Technologies

Support

Information support was provided by the IEEE Russia Section. Financial support was provided by the Russian Foundation for Basic Research. The conference was held in the framework of the RUDN University Competitiveness Enhancement Program "5-100."

Contents

Tools and Techniques for Applications in 5G Networks and Beyond

Tommi Mikkonen[1(✉)] and Yevgeni Koucheryavy[2]

[1] University of Helsinki, Helsinki, Finland
tommi.mikkonen@helsinki.fi
[2] Tampere University of Technology, Tampere, Finland
yk@cs.tut.fi

Abstract. Future telecommunications networks, going beyond 5G, introduce numerous opportunities for new applications. Increased flexibility implies that new tools and techniques will be needed to take the most out of the networks, as otherwise we will simply create replicas of today's networks, which potentially include the same bottlenecks. In this keynote, we discuss network topologies, application architectures, and adaptability options that eventually will help in building superior user experience in future telecommunication networks and their applications. This will pave the way towards the Internet of people where technology is simply an enabler for satisfying end-user needs, and technological underpinnings are selected such that they best serve these needs.

Keywords: 5G networks · Programmable world
Software architecture · Edge computing · Fog computing
Isomorphic software systems

1 Introduction

Ever since the introduction of 3G, telecommunication networks have become more dependent on software. In the beginning of this evolution, software was used to replace features that were hardwired in early telecommunications networks; later generations have accepted the role of software, and complex management systems have been introduced to properly control all parts of the network, starting from switches and ending at base stations. Each of the network elements have had their own, largely predefined roles stemming from applicable standards, and there has been little room to redefine where applications and functions included in the network are executed.

5G marks a culmination point of this evolution, where managing software applications and components that constitute them is no longer feasible to do manually. Instead, the system should be allowed to freely establish direct connections between computing nodes on-the-need basis, and software applications should be able to adapt to the newly established connections. Therefore, going beyond 5G (5G+) calls for total reconsideration of software features in order to

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 1–8, 2018.
https://doi.org/10.1007/978-3-319-99447-5_1

make the best out of new possibilities. Such self-adaptive, total optimization is the only way to implement software based features that call for time-sensitive networking, leading towards an increasingly programmable world [1,2].

In general, self-adaptivity enables a software system to adapt to its changing environment and internal operations. However, the self-adaptation capability of the system is typically limited by its designer's ability to foresee and design for future adaptation needs. This limitation can be overcome by (i) introducing a layer of creativity on top of the adaptivity features; and (ii) accepting that the software in question must be able to move from one computer to another, and continue its operations in the new environment, unharmed by the transition These depend on the (iii) general capabilities of the underlying infrastructure(s) that provide the potential for unrestricted innovation in the application space. Furthermore, they also enable taking into account application and technology specific restrictions, in particular those that are sensitive to timing, data access, and computing resources.

As a solution to (i), we propose a creativity layer enriched with bio-inspired computing that reflect the capabilities of human brain and its ability to react either fast or slow [10]. The speed of the reaction depends on the criticality, available energy resources, time to act, and so on. Furthermore, the layer shall be able to create new configurations, with different characteristics, as supported by the underlying software framework we call liquid software [6] which acts as the solution for (ii). A central aspect of a liquid computing experience is the ability to move fluidly from one device or node to another. With liquid software, applications and data can flow from one device to another seamlessly, allowing the users to roam freely from one device to another, no longer worrying about device management, not having their favorite applications or data, having to remember complex steps, or consider if they are interacting with the cloud directly or using a proxy that happens to be available at the edge of the cloud. As for (iii), we need flexible communication patterns that allow dynamically shifting between centrally controlled communications and peer-to-peer networking, as well as sensor systems to improve overall situational awareness, much in the same way as human brain does. Furthermore, security mechanisms that are applicable both in the central cloud as well as direct communication are needed.

In this keynote, we propose to combine steps (i), (ii), and (iii) from our previous research into a seamless framework that is optimized for network topologies going beyond 5G network technologies (cloud, fog, edge) and their evolution. The goal is to take 5G networks as main foundation for building software applications that use both cloud-based communication, edge and device-to-device (D2D) communication in parallel, depending on the situation.

In more detail, the envisioned system works as follows. When a device senses regular use situations, it collects data to be analyzed on the cloud, resembling the higher functions of the brain. This data can then be used as basis for optimizing the fashion applications are partitioned in the network as well as which communication techniques and channels to use in application-to-application as well as applications' internal communication. Data such as time-sensitiveness of

executions as well as system's total optimization are the key parameters in defining a configuration that meets specific needs of novel software based features, available only in networks that go beyond 5G. Furthermore, when faced with a mission critical situation, this device needs to share this information quickly to other devices around it in order to procure a fast reaction. While fog computing does offer a solution to perform time-sensitive computations near the edge, these computations are pre-programmed and static in nature. A more liquid computation or components will independently decide what goes where, allowing the computations benefit from the resources available at the edge infrastructure. The edge could eventually become well-suited for the repetitive computational needs of the devices through self-learning, thus resembling the hindbrain learned reactions learned over time.

2 The Cloud, the Edge, the Fog

The Cloud. Today's dominant design for most computing-intensive tasks is a cloud-based system where devices stream their data to a back-end and in return receive instructions on how to act. In fact, even IoT systems, where a large number of devices monitor and act on the field are often built using this approach, despite the fact that the devices would be capable of independent actions. Furthermore, it is difficult to determine what is computed locally, and which operations truly require backend services.

The Fog. This view is challenged when delays caused by communication with the back-end become an obstacle for certain types of applications. So-called Fog Computing approaches allow devices to communicate and to orchestrate their operations collectively on the fly close to people and the data's origin [3]. Furthermore, the amount of network functions can be optimized, depending on application needs, which in turn makes applications more flexible. This in turn means that the load introduced by the applications can be managed using architectures and topologies best fit for the task at hand.

The Edge. Edge computing is a method of optimizing cloud computing systems by taking the control of computing applications, data, and services away from some central nodes (the "core") to the other logical extreme (the "edge") of the Internet which is closer to the physical world [4]. While the terminology between edge and fog computing is not always very clear, in this paper we assume an interpretation that fog computing allows using network resources on the need basis, whereas edge computing purely means communication with users' devices and the computing devices they directly communicate with on the network side.

3 Application Development Approaches

We expect that creating flexible future networks requires isomorphic software systems that build on containerized architectures capable of running executable

code on any computing element. Therefore, the computing elements to be used can be selected based on time based strategies in 5G networks, thus offering flexible, software-driven capabilities that satisfy user and application-specific needs in an operator-friendly fashion.

Container Technologies. New software development approaches that rely on continuous updates and upgrades of software systems has changed the way software systems are deployed. Instead of thinking about fixed, rigid configurations where each subsystem has well-defined role, we today build systems out of individually deployable containers (e.g. Docker (https://www.docker.com/)). Furthermore, to deal with the ever-increasing complexity of container systems we use special software to manage them (e.g. Kubernetes (https://kubernetes.io/)). Today, these technologies are more meant for desktop and especially cloud environment, but analogous facilities can be implemented for resource constrained systems, such as those used in telecommunications networks.

Liquid Software. Liquid software refers to a style of workflow interaction of applications and computing services across multiple devices, such as computers, smartphones, and tablets [5,6]. The underlying concepts have long existed in computer science, such as in the notions of pervasive computing and ubiquitous computing. The fundamental goal is to include facilities that enable relocation of software with ease in applications, and several techniques exist to implement this function [7]. In general, such liquid techniques can be used to create new application configurations on-the-fly by migrating applications in the network, either towards the edge or the cloud, as implied by applications' real-time requirements, computational complexity, and data needs.

Isomorphic Software Architectures. Isomorphic software architectures allow running the same software packages in any computational element of the end-to-end system. While such architectures are not yet common, in 5–10 years we expect that devices, gateways and the cloud will have the ability to run the same software components and services. The benefit is flexible migration of code between any element in the overall system. In an isomorphic system architecture, there does not have to be any technical differences between software that runs in the backend or in the edge of the network. Rather, when necessary, software can freely "roam" between the cloud and the edge in a seamless, liquid fashion.

4 Network and Application Adaptivity

The increasing flexibility in network, associated topology, and application architectures means that new approaches are also needed to manage their configurations – today's somewhat static partitioning of responsibilities and operations simply will not satisfy new needs. To this end, we see opportunities in self-adaptation in its various forms.

Self-Adaptive Functions. The ever-increasing complexity of software systems demands a radical new thinking towards how we imagine and implement them. Furthermore, the environment and user needs are constantly changing or evolving and manually keeping up with those variables is both challenging and costly. Then, there are systems which are not within our reach all the time, for example, software components of a rover on Mars cannot be maintained through the traditional methods. Self-adaptive systems paradigm offers a relief in such scenarios where software systems are expected to take care of various own and user needs independently [8]. The needs could emerge from internal operations, environment or a change in user needs.

The awarenesses within the system collectively makes the 5G network self-perceiving. The network can observe itself and reflect upon various situations, and it can also have reconfigurational choices. Furthermore, the self-awareness and self-reflection forms a close loop which is typical of any self-adaptive system; the system monitors itself, adapts if a need arises and keeps on observing the changes. The loop is usually referred to as a MAPE loop [9], an abbreviation for Monitor-Analyse-Plan-Execute actions.

Bio-Inspired Computing. In many ways, future telecommunications networks operate like human brain – some actions require little attention and take place in an energy efficient, rapid fashion, whereas some other operations require careful consideration, which takes time and consumes considerable amount of energy [10]. So far, bio-inspired computing has been used for wireless network design [11], but we are far from expanding the approach to complete telecommunications networks. Still, we see tremendous potential in pushing this work to the next level, and therefore look forward to building experimental systems with this in mind.

5 Putting the Pieces Together: User Experience for Next-Generation Applications

Today, interaction with wireless networks is complicated by the fact that their structure is not visible to the naked eye. As a result, it becomes difficult to choose which object one wishes to interact with at any given moment. As one of the scenarios, we propose that the augmented humans could be enabled to see the sources of wireless signals and communicate with remote objects in the most natural way – by simply looking at them by means of beam forming.

Specifically, we aim to develop the technology and demonstrators with the ability to position the sources of wireless transmission in the augmented reality environment in real-time, such that the human user can interact with remote sensors and other smart objects through direct visual contact. The information received from the objects would then be positioned in the AR environment according to the actual location of the source of the radio signal. This has been illustrated in Fig. 1. For further discussion, the reader is referred to [12].

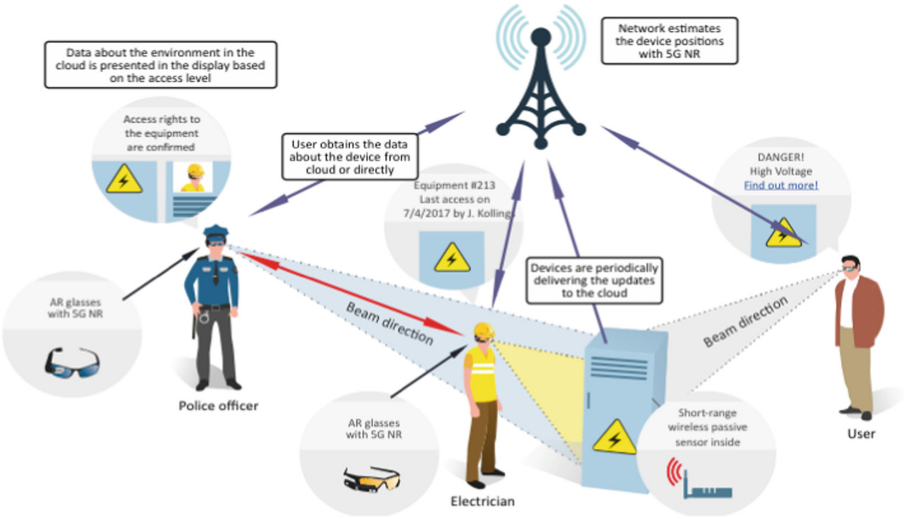


Fig. 1. Applications in a future telecommunications network illustrated.

The technology and network topology aspects required to locate the sources of transmissions with sufficient accuracy for overlaying them with AR in real-time, as well as to communicate with them selectively or by utilizing the caching information related to the sensor from cloud. This will require studies in the area of radiophysics and communications. Next, security aspect is to be studied, i.e. how to ensure that only the authorized users have access to restricted content, how to set up security contexts, etc. The main outcome will be a technology prototype, which would enable the user to see concealed wireless sensors in their actual locations based on the radio signal positioning based on 5G NR, while presenting themselves differently to different users. An example of such scenario is given in the figure below.

Here, three users of the augmented reality are present: a police officer, an electrician and, a common citizen. The object of interest (the electrical transformer box) is equipped with wireless sensor nodes that report to the cloud through the cellular NR interface. AR glasses have a scanning antenna array that can request a specific sensor to report, subject to security policy. The report is presented in the AR of each user differently: a common citizen would get basic description, while electrician would see the maintenance data and the schematics of the box. At the same time, a policeman would see the access logs and other information relevant to his/her case, while technically having access to all data. The sensor (or user) data is periodically transmitted to the cloud and could be obtained through NR link by the other user. The network has a self-learning mechanism allowing to precisely estimate the position of the devices for both static and mobile cases.

We believe that the augmentation of perception enabled by this technology will be the norm for the high-tech environments of the future, as the penetration of the wireless connectivity keeps on increasing. Further, such augmentation allows the user to interact better with other augmented humans by enabling them to visualize their peers for remote communications in the AR environments, thus simplifying social contact, as well as business and work relations. For example, finding a friend in a crowd will never become a problem again, as every augmented human would be able to home in on the radio signals of his/her friend's equipment. Finally, we believe that this technology is highly synergistic with other AR applications, thus supporting the entire AR market, which is one of the key drivers for human augmentation today.

6 Conclusion

In this keynote, we have addressed key technologies for application development in future telecommunications networks. The amount of flexibility and adaptability embedded in such networks allows unforeseen application configurations, and rapid reconfigurations as things change and new situations emerge. Furthermore, we believe that the networks will truly enable the internet of people service paradigm envisioned in [13], meaning that technology is simply an enabler for satisfying end-user needs, and that technological underpinnings are selected such that they best serve these needs.

In this keynote, we discuss network topologies, application architectures, and adaptability options that eventually will help in building superior user experience in future telecommunication networks and their applications. However, while many of the building blocks are readily available, their seamless integration requires a lot of future work. In addition, standardization activities are necessary to make applications roam across networks, a topic that also plays a role when defining and selecting implementation tools and techniques.

Acknowledgments. The authors wish to thank our research teams in Helsinki, Finland (Francois Christophe, Hadaytullah Hadaytullah, and Niko Mäkitalo) and Tampere, Finland (Sergey Andreev and Alex Ometov) for inspiration and ideas, as well as for the help in completing this paper.

References

1. Wasik, B.: In the Programmable World, All Our Objects Will Act as One. Wired (2013)
2. Taivalsaari, A., Mikkonen, T.: A roadmap to the programmable world: software challenges in the IoT era. IEEE Softw. **34**(1), 72–80 (2017)
3. Perera, C., Qin, Y., Estrella, J.C., Reiff-Marganiec, S., Vasilakos, A.V.: Fog computing for sustainable smart cities: a survey. ACM Comput. Surv. (CSUR) **50**(3), 32 (2017)
4. Garcia Lopez, P., et al.: Edge-centric computing: vision and challenges. ACM SIGCOMM Comput. Commun. Rev. **45**(5), 37–42 (2015)

5. Hartman, J.J., et al.: Joust: a platform for liquid software. Computer **32**(4), 50–56 (1999)
6. Taivalsaari, A., Mikkonen, T., Systä, K.: Liquid software manifesto: the era of multiple device ownership and its implications for software architecture. In: 38th Annual IEEE Computer Software and Applications Conference (COMPSAC), pp. 338–343. IEEE (2014)
7. Gallidabino, A., et al.: Architecting liquid software. J. Web Eng. **16**(5&6), 433–470 (2017)
8. Salehie, M., Tahvildari, L.: Self-adaptive software: landscape and research challenges. ACM Trans. Auton. Adapt. Syst. (TAAS) **4**(2), 14 (2009)
9. Kephart, J.O., Chess, D.M.: The vision of autonomic computing. Computer **36**(1), 41–50 (2003)
10. Kahneman, D., Egan, P.: Thinking, Fast and Slow, vol. 1. Farrar Straus and Giroux, New York (2011)
11. Christophe, F., Laukkarinen, T., Mikkonen, T., Massera, J., Andalibi, V.: Building wireless sensor networks with biological cultures: components and integration challenges. Int. J. Parallel Emergent Distrib. Syst. **32**(1), 56–73 (2017)
12. Mäkitalo, N., Ometov, A., Kannisto, J., Andreev, S., Koucheryavy, Y., Mikkonen, T., et al.: Safe, secure executions at the network edge. IEEE Softw. (2018)
13. Miranda, J., et al.: From the internet of things to the internet of people. IEEE Internet Comput. **19**(2), 40–47 (2015)

Evaluation of a Simulated Distributed Sensor- and Computational Network for Numerical Prediction Calculations

Ádám Vas(✉) and László Tóth

Faculty of Informatics, University of Debrecen,
Kassai str. 26, Debrecen 4028, Hungary
{vas.adam,toth.laszlo}@inf.unideb.hu

Abstract. In this paper we investigate a sensor network architecture with in-network computation capabilities. The computational tasks are based on numerical calculations, and the parallelization is provided on the lowest, mathematical layer. Since the real network nodes are equipped with atmospheric sensors, we chose a very simple weather prediction model based on the barotropic vorticity equation. To evaluate our network's capability of producing accurate prediction results, the simulations were executed on a European grid. We used recently measured publicly available data as initial values. Below the results of the distributed calculations are shown.

Keywords: Simulation · Sensor network · Distributed computing
Numerical calculations

1 Introduction

The main purpose of traditional computer networks is to provide information exchange between end nodes. Data packets or streams flow through intermediate nodes which are only responsible for forwarding the data towards the destination. With the appearance of sensor networks a different approach evolved, where the purpose of the nodes is to provide significant information about a parameter they measure with different kind of sensors attached to them. There are cases when each node is connected to a central server - this way they act only as end nodes. In other cases the information is passed and even modified between the nodes before reaching the central server.

In the latter case when peer-to-peer communication is already present in the network, the nodes can be used for performing calculations too. Nowadays they are usually equipped with a sufficiently powerful processing unit which performs signal processing and network communication tasks, but those rarely if ever cause 100% load, thence there remains enough processing time for other tasks.

Previously we succeeded in converting the original weather prediction algorithm created by CFvN [1] into a distributed form [2] and found practically

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 9–20, 2018.
https://doi.org/10.1007/978-3-319-99447-5_2

no difference between their results. The grid points and their initial values were taken from historical databases and the grid covered a large portion of the Northern hemisphere. Our final goal is to cover a European area with our sensor network. The focus of this paper is to test the distributed algorithm on this new, smaller grid using recent atmospheric measurements as initial parameters. The network nodes (Fig. 1) were simulated on a computer by individual threads communicating with each other over TCP/IP (Fig. 2). There are physical constraints of an operating sensor networks, such as small bandwidth, unstable network connection or limited power supply, which we do not consider here.

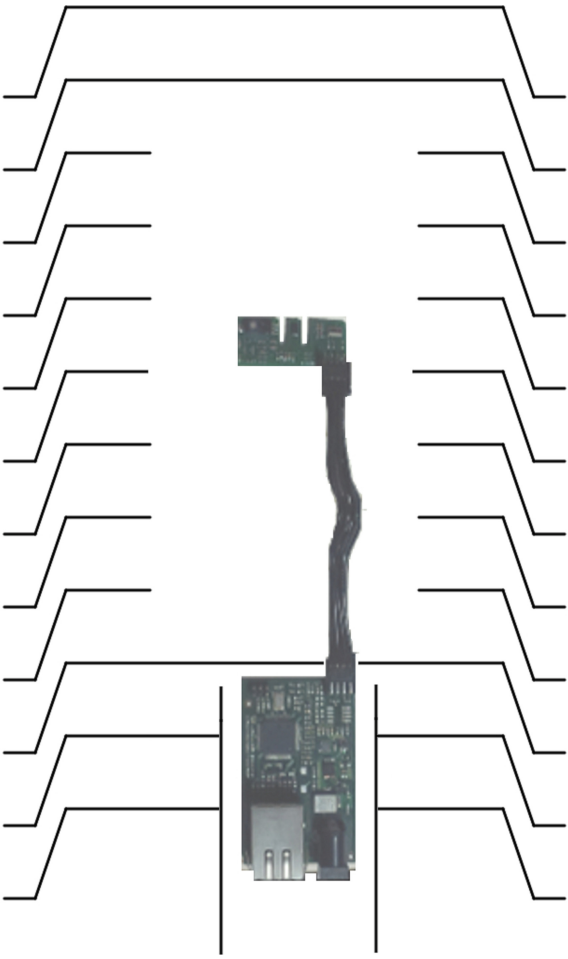


Fig. 1. Schematic of our custom weather station [5].

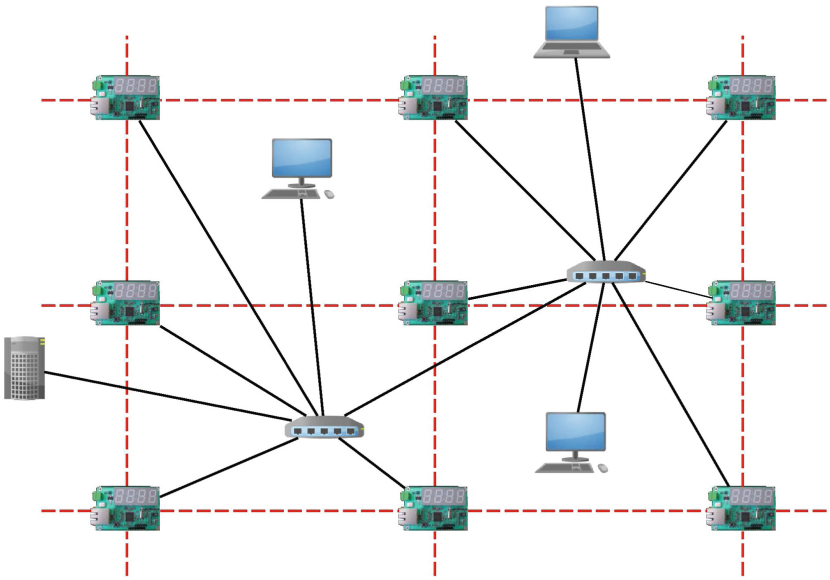


Fig. 2. The conception of DSN-PC sensor network with simulated sensor circuits forming a regular grid [2].

2 System and Model Description

Our future sensor network which covers a European area forms a 10×10 grid (see Fig. 3). To correspond with the original algorithm the polar stereographic map projection was used with the following grid parameters:

- Coordinates of the lower-left grid point: 39°N, 2.2°W
- Coordinates of the upper-right grid point: 53.689°N, 36.2161°E
- Grid step at North Pole: 300 km
- Central angle of the map: 0°

We previously covered the details of the distributed prediction algorithm based on CFvN [2]. To get the initial values of the prognostic variable (the geopotential height of the 500 mbar pressure level, z_{500}) we used the NOAA Integrated Global Radiosonde Archive (IGRA) which consists of recent and historical radiosonde and pilot balloon observations at over 2700 globally distributed stations [3]. We extracted the datasets of stations located between 33.3°N–65.5333°N and 14.4°W–44.5°E. The z_{500} values at the grid points is calculated by natural neighbor interpolation method [4]. The interpolation is performed on the (x,y) coordinate pairs of the stations according to polar stereographic projection [1].

After determining the initial parameters we ran the distributed CFvN algorithm on our 10×10 virtual sensor- and computational network. The forecasts

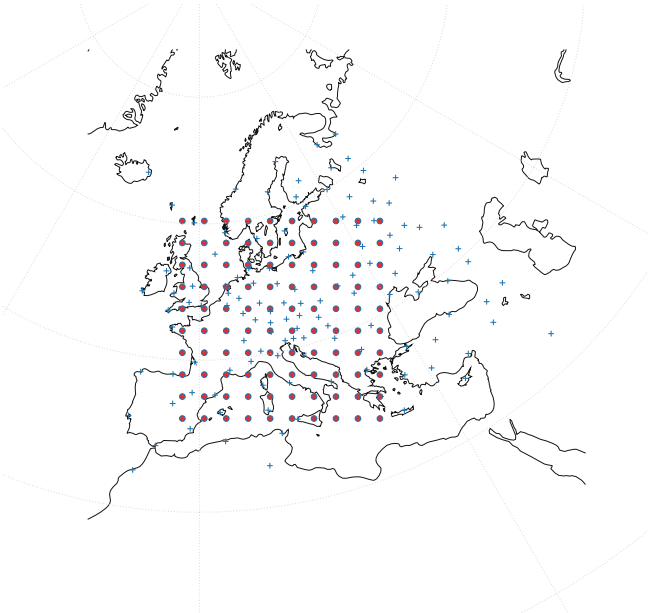


Fig. 3. The regular grid of our virtual sensor network and the location of weather balloons that are considered in our calculations.

were performed on measurements between 2008.01.01 and 2014.12.31. Measurements are available at 00:00 and 12:00 UTC for each day. There are some missing data, thence the total number of pressure maps is less than 730 in some years. The forecast length was 24 h, the applied time step was 0.1 h. We calculated the Mean Absolute Error (MAE) of the forecasts and compared it with the MAE values of the persistence method on a yearly basis to determine whether our model produces proper results. The MAE was calculated neglecting the boundary nodes:

$$MAE = \frac{1}{8 \cdot 8} \sum_{i=1}^{8} \sum_{j=1}^{8} |z_{\mathrm{i,j}} - z'_{\mathrm{i,j}}|, \tag{1}$$

where z'$_{i,j}$ is the calculated and z$_{i,j}$ is the real 500 hPa geopotential height measured after 24 h.

3 Simulation Results

On Figs. 4, 6, 8, 10, 12, 14 and 16 it is noticeable that the CFvN algorithm produces reasonable MAE values which are comparable to those produced by persistence method, although the model was originally designed for much larger areas. Below each scatter plot we show the result of the most accurate forecast for the given year (see Figs. 5, 7, 9, 11, 13, 15 and 17).

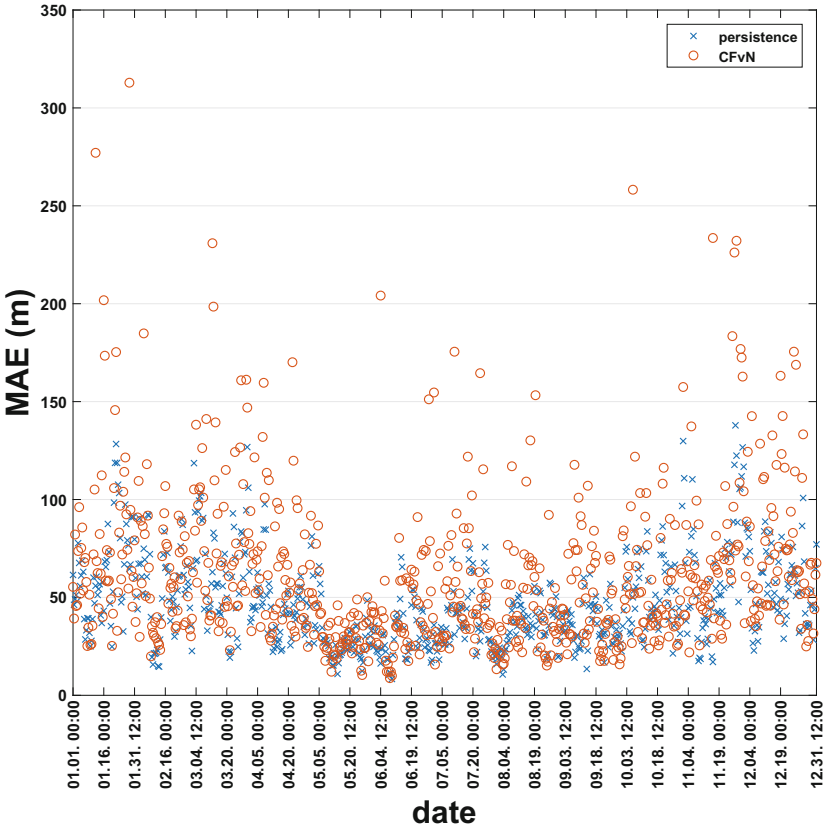


Fig. 4. The MAE values of the persistence and the CFvN methods applied to atmospheric measurements of 2008. 6 values are omitted because our algorithm produced incorrect (NaN) results.

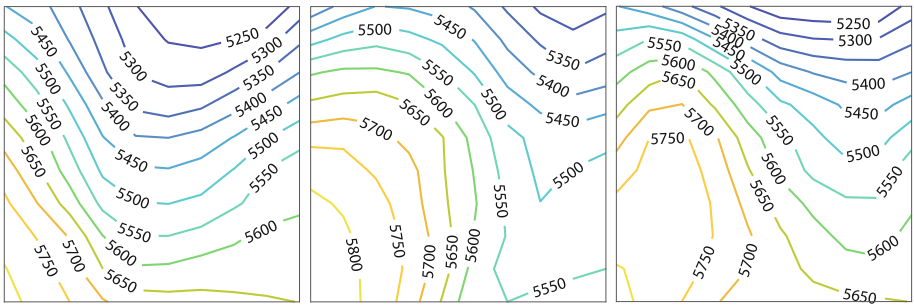


Fig. 5. The result of the forecast performed on data from 2008.01.22. 12:00 UTC, calculated by the simulated DSN-PC. Left: Initial height field. Center: Height field measured 24 h later. Right: Forecast height field.

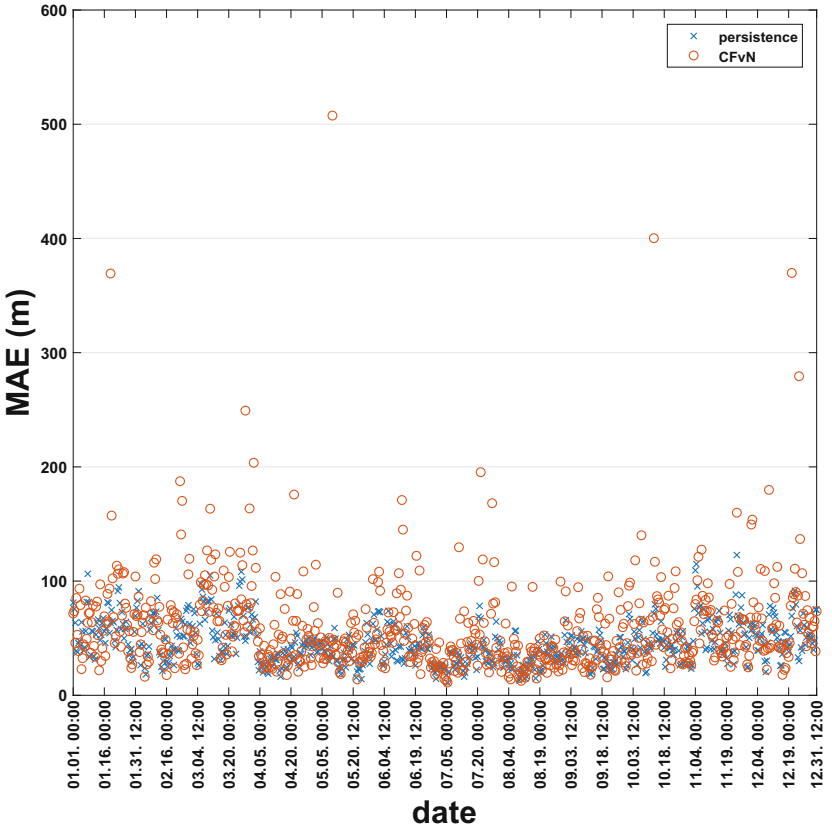


Fig. 6. The MAE values of the persistence and the CFvN methods applied to atmospheric measurements of 2009. 12 values are omitted because our algorithm produced incorrect (NaN) results.

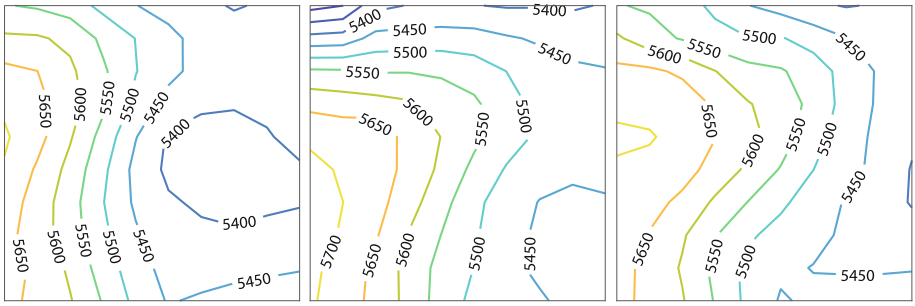


Fig. 7. The result of the forecast performed on data from 2009.02.24. 12:00 UTC, calculated by the simulated DSN-PC. Left: Initial height field. Center: Height field measured 24 h later. Right: Forecast height field.

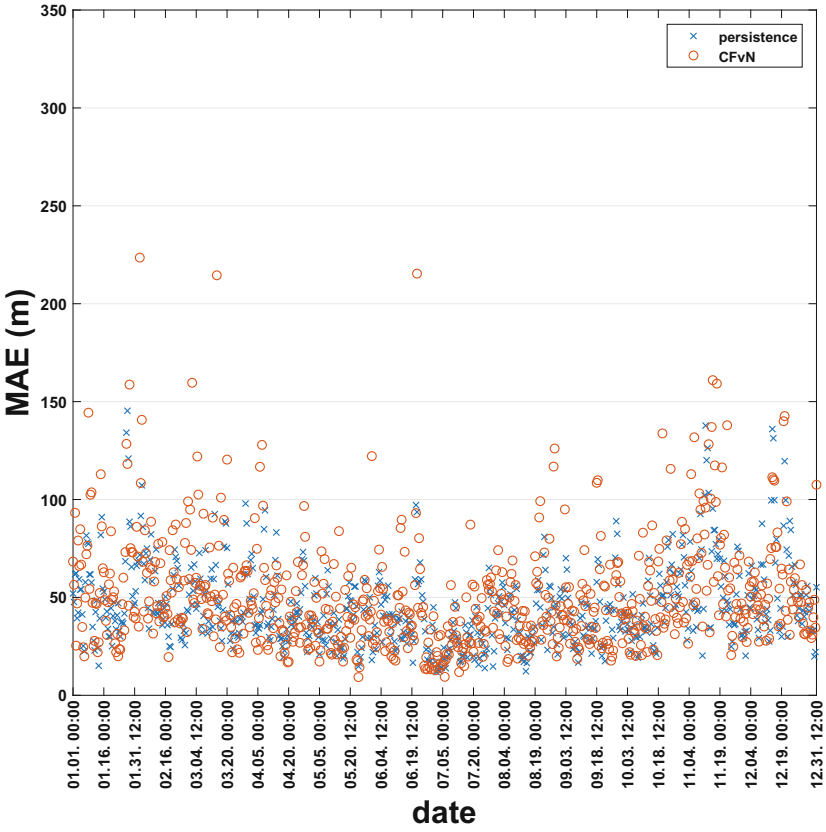


Fig. 8. The MAE values of the persistence and the CFvN methods applied to atmospheric measurements of 2010. 5 values are omitted because our algorithm produced incorrect (NaN) results.

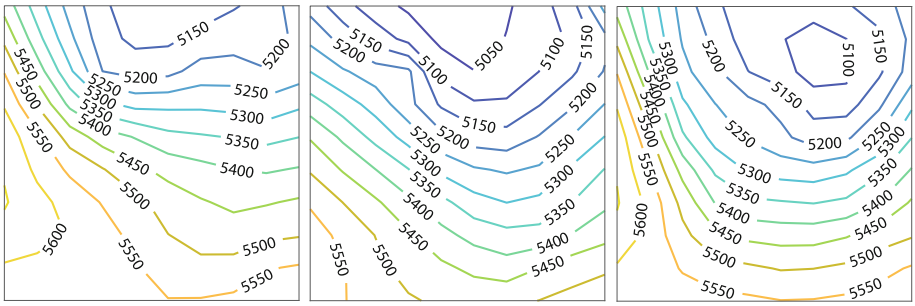


Fig. 9. The result of the forecast performed on data from 2010.01.28. 00:00 UTC, calculated by the simulated DSN-PC. Left: Initial height field. Center: Height field measured 24 h later. Right: Forecast height field.

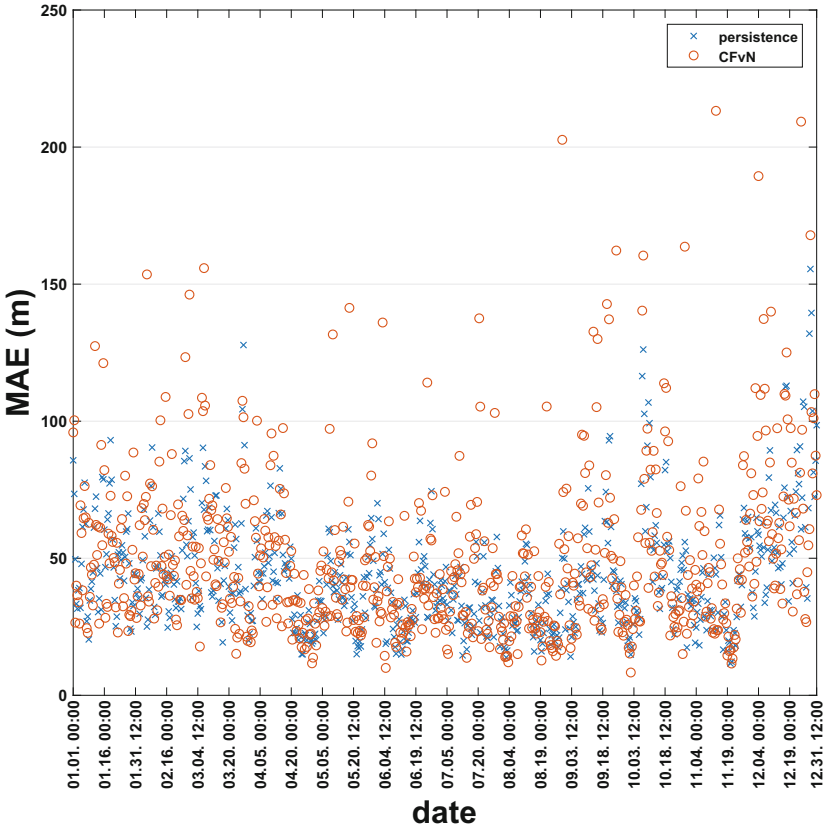


Fig. 10. The MAE values of the persistence and the CFvN methods applied to atmospheric measurements of 2011. 13 values are omitted because our algorithm produced incorrect (NaN) results.

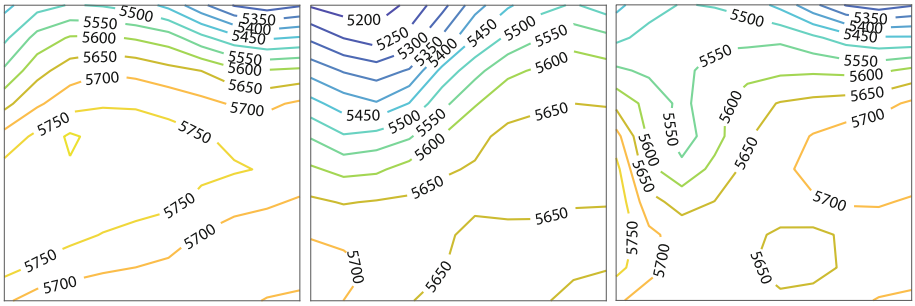


Fig. 11. The result of the forecast performed on data from 2011.12.28. 00:00 UTC, calculated by the simulated DSN-PC. Left: Initial height field. Center: Height field measured 24 h later. Right: Forecast height field.

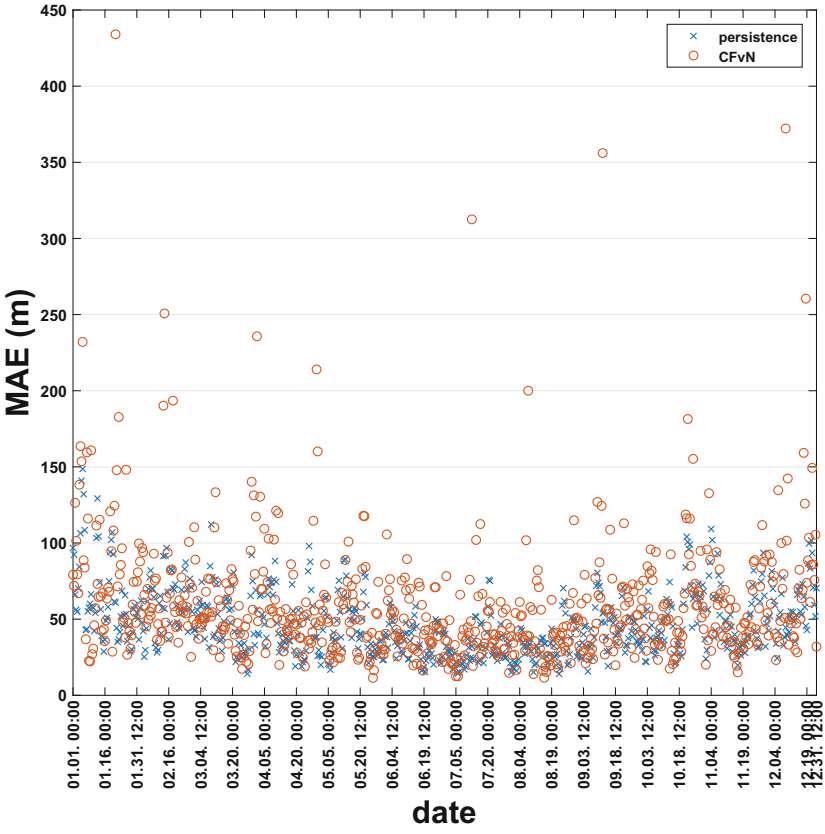


Fig. 12. The MAE values of the persistence and the CFvN methods applied to atmospheric measurements of 2012. 12 values are omitted because our algorithm produced incorrect (NaN) results.

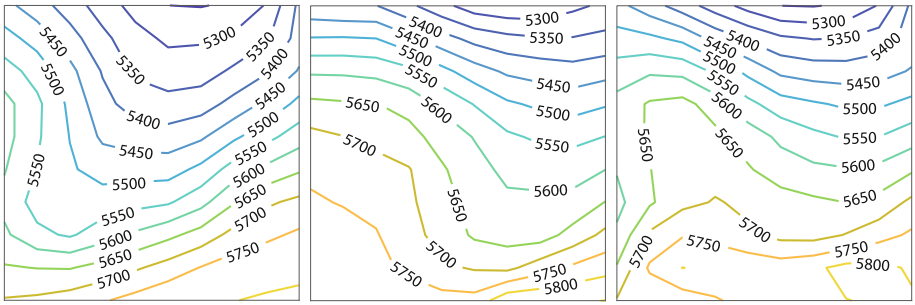


Fig. 13. The result of the forecast performed on data from 2012.11.06. 12:00 UTC, calculated by the simulated DSN-PC. Left: Initial height field. Center: Height field measured 24 h later. Right: Forecast height field.

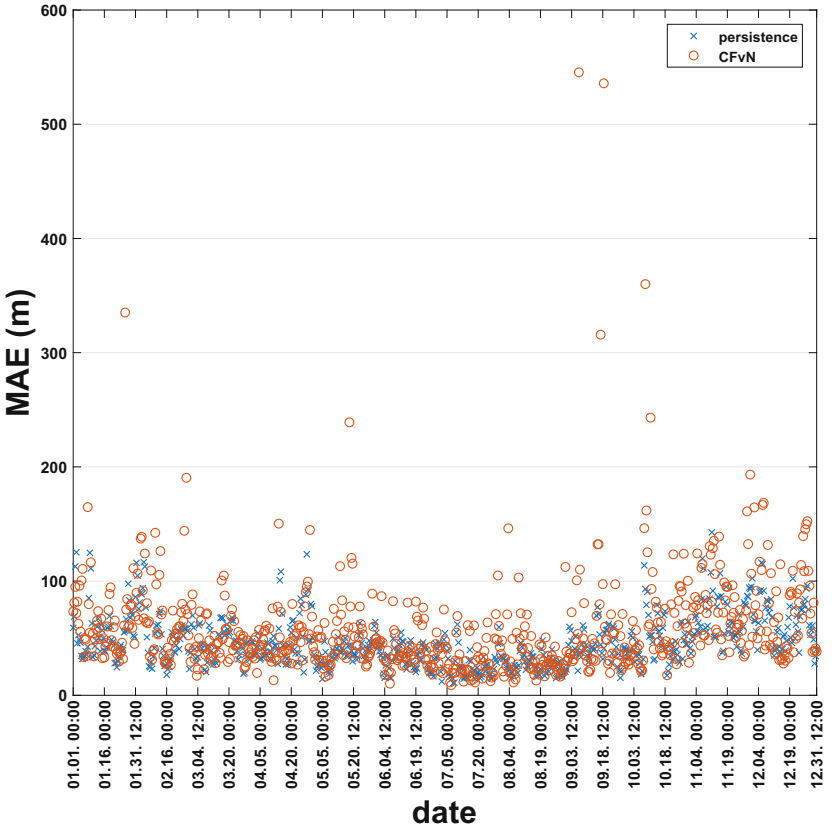


Fig. 14. The MAE values of the persistence and the CFvN methods applied to atmospheric measurements of 2013. 13 values are omitted because our algorithm produced incorrect (NaN) results.

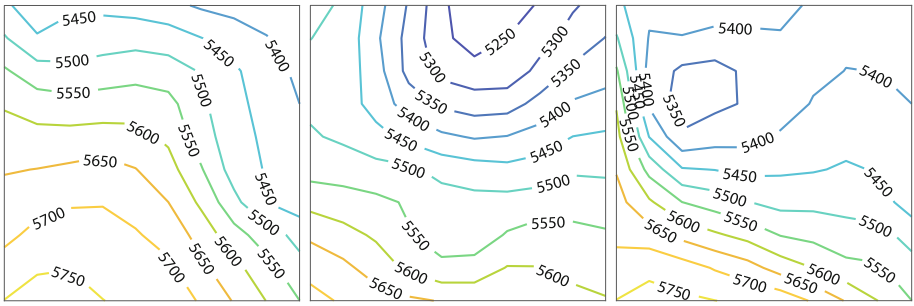


Fig. 15. The result of the forecast performed on data from 2013.01.09. 12:00 UTC, calculated by the simulated DSN-PC. Left: Initial height field. Center: Height field measured 24 h later. Right: Forecast height field.

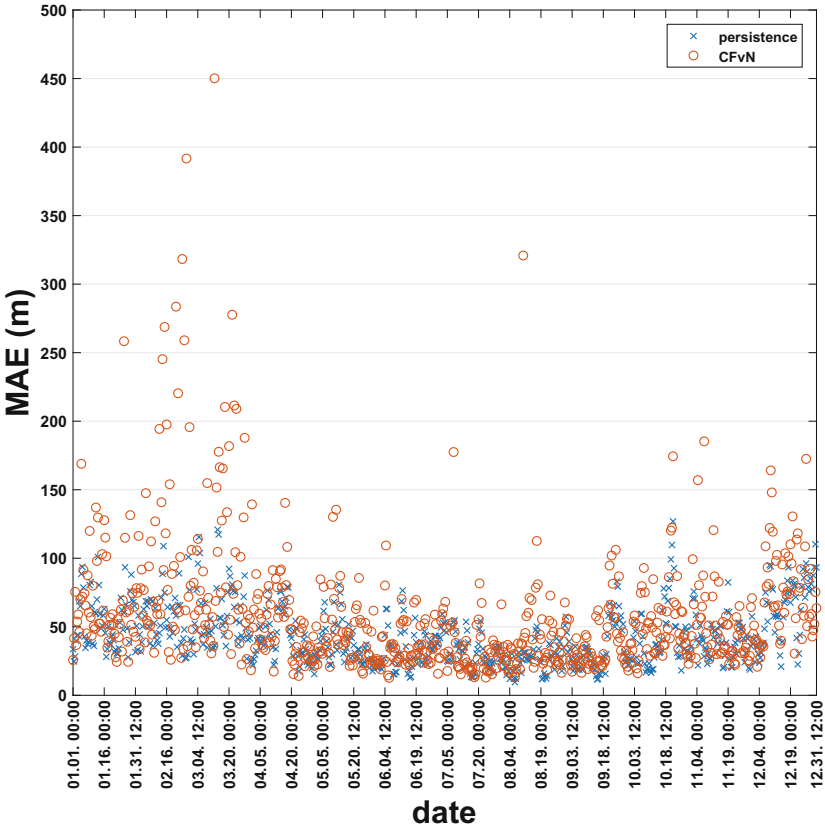


Fig. 16. The MAE values of the persistence and the CFvN methods applied to atmospheric measurements of 2014. 14 values are omitted because our algorithm produced incorrect (NaN) results.

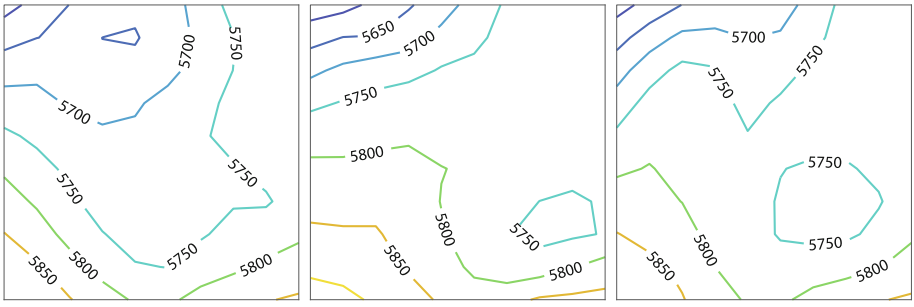


Fig. 17. The result of the forecast performed on data from 2014.07.14. 12:00 UTC, calculated by the simulated DSN-PC. Left: Initial height field. Center: Height field measured 24 h later. Right: Forecast height field.

4 Conclusion

Based on our conception we succeeded to apply the redesigned numerical weather prediction algorithm originally created by CFvN on a grid formed by the nodes of a virtual sensor network. The results are encouraging because they show the possibility of running distributed prediction calculations with real-time data over a European area. The next step will be a hybrid network including our custom weather stations [5] which can perform real-time measurements and calculations together with the virtual nodes. More complex prediction models can also be redesigned and implemented on the network to involve several atmospheric parameters and to improve the accuracy of the forecasts.

Acknowledgments. This work was supported by the construction EFOP-3.6.3-VEKOP-16-2017-00002. The project was co-financed by the Hungarian Government and the European Social Fund.

References

1. Charney, J.G., Fjørtoft, R., Von Neumann, J.: Numerical integration of the barotropic vorticity equation. Tellus **2**, 237–254 (1950)
2. Vas, Á., Fazekas, Á., Nagy, G., Tóth, L.: Distributed sensor network for meteorological observations and numerical weather prediction calculations. Carpath. J. Electron. Comput. Eng. **6**(1), 56–63 (2013)
3. Integrated Global Radiosonde Archive (IGRA). http://www.ncdc.noaa.gov/data-access/weather-balloon/integrated-global-radiosonde-archive
4. Sibson, R.: A Brief Description of Natural Neighbor Interpolation. In: Interpolating Multivariate Data, chap. 2, pp. 21–36. Wiley, New York (1981)
5. Vas, Á., Nagy, G., Tóth, L.: Networkable sensor station for DSN-PC system. Carpath. J. Electron. Comput. Eng. **8**(2), 37–40 (2015)

Estimation of a Heavy-Tailed Weibull-Pareto Distribution and Its Application to QoE Modeling

Udo R. Krieger[1(✉)] and Natalia M. Markovich[2]

[1] Fakultät WIAI, Otto-Friedrich-Universität,
An der Weberei 5, 96047 Bamberg, Germany
udo.krieger@ieee.org
[2] V.A. Trapeznikov Institute of Control Sciences, Russian Academy of Sciences,
Profsoyuznaya Str. 65, Moscow 117997, Russia
markovic@ipu.rssi.ru

Abstract. We model the end-to-end delay of advanced services in the Internet by means of a heavy-tailed Weibull-Pareto distribution (WPD). First we summarize the structural properties of the three-parameter WPD class and indicate its relation to the general Weibull-TX class. Then we present an effective estimation scheme to compute the parameters of a WPD distribution by a finite sample. Finally we show how a WPD distribution can be applied to determine the relevant QoE performance metric MOS of end-to-end delay dependent services in the Internet.

Keywords: Heavy-tailed distributions · Weibull-pareto distribution
Weibull-TX class · QoE modeling

1 Introduction

In recent years the fast evolution of new Web and multimedia applications has generated a rapidly changing load environment for the transport of data streams in modern high-speed networks. Considering the related traffic planning and quality-of-service (QoS) assessment in these dynamic wired and wireless access networks as well as backbone networks, important issues of load modeling and traffic characterization at different time scales by thorough statistical techniques were addressed as basic steps of teletraffic engineering in the last decades (see [3,4,6,11,13]). Measurements of the related traffic characteristics such as session durations of multimedia connections during the Internet access or the response times of content distribution network servers during Web sessions, have illustrated that the underlying random variables (rvs) are determined by long- and heavy-tailed distributions (cf. [12] and references therein). We known that the corresponding marginal distributions of the one-way delays between a client and a Web server and the round-trip times of related advanced multimedia services

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 21–30, 2018.
https://doi.org/10.1007/978-3-319-99447-5_3

are normally governed by a mixture of some body part of the distribution with a unimodal or multimodal shape and some Pareto shaped tail. This behavior is due to the accumulated waiting times in consecutive network nodes along the path of a connection in the Internet. It must be considered as a fundamental feature of delay modeling in teletraffic theory which cannot be ignored.

Roughly speaking, the underlying heavy-tailed distributions of those end-to-end (E2E) delays follow a cumulative distribution function (cdf) whose tail decays to zero at a slower rate than that one of an exponential distribution. The latter can be considered as the boundary between heavy and light tails. From a statistical perspective, a thorough mathematical definition of such behavior of heavy-tailed distributions has been provided in the corresponding literature, see, for instance, [5,20] and the examples therein.

In last decades the quality-of-service (QoS) of advanced multimedia and Web services in the Internet has been intensively studied by means of teletraffic theory (cf. [23] and references therein). It depends on the delay-loss-throughput profile induced by the underlying router and server infrastructures. Recently, a shift of the research focus towards the associated quality-of-experience (QoE) of these services that is perceived by a sampled user population has occurred (cf. [9,10]).

In this study we model the end-to-end delay of advanced Web and multimedia services in the Internet by means of a heavy-tailed Weibull-Pareto distribution (WPD) proposed in [1]:

$$G(x) = 1 - \exp\left(-\left(\beta \ln(\frac{x}{\theta})\right)^c\right), \; x \geq \theta > 0, c > 0 \tag{1}$$

First we summarize the structural properties of this three-parameter $WPD(c, \beta, \theta)$ class and indicate its relation to the general Weibull-TX class. Then we develop a modified estimation procedure for a three-parameter Weibull-Pareto distribution based on the maximum likelihood approach sketched in [1] and a tail-index estimator derived by extreme-value theory (cf. [2]). Finally, we look at the QoE performance of delay-dependent services in the Internet and show how a $WPD(c, \beta, \theta)$ model can be applied to determine the relevant metric MOS of such services.

The paper is organized as follows. In Sect. 2 we review the properties of a Weibull-Pareto distribution and illustrate its relationship to the Weibull-TX class of heavy-tailed distributions. We show how the parameters of a WPD model can be determined by a finite sample. In Sect. 3 we discuss the application of a WPD model to determine the MOS metric of a delay-dependent multimedia service. Finally, some conclusions are drawn.

2 The Weibull-Pareto Distribution and Its Properties

In recent years there have been numerous attempts to develop generalizations of the two-parameter Weibull and Burr distributions (cf. [15,18,19]). Among other requirements, the latter should be able to describe lifetimes with non-monotone hazard rate functions such as bathtub shaped hazard rates which

are often observed in practice. Three-parameter generalizations of the classical Weibull distribution

$$F(x) = \mathbb{P}\{Y \leq x\} = \begin{cases} 1 - \exp(-(x/\gamma)^c) & , x \geq 0 \\ 0 & , x < 0 \end{cases} \tag{2}$$

of a nonnegative random variable (rv) Y with the shape parameter $c > 0$ and a real scale parameter $\gamma > 0$ based on the exponentiation technique or the transformed-transformer (TX) methodology can fulfill these requirements (cf. [1,15,16,18]).

Following such concepts, Alzaatreth et al. [1] have considered a nonnegative random variable Y with cumulative distribution function (cdf) $F(x)$ and an independent nonnegative random variable T with cdf $R(t)$ and probability distribution function (pdf) $r(t)$, both defined on $[0, \infty)$. They determined a new nonnegative random variable X by the following transformed-transformer (TX) construction:

$$\begin{aligned} G(x) &= \mathbb{P}\{X \leq x\} = \mathbb{P}\{T \leq -\ln(1 - F(x))\} \\ &= \mathbb{P}\{1 - e^{-T} \leq F(x)\} \\ &= \int_0^{-\ln(1-F(x))} dR(t) = \int_0^{-\ln(1-F(x))} r(t)dt \\ &= R(-\ln(1 - F(x))) \end{aligned} \tag{3}$$

This transformed-transformer (TX) methodology [1] and the underlying Weibull-X family have been generalized recently to the more general Weibull-G family by Tahir et al. [22].

2.1 The Weibull-TX Model of a Weibull-Pareto Distribution

In the following, we show how the construction of the three-parameter Weibull-Pareto distribution $WPD(c, \beta, \theta)$ by Alzaatreth et al. [1] can be derived easily by general hazard rate construction principles developed by Gurvich et al. [7] and Pham and Lai [17].

We can formulate the derivation of a generalized transformed-transformer (TX) construction of the distribution function

$$G(x) = \mathbb{P}\{X \leq x\} = R(-\ln(1 - F(x)))$$

associated with a nonnegative rv X in terms of the cdf

$$R(t) = \mathbb{P}\{T \leq t\} = \int_0^t dR(\tau) = \int_0^t r(\tau)d\tau$$

of a nonnegative rv T with pdf $r(t)$ and the survival or reliability function

$$\overline{F}(y) = 1 - F(y) = \mathbb{P}\{Y > y\} = \int_y^\infty f(t)dt$$

of a nonnegative rv Y with density function $f(y)$. Then the corresponding hazard or failure rate function $h(y)$ and cumulative failure rate function $H(y)$ are defined by

$$h(y) = \frac{f(y)}{\overline{F}(y)} = \frac{f(y)}{1 - F(y)}$$
$$H(y) = \int_0^y h(t)dt \tag{4}$$

(cf. [17]). The cumulative failure rate function $H(y)$ satisfies the three conditions:

1. $H(y)$ is nondecreasing for all real $y \geq 0$.
2. H(0) = 0
3. $\lim_{y\to\infty} H(y) = \infty$

We know that the survival function of any rv $Y \geq 0$ can be represented by the cumulative failure rate function $H(y)$ in terms of

$$\overline{F}(y) = e^{-H(y)}, y \geq 0.$$

Regarding the survival function of a Weibull distribution, Gurvich et al. [7] have shown that the generalization of its exponential structure

$$\overline{G}(x) = 1 - G(x) = \mathbb{P}\{X > x\} = e^{-aF(x)}, a > 0,$$

in terms of a general monotonically increasing real function $F(x)$ can generate a large variety of generalized Weibull distributions with various types of hazard function behavior (cf. [14,17]). If we use the equivalence

$$\begin{aligned} &\overline{F}(x) = e^{-H_F(x)} \\ \Leftrightarrow\ &H_F(x) = -\ln(1 - F(x)) = -\ln(\overline{F}(x)) \end{aligned} \tag{5}$$

we can reformulate the TX-construction of Alzaatreth et al. [1] in terms of the general hazard rate principle based on a Weibull distribution by Gurvich et al. [7] as generalized Weibull construction

$$\overline{G}(x) = 1 - G(x) = \mathbb{P}\{X > x\} = e^{-H_F(x)} \tag{6}$$
$$G(x) = \mathbb{P}\{X \leq x\} = 1 - e^{-H_F(x)} \tag{7}$$

by identifying $H_F(x) = aF(x)$. Then the general TX-construction (3) looks as follows:

$$\begin{aligned} G(x) &= \mathbb{P}\{X \leq x\} = R(-\ln(1 - F(x))) \\ &= R(H_F(x)) \end{aligned} \tag{8}$$

Applying a Pareto distribution

$$F(y) = \mathbb{P}\{Y \leq y\} = 1 - \left(\frac{\theta}{y}\right)^{\alpha}, y \geq \theta > 0$$

with scale parameter $\theta > 0$ and the real tail index $\alpha > 0$ as shape parameter for a basic rv Y, we get the resulting failure rate function

$$h_F(y) = \alpha/y$$

and the cumulative failure rate function

$$H_F(y) = \alpha \ln(y).$$

We parametrize the latter by a positive scale parameter $\theta > 0$ in terms of

$$H_F(y, \theta) = \alpha \ln(\frac{y}{\theta}).$$

If we select a Weibull distribution of a nonnegative rv T

$$R(t) = \mathbb{P}\{T \leq t\} = 1 - \exp\left(-\left(\frac{t}{\gamma}\right)^c\right), \quad t \geq 0, \tag{9}$$

with the tail index $c > 0$ and the scale parameter $\gamma > 0$ as basic component, we get the corresponding Weibull-Pareto distribution

$$G(x) = R(H_F(x, \theta)) = 1 - \exp\left(-\left(\frac{\alpha}{\gamma} \ln(\frac{x}{\theta})\right)^c\right)$$

and after the substitution $\beta = \alpha/\gamma > 0$ the cdf

$$G(x) = 1 - \exp\left(-\left(\beta \ln(\frac{x}{\theta})\right)^c\right), \ x \geq \theta, \tag{10}$$

of a three parameter Weibull-Pareto distribution ($WPD(c, \beta, \theta)$). We see that we get the Pareto distribution with scale parameter θ and tail index β for the special case $c = 1$.

The cumulative failure rate associated with the $WPD(c, \beta, \theta)$ distribution function $G(x)$ is determined by

$$H_G(x) = \left(\beta \ln(\frac{x}{\theta})\right)^c$$

and yields

$$\overline{G}(x) = e^{-H_G(x)}.$$

A comparison with the cumulative failure rate function $H_W(t)$ of the Weibull distribution

$$H_W(t) = \int_0^t \frac{c}{\gamma}\left(\frac{\tau}{\gamma}\right)^{c-1} d\tau = \left(\frac{t}{\gamma}\right)^c$$

associated with T yields the resulting identities:

$$\frac{t}{\gamma} = \frac{\alpha}{\gamma} \ln(\frac{x}{\theta})$$
$$\Longleftrightarrow x = \theta \cdot e^{t/\alpha}$$

They indicate the underlying fundamental transformation

$$X = \Phi(T) = \theta \cdot e^{T/\alpha} \tag{11}$$

among the Weibull rv T and the Weibull-Pareto rv X.

2.2 Parameter Estimation of the Weibull-Pareto Model

Alzaatreth et al. [1] have derived a parameter estimation procedure for $WPD(c,\beta,\theta)$ distributions based on a modified MLE scheme. We will follow this procedure and integrate a modified Hill-type estimator of the tail index $c>0$ (cf. [2]).

Let $\mathcal{X}=\{x_1,\ldots,x_n\}$ be an iid sample of a rv X with $WPD(c,\beta,\theta)$ distribution of size $n\in\mathbb{N}$. We assume that its minimum is given by

$$x_{(1)}=x_{1,n}=\min\{x_i : i=1,\ldots,n\}$$

and occurs $n_1=\mid\{j|x_j\in\mathcal{X},x_j=x_{(1)}\}\mid<n$ times.

We are interested to estimate the parameters (c,β,θ) for the case of a heavy tailed underlying Weibull distribution with shape parameter $c>0$. As the support of the distribution class $WPD(c,\beta,\theta)$ is determined by the set $[\theta,\infty)$ a natural estimate of the scale parameter $\theta>0$ is determined by the minimal order statistics

$$\widehat{\theta}=x_{(1)}=\min_{i=1,\ldots,n}\{x_i\} \tag{12}$$

of the sample $\mathcal{X}$ (cf. [1]).

Assuming w.l.g. $\alpha=1$, we have realized in the previous section that the transformation

$$X=\theta\cdot e^Y$$

yields a Weibull-Pareto distribution $WPD(c,\beta,\theta)$ for a given rv Y with Weibull distribution and parameters $(c,1/\beta)$ (cf. [1]). Hence, we need to estimate the unknown tail index $c>0$ of the rv $Y=\ln(X/\theta)$. We can estimate it either in terms of the original sample points $\mathcal{X}=\{x_1,\ldots,x_n\}$ or the transformed data points $\{y_i=\ln\left(\frac{x_i}{x_{(1)}}\right):i=1,\ldots,n\}$. Here, we apply the scaled Hill-type estimator

$$\widehat{\tau}=\frac{1}{\widehat{c}}=\frac{\sum_{i=1}^{k_n-1}(\ln y_{n-i+1,n}-\ln y_{n-k_n+1,n})}{\sum_{i=1}^{k_n-1}\ln(\ln(n/i))-\ln(\ln(n/k_n))} \tag{13}$$

of the Weibull tail coefficient $c=1/\tau$ proposed by Beirlant et al. [2]. It is based on the logarithmic spacing of the upper k_n points of the transformed data y_i. In this estimator $\widehat{\tau}=1/\widehat{c}$ we use the ordered sample points

$$y_{1,n}\le y_{2,n}\le\ldots\le y_{n,n}$$

of these transformed data and an intermediate sequence $k_n\in\mathbb{N}$ satisfying $\lim_{n\to\infty}k_n=\infty$, $\lim_{n\to\infty}k_n/n=0$.

Using this estimate $\widehat{c}$ of the tail index c and the data points

$$y_i=\ln\left(\frac{x_i}{x_{(1)}}\right)\in[0,\infty),$$

the estimate $\widehat{\beta}$ of the shape parameter $\beta > 0$ is derived as solution of a modified system of MLE equations in the following way (cf. [1]):

$$\widehat{\beta} = \left(\frac{n - n_1}{\sum_{i:x_i \neq x_{(1)}} \left(\ln \left[\frac{x_i}{x_{(1)}} \right] \right)^{\widehat{c}}} \right)^{1/\widehat{c}} = \left(\frac{1}{n - n_1} \sum_{i:x_i \neq x_{(1)}} {y_i}^{\widehat{c}} \right)^{-(1/\widehat{c})} \tag{14}$$

Following a maximum likelihood estimation procedure, Alzaatreth et al. [1] have proposed to determine the tail parameter c by the solution of a corresponding set of nonlinear equations. Considering the computation of the estimate $\widehat{\eta} = 1/\widehat{c}$, an iterative fixed-point solver may be applied to calculate the value $\widehat{\eta}$ by the following linear transformation of the MLE problem:

$$\widehat{\eta} + \sum_{i:x_i \neq x_{(1)}} \ln \left[\ln \left(\frac{x_i}{x_{(1)}} \right) \right] = \frac{\sum_{i:x_i \neq x_{(1)}} \left[\ln \left(\frac{x_i}{x_{(1)}} \right) \right]^{1/\widehat{\eta}} \ln \left[\ln \left(\frac{x_i}{x_{(1)}} \right) \right]}{\sum_{i:x_i \neq x_{(1)}} \left[\ln \left(\frac{x_i}{x_{(1)}} \right) \right]^{1/\widehat{\eta}}}.$$

A direct formulation of the MLE approach with $y_i = \ln \left(\frac{x_i}{x_{(1)}} \right) \in \mathbb{R}$ reads as follows:

$$\begin{aligned} \widehat{c} &= \left[1 - \sum_{i:x_i \neq x_{(1)}} \frac{\ln(y_i)}{\sum_{i:x_i \neq x_{(1)}} \ln(y_i)} \cdot \frac{{y_i}^{\widehat{c}}}{\sum_{i:x_i \neq x_{(1)}} {y_i}^{\widehat{c}}} \right]^{-1} \cdot \frac{1}{\sum_{i:x_i \neq x_{(1)}} \ln(y_i)} \\ \widehat{\beta} &= \left(\frac{1}{n - n_1} \sum_{i:x_i \neq x_{(1)}} {y_i}^{\widehat{c}} \right)^{-(1/\widehat{c})} \end{aligned} \tag{15}$$

It yields a simple implementation of the fixed point method to determine the parameters $(\widehat{c}, \widehat{\beta})$.

We prefer to use the more effective estimation procedure based on (12), (13), (14). It determines the corresponding estimates of the unknown parameters (c, β, θ) of a Weibull-Pareto model in terms of a finite sample $\mathcal{X}$ by means of extreme-value theory combined with the MLE approach.

3 Estimation of a QoE Performance Metric

The quality-of-service (QoS) and the associated quality-of-experience (QoE) of advanced multimedia services in the Internet depend on the delay-loss-throughput profile induced by the underlying router and server infrastructures. It has been shown that heavy-tailed distributions such as a Gamma body enhanced by a Pareto tail model or a mixture of Weibull distributions can be applied to capture the basic delay characteristics of a transfer path that is used by packets of an advanced service process (cf. [8,12]).

Regarding the QoE performance of such end-to-end (E2E) delay dependent services, Hoßfeld et al. [10] have shown that the latter characteristic can be modelled in terms of a simple regression

$$M = f(T) = -\hat{a} \cdot \log_{10}(T + b) + d, \quad \hat{a} > 0, d > 0, \tag{16}$$

with the random delay or response time characteristic $T \geq 0$ and the QoE metric MOS described by ordinal data $M \in [0, 5]$ (see also [9]). It is reflecting the well-known Weber-Fechner law.

Rewriting the transformation (16) in terms of the function ln yields the basic transformation

$$m = f(t) = -a \cdot \ln(t + b) + d, \ a = \hat{a}/\ln(10), \tag{17}$$

among the monitored sample values $t = T(\omega)$ and the MOS metric $m = M(\omega)$ in an associated probability space $(\Omega, \mathcal{A}, \mathbb{P})$. We model the response time rv T in terms of a Weibull-Pareto distribution $WPD(c, \beta, \theta)$ with the cdf

$$\begin{aligned} G_T(t) &= \mathbb{P}\{T \leq t\} = R(H_P(t)) \\ &= 1 - \exp\left(-\left(\beta \ln(\frac{x}{\theta})\right)^c\right) \end{aligned} \tag{18}$$

(see also [8]). Then the QoE performance is described by the distribution of the MOS transformation

$$\begin{aligned} F_M(m) &= \mathbb{P}\{M \leq m\} \\ &= \mathbb{P}\{f(T) \leq m\} = \mathbb{P}\{T \leq f^{-1}(m)\} \\ &= G_T(f^{-1}(m)) = R(H_P(f^{-1}(m))) \end{aligned} \tag{19}$$

where H_P denotes the cumulative failure rate of the underlying scaled Pareto distribution of the $WPD(c, \beta, \theta)$ model and R the cdf of the Weibull model.

Inverting the transformation f in (17), we get

$$t = f^{-1}(m) = \left(e^{(d-m)/a}\right) - b \tag{20}$$

$$F_M(m) = 1 - \exp\left[-\left\{\beta \ln\left(\frac{1}{\theta}[e^{(d-m)/a} - b]\right)\right\}^c\right] \tag{21}$$

Let $p \in (0, 1)$. The quantile t_p belonging to the top $(100 \cdot p)$ % of the involved users of a sampled population with QoE MOS values $m = M(\omega)$ can be derived with

$$\lambda = 1 - p = G_T(t_p) \in (0, 1)$$

in terms of the associated quantile function

$$\begin{aligned} t_p = Q(\lambda) &= {G_T}^{-1}(1 - p) \\ &= \theta \cdot \exp\left(\frac{(-\ln p)^{1/c}}{\beta}\right) \end{aligned} \tag{22}$$

of an underlying distribution $WPD(c, \beta, \theta)$. It yields the MOS quantiles

$$m_p = -a \ln \left(\theta \cdot \exp \left(\frac{(-\ln p)^{1/c}}{\beta} \right) + b \right) + d, \quad p \in (0, 1). \tag{23}$$

The estimation of the required parameters (c, β, θ) of the Weibull-Pareto model can be derived from samples $\mathcal{T} = \{t_1, \ldots, t_m\}$ of the E2E delay characteristics based on the procedures (12)–(14) sketched in the previous section.

4 Conclusion

Considering new multimedia and Web services in the dynamically changing wired and wireless high-speed networks in the last decades, teletraffic engineering has focussed on traffic characterization at different time scales by thorough statistical techniques, traffic planning of IP flows and the quality-of-service assessment of new interactive services. Regarding the modeling of the end-to-end delay of these advanced user services, we may apply a heavy-tailed Weibull-Pareto distribution (WPD).

In this study we have first summarized the structural properties and the parameter estimation of the three-parameter WPD class. We have also indicated its relation to the general Weibull-TX class. Then we have applied the three parameter WPD distribution to determine the relevant QoE performance metric MOS regarding end-to-end delay dependent services.

In our future work we shall validate the proposed model and investigate its efficiency w.r.t. field data arising from advanced multimedia and groupware services (cf. [9,10,21]).

References

1. Alzaatreh, A., Famoye, F., Lee, C.: Weibull-pareto distribution and its applications. Commun. Stat. Theory Methods **42**(9), 1673–1691 (2013)
2. Beirlant, J., Teugels, J., Vynckier, P.: Practical Analysis of Extreme Values. Leuven University Press, Leuven (1996)
3. Bolotin, V.A., Levy, Y., Liu, D.: Characterizing data connection and messages by mixtures of distributions on logarithmic scale. In: Key, P., Smith, D. (eds.) Teletraffic Engineering in a Competitive World, vol. 3b, pp. 887–896. Elsevier, Amsterdam (1999)
4. Center for Applied Internet Data Analysis (CAIDA) (2018). http://www.caida.org
5. Embrechts, P., Klüppelberg, C., Mikosch, T.: Modelling Extremal Events: for Insurance and Finance. Springer, Berlin (1997). https://doi.org/10.1007/978-3-642-33483-2
6. Fasbender, A.: Messung und Modellierung der Dienstgüte paketvermittelter Netze. Ph.D. thesis, Informatik IV, RWTH Aachen, Germany (1998)
7. Gurvich, M.R., Dibenedetto, A.T., Rande, S.V.: A new statistical distribution for characterizing the random strength of brittle materials. J. Mater. Sci. **32**, 2559–2564 (1997)

8. Hernandez, J.-A., Phillips, I.W.: Weibull mixture model to characterise end-to-end Internet delay at coarse time-scales. IEE Proc. Commun. **153**(2), 295–304 (2006)
9. Hoßfeld, T., Schatz, R., Krieger, U.R.: QoE of YouTube video streaming for current internet transport protocols. In: Fischbach, K., Krieger, U.R. (eds.) Measurement, Modelling, and Evaluation of Computing Systems and Dependability and Fault Tolerance. LNCS, pp. 136–150. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-05359-2_10
10. Hoßfeld, T., Varela, M., Heegaard, P.E., Skorin-Kapov, L.: QoE Analysis of the Setup of Different Internet Services for FIFO Server Systems. In: German, R., Hielscher, K.-S., Krieger, U.R. (eds.) MMB 2018. LNCS, vol. 10740, pp. 234–248. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-74947-1_16
11. Krieger, U.R., Markovitch, N.M., Vicari, N.: Analysis of world wide web traffic by nonparametric estimation techniques. In: Goto, K., Hasegawa, T., Takagi, H., Takahashi, Y. (eds.) Performance and QoS of Next Generation Networking, pp. 67–83. Springer, London (2001). https://doi.org/10.1007/978-1-4471-0705-7_4
12. Markovitch, N.M., Krieger, U.R.: The estimation of heavy-tailed distributions, their mixtures and quantiles. Comput. Netw. **40**(3), 459–474 (2002)
13. Nabe, M., Murata, M., Miyahara, H.: Analysis and modelling of World Wide Web traffic for capacity dimensioning of internet access lines. Perform. Eval. **34**, 249–271 (1998)
14. Nadarajah, S., Kotz, S.: On some recent modifications of Weibull distribution. IEEE Trans. Reliab. **54**(4), 561–562 (2005)
15. Nadarajah, S., Kotz, S.: On the extended Burr XII distribution. Hydrol. Sci. J. **51**(6), 1203–1204 (2006)
16. Nadarajah, S., Cordeiro, G.M., Ortega, E.M.M.: The exponentiated Weibull distribution: a survey. Stat. Pap. **54**(3), 839–877 (2013)
17. Pham, H., Lai, C.-D.: On recent generalizations of the Weibull distribution. IEEE Trans. Reliab. **56**(3), 454–458 (2007)
18. Rinne, H.: The Weibull Distribution: a Handbook. Chapman & Hall/CRC Press, Boca Raton (2009)
19. Shao, Q., Wong, H., Xia, J., Ip, W.-C.: Models for extremes using the extended three-parameter Burr XII system with application to flood frequency analysis. Hydrol. Sci. J. **49**(4), 685–702 (2004)
20. Sigman, K.: Appendix: a primer on heavy-tailed distributions. Queueing Syst. **33**, 261–275 (1999)
21. Stadler, H., Großmann, M., Krieger, U.R.: Design of a secure mobile business communication platform utilizing next generation web technologies. In: Proceedings of the 13th IEEE International Conference on e-Business Engineering (ICEBE) 2016, Macau, China, 4–6 November 2016 (2016)
22. Tahir, M.H., Zubair, M., Mansoor, M., Cordeiro, G.M., Alizadeh, M., Hamedani, G.G.: A New Weibull-G Family of Distributions. Hacettepe University Bulletin of Natural Sciences and Engineering Series B: Mathematics and Statistics, March 2015
23. Tran-Gia, P., Vicari, N.: COST-257 - Impacts of New Services on the Architecture and Performance of Broadband Networks. COST-257 Final Report (2000). https://www.semanticscholar.org/paper/COST-257-Final-Report-Impacts-of-New-Services-on-of-Tran-Gia-Vicari/d80907855878b3c1737b61ffe8907589ab3af9f8

Communication Capabilities of Wireless M-BUS: Remote Metering Within SmartGrid Infrastructure

Pavel Masek[1(✉)], David Hudec[1], Jan Krejci[1], Aleksandr Ometov[2], Jiri Hosek[1], and Konstantin Samouylov[3]

[1] Department of Telecommunications, Brno University of Technology, Brno, Czech Republic
masekpavel@vutbr.cz
[2] Laboratory of Electronics and Communications Engineering, Tampere University of Technology, Tampere, Finland
[3] Applied Probability and Informatics Department, Peoples' Friendship University of Russia (RUDN University), Moscow, Russia

Abstract. In today's landscape of utility management, the contribution of Internet of Things (IoT) to smart grids has acquired extensive potential. IoT paves a way to virtually control every smart device in almost every domain of society. Contrariwise, the smart grid networks attracted the attention of the universal research community. The idea of merging IoT with smart grid together demonstrates enormous potential. In this work, we investigate the suitability of Wireless M-BUS communication protocol for possible adoption in remote metering by evaluating possible communication range and system stability in future housing estate represented by university campus made of steel and concrete – this living area acts well when it comes to wireless transmissions. Measurements were executed by means of constructed prototype sensor devices utilizing the frequency 868 MHz which is the frequency by far the most used by WM-BUS devices in Europe.

Keywords: Wireless M-BUS · Remote reading · M2M
Wireless communication · Industry 4.0

1 Introduction

In 2017, the number of smart devices capable to transmit data through the network infrastructure reached 8.4 billion. By 2020, the studies indicate about 30 billion smart objects capable to establish connection without the human interaction [1,2]. Following the definition given by European Commission in document Internet of Things – An action plan for Europe, Internet of Things (IoT) stands for "*a network of interconnected computers towards a network of interconnected objects*". In the context of Smart Grid (SG) landscape, IoT opens the doors for a promising future enabled by smart analytics. No doubt this is only feasible

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 31–42, 2018.
https://doi.org/10.1007/978-3-319-99447-5_4

to realize owing the analytics data provided from the end users towards utility provider(s). At the end of the day, users' data could potentially enhance the efficiency as well as reduce congestion in the Smart Grid networks [3].

Focusing on the electricity, the transformation of the legacy electric power grid into intelligent bidirectional communication systems had paved the way to the Smart Grid of the "future" where new intelligent grids are supposed to enable interconnections between the already implemented SCADA systems and future implementations. The utilization of smart meters in all residential, commercial, and industrial places allows the utility provider(s) to gain the knowledge of customers' behavior on daily basis [4]. Smart meters, in general, provide end-users with the ability to interact with the utilities and wirelessly monitor, e.g., power consumption providing the assistance to reach the goal of reducing the bills.

1.1 Technologies for Smart Grid

In 2018, we can list many IoT-based technologies which are ready to be utilized for the need for SG applications. One hand, many communication technologies are possible to use. On the other hand, the current situation does not offer guidelines towards the proper association between technology and SG application [5]. Focusing on IoT technologies, they are used mainly for long-range data transmissions [6]. The SG systems require advanced wireless technologies in comparison with the wired technologies, e.g., Power Line Communication (PLC), optical communication which takes place in case of scenarios when interference occurs or the metering devices are placed in "deep indoor", i.e., the signal attenuation of 60 dB.

Information streams of data within Smart Grid infrastructure can be treated in two ways: (i) the stream between all the smart meters connected in star topology where the Machine-Type Communication Gateway (MTCG) acts as the connecting point for all devices [7]; (ii) data streams between the MTCGs and remote control centers operated at the side of utility. In this work, we focus on the first type of communication where the smart device communicates directly with the MTCG. Owing to the tight and long-term cooperation with industry partners in Czech Republic and Austria, it has been recognized that Wireless M-BUS (WM-BUS) communication protocol is used very often as an alternative for PLC technology in case of mid-range communication indoor scenarios, i.e., for distances of up to 100 m[1].

WM-BUS is based on an open standard for automatic meter readout. The design of the protocol implies a battery-powered communicating device to operate for up to 10 years autonomously. This requirement can be achieved if the radio management module is switched to the low-power mode for as long as possible, and hence the awakening and transmitting procedures are managed and

[1] See the first complete smart metering project in Austria – a complete model solution for whole Austria, Kamstrup, 2017: https://www.kamstrup.com/en-us/case-stories/electricity-casestories/case-telekom-austria-system-at.

optimized – the summary of standardization activities related to Smart Grid is given in Table 1. Based on the WM-BUS standard, the communication initiation node is always the smart meter and never the concentrator, which is more suitable to extend the lifespan of the sensor battery [8,9].

Table 1. Standardization activities in support of Industrial IoT (IIoT)

<table>
<tr><th colspan="2">Protocol</th><th>DDS</th><th>CoAP</th><th>AMQP</th><th>MQTT</th><th>XMPP</th><th>HTTP, REST</th><th>SIP</th><th>WM-BUS</th></tr>
<tr><td colspan="2">Service Discovery</td><td colspan="4">mDNS</td><td colspan="3">DNS-SD</td><td rowspan="5">Proprietary</td></tr>
<tr><td rowspan="4">Infrastr. Protocols</td><td>Routing Protocols</td><td colspan="7">RLP</td></tr>
<tr><td>Network Layer</td><td colspan="5">6LoWPAN</td><td colspan="2">IPv4/IPv6</td></tr>
<tr><td>Link Layer</td><td colspan="7">IEEE 802.15.4, IEEE 802.11, IEEE 802.3</td></tr>
<tr><td>Physical Layer</td><td>LTE-A; NB-IoT</td><td colspan="2">EPC Global</td><td>IEEE 802.15.4</td><td>Z-Wave</td><td>IMS</td></tr>
</table>

1.2 Wireless M-BUS in IIoT

M-Bus (wired) was developed and first introduced in the early 1990s. It was further extended wireless in 2005 (when the first draft of the EN 13757-4 was published, approved a year later [10]), which is 5 years before the concepts of the IoT and Industry 4.0 started gaining popularity in 2011 and even longer before they captured attention of the mass market in 2014 [11]. Industrie 4.0 is a term coined by the German Federal Government to optimize industrial production and provide smart manufacturing solutions [12]. Accordingly, Wireless M-Bus represents a solid competitor to protocols and networks tailored just for the IIoT, such as Sigfox, LoRaWAN (public or private implementations), or Narrowband-IoT [13]. Since it sets very similar goals (sensor-independence, battery-longevity, meter-automation), but has a few years of advantage, it may be even better established, settled, and stabilized than most of its counterparts [14,15].

The WM-BUS network topology represents a star, where one or more measuring nodes transmit(s) the data to the aggregator acting as a server. The latter always senses the wireless medium for incoming connections and subsequent data collection. WM-BUS can operate in six communication modes representing specific applications detailed in Table 2. First three modes (i.e., S, T, and R) correspond to the transfer speeds, which are further divided into modes 1 and

2 for unidirectional or bidirectional communication. The remaining three modes (i.e., N, C, and F) are supported only by specific devices [9]:

- In frequent transmit mode (T), the meter sends data periodically or whenever a packet is available. Sub-mode $T1$ defines power saving operation, in which the device transmits to the aggregator and immediately enters power saving mode without waiting for the ACK.
- Stationary mode (S) is designed for unidirectional or bidirectional communication between the stationary or mobile devices. It has three sub-modes, $S1$, $S1M$, and $S2$. Sub-mode $S1$ is for unidirectional communication without the ACK from a server. This mode is primarily to handle the "daily" data transmissions. Sub-mode $S1M$ supports bidirectional communication in predefined cycles without the need for the device to wake up.
- In frequent receive mode (R), the meter is not sending the data periodically but instead is waiting for the aggregation request. Most of the time, the meter is in the power saving mode and awakens only over the predefined intervals for the packet reception. If no valid wake-up frame is received, the meter reenters the power saving mode.

Table 2. WM-Bus protocol transfer modes

	Transfer type	Frequency	Cod. scheme	Speed
S	Stationary	868 MHz	Manchester	32768 kbps
T	Frequent transmit	868 MHz	Manchester 3 out of 6	100 kbps
R	Frequent receive	868 MHz	Manchester	4.8 kbps
N	Narrowband	169 MHz	NRZ	
C	Compact	868 MHz	Manchester	50 kbps
F	Frequent transmit and receive	433 MHz	NRZ	-

Wireless M-Bus Frame Structure. After the protocol operation modes and the topology were introduced, we focus on the data frame structure to make the reader more familiar with the WM-Bus operational details. This section describes the WM-Bus communication phases (handshakes). In the first step, an over-the-top application at the application layer of WM-Bus sends its data to the RF module as a packet – as demonstrated below [9]:

1 Byte	1 Byte	n Bytes
Length	CI	AppLayer

In the next step, the radio module adds the following fields: (i) Control field; (ii) Manufacturer identification; (iii) Unique address based on parameters saved

1 Byte	1 Byte	2 Bytes	6 Bytes	1 Byte	n Bytes	1 Byte
Length	C	ManID	Address	CI	AppLayer	RSSI

in the memory of the modules; and (iv) Optional information about received signal strength (RSSI). Therefore, the packet now has the following headers:

This packet is further encrypted (with AES-128 by default) and transmitted. If the connection is implemented as tunneling (P2P connection) between two Wireless M-Bus modules, the address field, and the affiliated info is optional – thus allowing for simpler packet structure by sending only the RSSI:

1 Byte	1 Byte	n Bytes	1 Byte
Length	CI	AppLayer	RSSI

The AppLayer field is defined by the M-Bus application layer, which is used as a transition mechanism for the communication from link layer to higher layers. It uses the OMS 3.0.1 specification [16] derived from the EM 13757-4 standard for wireless communication [8]. In this work, we focus primarily on one of the WM-Bus equipped devices – IQRF radio modules. For our implementation, we selected the IQRF TR-72D-WMB module (see Fig. 1) [17]. This module is from the programmable IQRF technology line produced by MICRORISC that allows implementing the Wireless M-Bus or a similar protocol on the fly. It is further equipped with SPI and UART interfaces for communication with the master devices. Its block diagram is depicted in Fig. 2.

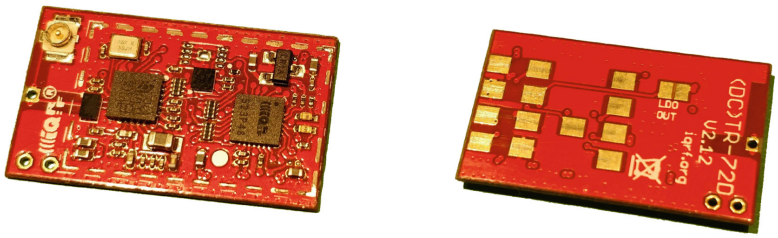


Fig. 1. Utilized IQRF TR-72DA-WMB module.

The said module supports following WM-BUS operating modes $S1$, $S2$, $T1$, and $T2$. The power voltage is in the range from 3.1 to 5.3 V with the maximal current of 1 µA in the sleep mode and 8–22 mA in the transmit mode. This value is based on the output power setting, which caps at 12.5 W. The module itself has support for 169, 433, and 868 MHz frequency bands, whereas the chip supports operation in one of the following modes [9]:

– *Meter*: The module can be connected via UART to the micro-controller, which serves as a data handler, i.e., it could be utilized to build the proprietary measurement devices based on WM-Bus protocol.
– *Multi-Utility Controller*: The module serves as the communication device for the meter data readout. Current firmware only supports bidirectional communication with meters in S and T modes and is still under development.
– *Sniffer*: The module captures all the available communication in the selected transmission mode. Owing to the implementation of the Wireless M-Bus protocol, it can also capture and decrypt the encrypted communication.

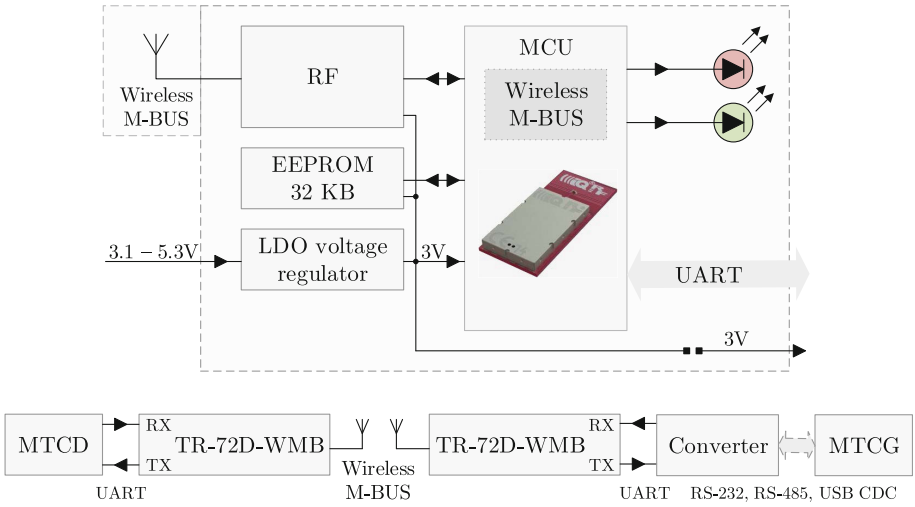


Fig. 2. Block diagram of TR-72DA-WMB module [9].

1.3 Main Contribution

In this paper, we expand our vision of Machine-to-Machine (M2M) communication for embedded devices basing on our previously developed industrial projects [9,18]. In particular, we consider the situation, when low-cost IQRF radio modules can be configured in the role of Wireless M-Bus receivers. Inspired by that, we analyze and implement a real-world scenario, where the IQRF TR-72DA-WMB module [17] becomes a part of the MTCG device and receives the Machine-Type Communication (MTC) data sent via the WM-BUS communication protocol from electricity meters. What has changed compared to our last trial is the software generator at the side of MTC which was implemented completely from scratch. It is a Java program, which can generate Wireless M-BUS data from both graphical and command-line interfaces. It allows for precise protocol data unit specification and can send these message definitions in the form of telegrams to the Wireless M-BUS network using a supported hardware transceiver. Two device variations supported by the application are a standalone Wireless M-BUS module of IQRF TR-72D-WMB and more complex solution

using the UniPi Neuron S103 board [19]. The software-hardware combination, created as a ready-to-use generator solution represents a powerful option in the area of testing Wireless M-BUS networks.

The remainder of this paper is structured as follows. In Sect. 2, we take a closer look at the measurement scenario where the Wireless M-BUS communication protocol is utilized. Going further, the Sect. 3 discusses the results obtained from designed test scenario as well as lessons learned originating from development phase.

2 Prototyped Industrial IoT Scenario

To demonstrate the functionality of previously created solution, see the [9], we have recently completed a full-scale implementation of WM-BUS. We investigated the suitability of Wireless M-BUS communication protocol for potential adoption in remote metering by evaluating possible communication range and connection stability in future housing estate represented by university campus made of steel and concrete – this living area acts well when it comes to wireless transmissions. Measurements were executed employing constructed prototype sensor devices utilizing the frequency 868 MHz, i.e., transmitting in Industrial, Scientific, and Medical (ISM) radio band.

2.1 Selected HW Devices

We have selected Raspberry Pi3 with I/O shield for our implementation. The connection between the Raspberry Pi and the mentioned shield is realized via the 26 pin board. The UART bus is escorted to the SIM slot on the shield, which is prepared for the connection. It is equipped with the IQRF TR-72D-WMB module. The said module supports WM-Bus S1, S2, T1, and T2 operating modes. The power voltage is in the range from 3.1 to 5.3 V with the maximal current of 1 uA in the sleep mode and 8–22 mA in the transmit mode. This value is based on the output power setting, which caps at 12.5 W. The module has support for 169, 433, and 868 MHz frequency bands, whereas the chip supports operation in one of the following modes: (i) Meter, (ii) Multi-Utility controller, and (iii) Sniffer.

2.2 Measurement Methodology

For the purpose of our trial, we concentrated on communication distance between two devices transmitting data indoor: (i) Wireless M-BUS transmitter running our proprietary software acting as M2M data generator, and (ii) universal WM-BUS USB Adapter AMB8465-M [20] in role of the receiver i.e., MTCG device – the integrated microprocessor controls the entire data communication as well as block- and checksum-creation. Data packets are built and transmitted according to EN13757-4. The USB-adapter is versatile configurable and supports all operating modes according to the wireless M-BUS specification. The quality of the radio link can be assessed by using the measured field strength (RSSI value).

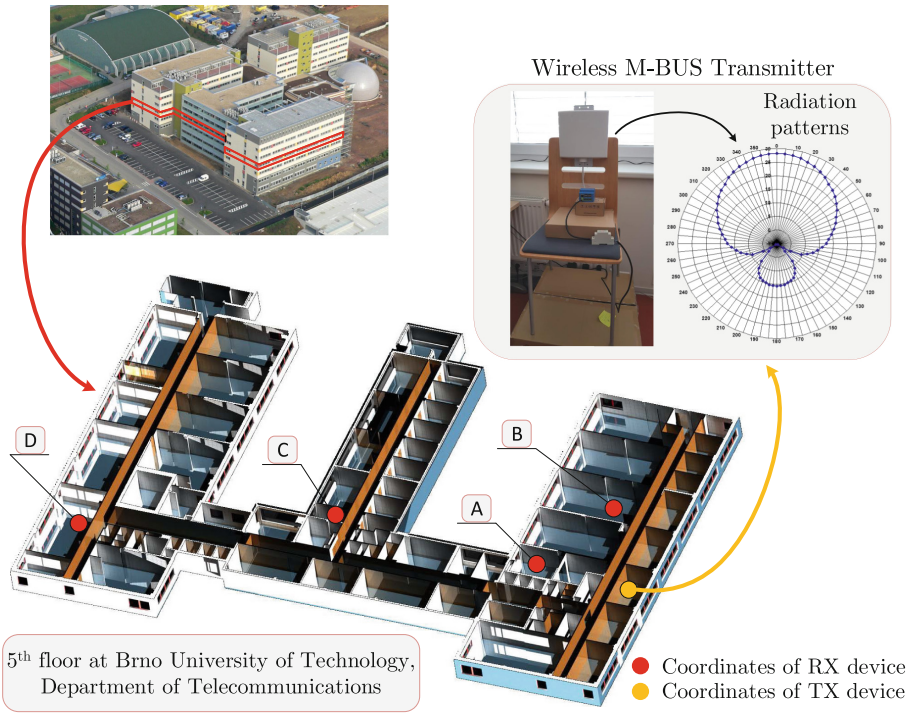


Fig. 3. Implemented WM-BUS scenario at Brno University of Technology, Czech Republic. (Color figure online)

The realized WM-BUS scenario is shown in Fig. 3 where all the important points are displayed. The "orange circle" represents the Wireless M-BUS data generator equipped by the external planar antenna. The frequency range of the antenna is 824 MHz to 896 MHz which suits well to our scenario where the devices communicate at 868 MHz. The gain added to the system by the antenna (RF part) is set to 6.7 dBi. Going further, the "orange circles" stand for the positions where we placed the RX device and tested one-by-one the parameters of the communication link. The list of possible combinations of power levels on both communicating sides is shown in Table 3. At each location (Location A, B, C, and D), all combinations of RF levels discussed in Table 3 were performed. The data transmission consisted of sending 15 telegrams in a row with the time interval set to 20 s.

Table 3. List of RF levels for both the WM-BUS transmitter and receiver.

	RF power levels [dBm]
IQRF TR-72D-WMB (Transmitter)	−30; −12; 0; +15
AMB8465-M (Receiver)	−5; 0; Max

3 Lessons Learned and Conclusions

During the development and implementation phases of our work, we have solved a number of challenges and drawbacks: (i) Raspberry Pi 3 uses different access to the serial interface. Hence, modifications on the boot level were needed; (ii) communication between wireless IQRF TR-72D-WMB module and the processing unit via UART had to be redesigned for the target case; (iii) the sniffed packets required re-encryption with the AES key of the IQRF module and were decrypted again to access the data (see our previous work [9] where the process of unencrypted and encrypted communication and following SW implementation is described in detail); (iv) the implementation of the data packets is not identical across the manufacturers and therefore for each device the sniffed data needed to be analyzed separately.

Table 4. Summary of configured RF power levels on both devices together with number of successfully received telegrams for all measured locations. The combinations of RF levels on the side of TX/RX which leads to data transmissions without packet loss are highlighted by grey.

	Location A			**Location B**			**Location C**			**Location D**		
RF Levels [dBm]	**TX**	**RX**	**No.**	**TX**	**RX**	**No.**	**TX**	**RX**	**No.**	**TX**	**RX**	**No.**
	-30	-5	6	-30	-5	0	-30	-5	0	-30	-5	0
	-12	-5	13	-12	-5	14	-12	-5	0	-12	-5	0
	0	-5	15	0	-5	15	0	-5	0	0	-5	0
	15	-5	15	15	-5	15	15	-5	8	15	-5	11
	-30	0	10	-30	0	0	-30	0	0	-30	0	0
	-12	0	15	-12	0	10	-12	0	0	-12	0	0
	0	0	15	0	0	15	0	0	0	0	0	0
	15	0	15	15	0	15	15	0	9	15	0	11
	-30	Max.	11	-30	Max.	0	-30	Max.	0	-30	Max.	0
	-12	Max.	15	-12	Max.	4	-12	Max.	0	-12	Max.	0
	0	Max.	14	0	Max.	15	0	Max.	0	0	Max.	0
	15	Max.	14	15	Max.	13	15	Max.	12	15	Max.	11

In the course of our development, we have performed practical measurements in all of the above-mentioned locations, see Fig. 3. As the measurements took place indoor, types of used materials play the critical role for the signal propagation. Owing to the possibility to use information from the drawing documentation of the building, the following materials are used: (i) reinforced concrete, (ii) clay block masonry, (iii) autoclaved aerated concrete, (iv) gypsum boards, (v) thermal insulation, and (vi) acoustic insulation.

Going towards the practical measurements, on top of the redesigned HW, an entirely new software layer has been introduced to generate WM-BUS data

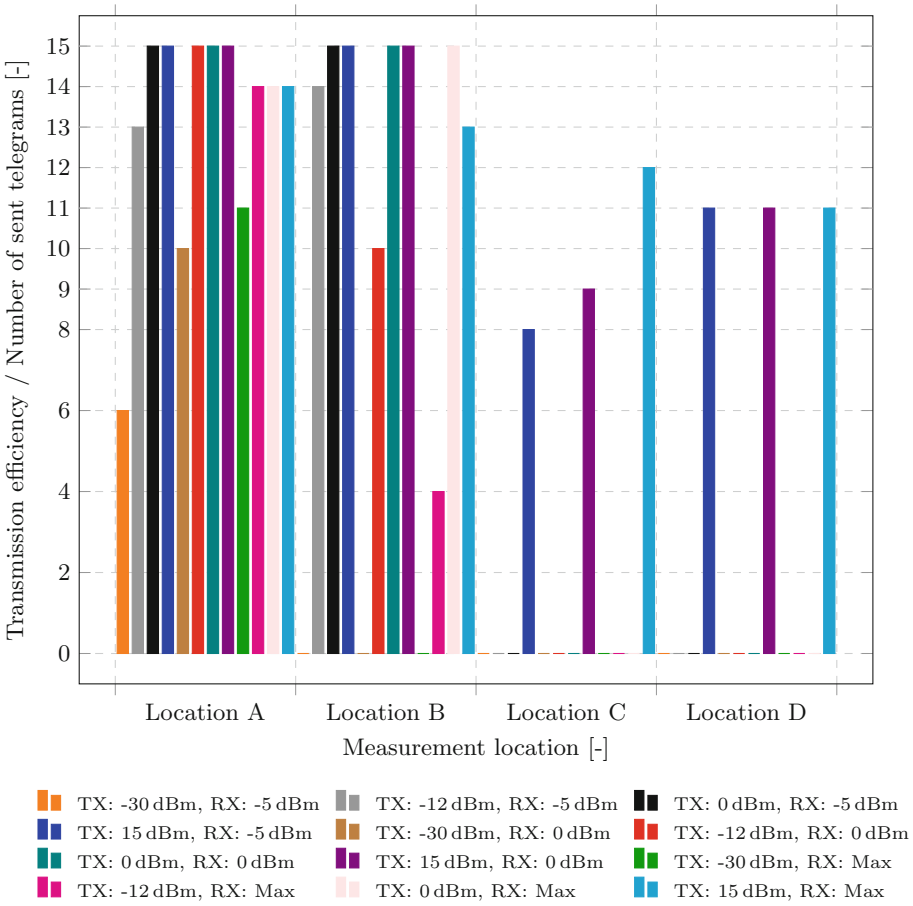


Fig. 4. Transmission efficiency – dependence on RX and TX output RF power levels.

traffic (in case of this trial, the data representing the electricity meter is sent periodically) at the side of the WM-BUS transmitter (in the role of a metering device; MTCD) towards the WM-BUS data concentrator (MTCG). The complexity of this solution is further highlighted by the fact that the transmitted data can also be encrypted; re-encryption with the AES key of the IQRF module is implemented.

The obtained data is shown in Table 4 and depicted in Fig. 4. One can see the results in case of some TX/RX RF level combinations do not follow the theoretical expectations. This behavior has two possible explanations: (i) the measurements were conducted during the working hours at the university. Therefore, the university staff and students influenced the signal propagation. On the other hand, those results stand for the real conditions expected to be met in case of remote metering, e.g., housing estate; (ii) the utilized frequency band is free to

use which together with the unique rooms acting as obstacles (EMC chamber, acoustic chamber, etc.) causes unexpected signal propagation while sending the data at 868 MHz.

As mentioned before, this paper was intended as a proof-of-concept hardware implementation that significantly reduces the cost of Wireless M-Bus based communication platform. In our future work, we are planning to expand the functionality of our platform by adding support for more smart-meter vendors, as well as to further work with smart and connected IIoT/Industry 4.0 enablers.

Acknowledgment. The described research was supported by the National Sustainability Program under grant LO1401. For the research, the infrastructure of the SIX Center was used. This work has been developed within the framework of the COST Action CA15104, The Inclusive Radio Communication Networks for 5G and beyond (IRACON).

References

1. Nordrum, A.: Popular internet of things forecast of 50 billion devices by 2020 is outdated. In: IEEE Spectrum, vol. 18 (2016)
2. Masek, P., et al.: Advanced wireless M-Bus platform for intensive field testing in industry 4.0-based systems. In: Proceedings of IEEE European Wireless Conference, vol. 1–6 (2018)
3. Bhatt, J., Shah, V., Jani, O.: An instrumentation engineer's review on smart grid: critical applications and parameters. Renew. Sustain. Energy Rev. **40**, 1217–1239 (2014)
4. Akpakwu, G.A., Silva, B.J., Hancke, G.P., Abu-Mahfouz, A.M.: A survey on 5G networks for the internet of things: communication technologies and challenges. IEEE Access **6**, 3619–3647 (2018)
5. Middleton, P., et al.: Forecast: internet of things-endpoints and associated services, worldwide, 2015. Gartner Inc., Stamford, CT, USA, Tech. Rep. G, vol. 290510, p. 57 (2015)
6. Ometov, A., et al.: System-level analysis of ieee 802.11ah technology for unsaturated MTC traffic. Int. J. Sensor Netw. **26**(4), 269–282 (2018)
7. Hosek, J., et al.: A SyMPHOnY of integrated IoT businesses: closing the gap between availability and adoption. IEEE Commun. Mag. **55**(12), 156–164 (2017)
8. European Committee for Standardization: EN 13757–4. Communication systems for meters and remote reading of meters - Part 4: Wireless meter readout (Radio Meter reading for operation in the 868–870 MHz SRD band) (2003). http://oldfjarrvarme.unc.se/download/1309/fj
9. Zeman, K., et al.: Wireless M-BUS in industrial IoT: technology overview and prototype implementation. In: Proceedings of 23th European Wireless Conference; Proceedings of European Wireless 2017, pp. 1–6. VDE (2017)
10. Arian, M., Soleimani, V., Abasgholi, B., Modaghegh, H., Gilani, N.S.: Advanced metering infrastructure system architecture. In: 2011 Asia-Pacific Proceedings of Power and Energy Engineering Conference (APPEEC), pp. 1–6. IEEE (2011)
11. Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., Ayyash, M.: Internet of things: a survey on enabling technologies, protocols, and applications. IEEE Commun. Surv. Tutor. **17**(4), 2347–2376 (2015)

12. Lade, P., Ghosh, R., Srinivasan, S.: Manufacturing analytics and industrial internet of things. IEEE Intell. Syst. **32**(3), 74–79 (2017)
13. Wang, H., Fapojuwo, A.O.: A survey of enabling technologies of low power and long range machine-to-machine communications. IEEE Commun. Surv. Tutor. **19**(4), 2621–2639 (2017)
14. Spinsante, S., Squartini, S., Gabrielli, L., Pizzichini, M., Gambi, E., Piazza, F.: Wireless m-bus sensor networks for smart water grids: analysis and results. Int. J. Distrib. Sensor Netw. **10**(6), 579271 (2014)
15. Kaloudiotis, E.: A 169 MHz and 868 MHz Wireless M-Bus Based Water and Electricity Metering System (2015)
16. OMS-Group: The Open Metering System specification (2016). http://oms-group.org/en/oms-group/about-oms-group/
17. IQRF - Technology for Wireless, TR-72D-WMB series (2016). http://www.iqrf.org/products/transceivers/tr-72d-wmb
18. Masek, P., et al.: Implementation of true IoT vision: survey on enabling protocols and hands-on experience. Int. J. Distrib. Sensor Netw. **12**(4), 8160282 (2016)
19. UniPi.technology, subsidiary of Faster CZ spol. s r.o., UniPi Neuron S103 (2016). https://www.unipi.technology/unipi-neuron-s103-p2
20. Wurth Elektronik eiSos GmbH & Co. KG: Wireless M-Bus USB Adapter 868 MHz (2018). https://www.amber-wireless.com/en/amb8465-m.html

On a Queueing System with Processing of Service Items Under Vacation and N-policy

V. Divya[1], A. Krishnamoorthy[2(✉)], and V. M. Vishnevsky[3]

[1] Department of Mathematics, N.S.S. College, Cherthala 688556, India
divyavelayudhannair@gmail.com
[2] Department of Mathematics, CMS College, Kottayam 686001, India
achyuthacusat@gmail.com
[3] V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, 65 Profsoyuznaya Street, Moscow 117997, Russia
vishn@inbox.ru

Abstract. In this paper we assume that the customers arrive at a single server queueing system according to Markovian Arrival process. When the system is empty, the server goes for vacation and produces inventory for future use during this period. The maximum processed inventory at a stretch is L. The inventory processing time follows phase type distribution. These are required for the service of customers-one for each customer. The server returns from vacation when there are N customers in the system. The service time follows two distinct phase type distributions depending on whether there is processed items or no processed item available at service commencement epoch. We analyse the distributions of time till the number of customers hit N or the inventory level reaches L, idle time, the distribution of time until the number of customers hit N and also the distribution of the number of inventory processed before the arrival of first customer. Also we provide the distribution of a busy cycle, LSTs of busy cycles in which no item is left in the inventory and at least one item is left in the inventory. We perform some numerical experiments to evaluate the expected idle time, standard deviation and coefficient of variation of idle time of the server.

Keywords: Vacation · Inventory · N-policy

V. Divya—Research is supported by the UGC, Govt. of India, under Faculty Development Programme (Grant No.F.No.FIP/12th Plan/KLKE008 TF 04) in Department of Mathematics, Cochin University of Science and Technology, Cochin-22.
A. Krishnamoorthy—Emeritus Fellow (EMERITUS-2017-18 GEN 10822(SA-II)), UGC, Govt. of India and Indo-Russian project: INT/RUS/RSF/P-15, funded by DST.
V. M. Vishnevsky—Research supported by the Russian Science Foundation and the Department of Science and Technology (India) via grant No. 16-49-02021 for the joint research project by the V.A. Trapeznikov Institute of Control Science and the CMS college Kottayam.

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 43–57, 2018.
https://doi.org/10.1007/978-3-319-99447-5_5

1 Introduction

In a vacation queueing system, the server may not be available for a period of time due to several reasons like server working on some supplementary jobs or doing some maintenance work, server's failure that interrupt customer service or simply taking a break. Levy and Yechiali [5] introduced the concept of server vacation. A considerable number of work in this area were surveyed by Doshi in [2]. More studies on vacation models could be found in Takagi [8] and in Tian and Zhang [9]. Kazimirsky [4] studied BMAP/G/1 queue with infinite buffer and service time distribution depending on number of processed items: When customers are absent in the system, the server begins to produce items that are put to the storehouse until the storehouse capacity is reached or a group of customers arrives. When a group of customers arrives, the item processing stopped and the service of the customers begins. Service time of a customer depends on the amount of items at the storehouse at the beginning of the service. After the service of the customer is completed, it departs from the systems and, if its service is begun with nonempty storehouse, the number of items at the storehouse decreases by one unit at a service completion epoch. In this queueing model, he considered the systems with the possibility of preliminary service. An efficient algorithm for calculating the stationary queue length distribution was proposed, and Laplace-Stieltjes transform of the sojourn time was derived. Also he proved Little's law and an associated optimization problem was analyzed.

The motivation for our work is a paper by Hanukov et al. [3]. In their model, they studied a single server queue in which the service consists of two independent stages. The first stage can be performed even in the absence of customers, whereas the second stage requires the customer to be present. When the system is devoid of customers, the server produces an inventory of first stage called 'preliminary' services, which is used to reduce customer's overall sojourn times. Hence in this model customer will not have to wait for the entire service to be carried out from the beginning, provided processed item is available at the time the customer is taken for service. Such customers have a shorter service time in comparison to those who encounter the system with no processed item when taken for service. Yadin and Naor [10] introduced the concept of N-policy in which the server turns on with the accumulation of N or more customers and turns off when the system is empty. This has the advantage that the length of a busy period becomes larger when server is activated on accumulation of N or more customers, thereby bringing down the expected cost incurred per unit time.

In this paper we consider a single server queueing system in which customers arrive according to Markovian Arrival process. When the system is empty, the server goes for vacation and produces inventory for future use during this period. The maximum inventory level is L. The inventory processing time follows phase type distribution. The server returns from vacation when there are N customers in the system. The service time follows two distinct phase type distributions according to whether there is no processed item or there are processed items

at the beginning of service. Each customer requires an item from inventory for service which is exclusively used for the service of that particular customer only.

The rest of the paper is arranged as follows. The mathematical formulation is given in Sect. 2. Section 3 provides steady state analysis of the model and also contain some important distributions. Some numerical results are discussed in Sect. 4.

Notations and abbreviations used in the sequel:

- $\mathbf{e}(a)$: Column vector of 1′s of order a.
- $\mathbf{e}$: Column vector of 1′s of appropriate order.
- $CTMC$: Continuous time Markov chain.
- I_a: identity matrix of order a.
- $\mathbf{e}_a(b)$: column vector of order b with 1 in the ath position and the remaining entries zero.
- MAP: Markovian Arrival Process.
- LST: Laplace-Steiltges Transform.
- $LIQBD$: Level Independent Quasi-Birth and-Death.

2 Model Description and Mathematical Formulation

We assume that customers arrive at a single server queueing system according to MAP with representation (D_0, D_1) of order n. When the system is empty, the server goes for vacation and produces inventory for future use during this period. The maximum inventory level permitted is L. The inventory processing time follows phase type distribution PH(α, T) of order m_1. These are required for the service of customers-one for each customer.The server returns from vacation when N customers accumulate in the system. The service time follows PH(β, S) of order m_2 when there is no processed item and it follows PH(γ, U) of order m_3 when there are processed items.

Let $Q^* = D_0 + D_1$ be the generator matrix of the type II arrival process and $\boldsymbol{\pi}^*$ be its stationary probability vector. Hence $\boldsymbol{\pi}^*$ is the unique (positive) probability vector satisfying

$$\boldsymbol{\pi}^* Q^* = 0, \ \boldsymbol{\pi}^* \boldsymbol{e} = 1.$$

The constant $\beta^* = \boldsymbol{\pi}^* D_1 \boldsymbol{e}$, referred to as *fundemental rate*, gives the expected number of arrivals per unit of time in the stationary version of the MAP. It is assumed that the arrival process is independent of the inventory processing and service process.

2.1 The QBD Process

The model described in Sect. 1 can be studied as a LIQBD process. First we introduce the following notations:

At time t:

$N(t)$: the number of customers in the system at time t,

$I(t)$: the number of processed inventory

$$J(t) = \begin{cases} 0, when\, the\, server\, is\, on\, vacation \\ 1, when\, the\, server\, is\, busy\, serving\, a\, customer \end{cases}$$

$K(t)$: the phase of the inventory processing/service process
$M(t)$: the phase of arrival of the customer.

It is easy to verify that $\{(N(t), I(t), J(t), K(t), M(t)) : t \geq 0\}$ is a LIQBD with state space: (i) no customer in the system
$l(0) = \{(0, i, 0, k_1, l) : 0 \leq i \leq L-1, 1 \leq k_1 \leq m_1, 1 \leq l \leq n\} \cup \{(0, L, 0, l) : 1 \leq l \leq n\}$.
(ii) when there are h customers in the system, for $1 \leq h \leq N-1$:
$l(h) = \{(h, i, 0, k_1, l) : 0 \leq i \leq L-1, 1 \leq k_1 \leq m_1, 1 \leq l \leq n\} \cup \{(h, L, 0, l) : 1 \leq l \leq n\} \cup \{(h, 0, 1, k_2, l) : 1 \leq k_2 \leq m_2, 1 \leq l \leq n\} \cup \{(h, i, 1, k_3, l) : 1 \leq i \leq L-N+h, 1 \leq k_3 \leq m_3, 1 \leq l \leq n\}$ (last part only when$L-N+h > 0$)
and for $h \geq N$:
$l(h) = \{(h, 0, 1, k_2, l) : 1 \leq k_2 \leq m_2, 1 \leq l \leq n\} \cup \{(h, i, 1, k_3, l) : 1 \leq i \leq L, 1 \leq k_3 \leq m_3, 1 \leq l \leq n\}$.

The infinitesimal generator of this CTMC is

$$\mathcal{Q}_1 = \begin{bmatrix} B_0 & C_0 & & & & & & & \\ B_1 & E_1 & F_1 & & & & & & \\ & B_2 & E_2 & F_2 & & & & & \\ & & \ddots & \ddots & \ddots & & & & \\ & & B_{N-2} & E_{N-2} & F_{N-2} & & & & \\ & & & B_{N-1} & E_{N-1} & F'_{N-1} & & & \\ & & & & B'_N & A_1 & A_0 & & \\ & & & & & A_2 & A_1 & A_0 & \\ & & & & & & \ddots & \ddots & \ddots \end{bmatrix}.$$

The boundary blocks $B_0, C_0, B_1, F'_{N-1}, B'_N$ are of orders $(Lm_1+1)n \times (Lm_1+1)n$, $(Lm_1+1)n \times \big((Lm_1+1)n+m_2n+(L-N+1)m_3n\big)$, $\big((Lm_1+1)n+m_2n+(L-N+1)m_3n\big) \times (Lm_1+1)n$, $\big((m_1+m_2)n+(L-1)(m_1+m_3)n+n\big) \times (m_2+Lm_3)n$, $(m_2+Lm_3)n \times \big((m_1+m_2)n+(L-1)(m_1+m_3)n+n\big)$ respectively. For $2 \leq h \leq N-1$, B_h is of order $\big((m_1+m_2)n+(L-N+h)(m_1+m_3)n+(N-h-1)m_1n+n\big) \times \big(m_1+m_2)n+(L-N+h-1)(m_1+m_3)n+(N-h)m_1n+n\big)$. For $1 \leq h \leq N-1$, E_h is of order $\big((m_1+m_2)n+(L-N+h)(m_1+m_3)n+(N-h-1)m_1n+n\big) \times \big((m_1+m_2)n+(L-N+h)(m_1+m_3)n+(N-h-1)m_1n+n\big)$. For $1 \leq h \leq N-2$, F_h is of order $\big((m_1+m_2)n+(L-N+h)(m_1+m_3)n+(N-h-1)m_1n+n\big) \times \big((m_1+m_2)n+(L-N+h+1)(m_1+m_3)n+(N-h-2)m_1n+n\big)$. A_0, A_1, A_2 are square matrices of order $(m_2+Lm_3)n$. Define the entries $B_{0_{(i_1,j_1,k_1,l_1)}}^{(i_2,j_2,k_2,l_2)}$, $C_{0_{(i_1,j_1,k_1,l_1)}}^{(i_2,j_2,k_2,l_2)}$, $B_{1_{(i_1,j_1,k_1,l_1)}}^{(i_2,j_2,k_2,l_2)}$, $E_{h_{(i_1,j_1,k_1,l_1)}}^{(i_2,j_2,k_2,l_2)}$, $B_{h_{(i_1,j_1,k_1,l_1)}}^{(i_2,j_2,k_2,l_2)}$, $F_{h_{(i_1,j_1,k_1,l_1)}}^{(i_2,j_2,k_2,l_2)}$, ${F'}_{N-1}{}_{(i_1,j_1,k_1,l_1)}^{(i_2,j_2,k_2,l_2)}$ and ${B'}_N{}_{(i_1,j_1,k_1,l_1)}^{(i_2,j_2,k_2,l_2)}$ as transition submatrices which contains transitions of the

form $(0,i_1,j_1,k_1,l_1) \rightarrow (0,i_2,j_2,k_2,l_2)$, $(0,i_1,j_1,k_1,l_1) \rightarrow (1,i_2,j_2,k_2,l_2)$, $(1,i_1,j_1,k_1,l_1) \rightarrow (0,i_2,j_2,k_2,l_2)$, $(h,i_1,j_1,k_1,l_1) \rightarrow (h,i_2,j_2,k_2,l_2)$, where $1 \leq h \leq N-1$, $(h,i_1,j_1,k_1,l_1) \rightarrow (h-1,i_2,j_2,k_2,l_2)$, where $2 \leq h \leq N-1$, $(h,i_1,j_1,k_1,l_1) \rightarrow (h+1,i_2,j_2,k_2,l_2)$, where $1 \leq h \leq N-2$, $(N-1,i_1,j_1,k_1,l_1) \rightarrow (N,i_2,j_2,k_2,l_2)$ and $(N,i_1,j_1,k_1,l_1) \rightarrow (N-1,i_2,j_2,k_2,l_2)$ respectively. Define the entries $A_{2_{(i_1,j_1,k_1,l_1)}}^{(i_2,j_2,k_2,l_2)}, A_{1_{(i_1,j_1,k_1,l_1)}}^{(i_2,j_2,k_2,l_2)}$ and $A_{0_{(i_1,j_1,k_1,l_1)}}^{(i_2,j_2,k_2,l_2)}$ as transition submatrices which contains transitions of the form $(h,i_1,j_1,k_1,l_1) \rightarrow (h-1,i_2,j_2,k_2,l_2)$, where $h \geq N+1$, $(h,i_1,j_1,k_1,l_1) \rightarrow (h,i_2,j_2,k_2,l_2)$, where $h \geq N$ and $(h,i_1,j_1,k_1,l_1) \rightarrow (h+1,i_2,j_2,k_2,l_2)$, where $h \geq N$ respectively. Since none or one event alone could take place in a short interval of time with positive probability, in general, a transition such as $(h_1,i_1,j_1,k_1,l_1) \rightarrow (h_2,i_2,j_2,k_2,l_2)$ has positive rate only for exactly one of h_1,i_1,j_1,k_1,l_1 different from h_2,i_2,j_2,k_2,l_2.

$$B_{0_{(i_1,j_1,k_1,l_1)}}^{(i_2,j_2,k_2,l_2)} = \begin{cases} T^o\alpha \otimes I_n & i_2 = i_1+1, 0 \leq i_1 \leq L-2; j_1 = j_2 = 0; 1 \leq k_1,k_2 \leq m_1; \\ & 1 \leq l_1,l_2 \leq n \\ T^o \otimes I_n & i_1 = L-1, i_2 = L; j_1 = j_2 = 0; 1 \leq k_1,k_2 \leq m_1; 1 \leq l_1,l_2 \leq n \\ T \oplus D_0 & i_1 = i_2, 0 \leq i_1 \leq L-1; j_1 = j_2 = 0; 1 \leq k_1,k_2 \leq m_1; \\ & 1 \leq l_1,l_2 \leq n \\ D_0 & i_1 = i_2 = L; j_1 = j_2 = 0; 1 \leq l_1,l_2 \leq n \end{cases}$$

$$C_{0_{(i_1,j_1,k_1,l_1)}}^{(i_2,j_2,k_2,l_2)} = \begin{cases} I_{m_1} \otimes D_1 & 0 \leq i_1 \leq L-1, i_1 = i_2; j_1 = j_2 = 0; 1 \leq k_1,k_2, \leq m_1; \\ & 1 \leq l_1,l_2 \leq n \\ D_1 & i_1 = i_2 = L; j_1 = j_2 = 0; 1 \leq l_1,l_2 \leq n \end{cases}$$

$$B_{1_{(i_1,j_1,k_1,l_1)}}^{(i_2,j_2,k_2,l_2)} = \begin{cases} S^0\alpha \otimes I_n & i_1 = i_2 = 0; j_1 = 1, j_2 = 0; 1 \leq k_1 \leq m_2, \\ & 1 \leq k_2 \leq m_1; 1 \leq l_1,l_2 \leq n \\ U^0\alpha \otimes I_n & 1 \leq i_1 \leq L-N+1; i_2 = i_1-1; j_1 = 1, j_2 = 0;; 1 \leq k_1 \leq m_3, \\ & 1 \leq k_2 \leq m_1; 1 \leq l_1,l_2 \leq n \end{cases}$$

For $1 \leq h \leq N-1$,

$$E_{h_{(i_1,j_1,k_1,l_1)}}^{(i_2,j_2,k_2,l_2)} = \begin{cases} T^0\alpha \otimes I_n & 0 \leq i_1 \leq L-2, i_2 = i_1+1; j_1 = j_2 = 0; \\ & 1 \leq k_1,k_2 \leq m_1; 1 \leq l_1,l_2 \leq n \\ T^0 \otimes I_n & i_1 = L-1, i_2 = L; j_1 = j_2 = 0; \\ & 1 \leq k_1 \leq m_1; 1 \leq l_1,l_2 \leq n \\ T \oplus D_0 & i_1 = i_2, 0 \leq i_1 \leq L-1; j_1 = j_2 = 0; 1 \leq k_1,k_2 \leq m_1; \\ & 1 \leq l_1,l_2 \leq n \\ S \oplus D_0 & i_1 = i_2 = 0, j_1 = j_2 = 1, 1 \leq k_1,k_2 \leq m_2, 1 \leq l_1,l_2 \leq n \\ U \oplus D_0 & i_1 = i_2, 1 \leq i_1 \leq L-N+h; j_1 = j_2 = 1, \\ & 1 \leq k_1,k_2 \leq m_3, 1 \leq l_1,l_2 \leq n \\ D_0 & i_1 = i_2 = L; j_1 = j_2 = 0; 1 \leq l_1,l_2 \leq n \end{cases}$$

For $2 \leq h \leq N-1$,

$$B_{h_{(i_1,j_1,k_1,l_1)}}^{(i_2,j_2,k_2,l_2)} = \begin{cases} S^0\beta \otimes I_n & i_1 = i_2 = 0; j_1 = j_2 = 1; 1 \leq k_1,k_2 \leq m_2; 1 \leq l_1,l_2 \leq n \\ U^0\beta \otimes I_n & i_1 = 1, i_2 = 0; j_1 = j_2 = 1; 1 \leq k_1 \leq m_3, \\ & 1 \leq k_2 \leq m_2; 1 \leq l_1,l_2 \leq n \\ U^0\gamma \otimes I_n & 2 \leq i_1 \leq L-N+h, i_2 = i_1-1; j_1 = j_2 = 1; \\ & 1 \leq k_1,k_2 \leq m_3; 1 \leq l_1,l_2 \leq n \end{cases}$$

For $1 \leq h \leq N-2$,

$$F_{h(i_1,j_1,k_1,l_1)}^{(i_2,j_2,k_2,l_2)} = \begin{cases} I_{m_1} \otimes D_1 & 0 \leq i_1 \leq L-1, i_1 = i_2; j_1 = j_2 = 0; 1 \leq k_1, k_2 \leq m_1; \\ & 1 \leq l_1, l_2 \leq n \\ I_{m_2} \otimes D_1 & i_2 = i_1 = 0; j_1 = j_2 = 1; 1 \leq k_1, k_2 \leq m_2, 1 \leq l_1, l_2 \leq n \\ I_{m_3} \otimes D_1 & i_2 = i_1, 1 \leq i_1 \leq L-N+h; j_1 = j_2 = 1; \\ & 1 \leq k_1, k_2 \leq m_3, 1 \leq l_1, l_2 \leq n \\ D_1 & i_1 = i_2 = L; j_1 = j_2 = 0; 1 \leq k_1, k_2 \leq m_1; 1 \leq l_1, l_2 \leq n \end{cases}$$

$$F'_{N-1}{}_{(i_1,j_1,k_1,l_1)}^{(i_2,j_2,k_2,l_2)} = \begin{cases} e(m_1) \otimes (\beta \otimes D_1) & i_1 = i_2 = 0; j_1 = 0, j_2 = 1; 1 \leq k_1 \leq m_1, \\ & 1 \leq k_2 \leq m_2; 1 \leq l_1, l_2 \leq n \\ I_{m_2} \otimes D_1 & i_2 = i_1 = 0; j_1 = j_2 = 1; 1 \leq k_1, k_2 \leq m_2, \\ & 1 \leq l_1, l_2 \leq n \\ I_{m_3} \otimes D_1 & i_2 = i_1, 0 \leq i_1 \leq L-1; j_1 = j_2 = 1;, \\ & 1 \leq k_1, k_2 \leq m_3, 1 \leq l_1, l_2 \leq n \\ e(m_1) \otimes (\gamma \otimes D_1) & 1 \leq i_1 \leq L-1; j_1 = 0, j_2 = 1; 1 \leq k_1 \leq m_1, \\ & 1 \leq k_2 \leq m_3; 1 \leq l_1, l_2 \leq n \\ \gamma \otimes D_1 & i_1 = i_2 = L; j_1 = 0, j_2 = 1; 1 \leq k_1 \leq m_1, \\ & 1 \leq k_2 \leq m_3; 1 \leq l_1, l_2 \leq n \end{cases}$$

$$B'_{N}{}_{(i_1,j_1,k_1,l_1)}^{(i_2,j_2,k_2,l_2)} = \begin{cases} S^0\beta \otimes I_n & i_1 = i_2 = 0; j_1 = j_2 = 1; 1 \leq k_1, k_2 \leq m_2, ; 1 \leq l_1, l_2 \leq n \\ U^0\beta \otimes I_n & i_1 = 1, i_2 = 0; j_1 = j_2 = 1; 1 \leq k_1 \leq m_3, \\ & 1 \leq k_2 \leq m_2; 1 \leq l_1, l_2 \leq n \\ U^0\gamma \otimes I_n & 2 \leq i_1 \leq L, i_2 = i_1 - 1; j_1 = j_2 = 1; 1 \leq k_1, k_2 \leq m_3; \\ & 1 \leq l_1, l_2 \leq n \end{cases}$$

$$A_{2(i_1,j_1,k_1,l_1)}^{(i_2,j_2,k_2,l_2)} = \begin{cases} S^0\beta \otimes I_n & i_1 = i_2 = 0; j_1 = j_2 = 1; 1 \leq k_1, k_2 \leq m_2; \\ & 1 \leq l_1, l_2 \leq n \\ U^0\beta \otimes I_n & i_1 = 1, i_2 = 0; j_1 = j_2 = 1; 1 \leq k_1 \leq m_3, 1 \leq k_2 \leq m_2; \\ & 1 \leq l_1, l_2 \leq n \\ U^0\gamma \otimes I_n & i_2 = i_1 - 1, 2 \leq i_2 \leq L; j_1 = j_2 = 1; 1 \leq k_1, k_2 \leq m_3; \\ & 1 \leq l_1, l_2 \leq n \end{cases}$$

$$A_{1(h,i_1,j_1,k_1,l_1)}^{(h,i_2,j_2,k_2,l_2)} = \begin{cases} S \oplus D_0 & i_1 = i_2 = 0, j_1 = j_2 = 1, 1 \leq k_1, k_2 \leq m_2, 1 \leq l_1, l_2 \leq n \\ U \oplus D_0 & i_1 = i_2, 1 \leq i_1 \leq L; j_1 = j_2 = 1, 1 \leq k_1, k_2 \leq m_3, 1 \leq l_1, l_2 \leq n \end{cases}$$

$$A_{0(i_1,j_1,k_1,l_1)}^{(i_2,j_2,k_2,l_2)} = \begin{cases} I_{m_2} \otimes D_1 & i_1 = i_2 = 0; j_1 = j_2 = 1; 1 \leq k_1, k_2 \leq m_2; 1 \leq l_1, l_2 \leq n \\ I_{m_3} \otimes D_1 & i_1 = i_2, 1 \leq i_1 \leq L; j_1 = j_2 = 1; 1 \leq k_1, k_2 \leq m_3; 1 \leq l_1, l_2 \leq n \end{cases}$$

3 Steady State Analysis

The stability condition for the system is given by

Lemma 1. *The system is stable iff* $\pi^* D_1 e < (\beta(-S)^{-1}e)^{-1}$.

Let $\boldsymbol{x}$ be the steady state probability vector of Q. We partition this vector as

$$\boldsymbol{x} = (\boldsymbol{x_0}, \boldsymbol{x_1}, \boldsymbol{x_2} \ldots),$$

where $\boldsymbol{x_0}$ is of dimension $(Lm_1+1)n$, $\boldsymbol{x_h}$, $1 \le h \le N-1$ are of dimension $(m_1+m_2)n+(L-N+h)(m_1+m_3)n+(N-h-1)m_1n+n$ and $\boldsymbol{x_N}, \boldsymbol{x_{N+1}} \ldots$ are of dimension $(m_2+Lm_3)n$. Under the stability condition, we have

$$\boldsymbol{x_{N+i}} = \boldsymbol{x}_N R^i, i \ge 1$$

where the matrix R is the minimal nonnegative solution to the matrix quadratic equation

$$R^2 A_2 + RA_1 + A_0 = 0$$

and the vectors $\boldsymbol{x_0}, \boldsymbol{x_1}, \cdots, \boldsymbol{x_N} \ldots$are obtained by solving the equations

$$\boldsymbol{x_0} B_0 + \boldsymbol{x_1} B_1 = 0 \tag{1}$$

$$\boldsymbol{x_0} C_0 + \boldsymbol{x_1} E_1 + \boldsymbol{x_2} B_2 = 0 \tag{2}$$

$$\boldsymbol{x_{i-1}} F_{i-1} + \boldsymbol{x_i} E_i + \boldsymbol{x_{i+1}} B_{i+1} = 0, \text{ for } 2 \le i \le N-2 \tag{3}$$

$$\boldsymbol{x_{N-2}} F_{N-2} + \boldsymbol{x_{N-1}} E_{N-1} + \boldsymbol{x_N} B_{N'} = 0 \tag{4}$$

$$\boldsymbol{x_{N-1}} F'_{N-1} + \boldsymbol{x_N}(A_1 + RA_2) = 0 \tag{5}$$

subject to the normalizing condition

$$\sum_{i=0}^{N-1} \boldsymbol{x_i e} + + \boldsymbol{x_N}(I-R)^{-1}\boldsymbol{e} = 1 \tag{6}$$

3.1 Distribution of Time till the Number of Customers Hit N or the Inventory Level Reaches L

We can study this by a phase type distribution $PH(\psi_1, V_1)$ where the underlying Markov process has state space $\{(h,i,j,k) : 0 \le h \le N-1, 0 \le i \le L-1, 1 \le j \le m_1, 1 \le k \le n\} \cup \{*_1\} \cup \{*_2\}$ where $*_1$ denotes the absorbing state indicating the inventory level hitting L and $*_2$ denotes the absorbing state indicating the number of customers reaching N. The infinitesimal generator is

$$\boldsymbol{V_1} = \begin{bmatrix} V_1 & V_1^{(0)} & V_1^{(1)} \\ \boldsymbol{0} & 0 & 0 \end{bmatrix} \text{ where, } V_1 = \begin{bmatrix} E & I_{Lm_1} \otimes D_1 & & \\ \ddots & \ddots & & \\ & E & I_{Lm_1} \otimes D_1 & \\ & & E & \end{bmatrix},$$

$$V_1^{(0)} = \begin{bmatrix} \boldsymbol{e}_L(L) \otimes (T^0 \otimes e(n)) \\ \vdots \\ \boldsymbol{e}_L(L) \otimes (T^0 \otimes e(n)) \end{bmatrix}, V_1^{(1)} = \begin{bmatrix} \boldsymbol{0} \\ \vdots \\ \boldsymbol{0} \\ \boldsymbol{e}(Lm_1) \otimes \delta \end{bmatrix}$$

with

$$E = \begin{bmatrix} T \oplus D_0 & T^0\alpha \otimes I_n & & \\ \ddots & \ddots & & \\ & T \oplus D_0 & T^0\alpha \otimes I_n & \\ & & T \oplus D_0 & \end{bmatrix} \text{ and } \delta = \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_n \end{bmatrix},$$

with δ_i representing the sum of ith row of the D_1 matrix.

The initial probability vector is

$$\psi_1 = (1/d_1)(x_{0,0,0,1,1}, \cdots, x_{0,0,0,1,n}, \cdots, x_{0,0,0,m_1,1}, \cdots, x_{0,0,0,m_1,n}, \mathbf{0})$$

where $d_1 = \sum_{l=1}^{n} \sum_{k=1}^{m_1} x_{0,0,0,k,l}$ and $\mathbf{0}$ is a zero matrix of order $1 \times \big((N-1)Lm_1n + (L-1)m_1n\big)$.

Thus we have the following lemma.

Lemma 2. *The expected duration of time till the inventory level reaches L before the number of customers hit N is given by* $\psi_1(-V_1)^{-2}V_1^{(0)}$ *and the expected duration of time till the number of customers hit N before the inventory level reaches L is given by* $\psi_1(-V_1)^{-2}V_1^{(1)}$.

3.2 Distribution of Idle Time

Case (i)

Suppose that the number of customers become N only after the inventory level hits L. The probability for this event is the probability for absorption of $PH(\psi_1, V_1)$ to $*_1$. In this case, we can study this conditional distribution by a phase type distribution $PH(\psi_2, V_2)$ where the underlying Markov process has state space $\{(h, L, 0, l) : 0 \le h \le N-1, 1 \le l \le n\} \cup \{*\}$ where $*$ denotes the absorbing state indicating that the number of customers hitting N. The infinitesimal generator is

$$\boldsymbol{V_2} = \begin{bmatrix} V_2 & V_2^0 \\ \mathbf{0} & 0 \end{bmatrix}, \text{ where, } V_2 = \begin{bmatrix} D_0 & D_1 & & \\ \ddots & \ddots & & \\ & D_0 & D_1 & \\ & & D_0 & \end{bmatrix}, V_2^0 = \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \delta \end{bmatrix}$$

where $\delta = \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_n \end{bmatrix}$ with δ_i representing the sum of ith row of the D_1 matrix. The initial probability vector is

$$\psi_2 = (1/d_2)(v_{0,L,0,1}, \cdots, v_{0,L,0,n}, \cdots, v_{N-1,L,0,1}, \cdots, v_{N-1,L,0,n})$$

where, for $0 \le h \le N-2, 1 \le l \le n$,

$$v_{h,L,0,l} = \sum_{k=1}^{m_1} \frac{\eta_k}{\sum_{l \ne l'} d_{ll'}^0 + \delta_l + \sum_{k \ne k'} T_{kk'} + \eta_k} x_{h,L-1,0,k,l}$$

and, for $h = N-1$, $1 \le l \le n$,

$$v_{N-1,L,0,l} = \sum_{k=1}^{m_1} \frac{\eta_k}{\sum_{l \ne l'} d_{ll'}^0 + \sum_{k \ne k'} T_{kk'} + \eta_k} x_{N-1,L-1,0,k,l},$$

with, $d_2 = \sum_{h=0}^{N-1} \sum_{l=1}^{n} v_{h,L,0,l}$.

Here, η_k represents the absorption rate from phase k in $PH(\alpha, T)$, $T_{kk''}$ represent the kk'th entry of T, $d^0_{ll'}$ represent the transition rates from the phase l to the phase l' without arrival and δ_l represent the lth row sum of D_1 matrix.

Case(ii)
Suppose that the number of customers become N before the inventory level hits L. The probability for this event is the probability for absorption of $PH(\psi_1, V_1)$ to $*_2$. In this case, the idle time = 0.

Thus we have the following theorem.

Theorem 1. *The LST of the distribution of the idle time is given by*

$$\left(\psi_2(sI - V_2)^{-1}V_2^0\right)\left(\int_{t=0}^{\infty} \psi_1 e^{V_1 t} V_1^{(0)} dt\right)$$

3.3 Distribution of Time Until the Number of Customers Hit N

We can study this by a phase type distribution $PH(\psi_3, V_3)$ where the underlying Markov process has state space $\{(h,i,j,k) : 0 \le h \le N-1, 0 \le i \le L-1, 1 \le j \le m_1, 1 \le k \le n\} \cup \{(h,L,k) : 0 \le h \le N-1, 1 \le k \le n\} \cup \{*\}$ where $*$ denotes the absorbing state indicating the number of customers reaching N. The infinitesimal generator is

$$\boldsymbol{V_3} = \begin{bmatrix} V_3 & V_3^0 \\ \mathbf{0} & 0 \end{bmatrix}, \text{ where } V_3 = \begin{bmatrix} F & I_{Lm_1+1} \otimes D_1 & & \\ \ddots & \ddots & & \\ & F & I_{Lm_1+1} \otimes D_1 & \\ & & F & \end{bmatrix}, V_3^0 = \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \boldsymbol{e}(Lm_1+1) \otimes \delta \end{bmatrix}$$

with

$$F = \begin{bmatrix} T \oplus D_0 & T^0\alpha \otimes I_n & & & \\ \ddots & \ddots & & & \\ & T \oplus D_0 & T^0\alpha \otimes I_n & & \\ & & T \oplus D_0 & T^0 \otimes I_n & \\ & & & D_0 & \end{bmatrix}$$

The initial probability vector is

$$\psi_3 = (1/d_1(x_{0,0,0,1,1}, \cdots, x_{0,0,0,1,n}, \cdots, x_{0,0,0,m_1,1}, \cdots, x_{0,0,0,m_1,n}, \mathbf{0})$$

where $d_1 = \sum_{l=1}^{n} \sum_{k=1}^{m_1} x_{0,0,0,k,l}$, where $\mathbf{0}$ is a zero matrix of order $1 \times \big((N-1)(Lm_1+1)n + ((L-1)m_1+1)n\big)$.

Thus we have the following lemma.

Lemma 3. *The distribution of time when the processing starts until the number of customers hit N is a phase type distribution with representation $PH(\psi_3, V_3)$.*

3.4 Distribution of Number of Inventory Processed Before the Arrival of First Customer

To compute the above distribution, first we find the following:

Distribution of Processing Time till the Arrival of First Customer. Consider the Markov process with state space $\{(i,j,k) : 0 \leq i \leq L-1, 1 \leq j \leq m_1, 1 \leq k \leq n\} \cup \{(L,k) : 1 \leq l \leq n\} \cup \{*\}$, where i denote the number of items in the inventory, j, the phase of inventory processing, k, the arrival phase of customer, *, the absorbing state indicating the arrival of a customer. The infinitesimal generator of the process is given by

$$\mathcal{V}_4 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ e(m_1)\otimes\delta & T\oplus D_0 & T^0\alpha\otimes I_n & 0 & 0 & 0 \\ e(m_1)\otimes\delta & 0 & T\oplus D_0 & T^0\alpha\otimes I_n & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ e(m_1)\otimes\delta & 0 & 0 & T\oplus D_0 & T^0\alpha\otimes I_n & 0 \\ e(m_1)\otimes\delta & 0 & 0 & 0 & T\oplus D_0 & T^0\otimes I_n \\ \delta & 0 & 0 & 0 & 0 & D_0 \end{bmatrix}.$$

The initial probability is given by

$$\psi_4 = \frac{1}{d_1}(x_{0,0,0,1,1,}, \ldots, x_{0,0,0,1,n}, \ldots, x_{0,0,0,m_1,1}, \ldots x_{0,0,0,m_1,n}, \mathbf{0})$$

where $\mathbf{0}$ is a zero matrix of order $1 \times ((L-1)m_1+1)n$.
Let Y denote the number of items processed before the first arrival and y_k be the probability that k items are processed before an arrival. Then y_k is the probability that the absorption occurs from the level k for the process. Hence y_k are given by

$$y_0 = -\psi_4(T\oplus D_0)^{-1}(e(m_1)\otimes\delta)$$

For $k = 1, 2, 3, \ldots L-1$

$$y_k = (-1)^{k+1}\psi_4\left((T\oplus D_0)^{-1}(T^0\alpha\otimes I_n)\right)^k(T\oplus D_0)^{-1}(e(m_1)\otimes\delta)$$

and

$$y_L = (-1)^{L+1}\psi_4\left((T\oplus D_0)^{-1}(T^0\alpha\otimes I_n)\right)^{L-1}(T\oplus D_0)^{-1}(T^0\otimes I_n)D_0^{-1}\delta$$

Thus we have the following lemma.

Lemma 4. *The distribution of number of inventory processed before the arrival of first customer is given by* $P(Y=k) = y_k$.

Definition 1. *Starting up with the epoch of departure of a customer leaving behind no customer in the system until the next epoch at which no customer is left at a service completion epoch is called a busy cycle.*

3.5 Distribution of Busy Cycle

First we assume that $L > N$.

The distribution of duration of busy cycle in which no item is left in the inventory can be studied by a continuous time Markov chain with state space $\{(h,i,0,k,l) : 0 \le h \le N-1, 0 \le i \le L-1, 1 \le k \le m_1, 1 \le l \le n\} \cup \{(h,L,0,l) : 0 \le h \le N-1, 1 \le l \le n\} \cup \{(h,i,1,k,l) : 1 \le h \le M, i = 0, 1 \le k \le m_2, 1 \le l \le n\} \cup \{(h,i,1,k,l) : 1 \le h \le N-1, 1 \le i \le L-N+h, 1 \le k \le m_3, 1 \le l \le n\} \cup \{(h,i,1,k,l) : N \le h \le M, 1 \le i \le L, 1 \le k \le m_3, 1 \le l \le n\} \cup \{*\}$, where $(h,i,0,k,l)$ denote the states that correspond to the server being in vacation with h customers in the system,i, items in the inventory, k, processing phase and l,the arrival phase, $(h,L,0,l)$ denote the states that correspond to the server being in vacation with h customers in the system, L, items in the inventory and l,the arrival phase, $(h,i,1,k,l)$ denote the states that correspond to the server being in normal mode with h customers in the system,i, items in the inventory, k, service phase and l,the arrival phase, $*$ denote the absorbing state indicating that the number of customers become zero by a service completion and M is chosen in such a way that $P\left(\sum_{h=0}^{M} x_h \boldsymbol{e} > 1-\epsilon\right) \to 0$ for every $\epsilon > 0$. Then the distribution of a busy cycle can be studied by a phase type distribution $PH(\phi, B)$, whose infinitesimal generator is given by

$$\mathcal{B} = \begin{bmatrix} B & B^0 \\ \mathbf{0} & 0 \end{bmatrix} \text{ where, } B = \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix}$$

Now,

$$B_{11} = \begin{bmatrix} F & I_{Lm_1+1} \otimes D_1 & & \\ & \ddots & \ddots & \\ & & F & I_{Lm_1+1} \otimes D_1 \\ & & & F \end{bmatrix}, \text{ with } F = \begin{bmatrix} T \oplus D_0 & T^0\alpha \otimes I_n & & \\ & \ddots & \ddots & \\ & & T \oplus D_0 & T^0\alpha \otimes I_n \\ & & & T \oplus D_0 \end{bmatrix}$$

$$B_{12} = e_N(N)e'_N(M) \otimes B'_{12},$$

where,

$$B'_{12} = \begin{bmatrix} e(m_1) \otimes (\beta \otimes D_1) & & & \\ & I_{L-1} \otimes (e(m_1) \otimes (\gamma \otimes D_1) & & \\ & & \ddots & \\ & & & \gamma \otimes D_1 \end{bmatrix}$$

$$B_{22} = \begin{bmatrix} E_1 & F_1 & & & & & & \\ G_1 & E_2 & F_2 & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & G_{N-2} & E_{N-1} & F_{N-1} & & & \\ & & & G_{N-1} & E_N & I \otimes D_1 & & \\ & & & & \ddots & \ddots & \ddots & \\ & & & & & G_{M-2} & E_{M-1} & I \otimes D_1 \\ & & & & & & G_{M-1} & E'_M \end{bmatrix}$$

For $1 \leq h \leq N-1$,

$$G_h = \begin{bmatrix} S^0\beta \otimes I_n & 0 \\ U^0\beta \otimes I_n & 0 \\ 0 & I_{L-N+h} \otimes (U^0\gamma \otimes I_n) \end{bmatrix}, E_h = \begin{bmatrix} S \oplus D_0 & \\ & I_{L-N+h} \otimes (U \oplus D_0) \end{bmatrix}$$

and

$$F_h = \begin{bmatrix} I_{m_2+(L-N+h)m_3} \otimes D_1 \ 0 \end{bmatrix}.$$

$$\text{For } N \leq h \leq M-1,\ E_h = \begin{bmatrix} S \oplus D_0 & \\ & I_L \otimes (U \oplus D_0) \end{bmatrix}$$

$$\text{For } h \geq N,\ G_h = \begin{bmatrix} S^0\beta \otimes I_n & 0 & 0 \\ U^0\beta \otimes I_n & 0 & 0 \\ 0 & I_{L-1} \otimes (U^0\gamma \otimes I_n) & 0 \end{bmatrix}$$

$$E'_M = \begin{bmatrix} E_{M0} & \\ & I_L \otimes (E_{M1}) \end{bmatrix}, \text{ where}$$

$E_{M0} = S \oplus D_0 - I_{m_2} \otimes \Delta$ and $E_{M1} = U \oplus D_0 - I_{m_3} \otimes \Delta$, with $\Delta = \begin{bmatrix} \delta_1 & & \\ & \ddots & \\ & & \delta_n \end{bmatrix}$.

and

$$B^0 = \begin{bmatrix} \mathbf{0} \\ B^{00} \end{bmatrix}, \text{ with, } B^{00} = \begin{bmatrix} B_{000} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}, \text{ where, } B^{000} = \begin{bmatrix} S^0 \otimes e(n) \\ e(L-N+1) \otimes (U^0 \otimes e(n)) \end{bmatrix}$$

The initial probability vector is

$$\phi = (\phi', \mathbf{0})$$

where, $\phi' = \frac{1}{d_3}(w_{0,0,0,1,1}, \cdots, w_{0,0,0,m_1,n}, \cdots, w_{0,L-1,0,1,1}, \cdots, w_{0,L-1,0,m_1,n}, \mathbf{0})$, where $d_3 = \sum_{i=0}^{L-1} \sum_{k'=1}^{m_1} \sum_{l=1}^{n} w_{0,i,0,k',l}$.

For $1 \le k' \le m_1$, $1 \le l \le n$,

$$w_{0,0,0,k',l} = \sum_{k=1}^{m_2} \frac{\sigma_k \alpha_{k'}}{\delta_l + \sum_{l \neq l'} d^0_{ll'} + \sigma_k + \sum_{k \neq k''} S_{kk''}} x_{1,0,1,k,l} + \sum_{k=1}^{m_3} \frac{\tau_k \alpha_{k'}}{\delta_l + \sum_{l \neq l'} d^0_{ll'} + \tau_k + \sum_{k \neq k''} U_{kk''}} x_{1,1,1,k,l}, \quad (7)$$

For $1 \le i \le L-1$, $1 \le k' \le m_1$, $1 \le l \le n$,

$$w_{0,i,0,k',l} = \sum_{k=1}^{m_3} \frac{\tau_k \alpha_{k'}}{\delta_l + \sum_{l \neq l'} d^0_{ll'} + \tau_k + \sum_{k \neq k''} U_{kk''}} x_{1,i+1,1,k,l},$$

where σ_k, τ_k represent the absorption rates from service phase k in $PH(\beta, S)$ and $PH(\gamma, U)$ respectively, $S_{kk''}$, $U_{kk''}$ represent the kk''th entry of S and U respectively, $\alpha_{k'}$ represents the probability that the processing of item starts in phase k', $d^0_{ll'}$ represent the transition rates from the phase l to the phase l' without arrival and δ_l represent the lth row sum of D_1 matrix.

From the above discussions we have the following.

Theorem 2. *The LST of the distribution of a busy cycle in which no item is left in the inventory is given by*

$$\hat{B_{C_1}}(s) = \phi(sI - B)^{-1} I'(B^0)'$$

where, I' denote the columns of identity matrix corresponding to the 1 customer level with number of items in the inventory 0 and 1 and

$$(B^0)' = \begin{bmatrix} S^0 \otimes e(n) \\ U^0 \otimes e(n) \end{bmatrix}$$

Theorem 3. *The LST of the distribution of a busy cycle in which atleast one item is left in the inventory is given by*

$$\hat{B_{C_2}}(s) = \phi(sI - B)^{-1} I''(B^0)''$$

where, I'' denote the columns of identity matrix corresponding to 1 customer level with number of items in the inventory > 1 and

$$(B^0)'' = e(L - N) \otimes (U^0 \otimes e(n))$$

Theorem 4. *For stationary MAP, the expected number of busy cycles in which at least one inventory left in an interval of length t is given by*

$$\left(t/(\phi(-B)^{-1}\boldsymbol{e})\right) \left(\hat{B_{C_2}}'(0) / \left(\hat{B_{C_1}}'(0) + \hat{B_{C_2}}'(0)\right)\right)$$

4 Numerical Results

We fix $\alpha = \begin{bmatrix} 1 & 0 \end{bmatrix}$, $\beta = \begin{bmatrix} 1 & 0 \end{bmatrix}$ and $\gamma = \begin{bmatrix} 0.8 & 0.2 \end{bmatrix}$, $T = \begin{bmatrix} -3 & 3 \\ 0 & -3 \end{bmatrix}$, $S = \begin{bmatrix} -4 & 4 \\ 0 & -4 \end{bmatrix}$, $U = \begin{bmatrix} -2 & 2 \\ 0 & -2 \end{bmatrix}$ and $D0 = -1$, $D1 = 1$

For these input parameters we get the system characteristics as given in Table 1. The behaviour of the performance characteristics is on expected lines. Let E denote Expected Idle time, SD, standard deviation of Idle time, CV, Coefficient of Variation of Idle time.

Table 1. Mean/Standard deviation/Coefficient of variation of idle time of the server

$L \downarrow N \rightarrow$	2			3			4		
	E	SD	CV	E	SD	CV	E	SD	CV
2	0.90	1.20	1.33	1.47	1.52	1.03	2.00	1.79	0.90
3	0.63	1.07	1.71	1.15	1.43	1.25	1.78	1.77	1.00
4	0.42	0.92	2.19	0.86	1.31	1.52	1.44	1.68	1.17
5	0.27	0.76	2.80	0.63	1.16	1.86	1.12	1.56	1.39

5 Conclusion

In this paper, we considered a MAP/PH/1 queue with processing of service items under Vacation and N-policy. We analysed the distribution of time till the number of customers hit N or the inventory level reaches L, distribution of idle time, the distribution of time until the number of customers hit N and also the distribution of number of inventory processed before the arrival of first customer. Also we provided the distribution of a busy cycle, LSTs of busy cycles in which no item is left in the inventory and atleast one item is left in the inventory. We performed some numerical experiments to evaluate the expected idle time, standard deviation and coefficient of varaiation of idle time of the server. We propose to extend the model to the case in which the customers are impatient. Also we will find individual optimal strategy, server's maximum revenue and social optimal strategy by numerical experiments in a future study.

References

1. Chakravarthy, S.R.: The batch Markovian arrival process: a review and future work. In: Krishnamoorthy, A., et al. (eds.) Advances in Probability Theory and Stochastic Processes, pp. 21–49. Notable Publications Inc., New Jersey (2001)
2. Doshi, B.T.: Queueing systems with vacations-a survey. Queueing Syst. **1**, 29–66 (1986)

3. Hanukov, G., Avinadav, T., Chernonog, T., Spiegal, U., Yechiali, U.: A queueing system with decomposed service and inventoried preliminary services. Appl. Math. Model. **47**, 276–293 (2017)
4. Kazimirsky, A.V.: Analysis of BMAP/G/1 queue with reservation of service. Stochast. Anal. Appl. **24**(4), 703–718 (2006)
5. Levi, Y., Yechiali, U.: Utilization of idle time in an M/G/1 queueing system. Manag. Sci. **22**, 202–211 (1975)
6. Neuts, M.F.: Matrix Geometric Solutions in Stochastic Models: An Algorithmic Approach. The Johns Hopkins University Press, Baltimore (1981)
7. Sreenivasan, C., Chakravarthy, S.R., Krishnamoorthy, A.: MAP/PH/1 queue with working vacations, vacation interruptions and N-policy. Appl. Math. Model. **37**(6), 3879–3893 (2013)
8. Takagi, H.: Queueing Analysis, Volume I: Vacation and Priority Systems, Part 1. North Holland, Amsterdam (1991)
9. Tian, N., Zhang, Z.G.: Vacation Queueing Models, Theory and Applications. Springer, New York (2006). https://doi.org/10.1007/978-0-387-33723-4
10. Yadin, M., Naor, P.: Queueing systems with a removable server. Oper. Res. **14**, 393–405 (1963)

Flying Network for Emergencies

Truong Duy Dinh[1(✉)], Van Dai Pham[1], Ruslan Kirichek[1,2], and Andrey Koucheryavy[1]

[1] The Bonch-Bruevich Saint - Petersburg State University of Telecommunications, 22/1 Prospekt Bolshevikov, Saint-Petersburg 193232, Russian Federation
din.cz@spbgut.ru, daipham93@gmail.com, kirichek@sut.ru, akouch@mail.ru
[2] Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St., Moscow 117198, Russian Federation
http://www.sut.ru

Abstract. The article presents an approach to the organization of a flying network among mobile communication subscribers based on WiFi (VoWiFi) technology in a disaster area where telecommunication infrastructure is completely or partially destroyed. The flying network is organized on the basis of unmanned aerial vehicles (UAVs), which interact with each other based on IEEE 802.11p wireless technology and with mobile subscribers based on IEEE 802.11n/ac wireless technologies. The interaction process between subscribers and UAVs is presented as a queuing system. Based on the developed model were measured and obtained network delay parameters and its value did not exceed 100 ms. The fulfillment of this condition was achieved by varying the number of UAVs and the channel load parameters. A series of numerical experiments showed the permissible number of UAVs to provide an acceptable quality of voice transmission between subscribers that are in the UAV coverage area.

Keywords: Flying network · UAVs · VoWi-Fi · IEEE 802.11p

1 Introduction

In the last years, global climate change has led to an increase in the frequency and severity of natural disasters, such earthquakes, wildfire, tsunamis, etc. Consequently, these natural disasters result in the complete or partial destruction of telecommunication infrastructure. In this regard, the implementation of rescue operations is very difficult due to the lack of communication between emergency services, as well as the connection between emergency services and the sufferers. Considering that the deployment of a wireless communication network between emergency services becomes a priority, a connection needs to be quickly implemented by using advanced technologies.

The research of flying networks has been devoted to many research works, which deal with UAVs interaction without connection to the base station of the communication operator [1–4]. To solve this problem, a concept of a rapidly deployable flying network for emergencies was introduced [5–8]. According to

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 58–70, 2018.
https://doi.org/10.1007/978-3-319-99447-5_6

the proposed concept, to ensure UAVs interaction in order to provide maximum coverage of the destroyed area it is necessary to organize a flying network that support network mesh topology. In this network, each UAV can be considered as a mobile heterogeneous gateway [9,10]. A mobile heterogeneous gateway is a network device or relay system designed to provide interoperability between two information networks that have different characteristics, use different protocol sets, and support various data transmission technologies. Thus, each UAV is supposed to set a heterogeneous gateway that allows encapsulating data from/to mobile phones (IEEE 802.11n/ac) into the data transferred between the UAVs (IEEE 802.11p).

2 Related Work

During the last few years, a number of approaches and contributions about communication and networking in the flying network are proposed.

In this paper [11], the authors announced that in the near future, flying network will play an important role in everyday life. They found that it is really difficult to operate and manage the air traffic due to the great number of unmanned aerial vehicles (UAVs). Reliable communication links is necessary to support the operators and the unmanned traffic management system. It also depends on the unmanned traffic management system structure and the number of UAVs. The paper mentioned that there are two types of flying network architectures, which are cellular and ad-hoc; UAV-to-UAV direct links are needed to allow real-time information exchange and suggested IEEE 802.11p standard is a real possible choice for UAVs communication in flying network.

The comparison of different communication network architectures for flying network was considered in the paper [12]. The authors discussed advantages and disadvantages of each one. They made a review of legacy and next-generation data link systems for communications between the ground station and a UAV, between UAVs. Additionally, the authors mentioned that next-generation data link systems will be used for decentralized communications for UAV networks. They concluded that a UAV ad-hoc network is suitable to network groups of UAV and a multi-layer UAV ad-hoc network is more suitable to multiple groups of heterogeneous UAV.

The communication requirements for applications in micro UAV networks are discussed in the paper [13]. The authors mentioned that a set of candidate wireless technologies can be exploited for micro UAV networks such as IEEE 802.15.4, IEEE 802.11, 3G/LTE, and infrared. They considered a combination of IEEE 802.11 and XBee-Pro is suggested as a communication link for small networks but it cannot provide reliable, time-critical communication for large networks.

In [14] the role of meshed airborne communication networks in the operational performance of small UAV was proposed. The authors mentioned that only meshed ad-hoc networking, where nodes in the network are able to self-organize to act as relays, can meet the communication demands for the large number

of small UA expected to be deployed in future. They presented experimental results to show the feasibility of meshed airborne communication using the heterogeneous unmanned aircraft system and a net-centric operation of multiple cooperating UAV over mesh network is possible.

The authors in [15] developed a framework called UAVNet for the autonomous deployment of a flying Wireless Mesh Network using small quadrocopter-based UAVs as a flying node. They assumed that the flying wireless mesh nodes can be interconnected to each other and building an IEEE 802.11s wireless mesh network. In order to communicate with ground stations, i.e. notebooks, the flying wireless mesh nodes use IEEE 802.11b/g wireless standard. The authors aim at interconnecting two end systems by setting up an airborne relay, consisting of one or several flying wireless mesh nodes. At the end, they mentioned that UAVNet can be used to deploy a complete communication network in emergency and disaster recovery scenarios.

The authors in [16] considered the use of LTE for transferring data from and to UAV in a suburban environment. By means of measurements and simulations, the article analyzed the impact of interference and path loss when transmitting data to and from the UAV. They archived results that interference is a major limiting factor and that LTE might not be used effectively in the flying network.

A research on long range data transmission on flying sensor network was considered in [10]. The authors analyzed the problem of data delivery with the terrestrial segment of the flying sensor network over long distances using UAVs such as repeater chain. In the terrestrial segment IEEE 802.15.4 (6LoWPAN protocol) was used for interaction nodes and in the flying segment IEEE 802.15.4g (LoRaWAN protocol) is used for UAV interaction. At the end, they found delay and packet loss, occurring in the all stations of transmission network at different data rates and also the optimal data rate of network was found.

In summary, the articles reviewed above show that the use of UAV as a gateway in connection to ground stations is feasible and necessary. However, when natural disasters occur, telecommunication infrastructure in the disaster area is completely or partially destroyed, how emergency services can contact the victim is not mentioned. In this paper, we propose an approach to the organization of a flying network among mobile communication subscribers based on WiFi (VoWiFi) technology in a disaster area where telecommunication infrastructure is completely or partially destroyed, which will be discussed in the next sections.

3 Data Transmission Technologies for Flying Networks

Currently, with the transition to five generation communication networks 5G/IMT-2020, a number of technologies will play an important role in supporting of emergency services. One of the fundamental and widely used radio technologies at network access layer is wireless local area networks Wi-Fi.

As mentioned above, we assume that all telecommunication infrastructure in the disaster area is destroyed. Traditional GSM networks, which are used for voice communication, proposed the use of base stations whose weights and

dimensions do not allow their implementation on UAVs. In addition, the delivery of new base stations to the disaster area, is a logistical task. Thus, the organization of communication among mobile subscribers cannot be solved on the basis of GSM technology. Today, one of the most effective solutions is the organization of interaction with mobile subscribers of GSM networks via Wi-Fi based on voice over Wi-Fi (VoWi-Fi) applications [17–19]. To implement this approach, it is necessary that the entire disaster area be fully covered with a Wi-Fi radio signal. In order to achieve this goal, we propose the use of a flying network that consists of UAVs, which are mobile access points and relaying the received/transmitted data to a base station that operates in normal mode [9,20,21]. This approach allows organizing a hierarchical wireless ad-hoc network with mobile nodes. The role of the base station for subscribers will be performed by a Wi-Fi access point on board of a UAV that supports IEEE 802.11n or IEEE 802.11ac. Due to fact that VoWi-Fi technology has spread in a large number of different mobile phones models, it can be assumed that this approach will allow making calls over Wi-Fi in the organization of a flying network supporting this technology. It is also worth noting that all calls are made through the operator with the numbering and identification of mobile network subscribers.

Figure 1 shows the architecture of the flying network for emergencies, in which one or more UAVs are assumed to be used in the flying segment. This figure reflects the organization concept of communication and interaction of all elements of the system.

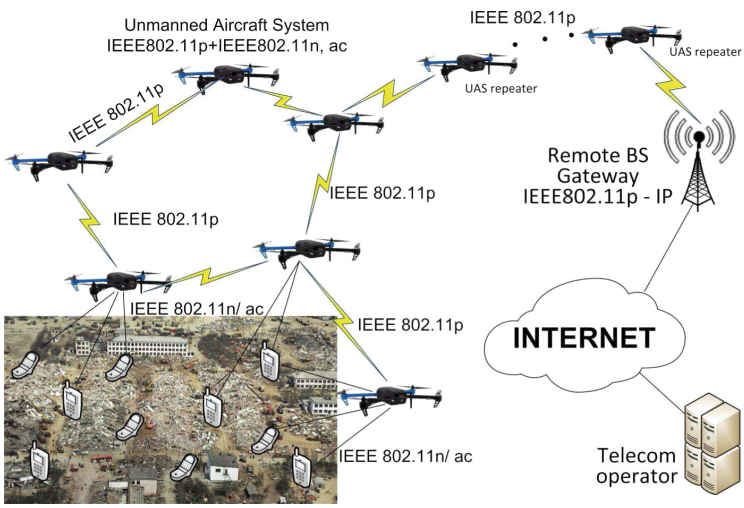


Fig. 1. The architecture of the flying network for emergencies

According to Fig. 1, the architecture of the voice service over WiFi is represented by the following segments:

- Terrestrial segment: This segment includes subscriber terminals (telephones) that have some network interfaces, such as Bluetooth, WiFi, GSM, LTE/4G, etc. At an acceptable distance, subscriber terminals can interact directly with each other, such as device-to-device communication (D2D Communication). Currently, some wireless technologies are known, such as Bluetooth, WiFi Direct, WiFi Hotspot, which allow data exchange between two devices as D2D communication. When using such technologies, the communication distance is limited. Therefore, for providing communication between two subscribers, which are at a large distance, active communication between several relays is proposed, in our case relays are UAVs with various wireless interfaces. In this case, subscriber terminals with the supporting IEEE 802.11n/ac technologies are WiFi-stations. Thus, IEEE 802.11n/ac technologies provide connection of subscriber terminals to the flying segment.
- Flying segment: This segment is built on the basis of UAV networks, which also support several network interfaces, such as IEEE 802.11p, IEEE 802.11n, IEEE 802.11ac, and wireless telemetry. In this architecture, the flying segment must provide the communication with the terrestrial segment, the connection between the UAV, and the connection with the Operator Center. Figure 1 shows that the terrestrial segment and the flying segment are connected by using the IEEE 802.11n/ac technology, and the connection of UAVs uses IEEE 802.11p technology.
- Operator segment: In this segment, there is a connection between the UAV group and the Operator Center, which determines the access of subscriber terminals to the service, i.e. whether there is a possibility of communication with another subscriber. Each subscriber terminal must be identified in the Operator Center. One of the UAV group is accessible to the base station via wireless technology IEEE 802.11p, as shown in Fig. 1.

4 Emergency Flying Network Queuing Model

Assume that in the disaster area subscriber 1 wants to call subscriber 2 via VoWi-Fi using the UAV group. An example of such a call may be the connection of an emergency service officer with subscribers in the disaster zone. According to mobile phones functioning algorithms, in the absence of communication with the base station, the phones switch to scanning mode of available networks. Scanning in the area of a natural disaster will help discover subscribers who potentially can be under the rubble waiting for help.

At the beginning, the call connection process between two subscribers is carried out by an operator, i.e. each UAV is connected to a mobile network operator, which allows the collection of subscriber data. According to the interworking scheme, a call can be made only after the operator sent a connection confirmation. After that, a call between two subscribers will be performed through a chain of UAVs interacting with each other. In this paper, we consider the process of voice traffic transmission from subscriber 1 to subscriber 2 after the connection is established. In order to make a call with an acceptable quality of perception,

it is necessary to ensure the voice transmission delay is not more than 100 ms [22].

Hence, the proposed service model requires the delivery time of voice traffic not more than 100 ms, as one of the basic parameter of quality of service. When considering the proposed architecture (Fig. 1), obviously, that the delivery delay of a packet from the first subscriber to the second subscriber is the total time that passes through the terrestrial segment and the flying segment in the condition of the established connection. When implementing a specific communication network in each segment, their effects on the delivery delay in more detail are discussed. With this architecture, obviously, that the delivery time depends on the conditions of the terrestrial segment, such as subscriber terminals, network interfaces at terminals, the degree of breaking in place, and etc. And the delivery time depends on the conditions of the flying segment, such as the method of UAV networks organization, the data transmission technologies. In this paper, we consider the requirement of the number of UAVs for voice delivery from the first subscriber to the second subscriber. We describe the processing of voice delivery by a multiphase queuing system model. In the connection between two subscribers, each UAV is presented by a single-phase queuing system.

A flying network consisting of UAVs is represented as a multiphase queuing system [6,7,10,23], which is shown in Fig. 2. Each UAV receives, processes and sends subscriber data and voice traffic to the next UAV node based on the service message exchange. The voice traffic, which is generated by the subscriber terminals, go into every UAV node. Because each UAV node is a single-phase queuing system, voice traffic is waited in queues in the path of departure to the subscriber terminal of destination. Therefore, the choice of the queuing models, which is used in each UAV node, significantly affects the delay in the transmission of voice between two subscribers.

It is assumed that the incoming streams to each UAV have the same properties. Accordingly, it is possible to compute the average delivery time for each subsystem of the multiphase queuing system for the models under consideration. For simplicity, we consider 2 types of queuing system models M/M/1 and G/G/1. The multiphase queuing system model with n queuing phases is shown in Fig. 3. The multiphase queuing system is understood that each UAV receives processes and sends subscriber's information and voice to the next UAV node.

Figure 3 shows that the voice transmission delay from subscriber 1 to subscriber 2 is represented by the sum of all average interarrival times in all phases, which is represented by formula (1).

$$\overline{T} = \overline{T_1} + \sum_{j=1}^{n} \overline{T_j} + \overline{T_n} \tag{1}$$

where:

- $\overline{T}$ – the sum of average interarrival time in all phases,
- $\overline{T_1}$ – the average interarrival time in the first phase (between subscriber 1 and UAV 1) (ms),

– $\overline{T_j}$ – the average interarrival time between UAVs (ms),
– $\overline{T_n}$ – the average interarrival time in the last phase (between UAV n and subscriber 2) (ms),
– n – the number of UAVs.

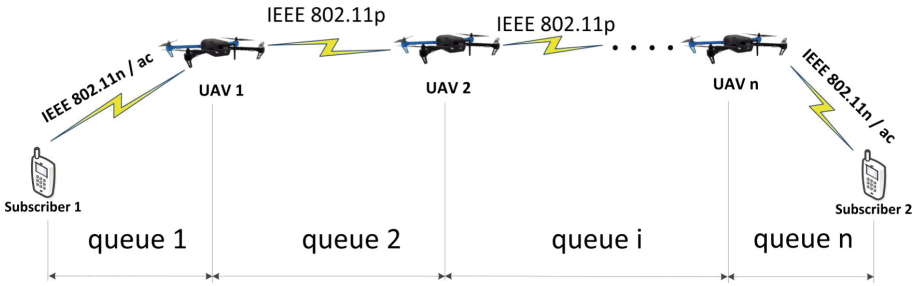


Fig. 2. Multiphase queuing model of a flying network for voice transmission

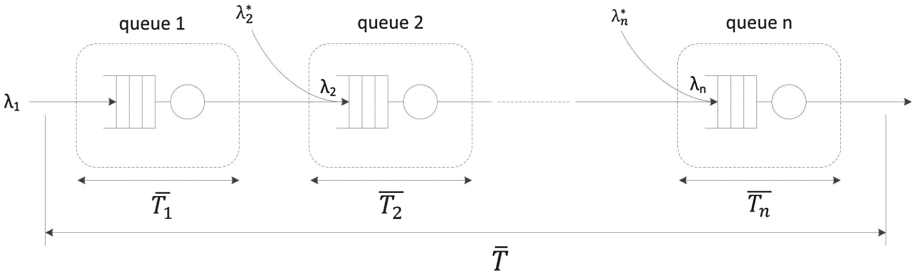


Fig. 3. Multiphase queuing model

According to the formula (1), the number of UAVs, which provide the transmission time of voice traffic between two subscribers with a delay not exceeding 100 ms, can be found. It means that $\overline{T} = \overline{T_1} + \sum_{j=1}^{n} \overline{T_j} + \overline{T_n} \leq 100$ ms. We assume that the average interarrival time between subscribers and UAVs are the same ($\overline{T_1} = \overline{T_n}$) and the average interarrival time between UAVs are also the same. Consequently, the number of UAVs can be found by expression (2):

$$2 * \overline{T_1} + (n-1)\overline{T_j} \leq 100 \Rightarrow n \leq \frac{100 - 2 * \overline{T_1}}{\overline{T_j}} + 1 \tag{2}$$

System load intensity can be found by formula (3):

$$\rho_i = \frac{\lambda_i}{\mu_i}(Erlang) \tag{3}$$

where:

– λ_i – The average arrival rate (packet/ms),
– μ_i – The average service rate (packet/ms),
– $\overline{t_i} = \frac{1}{\mu_i}$ – The average service time (ms).

When considering the two types of queuing systems M/M/1 and G/G/1, we will use formulas to calculate the packet processing time [17]. Formulas (2) and (3) show the average time of delivery, passing through the multiphase queuing system i for the models M/M/1 and G/G/1. For the M/M/1 system, the average delivery time can be calculated using formula (4):

$$\overline{T_i} = \overline{w_i} + \overline{t_i} = \frac{\overline{t_i}}{1 - \rho_i} \tag{4}$$

where:

– $\overline{T_i}$ – the average interarrival time in phase i (ms),
– $\overline{w_i}$ – the average time spent waiting in the queue (ms),
– $\overline{t_i}$ – the average service time (ms).

For the G/G/1 system, the average delivery time can be calculated using formula (5):

$$\overline{T} = \overline{w_i} + \overline{t_i} = \frac{\overline{t_i} * \rho_i}{1 - \rho_i} * \left(\frac{\overline{\sigma_a}^2 * \overline{\sigma_b}^2}{\overline{t_i}^2} \right) * \left(\frac{\overline{t_i}^2 * \overline{\sigma_b}^2}{a^2 + \overline{\sigma_b}^2} \right) \tag{5}$$

where:

– $\overline{T_i}$ – the average interarrival time in different phases (ms),
– $\overline{w_i}$ – the average time spent waiting in the queue (ms),
– $\overline{t_i}$ – the average service time (ms),
– $\overline{\sigma_a}^2$ – the variance of the time interval between arrivals,
– $\overline{\sigma_b}^2$ – the variance of service time,
– a – the average time interval size between arrivals.

As mentioned above, the voice transmission between subscribers and UAVs can be achieved by IEEE 802.11n data transmission standard (with data rate $b_n = 300$ Mbps) or IEEE 802.11ac (with data rate $b_{ac} = 650$ Mbps) and for the voice transmission between UAVs, IEEE 802.11p data transmission standard (with data rate $b_p = 12$ Mbps). The data rates are given as the average values after deduction of the exchange of service messages. Consider the average packet size $L = 1000$ bytes (8000 bits). As is known, the average service time is still considered using formula (6):

$$\overline{t_i} = \frac{L}{b_i}(ms) \tag{6}$$

where:

– L – the average packet size (bit),
– b_i – data rate (bit/ms).

5 Numerical Results Based on Mathematical Models

Using the formulas (1), (4), (5) and the parameters presented in Table 1, we get the voice delivery time with a change in the number of UAVs. The variation of the number of UAVs is carried out until the voice transmission time exceeds 100 ms, this is represented as an unacceptable quality of service for the voice transmission.

Table 1. Parameters for model G/G/1

	Communication between subscribers and UAVs (IEEE 802.11n)	Communication between subscribers and UAVs (IEEE 802.11ac)	Communication between UAVs (IEEE 802.11p)
$\overline{\sigma_a}^2$	0.2844	0.01	1.778
$\overline{\sigma_b}^2$	0.2	0.1	0.5
$\overline{t_i}$ (ms)	0.0267	0.0123	0.6667
a	0.053	0.025	1.333

The results of calculating the voice transmission time in this case are presented in Table 2.

As well as with the change in the load factor according to formulas (1), (2), (4), (5) and the parameters presented in Table 1, we get the voice transmission time between subscribers and the number of UAVs that can provide communication between them. The results in this case are presented in Tables 3 and 4.

6 Analysis of Results

Table 2 shows the voice transmission time between subscribers with the change of the number of UAVs for the two models when the load factor of the whole system is 0.5. Accordingly, the number of UAVs necessary to cover the disaster area, is found. When considering the two models M/M/1 and G/G/1, there is a big difference in the maximum number of UAVs necessary to provide the voice transmission delay less than 100 ms. Using M/M/1 model, we got the number of UAVs (50 pcs), which is twice number of UAVs (24 pcs) when using the G/G/1 model. Therefore, the service area with the increase in the number of UAVs also expands. With the same number of UAVs, the voice transmission delay using the M/M/1 model is less than using the G/G/1 model. Tables 3 and 4 show the maximum required number of UAVs with increasing system load factor. From these results we can see that there is a difference in the delay when using IEEE 802.11n and IEEE 802.11ac technologies, which provide the communication between subscribers and UAVs. The number of UAVs decreases with increasing system load factor. Consequently, the service area for two subscribers

Table 2. The voice transmission time between two subscribers, using different data transmission standards and multiphase queuing models

Number of UAVs, n	Voice transmission time between subscribers and UAV using IEEE 802.11n between the UAV using IEEE 802.11p		Voice transmission time between subscribers and UAV using IEEE 802.11ac between the UAV using IEEE 802.11p	
	M/M/1	G/G/1	M/M/1	G/G/1
1	0.16	2.286	0.074	1.578
2	2.16	6.421	2.074	5.713
3	4.16	10.556	4.074	9.848
4	6.16	14.691	6.074	13.983
5	8.16	18.826	8.074	18.118
...	...	...	...	..
23	44.16	93.256	44.074	92.548
24	46.16	**97.391**	46.074	**96.683**
25	48.16	101.526	48.074	100.818
...	...	...	...	...
48	94.16	196.631	94.074	195.923
49	96.16	200.766	96.074	200.058
50	**98.16**	204.901	**98.074**	204.193
51	100.16	209.036	100.074	208.328

Table 3. Results of the voice transmission time and number of UAVs, using IEEE 802.11n for communication between subscribers and UAVs and IEEE 802.11p for communication between UAVs

Load factor, ρ	M/M/1		G/G/1	
	Voice transmission time T (ms)	Number of UAVs	Voice transmission time T (ms)	Number of UAVs
0.1	1.52	71	1.422	95
0.2	1.62	67	2.3	65
0.3	1.75	62	3.429	46
0.4	1.92	57	4.935	33
0.5	2.16	50	7.041	24
0.6	2.52	43	10.201	17
0.7	3.12	35	15.469	11
0.8	4.32	25	26.005	7
0.9	15.12	8	57.609	3

Table 4. Results of the voice transmission time and number of UAVs, using IEEE 802.11ac for communication between subscribers and UAVs and IEEE 802.11p for communication between UAVs

Load factor, ρ	M/M/1		G/G/1	
	Voice transmission time T (ms)	Number of UAVs	Voice transmission time T (ms)	Number of UAVs
0.1	1.46	72	1.25	95
0.2	1.56	67	1.946	65
0.3	1.68	62	2.843	47
0.4	1.84	57	4.039	34
0.5	2.07	50	5.713	24
0.6	2.42	43	8.223	17
0.7	2.99	35	12.41	12
0.8	4.15	25	20.77	7
0.9	14.52	8	45.88	3

becomes narrower or the distance between them is extremely short. According to the data in the tables, we can see that when using the M/M/1 model, 25 UAVs are required with the load factor equals to 0.8 while when using the G/G/1 model requires 24 UAVs with the load factor equals to 0.5.

7 Conclusion

The article considered the architecture of a network model for connecting mobile subscribers in the disaster area when telecommunications infrastructure are destructed. The network model is organized on the basis of a flying network, in which IEEE 802.11p technology is used for UAVs communication, and IEEE 802.11n/ac technology for communication between UAVs and mobile phones. The article analyzed the models of multiphase queuing system type M/M/1 and G/G/1, which are considered for UAVs communication, as well as for communication between mobile phones and UAVs. For each model, we calculated the voice transmission delay and the number of UAVs, at which permissible quality of service of calls in the disaster zone can be guaranteed. The results show that it is possible to establish the number of UAVs needed to cover the disaster zone in various cases. This can be of considerable assistance in the search and rescue of victims of a natural disaster.

References

1. De Freitas, E.P., et al.: UAV relay network to support WSN connectivity. In: Proceedings of the International Congress on Ultra Modern Telecommunications and Control Systems, Moscow, pp. 309–314. IEEE (2010)
2. Orfanus, D., Eliassen, F., de Freitas, E.P.: Self-organizing relay network supporting remotely deployed sensor nodes in military operations. In: 6th International Congress on Ultra-Modern Telecommunications and Control Systems and Workshops (ICUMT), St. Petersburg, pp. 326–333. IEEE (2014)
3. Vasiliev, D.S., Meitis, D.S., Abilov, A.: Simulation-based comparison of AODV, OLSR and HWMP protocols for flying ad hoc networks. In: Balandin, S., Andreev, S., Koucheryavy, Y. (eds.) NEW2AN 2014. LNCS, vol. 8638, pp. 245–252. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10353-2_21
4. Bekmezci, I., Sahingoz, O.K., Temel, S.: Flying ad-hoc networks: a survey. Ad Hoc Netw. **11**, 1254–1270 (2013)
5. Koucheryavy, A., Vladyko, A., Kirichek, R.: State of the art and research challenges for public flying ubiquitous sensor networks. In: Balandin, S., Andreev, S., Koucheryavy, Y. (eds.) ruSMART 2015. LNCS, vol. 9247, pp. 299–308. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23126-6_27
6. Kirichek, R., Paramonov, A., Koucheryavy, A.: Flying ubiquitous sensor networks as a queuing system. In: Proceedings of the 17th ICACT, pp. 127–132 (2015)
7. Kirichek, R., Paramonov, A., Koucheryavy, A.: Swarm of public unmanned aerial vehicles as a queuing network. In: Vishnevsky, V., Kozyrev, D. (eds.) DCCN 2015. CCIS, vol. 601, pp. 111–120. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30843-2_12
8. Kirichek, R., Vladyko, A., Paramonov, A., Koucheryavy, A.: Software-defined architecture for flying ubiquitous sensor networking. In: International Conference on Advanced Communication Technology, ICACT, pp. 158–162 (2017)
9. Shilin, P., Kirichek, R., Paramonov, A., Koucheryavy, A.: Connectivity of VANET segments using UAVs. In: Galinina, O., Balandin, S., Koucheryavy, Y. (eds.) NEW2AN/ruSMART -2016. LNCS, vol. 9870, pp. 492–500. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46301-8_41
10. Kirichek, R., Kulik, V.: Long-range data transmission on flying ubiquitous sensor networks (FUSN) by using LPWAN protocols. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds.) DCCN 2016. CCIS, vol. 678, pp. 442–453. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-51917-3_39
11. Schalk, L.M.: Communication links for unmanned aircraft systems in very low level airspace. In: Integrated Communications, Navigation and Surveillance Conference (ICNS), Herndon, p. 6B2-1. IEEE (2017)
12. Li, J., Zhou, Y., Lamont, L.: Communication architectures and protocols for networking unmanned aerial vehicles. In: Globecom Workshops (GC Wkshps), Atlanta, pp. 1415–1420. IEEE (2013)
13. Andre, T., et al.: Application-driven design of aerial communication networks. IEEE Commun. Mag. **52**(5), 129–137 (2014)
14. Frew, E.W., Brown, T.X.: Airborne communication networks for small unmanned aircraft systems. Proc. IEEE **96**(12), 2008–2027 (2008)
15. Morgenthaler, S., Braun, T., Zhao, Z., Staub, T., Anwander, M.: UAVNet: a mobile wireless mesh network using unmanned aerial vehicles. In: Globecom Workshops (GC Wkshps), Anaheim, pp. 1603–1608. IEEE (2012)

16. Van der Bergh, B., Chiumento, A., Pollin, S.: LTE in the sky: trading off propagation benefits with interference costs for aerial nodes. IEEE Commun. Mag. **54**(5), 44–50 (2016)
17. Da Conceicao, A.F., Li, J., Florencio, D.A., Kon, F.: Is IEEE 802.11 ready for VoIP? In: IEEE 8th Workshop on Multimedia Signal Processing, pp. 108–113 (2006)
18. Chagh, Y., Guennoun, Z., Jouihri, Y.: Voice service in 5G network: towards an edge-computing enhancement of voice over Wi-Fi. In: 39th International Conference on Telecommunications and Signal Processing (TSP), pp. 116–120 (2016)
19. Ngongang, S.F.M., Tadayon, N., Kaddoum, G.: Voice over Wi-Fi: feasibility analysis. In: Advances in Wireless and Optical Communications (RTUWO), pp. 133–138 (2016)
20. IEEE Standards Association. Part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications. IEEE std, 802 (2012)
21. Kim, S.-Y., Ro, J.-H., Song, H.-K.: Channel estimation scheme for the enhanced reliability in the flying ad-hoc network. Int. J. Eng. Res. Appl. **7**(4), 63–66 (2017)
22. ITU-T Recommendation G.114. One-way transmission time (2003)
23. Kleinrock, L.: Queueing Systems. Volume 2: Computer Applications, vol. 66. Wiley, New York (1976)

Statistical Clustering of a Random Network by Extremal Properties

Natalia M. Markovich[1(✉)], Maxim S. Ryzhov[1], and Udo R. Krieger[2]

[1] V.A. Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Profsoyuznaya Str. 65, 117997 Moscow, Russia
markovic@ipu.rssi.ru

[2] Fakultät WIAI, Otto-Friedrich-Universität, An der Weberei 5, 96047 Bamberg, Germany
udo.krieger@ieee.org

Abstract. We propose the new EI-clustering method for random networks. Regarding the underlying graph of a random network, EI-clustering is an advanced statistical tool for community detection and based on the estimation of the extremal index (EI) associated with each node. The EI metric is estimated by samples of indices of the node influences. The latter quantities are determined by the PageRank and a Max-Linear Model. The EI values of both models are estimated by a blocks estimator for each node which is considered as the root of a Thorny Branching Tree. Generations of descendant nodes related to the root node of the tree are used as blocks. The reciprocal of the EI value indicates the average number of influential nodes per generation containing at least one influential node. In the context of random graphs the EI metric indicates the ability of a randomly selected node to attract highly ranked nodes in its orbit. Looking at the changing shape of a plot of the EI metric versus the node number, the node communities are detected. The EI-clustering method is compared with the conductance measure regarding the data set of a real Web graph.

Keywords: Clustering · Node influence · PageRank
Max-Linear Model · Extremal index · Web graph

1 Introduction

Random graphs are accepted as realistic models of real world networks such as technological networks, e.g. Internet, P2P, and transportation networks, social and information networks, [6,7,15]. Given the massive amount of data about real networks, the determination of the importance of nodes, a fast finding of the most influential nodes in a graph and the efficient detection of communities, i.e. clustering of nodes that are similar in some sense, constitute important research problems.

Clustering tools for random graphs such as the "null model" or the "configuration model" (see [6,15] for a survey) do not take into account the distributions

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 71–82, 2018.
https://doi.org/10.1007/978-3-319-99447-5_7

of extremes regarding the influence indices of nodes and their dependency. The intuition of those clustering methods is mostly based on the idea that nodes which are interconnected by a large number of edges are likely belonging to the same community. Then one has to find sets of nodes with a high internal connectivity that are highly disconnected between each other by calculating the number of edges of all the vertices in the network. Such approaches avoid the consideration of random graphs as "locally tree-like" structures due to loops and possible edges between vertices belonging to the same generation of a root node. A random graph with such a tree structure is called Thorny Branching Tree (TBT), [4].

These sketched clustering methods are based on an a-posteriori state of the network and do not allow to take into account random changes of links within the network and an on-line detection of structures. Moreover, a sudden explosion of edges of a node caused by local or temporary rare events like catastrophes in markets or sport events in social networks, that may generate new giant clusters, cannot be predicted by such tools. Those features require a new methodology concerning extremes of nodes in random graphs. It was theoretically derived in [11] that the tail index and extremal index of the PageRank (PR) and the Max-Linear Model (MLM) that are used as influence indices of the nodes in random graphs coincide. In [13] we have checked this property by means of real data of a Web graph.

In this paper it is our first objective to develop a clustering algorithm for the detection of communities of similar nodes in a random network. To partition a random network into (weakly dependent) communities around highly ranked nodes, we propose to use the extremal index (EI). The latter is a measure of dependence of extremes, [2,8]. The reciprocal of the EI approximates the mean cluster size of a structure. In this context a cluster is determined as a block of data with at least one exceedance over a given threshold. In a random graph the EI metric indicates the ability of a node u to perform clustering or, in other words, to attract influential nodes following u in its orbit.

In [1] semi-supervised learning methods are proposed for weighted similarity graphs. Two sets of nodes are determined as similar if they are connected by an edge and the weight of the edge indicates the strength of the similarity. We determine a community as a galaxy or a cluster of nodes related to a node with a large influence index, i.e. we detect extremes of the network and node followers of an extremal node as its community. Our approach is somewhat similar to [1], where local learning sets of similar nodes are used to build classifiers. We estimate the EI value of each node considering this entity as the root of a TBT, i.e. a branching tree with possible loops.

Our second objective is to compare the clustering approach that is based on the change of the conductance of a graph proposed in [10] by our EI-clustering based on the EI indices of the nodes.

The paper is organized as follows. In Sect. 2 basic results related to the detection of node communities and EI-estimation methods are sketched. Section 3 presents the EI-clustering algorithm and its comparison with the clustering based

on the conductance measure proposed in [10] by means of a real Web data set. The exposition is finalized by some conclusions.

2 Fundamental Properties of Random Networks

2.1 Community Detection by Clustering Methods

Surveys about graph partitioning techniques are presented in [6,10]. They reveal that the conductance metric and the clustering coefficient are important characteristics of random graphs.

The conductance measure of a set S of nodes in a random graph is calculated as

$$\phi = s/v, \quad \text{or} \quad \phi = s/(s + 2e). \tag{1}$$

Here s is the number of edges with one endpoint in S and one endpoint in $\overline{S}$, where $\overline{S}$ denotes the complement of S. v is the sum of degrees of nodes in S, and e is the number of edges with both endpoints is S, [10]. Small conductance of the set means that it is densely linked inside itself. It is remarkable that shape changes of the plot of ϕ against the number of nodes, called the Network Community Profile plot, indicate disconnected clusters.

The clustering coefficient C of a random network [15] is determined as

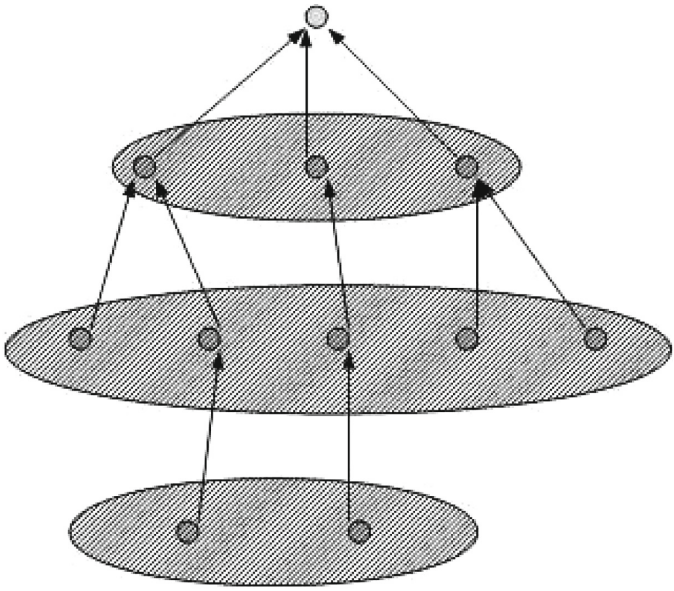
$$C = \frac{3 \cdot \text{number of triangles in the network}}{\text{number of connected triples}},$$

where a connected triple implies a single vertex connected by edges to two others. The triangle of nodes is considered as a basic social community. $C = 0$ means the lack of triangles. Then a part of the network associated with some node can be represented as a branching tree (see Fig. 1(a)). Descendants of each generation of the node taken as the root of a TBT are not linked and thus, independent. In reality, due to links between descendants within and between generations (see Fig. 1(b)), the dependence is determined in [15] by means of triangles.

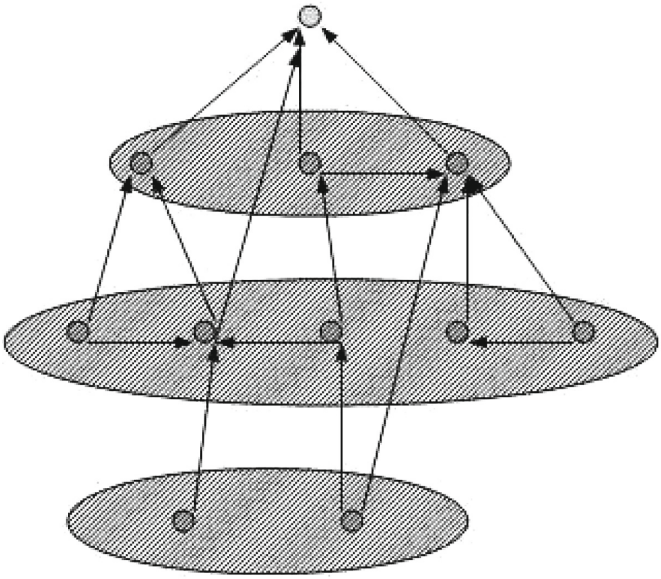
Our approach is to estimate the extremal index (EI) of each node influence in the random network. The EI value shows the ability of a node to create a cluster. It reveals the community built around the underlying node as a part of the Thorny Branching Tree associated with this node as its root.

2.2 Influence Indices of a Node

In fact, the in- and out-degrees of nodes are the only statistics gathered about a random network, [10]. The influence of a node may be determined by its in-degree, its PageRank (PR) and the Max-Linear Model (MLM), [12,13]. Both latter statistics are calculated by the in-degree and out-degree of a node and the PR values of those nodes that point to it. The PR of a node, e.g. a Web page, grows with the PRs of those nodes pointing to it and with the in-degree of the node. It can be calculated by different methods, see for instance [3,4,14].



(a)



(b)

Fig. 1. Branching tree corresponding to a cluster coefficient $C = 0$ (a); Branching tree with short loops containing triangles and single edges corresponding to $0 < C < 1$ (b); Generations of nodes following the root node are shown by grey ellipses.

2.3 The Extremal Index

Definition 1. ([8, p. 53])
The stationary sequence $\{R_n, n \geq 1\}$ *is said to have extremal index* $\theta \in [0,1]$ *if for each* $0 < \tau < \infty$ *there is a sequence of real numbers* $u_n = u_n(\tau), n \in \mathbb{N}$, *such that it holds*

$$\lim_{n\to\infty} n\overline{F}(u_n) = \lim_{n\to\infty} n(1 - F(u_n)) = \tau, \quad \lim_{n\to\infty} P\{M_n \leq u_n\} = e^{-\tau\theta}, \tag{2}$$

where $M_n = \max\{R_1, ..., R_n\} = \bigvee_{j=1}^{n} R_j$ *holds.*

Conditions (2) determine the thresholds u_n in such a way that the probability to exceed it are very small. This feature corresponds to high quantiles of $\{R_n\}$. For independent r.v.s $\theta = 1$ holds, but the converse is not true. $\theta \approx 0$ implies a strong dependence. $\theta = 0$ implies that the maximum M_n likely does not exceed a sufficiently high threshold $u \in \mathbb{R}$. As it holds [2]

$$\theta = \lim_{n\to\infty} \frac{P\{M_{r_n} > u_n\}}{r_n(1 - F(u_n))},$$

where $r_n = o(n)$ as $n \to \infty$, $P\{M_{r_n} > u_n\}$ implies the probability that a data block of size r_n contains at least one exceedance over the threshold u_n (such block is called a cluster) and $r_n(1 - F(u_n))$ shows the fraction of exceedances in the sample. Hence,

$$1/\theta \approx \frac{\text{number of exceedances}}{\text{number of clusters}} \tag{3}$$

approximates the mean cluster size.

With regard to a random graph the EI value of a node influence implies that each generation of descendants of the node that are accepted as blocks has on average $[1/\theta]$ nodes with high influence values. As the influence of a root node decreases over generations, one can use a truncated TBT such that its leaves do not impact significantly on the PR of the root. Such a truncation rule is specified in [13].

$\theta = 1$ or $\theta \approx 1$ mean that followers of a node which have links to a root node have independent or approximately independent influence characteristics. $\theta \approx 0$ implies a condensed cluster with regard to dependent influence values of followers.

2.4 Bias-Reduced Estimation of the Extremal Index

Nonparametric estimators of the extremal index (EI), like the blocks, runs, or intervals estimator, [2], are calculated by sequences of an underlying node characteristic. These estimators are usually calculated by random sequences or time series. They are distinguished by the definitions of the cluster. In this respect their application to random graphs constitutes further problems. The EI-estimators require one or two parameters. To apply the blocks estimator, for

instance, one can split the sample $\{X_1, ..., X_n\}$ of size n into $m_n = \lfloor n/r_n \rfloor \in \mathbb{N}$ blocks of length $r_n \in \mathbb{N}$. Then the blocks estimator is determined by

$$\widehat{\theta}_n = \frac{\sum_{j=1}^{m_n} \mathbb{1}\{\max_{(j-1)r_n < i \le jr_n} X_i > u_n\}}{\sum_{j=1}^{m_n} \sum_{i=(j-1)r_n+1}^{jr_n} \mathbb{1}\{X_i > u_n\}} \tag{4}$$

for a sequence of thresholds $u_n \in \mathbb{R}$ satisfying $r_n \overline{F}(u_n) \to 0$, but $n\overline{F}(u_n) \to \infty$.

Dealing with graphs and to create a sequence, we can only use the blocks estimator and its modifications to avoid the numeration of nodes in the graph. The blocks estimator requires the block size $r \in \mathbb{N}$ and the threshold $u \in \mathbb{R}$ as its parameters.

There are bias-reduced estimators of the EI which avoid a selection of u, [5,16]. These estimators have the advantage that their plots $\widehat{\theta}_n$ against u are stable. The stable plateau that is close to a constant indicates the estimated EI-value. Hence, one has to select only an appropriate parameter r.

Let $\{X_i, 1 \le i \le n \in \mathbb{N}\}$ be an univariate stationary time series with distribution function F and extremal index $\theta \in (0, 1]$. One may use the following bias-reduced estimator [5]

$$\widehat{\theta}_{n,t} = \frac{\sum_{j=1}^{m_n} \mathbb{1}\{\max_{(j-1)r_n < i \le jr_n} X_i > X_{n-\lceil nv_n t \rceil, n}\}}{\sum_{j=1}^{m_n} \sum_{i=(j-1)r_n+1}^{jr_n} \mathbb{1}\{X_i > X_{n-\lceil nv_n t \rceil, n}\}}, \quad t \in (0, 1], \tag{5}$$

where the sequence $\{v_n, n \in \mathbb{N}\}$ is such that $r_n v_n \to 0$, $nv_n \to \infty$ as $n \to \infty$.

In [12,13] generations of descendants of the TBT root node (see Fig. 1(b)) were proposed as blocks. Due to loops these blocks can be overlapping since the same node can be assigned to different generations. Such node may appear in one of the blocks or in all blocks which contain it. One can consider sets of nodes located on a path with m links (edges) from the root as the mth generation. To find the block size automatically and to build confidence intervals of the EI-estimate we use the bootstrap method described in [13].

3 The EI-Clustering Method

3.1 The EI-Clustering Algorithm

We study the clustering of n nodes in a random network using the extremal index of an influence characteristic of the nodes.

Algorithm 31

1. *Estimate the PR and the MLM values of each node by one of the recurrent methods in [13].*
2. *Estimate the EI values using samples of the PRs and MLMs of the lengths $k \in \{1, \dots, n\}$ by means of the blocks estimator (4) or by its bias-reduced modification (5). Find a threshold u by the bootstrap method [13] for a given block size r in the case (4). In the case (5) r can be found by the bootstrap method for a given parameter $0 < t \le 1$ (u is not required).*

3. *Given the EIs of the samples and enlarging the lengths* $k, k+1, \ldots, n$, *one calculates the average EI among the nodes of each of those samples.*
4. *Partition the nodes into clusters according to changes of the shape regarding the curves (node number, PR) or (node number, MLM).*

3.2 The Bootstrap Algorithm

Let us further interpret k and k_1 as the total numbers of exceedances over the threshold $u \in \mathbb{R}$ in the sample $\{X_i, 1 \leq i \leq n \in \mathbb{N}\}$ and in the bootstrap re-sample $\{X_i{}^*, 1 \leq i \leq n_1 \in \mathbb{N}\}$ of a smaller size $n_1 < n$, respectively, i.e.

$$k = \sum_{i=1}^{n} \mathbb{1}(X_i > u), \quad k_1 = \sum_{i=1}^{n_1} \mathbb{1}(X_i{}^* > u). \tag{6}$$

Then one can find u corresponding to the selected k and find the estimate of the extremal index $\widehat{\theta}(u)$.

Algorithm 32

1. *Generate* B *re-samples* $\{X_1^*, \ldots, X_{n_1}^*\}$ *of size* $n_1 < n$ *with replacement from the original observations* $\{X_i, i = 1, \ldots, n\}$, *where* n_1 *is defined as*

$$n_1 = n^{\beta_b}, \qquad 0 < \beta_b < 1.$$

The number of the largest order statistics $k_1 \in \{1, ..., n_1 - 1\}$ *corresponding to any re-sample relates to* k *and* n *by*

$$k = k_1 \left(\frac{n}{n_1}\right)^{\alpha_b}, 0 < \alpha_b < 1. \tag{7}$$

2. *Estimate* B *values* $\widehat{\theta}_{n_1}$ *using the blocks estimator (4) by each of* B *re-samples.*
3. *Calculate the mean squared error (MSE) by the re-samples,*

$$MSE\,(n_1, k_1) = (bias\,(n_1, k_1))^2 + var\,(n_1, k_1)\,, \tag{8}$$

where the bias and variance are the following quantities

$$bias\,(n_1, k_1) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\theta}_{n_1} - \widehat{\theta}_n,$$

$$var\,(n_1, k_1) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}_{n_1} - \widehat{\theta}_{n_1}\right)^2,$$

and find a minimal $MSE\,(n_1, k_1)$ *among different* $k_1 \in \{1, ..., n_1 - 1\}$.
4. *Using the obtained* k_1 *find the optimal* k *by (7) and then the corresponding estimate* $\widehat{\theta}_n$ *by (4).*

In this case the values α_b and β_b are not precisely known due to the lack of theory and we may take $\alpha_b = 2/3$ and $\beta_b = 1/2$ similar to the tail index estimation [13].

3.3 Comparison of Clustering Approaches

Similar to [13] we study a Web graph of the Berkeley-Stanford dataset [10]. Therein, nodes represent Web pages and edges represent hyperlinks between those pages, [9]. The selected graph contains 685230 nodes and 7600595 edges.

Figures 2 and 3 visualize the clustering of the 10000 most influential nodes regarding their influence indices based on PR and the MLM as well as the EI metrics that are estimated by the MLM values of each node by the blocks estimator (4) combined with the bootstrap method.

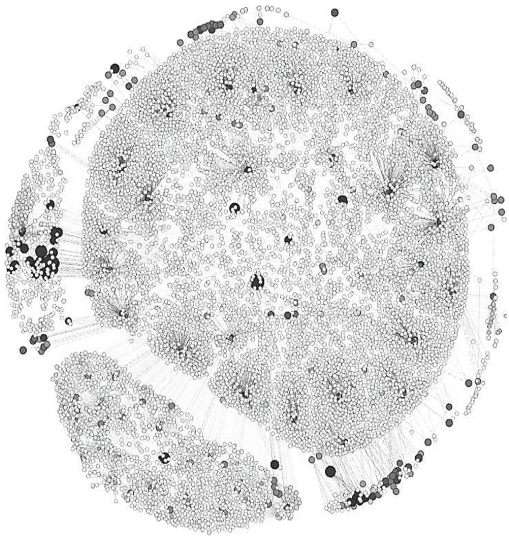


Fig. 2. Clustering of the sub-graph of the 10000 most influential nodes with regard to PR and the MLM from the Berkeley-Stanford dataset: the more intensive colour reflects the larger value of PR and, the bigger the circle is the larger is the MLM.

We apply Algorithm 31 to partition the network nodes into clusters both by the EI-values and conductance metrics for all nodes of the considered sub-graph. Figures 4 and 5 show that the changing shape of the conductance metrics (1) is close to those ones arising from the EI plots of the PR and the MLM of corresponding nodes. The EI-plot of PR is more sensitive than that one of the MLM regarding the community detection. The EI-values of PR and MLM tend to be similar for larger values of k which is in the agreement with the theoretical conclusions [11].

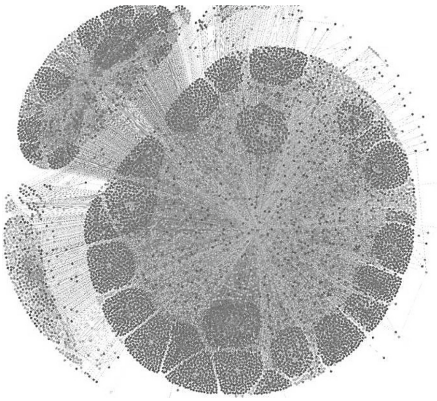


Fig. 3. Clustering of the same graph as in Fig. 2 with regard to EI values of the MLM: the more and the less intensive colours imply the EIs that are close to 1 and to 0, respectively.

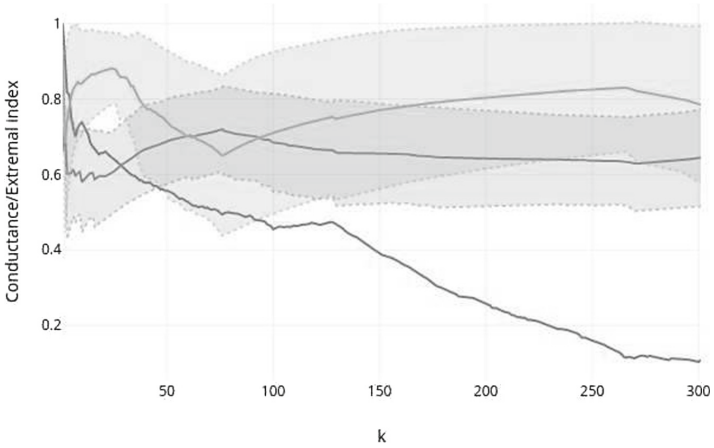
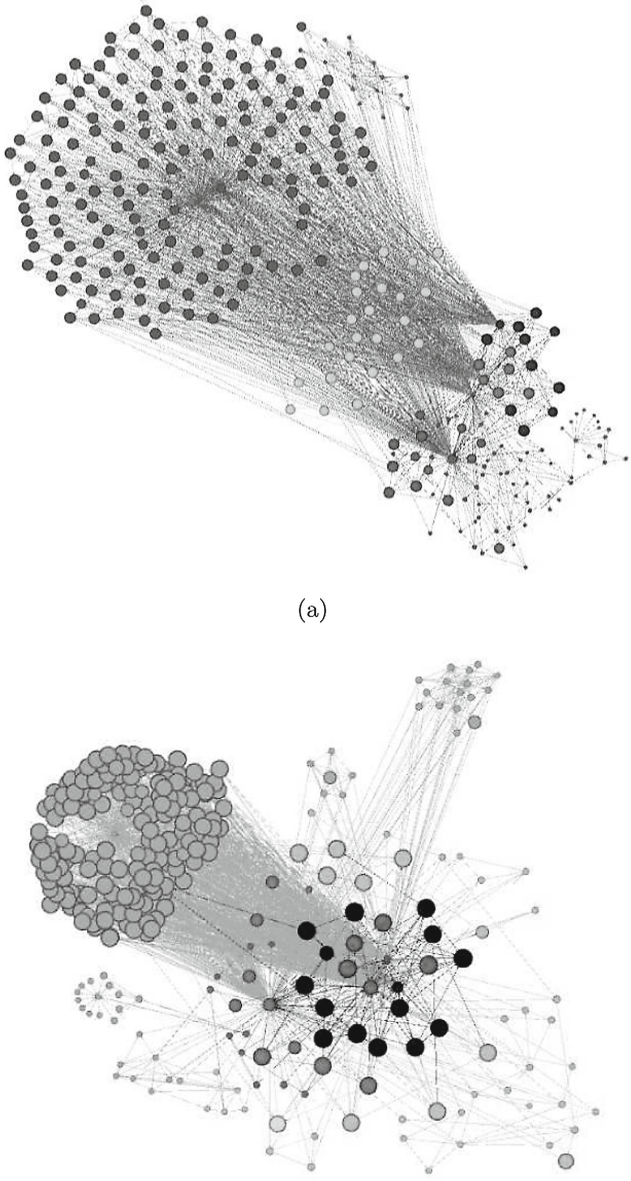


Fig. 4. Conductance (lower line at $k = 300$), the blocks estimates of the EI values of PR (upper line at $k = 300$) and the MLM (middle line at $k = 300$) of nodes with their 95% bootstrap confidence intervals against the numbers of nodes k.



(a)

(b)

Fig. 5. Clusters of nodes corresponding to conductance steps (a) and to steps of the EI-plot of the PR (b); Grey circles at left-hand side on top correspond to the smallest conductance values and EIs equal to 0.7.

4 Conclusions

We have proposed the EI-clustering algorithm for random networks. It is a new tool for community detection in random graphs based on the estimation of the extremal index (EI) of each node from an underlying vertex set. In contrast to known approaches, the EI index provides the dependence of extremes in the graph and shows the ability of a randomly selected node to attract highly ranked nodes in its orbit.

The EI-plots built by the PageRanks and the Max-Linear Model of nodes are compared with the conductance plot for a real data set. Finally, communities of nodes are partitioned corresponding to the changes of the shapes regarding the latter plots.

Our future study will elaborate on an on-line clustering algorithm for random networks.

References

1. Avrachenkov, K., Gonçalves, P., Sokol, M.: On the choice of kernel and labelled data in semi-supervised learning methods. In: Bonato, A., Mitzenmacher, M., Prałat, P. (eds.) WAW 2013. LNCS, vol. 8305, pp. 56–67. Springer, Cham (2013). https://doi.org/10.1007/978-3-319-03536-9_5
2. Beirlant, J., Goegebeur, Y., Teugels, J., Segers, J.: Statistics of Extremes: Theory and Applications. Wiley, Chichester (2004)
3. Borkar, V.S., Mathkar, A.S.: Reinforcement learning for matrix computations: pagerank as an example. In: Natarajan, R. (ed.) ICDCIT 2014. LNCS, vol. 8337, pp. 14–24. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-04483-5_2
4. Chen, N., Litvak, N., Olvera-Cravioto, M.: Pagerank in scale-free random graphs. In: Bonato, A., Graham, F.C., Prałat, P. (eds.) WAW 2014. LNCS, vol. 8882, pp. 120–131. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13123-8_10
5. Drees, H.: Bias correction for estimators of the extremal index. arXiv:1107.0935 (2011)
6. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**(3–5), 75–174 (2010)
7. van der Hofstad, R.: Random Graphs and Complex Networks. Cambridge University Press, Cambridge (2016)
8. Leadbetter, M.R.: Probability Theory and Related Fields. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **65**(2), 291–306 (1983)
9. Leskovec, J., Krevl, A.: SNAP Datasets: Stanford Large Network Dataset Collection (2014). http://snap.stanford.edu/data
10. Leskovec, J., Lang, K. J., Dasgupta, A., Mahoney, M. W.: Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. arXiv:0810.1355 (2008)
11. Markovich, N.M.: Extremes in Random Graphs Models of Complex Networks. arXiv:1704.01302v1 [math.ST], 5 April 2017
12. Markovich, N.M.: Analysis of clusters in network graphs for personalized web search. IFAC-PapersOnLine **50**(1), 5178–5183 (2017)
13. Markovich, N.M., Ryzhov, M., Krieger, U.R.: Nonparametric analysis of extremes on web graphs: pagerank versus max-linear model. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds.) DCCN 2017. CCIS, vol. 700, pp. 13–26. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66836-9_2

14. Nazin, A.V., Polyak, B.T.: Randomized algorithm to determine the eigenvector of a stochastic matrix with application to the PageRank problem. Autom. Remote Control **72**(2), 342–352 (2011)
15. Newman, M.E.J.: Random graphs with clustering. Phys. Rev. Lett. **103**, 058701 (2009)
16. Sun, J., Samorodnitsky, G.: Estimating the extremal index, or, can one avoid the threshold-selection difficulty in extremal inference? Reports of Cornell University (2010)
17. Volkovich, Y., Litvak, N.: On the exceedance point process for a stationary sequence. Adv. Appl. Prob. **42**(2), 577–604 (2010)

On Proximity-Based Information Delivery

Dmitry Namiot[1] and Manfred Sneps-Sneppe[2(✉)]

[1] Faculty of Computational Mathematics and Cybernetics,
Lomonosov Moscow State University,
GSP-1, 1-52, Leninskiye Gory, Moscow 119991, Russia
dnamiot@gmail.com
[2] Ventspils International Radio Astronomy Centre, Ventspils University College,
Inzenieru 101a, Ventspils 3601, Latvia
manfreds.sneps@gmail.com

Abstract. In this paper, we propose and discuss one approach to a data sharing among mobile subscribers. Our idea is to use the identification of wireless networks to simulate some analogue for a peer-to-peer network that will work in the absence of telecommunications infrastructure. A single mobile phone (smartphone) will be sufficient both for creating a node of such telecommunications network and for publishing (disseminating) information. Our proposal is the further development of ideas related to context-aware systems based on network proximity principles. The proposed model allows mobile users to create information hubs directly at the location of the mobile phone of the publisher, which will distribute information for mobile subscribers in the immediate vicinity of it.

Keywords: Bluetooth · Network proximity · BLE · Services

1 Introduction

Mobile networks that can function without telecommunications operators (without their infrastructure) are attracting increasing attention. One of the most famous examples is FireChat messenger [1]. Our early work [2] contains several other examples of such applications. It describes so-called mesh networks. Mesh networking is a network topology in which a device (a network node) transmits its own data and at the same time serves as a gateway for data transfer of other nodes. In other words, all nodes of the network interact during the message transfer process. The word "mobile" in the name of this type of network indicates the mobility of its nodes [3].

As per the classification of such systems (see, for example, our paper [2]), the following classes were distinguished: Mobile Ad-hoc NETworks (MANET), Wireless Mesh Networks (WMN), and Social Mesh Networks (SMN).

MANET is a self-configuring network of mobile devices. The term 'ad-hoc' for networks usually refers to a network connection established for one session. For example, the standards of wireless networks (Bluetooth, Wi-Fi, etc.) allow

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 83–94, 2018.
https://doi.org/10.1007/978-3-319-99447-5_8

direct connections of devices (at an achievable range). In MANET, neighboring devices interact with each other in ad-hoc mode, not only to exchange their own data, but also to transfer data from other nodes that cannot communicate directly [4].

VANET (Vehicular Ad hoc NETwork) uses cars as mobile nodes in MANET [5]. Wireless mesh network (WMN) is a mesh network of static routers that directly serve their customers. In other words, there is no direct connection between customers.

Social Mesh Networks (SMN) use existing equipment (mobile phones) to transfer data in the absence of network infrastructure. This is not about transferring data in point-to-point mode (or not only about it), but about so-called multi-hop connections, where data can be transmitted through several intermediate nodes.

Another interesting term (abbreviation) that is used in this connection is Mobile Networking in Proximity - MNP, which denotes the exchange of data between devices without available Internet connections. We would like to note in this regard an interesting patent from Facebook [6]. It describes a system that allows individuals and advertisers to connect directly to physically close users (devices) that have similar interests or are open to receiving certain advertisements.

The areas of application for the described networks differ. MANET is referred often to as military communications. Also, ad hoc networks are often used in emergency situations. In the field of personal communications, there are some communities created on the fly, groups for discussion, etc. For SMN, a typical application is the delivery of any personalized messages (e.g., it is a delivery of coupons with discounts to mobile subscribers who found themselves in physical proximity to a particular store, restaurant, etc. [6]).

The rest of the article is structured as follows. In Sect. 2, we describe similar works. Section 3 is devoted to the actual proposed system of data dissemination (publication). In Sect. 4, we discuss the technical details.

2 On Related Works

In this section, we want to focus on works and projects relevant to this topic. In its most general form, the idea is to use the identification of wireless networks for the dissemination of information.

Technically, network proximity could be defined with Wi-Fi, Core Bluetooth or Bluetooth Low Energy (BLE). All protocols are supported by the modern smartphones [7]. For Wi-Fi proximity (Bluetooth proximity), we can detect addresses (mac-address) and signal strength (RSSI) for access points (wireless nodes). Note, that it could be an existing wireless node and/or access point especially created for proximity measurements. Comparing with Bluetooth proximity, Wi-Fi based network proximity could be used on the bigger distance (so-called Wi-Fi distance versus so-called Bluetooth distance). The biggest disadvantage for Wi-Fi proximity is the lack of possibility to create Wi-Fi access point programmatically. In other words, the special node for proximity measurement could

not be created dynamically in case of Wi-Fi. Alternatively, Bluetooth proximity works on the smaller distance (Bluetooth distance) but it is possible to create Bluetooth objects (tags) programmatically.

For Core Bluetooth, we need to switch a Bluetooth device in so-called discoverable mode. In this mode, other devices can see device name and address, as well as obtain RSSI (the signal strength). Note that network proximity in general (and Bluetooth proximity too) has nothing to do with the connectivity [7].

In general, this topic refers to the so-called context-dependent programming (Context-Aware programming [8]) and mobile intelligence systems (Ambient mobile intelligence [9]). These articles are typical works in which the idea of network proximity is promoted. Based on network proximity, the location information (geo-location) is replaced by the availability ("visibility") of wireless networks.

Of course, iBeacon from Apple can be mentioned among other technologies. The basic element here is a tag based on Bluetooth Low Energy technology [10]. Each such tag translates (for a limited distance - the so-called Bluetooth distance) some unique identifier (UUID) and two integer values (this is configurable). This data is made available to mobile devices in the vicinity of the tag. Accordingly, these values can be used as keys to search for information. In this way, we can organize, for example, a local distribution of news information.

In our opinion, the biggest limitation for iBeacons (at least, in its original form) is the need to pre-configure tags for a particular application. The mobile application should statically declare UUIDs for the tags in questions. In the same time, Android OS is free from this limitation, and applications on Android can scan translated data from all nearby tags. A similar to iBeacons solution from Google directly distributes (translates) some URL instead of an abstract digital identifier.

Google Physical Web project is an example of integration Web technologies and physical world. As per Google's vision, the Physical Web is an example of discovery service. In this model, a smart Physical Object broadcasts relevant URLs that any nearby device can receive. It is very important also that user's phone can obtain advertised URL without connecting to the tag [8]. If this URL refers, for example, to an HTML5 web application, then it potentially can work using its cache without a network. Eddystone-URL is an URL in a compressed encoding format. Once decoded, the URL can be used by any client with access to the Internet. Google Eddystone can also be used as an engine for local news distribution [11].

In our previous paper [7], we have mentioned the biggest limitation (in our opinion, of course) for tag's based solutions (Apple and Google). The generic model for data delivery is based on push notifications. A mobile application (mobile user) will be notified if a tag with a given ID comes within range (or goes out of range) of the device. By our experience, push notification could be a quite disruptive delivery method: too often, these messages are delivered at the wrong time in terms of a model of user's behavior. In our opinion, the browsing

model is preferred here. In the case of a pull model, a mobile user should directly demonstrate the intention to get (request) some information. It is up to him to decide when and where to receive data (Fig. 1).

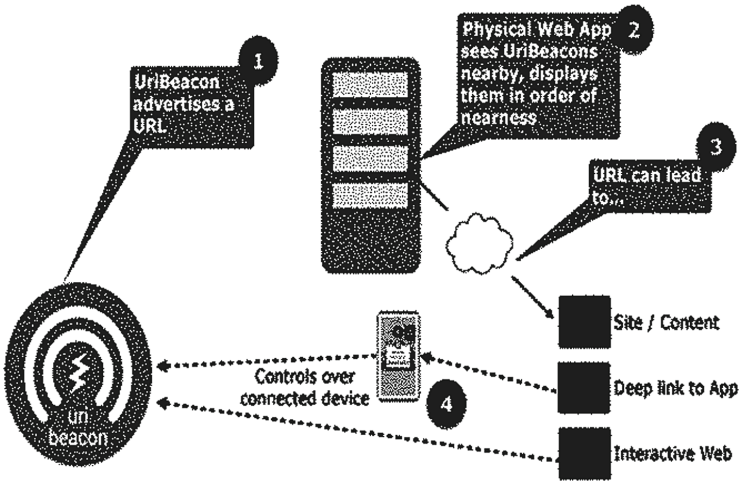


Fig. 1. The Physical Web: how it works.

The next idea that influenced our work is the beacon stuffing approach [12]. This is a low-bandwidth communication protocol for IEEE 802.11 networks that allows mobile customers to communicate with Wi-Fi access points without establishing Wi-Fi connections. Beacon stuffing allows customers to receive information from the nearest access points even if they are disconnected or when they connect to another access point. The authors proposed a scheme that supplements the 802.11 standards and works by overloading the control frames of the 802.11 protocol without violating the standard. The idea is that according to the Wi-Fi specifications, the mobile device and the access point exchange service packets (the so-called Probe Request). In the original form, it broadcasts information about the possible connection parameters. The use of this service information, in particular, is based on passive monitoring of Wi-Fi devices. Beacon stuffing reuses these service frames for data transmission. An obvious drawback is that we need to change the software that manages the access points. Nevertheless, this approach is used in a fairly large number of works. For example, in [13], the authors offer an Open Source solution for beacon stuffing. In the paper [14], authors use data coding in the access point identifier (SSID) to transmit information about moving objects. Override SSID (most often - in manual mode) is a widely used approach for sharing any information to mobile device owners in the immediate environment of the access point. In the paper [15], numerous examples of the use of political slogans (expressions of support for a certain candidate) in US political companies are given. In the note [16], this approach is

illustrated by advertising and marketing. Other examples are provided in web resources [17,18]. In the paper [19], the authors propose to use SSID as a key for forming an URL with contact information. SSID acts as the digital business card of the owner of the access point. In the paper [20], authors encode location information within SSID (Fig. 2).

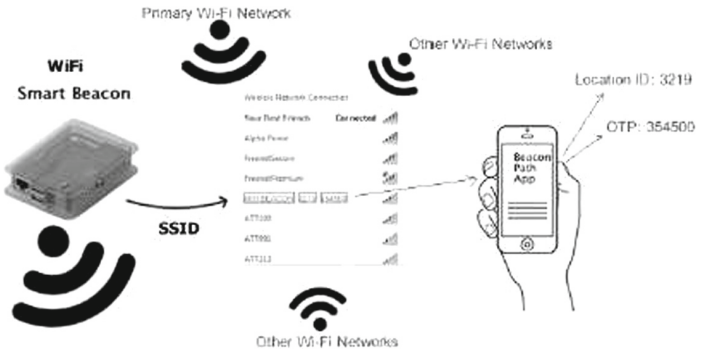


Fig. 2. Encoding location info [21].

All SSIDs follow the predefined format: "WIFIBEACON_XXX_YYY_ZZZ". The prefix to look for is WIFIBEACON, XXX– is the system ID, YYY– the location ID, and ZZZ is the one–time password.

3 On Blueshare Services

In our work on designing a local information distribution system, we proceeded from the following basic assumptions (requirements):

- the proposed model should work with the basic software, without modifying the software in the nodes of the potential network;
- the process of creating (publishing) information, its distribution, and receipt must allow completely software processing (without the direct participation of the mobile subscriber).

The general, our idea is the following: we propose a system that allows mobile users to programmatically describe (set) the identification of wireless nodes by encoding the necessary information in the device name (SSID). For the dissemination of this information (making it available/visible to surrounding users or programs), we use the standard mechanisms for announcing the identity of wireless nodes. For all data processing stages (publishing data, receiving data), mobile phone (application) on the Android platform is the only required equipment. The proposed approach is a further development of the ideas described

earlier in our papers [21,22]. Because the Android platform cannot programmatically create a Wi-Fi access point (there is no corresponding API), Bluetooth was used as the basic wireless network.

On the Android platform (with the user's permission, of course), Bluetooth can be enabled programmatically:

BluetoothAdapter mBluetoothAdapter = BluetoothAdapter.getDefaultAdapter (); MBluetoothAdapter.enable ();

And the Bluetooth device can be switched to visibility mode (Android: discoverable)

Intent myIntent = new Intent (ACTION _REQUEST_DISCOVERABLE); MyIntent. putExtra (BluetoothAdapter. EXTRA_DISCOVERABLE_DURATION, 0); StartActivity (myIntent);

The second call's parameter specifies the time that this Bluetooth node will be available (visible) to another. Parameter 0 corresponds to unlimited time. We should note that in our experiments, we did not register high energy consumption due to Bluetooth node existence. Actually, energy consumption is associated with the connectivity (pairing) and data transfer via Bluetooth. Because network proximity does not assume connectivity and data transfer it is an energy-safe operation. Our experiments with different models of Android devices show that without connection and real data transmission, the energy consumption remains low. Once a node is declared to be discoverable, its identification (SSID) is available for viewing (scanning) by other devices. This scan can be performed both directly by the user in manual mode using the standard tools of Android OS, or it could be done programmatically, by a special application. Of course, the value of the SSID (it is a message now) can be set both in manual and automatic (programmatically) mode.

The typical model of the whole process is as follows. The author of the message (publisher) types the message text in the mobile application on his phone. The application creates a public Bluetooth node on the phone, using the message text as SSID (node's identity). If such a Bluetooth node already exists, it simply changes its identification. The reader (the recipient of the message) uses the mobile application, scans the available (visible) Bluetooth nodes and extracts the text of the message from their identification. Obviously, receiving (and processing messages) can be performed completely programmatically.

4 On Technical Details

The SSID is a 32-byte string. In the current version, the first three bytes are reserved for the standard prefix (it is BDP). Accordingly, when scanning, only Bluetooth nodes with such prefixes are selected. Also, a scanner automatically deletes duplicate names (SSIDs). As the next step, the scanner can preprocess obtained messages. At this stage, the scanner extracts from the message (from the identifier) some standard objects that can be described using regular expressions. In this version, these include:

– phone number
– URL
– Links to Twitter accounts (@ name)
– Email address

Accordingly, the mobile application displays it as a clickable link when displayed (so, it is possible to call the phone, follow the hyperlink, etc.)

If the message size exceeds 29 bytes, then it is divided into parts. Each part ends with a special sign (=). The scanner when receiving such a message will "glue" it with previously received data (data from the Bluetooth node with the same MAC address). The time during which a separate part will act as an identifier is specified in the program settings (options).

Also, any node on the phone can work in relay mode. In this mode, obtained messages will be retranslated. As soon as a scanner application receives a message that ends with the plus sign (+), it publishes the message on its own behalf. Accordingly, it becomes available to another group of mobile users (users, who are nearby a receiver excluding an originator). Relay modes could be described in the application settings: to allow relaying messages from all publishers (from all MAC addresses), to allow relaying only from the specified MAC addresses, to forbid relaying. If there are multiple relay requests, the scanner organizes the queue. For example, Phone 1 declares Bluetooth node with own name (SSID, messages). Phone 2 obtains visible SSID (message) from Phone 1 and declares own Bluetooth node with the same SSID (translates a message to the own nearby devices). Note that the Phone 1 will also "see" the message from Phone 2, but it can simply ignore it, because it coincides with the previously "sent" message.

If we consider this as the transmission of data on the network, the following figures can be cited. Changing the name of the Bluetooth node and recognizing the new name by surrounding devices takes 500–1000 ms. Accordingly, from the point of view of information transfer, this corresponds to a transmission rate of 32 bytes per second. This can be called the main drawback of this approach. But such a low speed is a payment for practically 100% compatibility with different models of mobile devices.

As the typical use cases, we can present the following examples.
Client to Client (C2C) - the user announces his Facebook ID to those who are nearby. It could be a badge at the conference, for example.
Business to Clients (B2C) - the announcement of a free taxi, for example. Another real example is a trader on the street, who sent out (to the nearby mobile users) the code for a discount.
Business to Business (B2B) - Bluetooth node is in the car (e.g., multimedia panels there are Bluetooth nodes). The MAC address could be used to identify the car (customer). And the SSID could be used to identify the cargo, for example.

Of course, the scanning and its processing could be wrapped into some application programming interfaces (API) and this model could be used in business-to-business transactions too. It is a part of the future development.

The next type of services is context-aware communications. The idea is to provide a chat (messenger, discussion board) for the mobile users in the

proximity. Thus, Bluetooth tag (including any dynamically created tag on the mobile phone) defines own communication point (communication channel). Speaking about the possible extensions, we would like to draw attention to the fact that the proposed model could help develop a system of customized check-ins [22]. A check-in record in a social network is some message (post, status, etc.) linked to the particular location (to the particular geographical place). In other words, it is some geo-located message, presented on a social network. Traditionally, "places" in the social networks are described via classical geo-location info (latitude and longitude pair). Of course, it is already problematic for indoor applications, for example, when many places within a building could be in the same geo position. With BlueShare model, we can remove "places" from the social networks and make them completely dynamic. A new "place" could be just a network fingerprint (in our case, it is a set of "visible" Bluetooth nodes). A mobile application can ask a user about his identity in the social network (for example, it could be a Facebook login) and record (outside of the social network, in the own database) the proximity information with user's ID. As soon as we will get a new check-in with the similar fingerprint, we can show the participants their IDs. Such application can show all IDs for users in the proximity. Of course, this kind of check-ins could be customized too. Some business, for example, can show (deliver) special offerings for mobile visitors in exchange for check-in, etc.

5 On Device Discovery

In fact, what is described above suggests using the device recognition process in a wireless network to transmit useful information. The approaches described above offer some maximally portable solution that will work on most mobile systems. However, if we provide a deeper "intrusion" in the process of recognizing devices in wireless networks, we can get more interesting results. The idea remains the same - to ensure a simple distribution of information in a local area for mobile devices (mobile subscribers) located in this very area. At the same time, there is no talk about transferring (fast transferring) large amounts of data (e.g., similar to D2D in 5G networks). The key point here is just a simple detection (discovery) of nearby devices (mobile users), without the need for users to perform any special actions. That is why the process of searching (discovering) for devices is an ideal candidate.

Here we can consider the following technologies: Bluetooth 4.0, Bluetooth 5.0, Proximity based Services (ProSe) for LTE, and D2D in 5G networks.

A Bluetooth 4.0 device may operate in three different modes depending on required functionality: advertising, scanning and initiating. It is illustrated in Fig. 3. A device in advertising mode, named advertiser, periodically transmits advertising information in three channels [23].

So, there is a a special Packet Data Units (PDUs) ADV_IND, which could be used (overloaded) with custom data. The packet length is defined by the length of the payload, ranging from 0 to 37 bytes. Note, that Android source code contains time related constants for Bluetooth 4.0 advertising (It is from Android source code base for version 6.0):

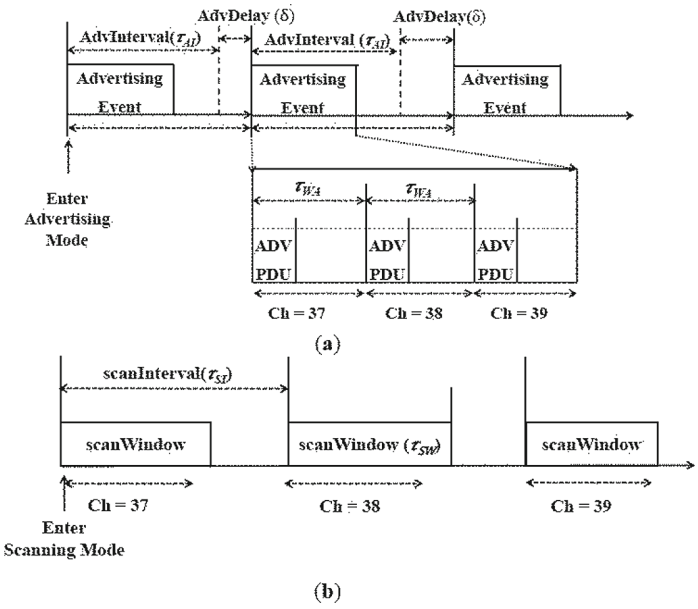


Fig. 3. On advertising (a) and scanning (b) in Bluetooth 4.0 [24]

```
/**
Scan params corresponding to regular scan setting
**/
private static final int SCAN_MODE_LOW_POWER_WINDOW_MS = 500;
private static final int SCAN_MODE_LOW_POWER_INTERVAL_MS = 5000;
private static final int SCAN_MODE_BALANCED_WINDOW_MS = 2000;
private static final int SCAN_MODE_BALANCED_INTERVAL_MS = 5000;
private static final int SCAN_MODE_LOW_LATENCY_WINDOW_MS = 5000;
private static final int SCAN_MODE_LOW_LATENCY_INTERVAL_MS =
5000;
```

But for Bluetooth 4.0 we should keep in mind the following architectural change. Since Android 5.0 (it is API Level 23), we can only get a randomized MAC address of external devices via Bluetooth Low Energy scan, and this MAC changed in each advertising packet. So, we can not use MAC addresses to identify data chunks from the same user like we did in Classic Bluetooth. So, we should add, for example, a randomized user-related UUIDs in advertising packets instead of MAC address.

Bluetooth 5.0 offers customizable device discovery process [25]. In Bluetooth 4.0 there is no way for the advertiser to notify the host that it had been discovered. In Bluetooth 5.0 this possibility has been introduced. As per Bluetooth 5.0 specification, the advertising channels is presented with two types: primary advertising channel and secondary advertising channel.

(1) The primary advertising channel is used to transmit the Advertising Event such as in Bluetooth 4.0.
(2) The secondary advertising channel reuses the 37 fixed channels previously reserved for data in order to transmit Extended Advertising Event. The advertising data payload in Extended Advertising packet (ADV_EXT_IND PDU) is 254 bytes. If we want to advertise data, which is more than 254 bytes, we can chain Extended Advertising packets (so-called AUX_CHAIN_IND PDUs). It offers the way to advertise non-text data too.

The Release 12 of the 3GPP specifications introduced ProSe (Proximity Services). It is a D2D (Device-to-Device) technology that allows LTE devices to detect each other and to communicate directly. It relies on multiple enhancements to existing LTE standards including new functional elements and a "sidelink" air interface for direct connectivity between devices [26]. Peer discovery has a similar functionality as that of cell search in LTE by which the mobile device determines the time and frequency parameters that are necessary to communicate with the network. The modes in this process are divided between whether the network core is involved in device discovery or not.

Here it is necessary to note the following. The purpose of this paper is not to establish a connection between two devices. As for the transfer of information, the connection (in fact) will be established between two of some similar applications. And our main goal is to avoid the need for these applications. The goal is to use the identification information, which is inherent in the search for devices, for the transfer of user data. It can be said that this point has not yet been considered for LTE. The current documents talk about ProSe enabled devices that should be subscribed to an operator service in order to be authorized to run ProSe enabled applications on it. The registration of the device occurs in the Home Public Land Mobile Network (HPLMN) where the subscriber's profile is held in a logical function, named ProSe Function [27]. And after that a mobile user can download a ProSe application (it is an application offering services to other ProSe-enabled devices) and activate the ProSe features on that application. As per the model, the user should also authenticate and authorize it through the ProSe Function that caches a list of all the IDs of the applications allowed to use ProSe features along with their corresponding authorized range class [27]. In other words, this model completely opposes to the dynamic nature of the above-described approach with user data incorporated into identification packages. In our opinion, there are prospects here in works on so-called proximity-aware networking models.

The situation with identification for D2D interactions in 5G networks is in many respects similar. The basic models consider, first of all, the search models of neighboring devices involving the network. The issues of customization of identification information are not considered, and this, in our opinion, is a very promising direction.

6 Conclusion

In this paper, we propose a system for local information delivery. The system is based on programmatically created Bluetooth wireless network nodes and uses standard mechanisms for announcing network nodes for information dissemination. In our opinion, the proposed approach is one of the simplest ways, in the sense of its implementation and its compatibility with various devices, mobile messages could be distributed between mobile devices without any telecommunication infrastructure. The possible use cases include personal communications (C2C model), business to consumers (B2C), and business to business (B2B) applications. The proposed model is based on the use of the process of discovering devices in a wireless network for the transfer of user data. In this regard, in addition to the Classical Bluetooth, we discuss the use of the proposed model on Bluetooth 4.0, Bluetooth 5.0, LTE, and 5G.

References

1. Bland, A.: FireChat-the messaging app that's powering the Hong Kong protests. Guardian **29** (2014)
2. Namiot, D.: On mobile mesh networks. Int. J. Open Inf. Technol. **3**(4) (2015)
3. Mobile & Ad Hoc Network. https://www.slideshare.net/cprakash2011/lecture-1-mobile-and-adhoc-network-introduction. Accessed Apr 2018
4. Kopekar, S., Kumar, A.: A study of ad-hoc wireless networks: various issues in architectures and protocols. Int. J. Comput. Appl. **122**(6) (2015)
5. Yousefi, S., Mousavi, M.S., Fathy, M.: Vehicular ad hoc networks (VANETs): challenges and perspectives. In: 2006 6th International Conference on ITS Telecommunications Proceedings. IEEE (2006)
6. Bill, D.S.: Wireless social networking. US Patent No. 7,720,037 18 May 2010
7. Namiot, D., Sneps-Sneppe, M.: On Bluetooth proximity models. In: Advances in Wireless and Optical Communications (RTUWO). IEEE (2016)
8. Namiot, D., Sneps-Sneppe, M.: Context-aware data discovery. In: 16th International Conference on Intelligence in Next Generation Networks (ICIN). IEEE (2012)
9. Sneps-Sneppe, M., Namiot, D.: On physical web models. In: 2016 International Siberian Conference on Control and Communications (SIBCON). IEEE (2016)
10. Koühne, M., Sieck, J.: Location-based services with iBeacon technology. In: 2014 2nd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS). IEEE (2014)
11. This Google-backed startup wants to make millennials read local news. https://thenextweb.com/insider/2017/07/25/are-bluetooth-beacons-the-savior-of-local-news/. Accessed Apr 2018
12. Chandra, R., et al.: Beacon-stuffing: Wi-fi without associations. In: Eighth IEEE Workshop on Mobile Computing Systems and Applications, HotMobile 2007. IEEE (2007)
13. Zehl, S., et al.: LoWS: a complete open source solution for Wi-Fi beacon stuffing based Location-based Services. In: 2016 9th IFIP Wireless and Mobile Networking Conference (WMNC). IEEE (2016)

14. Liu, Z., et al.: SenSafe: a smartphone-based traffic safety framework by sensing vehicle and pedestrian behaviors. In: Mobile Information Systems 2016 (2016)
15. Using a Wi-Fi Network's Name to Broadcast a Political Message. http://www.npr.org/sections/alltechconsidered/2017/02/07/513240428/using-a-wi-fi-networks-name-to-broadcast-a-political-message. Accessed Apr 2018
16. Cyber graffiti with WiFi network names as advertising. http://www.webinknow.com/2011/02/cyber-graffiti-with-wifi-network-names-as-advertising.html. Accessed Apr 2018
17. Clever SSIDs that Scare Off Leeches or Send a Message. https://www.lifehacker.com.au/2011/10/clever-ssids-that-scare-off-leeches-or-send-a-message/. Accessed Apr 2018
18. Scare your neighbors with a spooky Halloween network name. https://arstechnica.com/information-technology/2014/10/scare-your-neighbors-with-a-spooky-halloween-network-name/. Accessed Apr 2018
19. Let Others Contact You Through Your Own Wi-Fi Network. https://www.labnol.org/internet/share-wifi-with-neighbors/21024/. Accessed Apr 2018
20. Huseynov, E., Seigneur, J.-M.: Beacon authpath: augmented human path authentication. In: 2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT). IEEE (2016)
21. Namiot, D., Sneps-Sneppe, M.: CAT-cars as tags. In: 2014 7th International Workshop on Communication Technologies for Vehicles (Nets4Cars-Fall). IEEE (2014)
22. Namiot, D., Sneps-Sneppe, M.: Customized check-in procedures. In: Balandin, S., Koucheryavy, Y., Hu, H. (eds.) NEW2AN/ruSMART -2011. LNCS, vol. 6869, pp. 160–164. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22875-9_14
23. Liu, J., Chen, C., Ma, Y.: Modeling and performance analysis of device discovery in bluetooth low energy networks. In: Proceedings of the IEEE on Global Communications Conference (GLOBECOM), Anaheim, CA, USA, 3–7 December 2012, pp. 1538–1543 (2012)
24. Liu, J., Chen, C., Ma, Y.: Modeling neighbor discovery in bluetooth low energy networks. IEEE Commun. Lett. **16**, 1439–1441 (2012)
25. Hernández-Solana, Á., Perez-Diaz-de-Cerio, D., Valdovinos, A., Valenzuela, J.L.: Proposal and evaluation of BLE discovery process based on new features of bluetooth 5.0. Sensors 17(9), 1988 (2017)
26. 3GPP TR 22.803: Feasibility study for Proximity Services (ProSe) (Release 12), v. 12.2.0, June 2013
27. Doumiati, S., Artail, H., Gutierrez-Estevez, D.M.: A framework for LTE-A proximity-based device-to-device service registration and discovery. Procedia Comput. Sci. **34**, 87–94 (2014)

Enabling M2M Communication Through MEC and SDN

Ammar Muthanna[1,2(✉)], Abdukodir Khakimov[1], Abdelhamied A. Ateya[1], Alexander Paramonov[1], and Andrey Koucheryavy[1]

[1] The Bonch-Bruevich Saint-Petersburg State University of Telecommunications, 22/1 Prospekt Bolshevikov, 193232 Saint-Petersburg, Russian Federation
ammarexpress@gmail.com, khakimov.a@sdnlab.ru, a.ashraf@zu.edu.eg, alex-in-spb@yandex.ru, akouch@mail.ru

[2] Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Street, 117198 Moscow, Russian Federation
http://www.sut.ru

Abstract. Machine-to-Machine (M2M) traffic is growing fast on the Internet, this trend is manifested in the congestion of networks at different levels. The edge cloud unit is used for reducing number of intermediate nodes involved in the communication process and for offloading. The offloading can be seen as three parts. The first is the base station (BS) offloading, as the cloud unit helps the BS in some tasks related to call imitation such as resource allocations. The second is the workload passed from sensor nodes which represents the M2M traffic offloading and the third part is the cellular data offloading. In this paper, we propose an algorithm for offloading the core network using mobile edge computing, which allow devices to exchange traffic with the nearest mobile edge computing for improving QoS and offload channels in the core of the network.

Keywords: M2M · Traffic · SDN · Offloading · MEC

1 Introduction

Currently the infocommunication system development is characterized by an intensive growth of automatic devices connected to the communication network. Such devices can be various sensors, control devices, alarm devices, monitoring devices, etc. These devices generate traffic, which are not directly dependent on human behavior and are determined by the algorithms of the automation operation. The class of communications between such devices is called the Machine-to-Machine (M2M). This class of communications attracts more attention due to the wide spread of M2M technologies. This fact is confirmed by the publications of ITU-T [1], the leading manufacturers of telecommunications equipment [3] and the results of some studies [2]. An essential part of these devices is devices for monitoring various processes: production, changes in the state of the environment, condition of buildings, technical engineering systems, possibly the state of

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 95–105, 2018.
https://doi.org/10.1007/978-3-319-99447-5_9

vital signs of the human body, etc. The intensive dissemination of M2M systems and the growth of their traffic leads to researches, aimed at understanding the structure of the system for offloading this type of traffic.

Mobile edge computing (MEC) is a new communication paradigm introduces for the fifth generation of cellular system (5G) and next generation of communication systems [14]. MEC provides the computing resources and capabilities at the edge of the radio access network (RAN) of the cellular system. This achieves various benefits to the cellular network and other integrated systems, especially for latency sensitive systems. With the dramatic increase of sensor devices and the urgent demand of introducing reliable systems for M2M communication, MEC represents a vital solution. MEC provides an offloading way to the M2M traffic instead of forwarding this massive traffic to the core network. A summary of valuable benefits introduced by the MEC based systems can be founded in [12], and a frame work for the IoT devices are presented in [13,16,17].

Using MEC means moving away from centralized large data centers to distributed small data centers with limited capabilities compared to centralized units. Thus, you can achieve great success and benefit. Mobile cloud computing is undoubtedly a trend in the sphere of cloud computing because all calculations are transferred to the mobile portable device.

Software defined networking (SDN) is another communication paradigm that enables the separation of data plane and control plane of the SDN enabled networks [15]. The network part responsible for the traffic handling is the control plane, which consists of a control scheme. The control scheme is either centralized by deploying a single centralized controller or distributed scheme by deploying distributed controllers. The centralized scheme may be efficient for small scale networks with limited number of traffic. The distributed scheme is an efficient paradigm for large scale networks with massive traffic.

The data plane is the part responsible for traffic forwarding. SDN achieves various benefits to the SDN enabled networks and enables the introduction of new services. Network operators are able to configure, manage and control the whole network through special purposes customized software referred to as application programming interfaces (APIs). SDN provides networks with the required level of flexibility and simplifies the network hardware.

With the dramatic increase of wireless devices and the number traffic, next generation networks should deploy new technologies such as the MEC and SDN. The new paradigms enable the integration of these devices with the existing systems and facilitate the control and management of this enormous number of traffic. In this article, an SDN based algorithm with MEC enabled is proposed for M2M traffic offloading. The proposed algorithm employs the edge computing in multi-levels. In (Sect. 2) the background and related works to the proposed algorithm are introduced. Section 3 provides the proposed offloading algorithm and Sect. 4 provides the experimental results.

2 Related Work

Deploying SDN and MEC for M2M communication based networks is a vital solution to enable the massive number of devices and achieves various benefits. MEC provides an offloading way for M2M traffic and thus, introducing a reliable and efficient offloading algorithm is urgent. There are many studies introduced for developing offloading algorithms for MEC and SDN based networks. One main issue that is more considered for the MEC based M2M networks is the load balancing of MEC servers. In this part we consider the most related works to the proposed algorithm.

In [4], a dynamic and integrated resource scheduling algorithm (DAIRS) is introduced for load balancing of cloud datacenters. DAIRS considers many factors for balancing load among cloud datacenters, which differs from the past works that consider the traditional load balancing by considering only one factor dedicated with the physical machines of datacenters. These factors are the network bandwidth and memory dedicated with both virtual machines and physical machines of datacenters. The work provides four different scheduling algorithms, which are simulated and compared. Simulation results show that the algorithm is inefficient in time, as it sorts the machines of their use during resource allocation.

In [5], a real-time load balancing method for cost efficient hosting of Massively Multiplayer Online Games (MMOG) was developed. Simulation results show that the introduced method reduces the hosting costs by factor between two and five. The proposed method is experimentally tested for a fast-paced game and the proposed algorithm is able to reduce the number of game servers and achieve load distribution with the required quality of service (QoS).

In [6], authors introduced an automatic control algorithm for achieving load balancing and load leveling of M2M networks. The proposed system deploys the network without medium servers. The algorithm is simulated for different operation parameters, and results showed that the algorithm effectively achieves both load balancing and load leveling for different values of operating parameters.

The SDN approach demonstrates the possibility of improving the technique of load balancing [7,8]. In [9], authors propose an M2M improved architecture using SDN for dynamic flexible network management. The introduced architecture is validated via a simulation process; simulation results showed that SDN represents a powerful solution for the M2M networks.

In this work, an offloading algorithm is developed for load balancing of M2M traffic. The algorithm considers the deployment of SDN and MEC in multi-levels.

3 System Structure

This paper proposes a system for offloading M2M traffic using cloud computing [10,18]. The main task of mobile cloud computing is to provide a user interface for using these applications. Cloud computing in a mobile device can be classified into four types of service model [11]: service customer, service provider, broker service, and as a service representative. Each user in the cell can be considered

as one of these types. As a consumer, the mobile device does not perform any tasks, and the storage and use of all calculations occurs directly in the cloud or on other devices. This type is the most common. In the event that the device has the ability to separate the processing and storage of data itself, the mobile device belongs to the suppliers. In this example, the device collects information from its built-in sensors (camera or GPS function), and then transfers the data to other users. The broker-service works in much the same way as the supplier, the difference is only in the ability to organize networks and redirects. This means that mobile devices as a broker service can play the role of a gateway for other devices.

An entire network can be represented as a core network and distributed cells. Figure 1 illustrates the end-to-end structure of the proposed system. The structure of our system consists of user devices (the M2M traffic generator), RAN, cloud blocks, access switches, switches based on OpenFlow, Middleboxes, and finally, the SDN controller. Each base station is connected to the network through access switches that perform packet classification for traffic from user devices. Access switches are software switches, such as OpenvSwitch.

The entire network is connected through OpenFlow switches that manage data packets and forward traffic based on the thread tables. Middleboxes is an equipment that allows network operators to add additional functions, such as a firewall and network address translation. The main requirements for the functions and services implemented by these Middleboxes are the effective use of resources and the protection of the system from attacks. All these elements represent the plane of the network data.

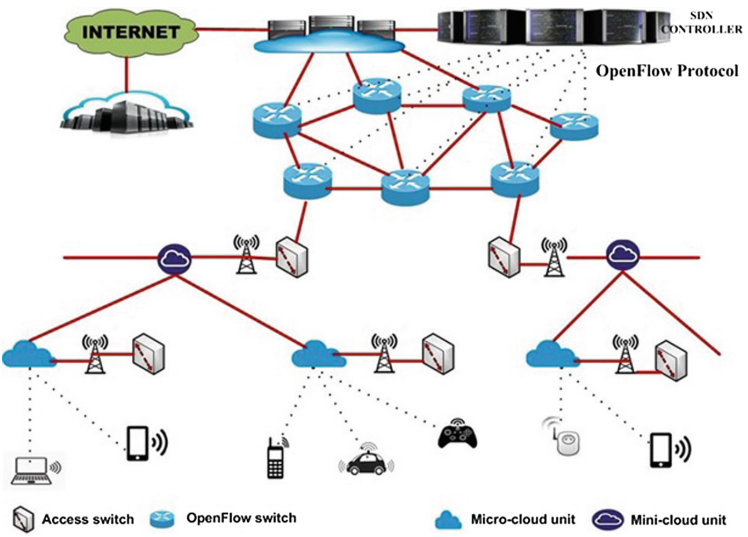


Fig. 1. End-to-end system structure

The cloud module is used to reduce the number of transitional nodes involved in the communication process and to offload the system. The offloading can be divided into three parts, as shown in Fig. 2. The first is the offloading of the base station, since the cloud module helps the base stations in some tasks related to allocating resources. The second is the workload transmitted from sensor nodes, which is the offloading of M2M traffic, and the third part is the offloading of cellular data.

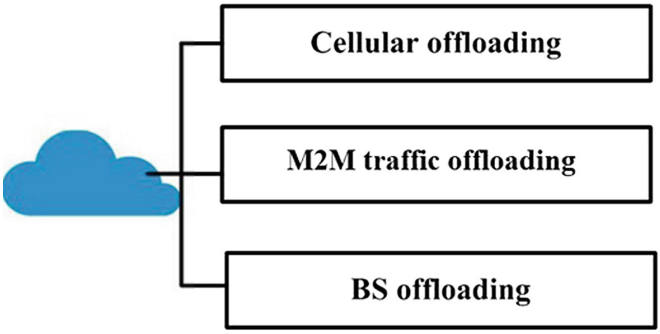


Fig. 2. Network offloading using the cloud module

In the other hand, the proposed system provides three main levels of offloading to the end user traffic and M2M data. Figure 3 illustrates the main offloading levels provided by the proposed structure. The first offloading level is the cloudlet offloading, which is represented by high computing capabilities user devices. These devises can offer computing resources to nearby devices and perform computing tasks. This introduced level can handle tasks that need limited computing resources, and thus reduce the number of data traffic offloaded to the cellular base station and hence to the core network.

The second offloading level is represented by Micro-cloud edge units, which is a small data center with limited computing capabilities. Micro-cloud units offer computing resources to tasks that can't be handled by the cloudlet. This represents an offloading way to end user data and offer resources to base station. The third offloading level is represented by Mini-cloud units, which handle tasks that can't be handled by Micro-cloud units. Mini-cloud unit is an edge server with higher computing capabilities than that of Micro-cloud unit. Each group of Micro-cloud units is connected to a Mini-cloud unit, which manage and control them.

Diagram below shows the logical scheme for uploading a channel from the remote cloud segment to the local access network. From the point of view of economic and technical indicators, this principle of network operation is optimal for use by operators. It can be seen that the width of the segment A is reduced due to the localization of the application to the communication node.

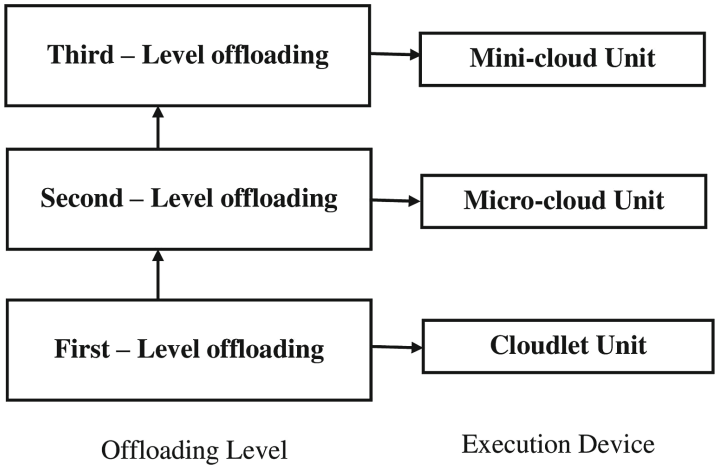


Fig. 3. The main offloading levels provided by the proposed structure

Obviously, in this structure, the traffic intensity at point A is less or equal to the traffic intensity at point B. If the probability that the traffic will be served by the MEC is p and the traffic intensity at point B is λ_B, then the traffic intensity at point A:

$$\lambda_A = (1 - p)\lambda_B \tag{1}$$

The probability of servicing in MEC p determines the effect of its application. Suppose that there is some conditional cost of servicing the traffic unit C_A at the point and the cost of the MEC C_M organization per unit of serviced traffic (Fig. 4).

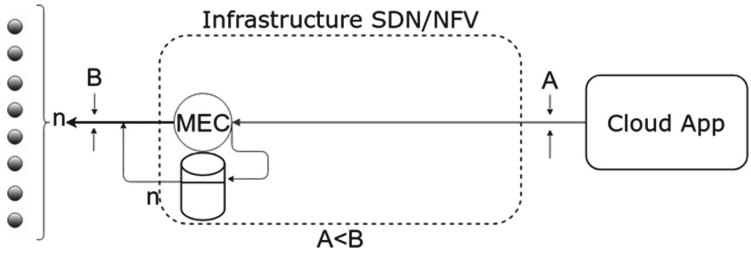


Fig. 4. The efficiency scheme of NFV-MEC

In view of the above, the positive effect of using MEC is achieved when:

$$p\lambda_B C_M < \lambda_A C_A \tag{2}$$

Taking (1) into account, we obtain condition:

$$pC_M < (1 - P)C_A \tag{3}$$

The above expressions (1)–(3) should take into account the conditional cost, which is related to the procedure for creating or removing the MEC. Of course, this component exists, it is caused by the need to perform a certain amount of work by network elements, it can be included in the value of C_M.

It should be noticed that any management decision, in particular, the decision to create the MEC, is taken from the calculation for the future. This means that the values of λ_B and p are values that will take place in the perspective, i.e. this is the predicted value.

$$\lambda_B, p = f(\lambda_B, p, t) \tag{4}$$

where $f(\lambda_B, p, t)$ – a model for predicting traffic values and the probability of its local maintenance, λ_B, p – statistical data for the previous period.

4 The Proposed algorithm

When transferring M2M traffic, the orchestrator monitors statistics and starts to predict. The monitoring criteria is as follows (Fig. 5):

- State of the CPU,
- RAM,
- Functional capacity of interface,
- Statistics of traffic flows Based on the data, orchestrator decides to create or not a platform for MEC. The orchestrator creates the MEC if it needs, and then loads the processing application into this network node. Next, the devices (node) begin to exchange data with the network node on which the MEC is located.

M2M traffic is growing rapidly on the Internet. This trend is manifested in the congestion of networks at different levels. In our work, we propose an algorithm for offloading the core of the network using MEC.

When devices start exchanging M2M traffic with a remote server, the orchestrator monitors traffic and predicts further network congestion.

In case the m2m traffic exceeds the conditional volume of the network allocated for it, the orchestrator creates the MEC space and loads the M2M traffic processing application into it. Then devices can exchange traffic with the nearest MEC, which will give better QoS and offload channels in the core of the network.

In this paper, traffic generators were used as M2M traffic.

The experiment testbed (Table 1):

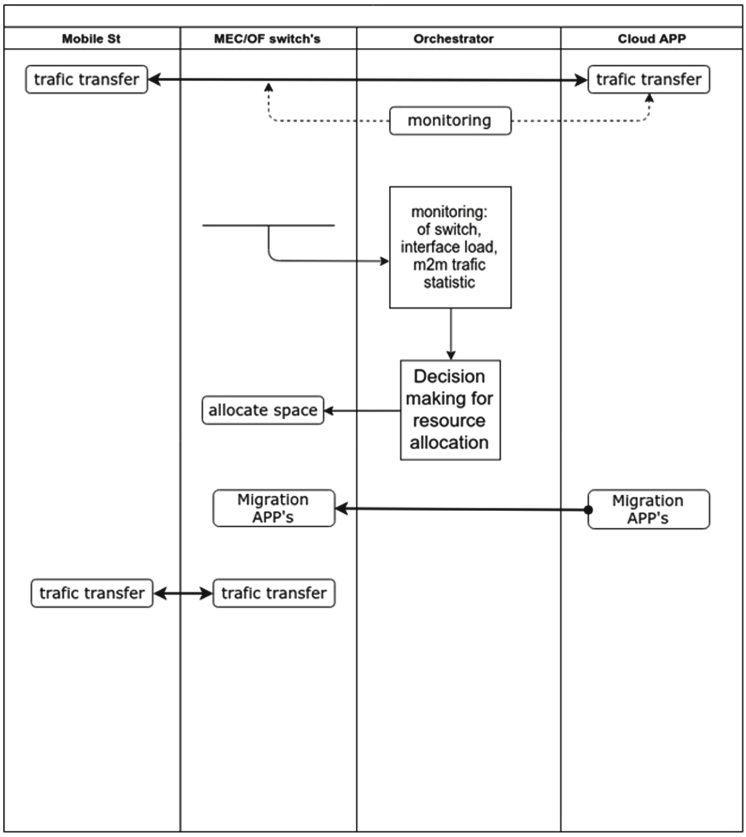


Fig. 5. The algorithm for offloading traffic M2M

Table 1. Experiment testbed parameters

Humansense	Timeconstant
AppCloud CPU	E5-2650 v4 @ 2.20 GHz
AppCloud Core	32,
AppCloud RAM	48 GB
Orchestrator/Brain4Net ServicePlatform	
OF Switch CPU	E5-2650 v4 @ 2.20 GHz
OF Switch Core	12
OF Switch RAM	40 GB
Node	Raspberrypi 3 (30 devices)

5 Results

At the end of the experiment, we obtained the following results (Figs. 6 and 7). The load of 30 nodes to the server decreased to an average of 3. That allowed to reduce the workload of the network, and to reduce delays. Also, the average time of applications for a remote server is reduced by 10 times distributing applications between localized MEC.

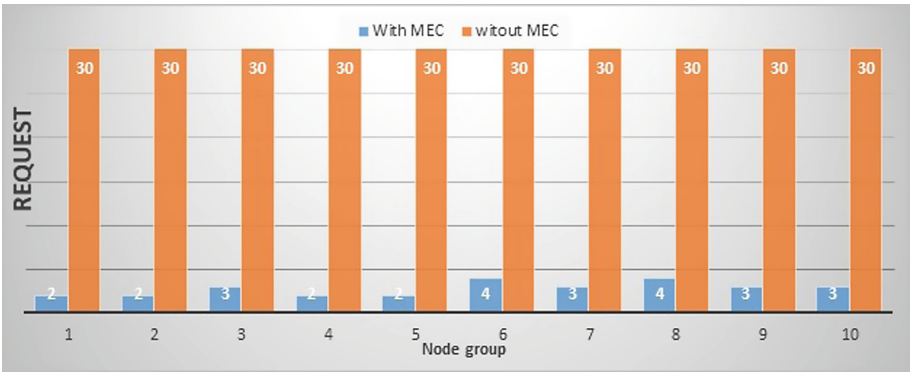


Fig. 6. Requests for loading a conditional application update

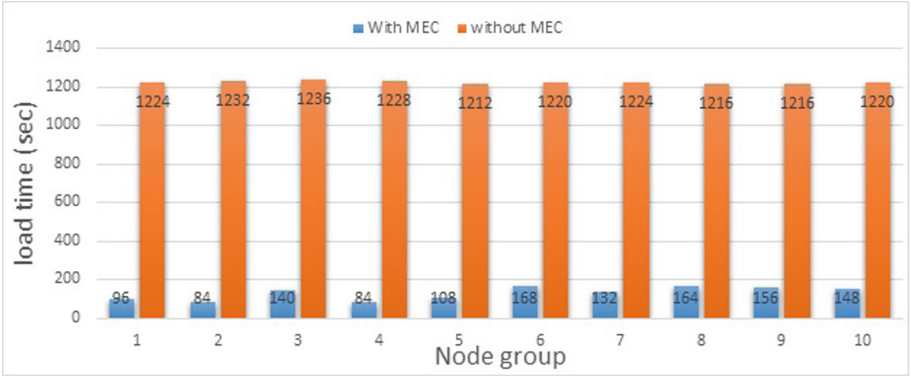


Fig. 7. M2M traffic request/offloading time

6 Conclusion

Proposed system can efficiently manage and establish an efficient and flexible routing path between any two end points. Thus it can reduce the number of

intermediate nodes involved in communication process. The proposed structure uses SDN in the core of the network, which means it is helpful and effective for offloading M2M traffic.

Acknowledgments. The publication has been prepared with the support of the "RUDN University Program 5-100" and funded by RFBR according to the research projects No. 17-07-00845 and No. 18-07-00576

References

1. ITU-TRecommendations. http://www.itu.int/en/ITUT/publications/Pages/recs.aspx
2. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012–2017
3. Zubair Shafiq, M.: A first look at cellular machine-to-machine traffic - large scale measurement and characterization. In: International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS), London, UK, June 2012
4. Tian, W., Zhao, Y., Zhong, Y., Xu, M., Jing, C.: A dynamic and integrated load-balancing scheduling algorithm for cloud datacenters. In: IEEE International Conference on Cloud Computing and Intelligence Systems, pp. 311–315 (2011)
5. Nae, V., Prodan, R., Fahringer, T.: Cost-efficient hosting and load balancing of massively multiplayer online games. In: IEEE/ACM International Conference on Grid Computing, pp. 9–16 (2010)
6. Kitagami, S., Kaneko, Y., Suganuma, T.: Method of autonomic load balancing for long polling in M2M service system. In: International Conference on Advanced Information Networking and Applications Workshops, pp. 294–299 (2012)
7. Muhizi, S., Shamshin, G., Muthanna, A., Kirichek, R., Vladyko, A., Koucheryavy, A.: Analysis and performance evaluation of SDN queue model. In: Koucheryavy, Y., Mamatas, L., Matta, I., Ometov, A., Papadimitriou, P. (eds.) WWIC 2017. LNCS, vol. 10372, pp. 26–37. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61382-6_3
8. Vladyko, A., Muthanna, A., Kirichek, R.: Comprehensive SDN testing based on model network. In: Galinina, O., Balandin, S., Koucheryavy, Y. (eds.) NEW2AN/ruSMART -2016. LNCS, vol. 9870, pp. 539–549. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46301-8_45. ISSN: 0302-9743
9. Park, S.M., Ju, S., Kim, J., Lee, J.: Software-defined-networking for M2M services. In: IEEE International Conference on ICT Convergence (ICTC), pp. 50–51 (2013)
10. Ateya, A.A., Muthanna, A., Gudkova, I., Abuarqoub, A., Vybornova, A., Koucheryavy, A.: Development of intelligent core network for tactile internet and future smart systems. J. Sens. Actuator Netw. **7**(1) (2018)
11. Ateya, A., Muthanna, A., Gudkova, I., Vybornova, A., Koucheryavy, A.: Intelligent core network for Tactile Internet system. In: International Conference on Future Networks and Distributed Systems (2017)
12. Ateya, A., Vybornova, A., Kirichek, R., Koucheryavy, A.: Multilevel cloud based Tactile Internet system. In: IEEE-ICACT2017 International Conference, Korea, February 2017
13. Ateya, A., Muthanna, A., Koucheryavy, A.: 5G framework based on multi-level edge computing with D2D enabled communication. In: 2018 20th International Conference on Advanced Communication Technology (ICACT), pp. 507–512. IEEE, February 2018

14. Muthanna, A., et al.: Analytical evaluation of D2D connectivity potential in 5G wireless systems. In: Galinina, O., Balandin, S., Koucheryavy, Y. (eds.) NEW2AN/ruSMART -2016. LNCS, vol. 9870, pp. 395–403. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46301-8_33
15. Volkov, A., Khakimov, A., Muthanna, A., Kirichek, R., Vladyko, A., Koucheryavy, A.: Interaction of the IoT traffic generated by a smart city segment with SDN core network. In: Koucheryavy, Y., Mamatas, L., Matta, I., Ometov, A., Papadimitriou, P. (eds.) WWIC 2017. LNCS, vol. 10372, pp. 115–126. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61382-6_10
16. Khakimov, A., Muthanna, A., Kirichek, R., Koucheryavy, A., Muthanna, M.S.A.: Investigation of methods for remote control IoT-devices based on cloud platforms and different interaction protocols. In: Proceedings of the 2017 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), pp. 160–163 (2017)
17. Masek, P., Fujdiak, R., Zeman, K., Hosek, J., Muthanna, A.: Remote networking technology for IoU: cloud-based access for Alljoyn-enabled devices. In: Proceedings of the 18th Conference of Open Innovations Association FRUCT and Seminar on Information Security and Protection of Information Technology, pp. 200–205 (2016)
18. Khakimov, A., Muthanna, A., Muthanna, M.S.A.: Study of fog computing structure. In: IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), pp. 51–54 (2018)

Queuing Management with Feedback in Cloud Computing Centers with Large Numbers of Web Servers

A. Z. Melikov[1(✉)], A. M. Rustamov[2], and J. Sztrik[3]

[1] Institute of Control Systems, ANAS, B. Vahabzade 9, AZ1141 Baku, Azerbaijan
agassi.melikov@gmail.com
[2] Baku Engineering University, H. Aliyev 120, Khirdalan, AZ0101 Baku, Azerbaijan
anrustemov@beu.edu.az
[3] University of Debrecen, Debrecen 4032, Hungary
sztrik.janos@inf.unideb.hu

Abstract. The objective of the paper is to analyzed QoS metric of cloud computing with large-scale of web server from queue management perspective. We propose a new method in order to effectively calculate the steady-state probabilities of cloud system with a large number of web servers. Numerical results showed that the proposed algorithms have higher accuracy and negligible computation time. Taking into account that in the cloud computing repair of server is taking some hours and response time is handling within some seconds, we can correctly apply space-merging algorithm.

Keywords: Queuing system · Cloud computing · Cloud technology

1 Introduction

Last decades, the optimization of applications and backend processing in the computing and virtualization technologies is becoming the integral part of the nowadays Internet applications. The main indicators of these technologies are still remaining the same: (1) high service rate and (2) cost-effectiveness. The main paradigm of the cloud technologies is to share computing resources among application within internal network or in the global network (the Internet) by providing the main two indicators mentioned above. All public cloud environments IaaS (Infrastructure-as-a-Service), PaaS (Platform-as-a-Service), and SaaS (Software-as-a-Service) should be arranged that any computation should be serviced with low resources [1,2].

Technically, cloud computing is set of servers with different specification, switches and routers. And servers are mainly considered as set of the hardware, software, environment, and human factors.

In cloud computing main servers that accepts the request from the users are called web servers. Web servers are one of the main factors that affect QoS metric

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 106–119, 2018.
https://doi.org/10.1007/978-3-319-99447-5_10

of the each request coming from users. From general point of view, decrease in latency and response times can be caused by *client's attitude, network,* and *server side* in the transaction in the client/server paradigm.

As we see, cloud computing has been deeply investigated by many scholars within the last decade from different perspective, such as network optimization, server utilization, QoS metric optimization, software architecture and etc. Among these directions, QoS metric of the servers (mainly web servers) is also studied by many scholar taking into account different QoS metric parameters.

Investigation shows that end-users are patient for the maximum delay in each request up to eight-ten seconds. That's way most of the researches are concentrating on providing service rate up to maximum eight-ten seconds, in ideal case one-three seconds [3].

Network services and system architecture planning are heavily dependent on the prediction of demand of the requested services. Nowadays, enormous increase of data size (video, picture, sound files and streaming) make scholars to find optimal solution for the serving them in short time within the minimum computation, that is the main constraint of the internet and cloud computing. In that case Poissonian properties of the arrival packets are little bit challenge to be modeled taking into account many attributes. Because of its tremendous analytical qualities, Poisson processes are widely used in network and system architecture planning and analysis. By implementation that kind of solution may result in non-negligible ramifications for the QoS offered, such as network optimization, queue management, platform optimization and etc.

In this paper, we will consider queuing management with feedback in cloud computing centers with large numbers of web servers. Our contribution in this kind research is feasibility of implementation of state merging algorithm because of unit difference in the server repair time and interarrival times of jobs. Namely, server repair time is calculated in hours, where interarrival times of jobs is with seconds.

The rest of the paper is organized as follows. In Sect. 1 detailed literature review about implementation of queuing system in the cloud computing is analyzed. In the Sect. 2 physical model of the proposed queue management in the cloud system is given, and both exact and approximate methods are detailed explained. Sections 3 and 4 are about the numerical results, and conclusion and future works, respectively.

2 Related Work

Modern web servers are extremely distinguished from small server to the giant computing system depending on services they're performing and number of users [4]. Because of network capabilities some arrival packet may be dropped because of no idle place in the processing queue, even though the server is under fairly light load [5]. Erramilli et al. [6] investigated dependence of traffics that required long time of service on queue management system, and found positive correlation between feasibility and practical impact on them. Authors showed that the key parameters of the cloud computing: buffer sizing, admission control and rate control have significant affect on the number of arrival and reserving packets.

In [7] author modeled software for apache web server where it can manage several parallel threads at the same time working independently or dependently with each other. They mentioned that implementation of static queue management system, such as FCFS is not proper for the web servers, further more simple queuing models don't satisfy modern web servers requirement. Today's cloud computing implementation diverse significantly depending on the their application area. Thus web servers should management the queue and serving mechanism so that minimum computation is done. Processes running at once, several threads running once, or with a combination. They also showed that neither service nor arrival processes is necessarily Poissonian. They implemented the M/G/1/K/PS queue management model in their research. The objectives of this study was to investigate the probability of blocking of the server within the given model. They have simulated arrival packets similar to the HTTP packets with Poisson and General distribution function. Congestion in the network and the probability of blocking in the web server side extremely affect the entire performance of the cloud computing. Because the web server as assumed as gateway to the cloud system. Mei et al. [5] deeply analyzed the effects of congestion in networks for response time and the probability of blocking in the web server side and found positive correlation between them.

In [8–13] some other parameters (such as response time, throughput and network utilization) of QoS metric in cloud computing were investigated. In [8], the authors applied a classic M/M/m model in order to get the response time distribution of a cloud system. Inter-arrival and service time were taken by exponential distribution function. In [12], Karlapudi's developed performance tool in order to predict web server demand on different response scenarios. In [13], Mei analyzes QoS metric in VoIP services within the cloud system.

In [14] the authors showed that the servers within the cloud system that utilize 60% of their resources in the peak power, can run on 20–30% utilization. They found PowerNap solution that save the energy about 23% by optimization queue in the system.

In [15], authors touched the same problem mentioned in [14]. They found that in the cloud systems with thousand of servers there is a 7–16% gap between real utilization of resources and suggested by the manufacturer. They also mentioned that each data center within the cloud system must have their own queue and resource management system based the real requirements. QoS metric performance evaluation of cloud technology through the simulation sometimes doesn't give precise result as we expected [16,17].

A cloud computing with multi-server system models many times faces with the server failures. The main reason of these failures comes from the inaccurate and imprecise queueing management model. This factor severely affect the entire performance of the cloud computing [18–27].

Enver [28] analyzed the ability to deliver acceptable levels of QoS for the cloud systems, and requirements for the performance of the system. Author proposed the analytical solution approach in order to solve QoS metric calculation problem within the large-scale systems in cloud computing with the presence

of failures and repairs. It is mentioned that there're significant difference in the final simulation results. Author's solution mainly focused on level decomposition approach.

As we see different models of queuing management system were implemented in order to analyze cloud computing, mainly web servers performance. However, to the best of our knowledge, so far feedback queuing model didn't applied to the cloud system. Similar approach is done by Chakka [29] for large-scale of networks.

In this paper we analyzed QoS metric of cloud computing with large-scale of web server from queue management perspective. Unlike [28] we take into account of being impatient of call when there isn't any operative server in the system. Moreover, we propose a new method in order to effectively calculate the steady-state probabilities of cloud system with a large number of web servers. By implementation of these steady-state probabilities, it's possible to calculate desired QoS metrics with low computation. Numerical results showed that the proposed algorithms have higher accuracy and negligible computation time. Taking into account that in the cloud computing repair of server is taking some hours and response time is handling within some seconds, we can correctly apply space-merging algorithm [30].

3 Exact and Approximate Methods of the Proposed Model

The physical model of the proposed queue management system in the cloud computing is given in Fig. 1. Clients are simultaneously sending requests. When the requests arrive to the web server, they have to wait for a while in the queue. The main point here is that a number of active thread in the web server. By default web server are generation one thread per request, and it results in increase in utilization and memory. Web servers must optimally handle the number of thread and buffer management, so that, CPU utilization and RAM usage should be below some thresholds. In our model request in any case (failure, not accepted, not validated and etc.) should return the client.

Jobs arrive to the servers from different users independently according to Poisson distribution with rate of λ. The service times for each job are independent and identically distributed (i.i.d.) random variable (r.v.) with exponential distribution with the rate μ. There are N number of servers in the cloud system and the total capacity of the system is R, i.e. the buffer size for waiting of jobs is equal to $R - N$. The queue is handled by arbitrary conservative discipline which means that if there are jobs in system then server does not idle. In other words, when a user has a job from the cloud system, in case at least one of the servers are idle, the job will be handled by one of the idle servers; otherwise, if all the servers are busy and the queue is sufficient to accept the incoming job, it will join the queue. Unlike the [14] here it is assumed that jobs in queue are impatient, i.e. every jobs in queue independently other ones waits in the queue

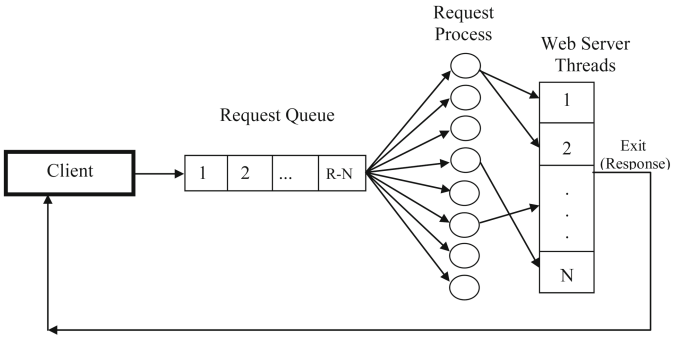


Fig. 1. The physical model of proposed queue management in cloud system

during random time, which has exponential distribution with mean α^{-1}. If the queue is full, the incoming job will be lost (rejected).

The servers considered are prone to failures where operative periods (i.e. period of good works) are exponentially distributed with a mean ϵ^{-1}. We consider two schemas: in schema I it is assumed that server might be failure in both cases, i.e. when server in working status and when server is idle; in schema II it is assumed that server might be failure only in case when server is idle. In both schemas, when server fails, it requires an exponentially distributed repair time with mean η^{-1}.

As in [28] it is assumed that services that are interrupted by failures are eventually resumed from the point of interruption or repeated with re-sampling. All inter-arrival, service, operative periods and repair times are independent of each other.

The main objectives of the paper is to develop efficient computational algorithm for calculation of the steady-state probabilities of the cloud system. By solving this problem, we can also easily obtain QoS metrics of the cloud system that are an average values of following quantities: number of servers in working status, number of jobs in system, throughput and response time. Probability of loss of jobs due to buffer overflow is represent interest in given study as well.

4 Generation of Q-Matrix

Firstly we consider Schema I. State of the system is defined by the two-dimensional vector (i, j), where i is the number of operative servers and j is the number of jobs in the system, respectively. Based on the distribution function of the random variables involved in the formation of the model, we determine that the two-dimensional Markov chain (2-D MC) describes the studied system. The set of all possible states of the system, i.e., state space of given 2-D MC is defined as lattice $S = \{0, 1, ..., N\} \times \{0, 1, ..., R\}$ (see Fig. 2).

The transition intensity from the state (i_1, j_1) to the state (i_2, j_2) is denoted as $q((i_1, j_1), (i_2, j_2))$. The combination of these values involves Q-matrix of given 2-D MC and are determined from the following relations (see. Figure 2):

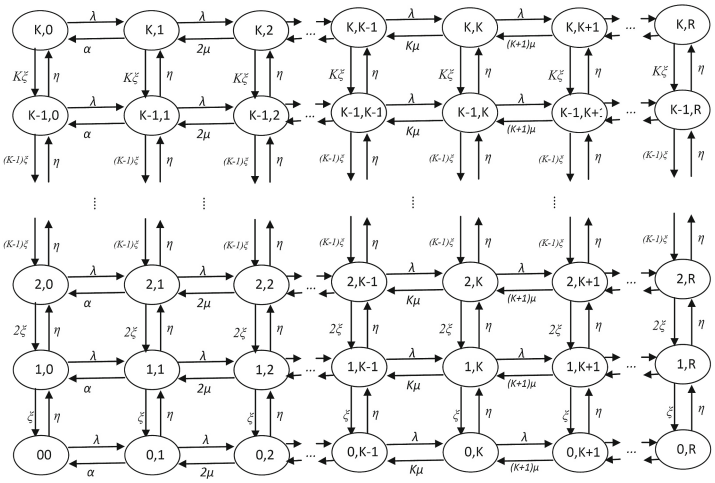


Fig. 2. State diagram of the proposed model

For the case $i_1 = 0$:

$$q((0, j_1), (i_2, j_2)) = \begin{cases} \lambda, & \text{if } (i_2, j_2) = (0, j_1 + 1), \\ \eta, & \text{if } (i_2, j_2) = (1, j_1), \\ j_1\alpha, & \text{if } (i_2, j_2) = (0, j_1 - 1), \\ 0, & \text{in other cases} \end{cases} \tag{1}$$

For the case $i_1 > 0$:

$$q((i_1, j_1), (i_2, j_2)) = \begin{cases} \lambda, & \text{if } (i_2, j_2) = (i_1, j_1 + 1), \\ \eta, & \text{if } (i_2, j_2) = (i_1 + 1, j_1), \\ i_1\xi, & \text{if } (i_2, j_2) = (i_1 - 1, j_1), \\ \min(i_1, j_1)\mu, & \text{if } (i_2, j_2) = (i_1, j_1 - 1), \\ 0, & \text{in other cases} \end{cases} \tag{2}$$

Note that the given finite 2-D MC is an irreducible, so there exists a stationary mode. Let $p(i, j)$ means the stationary probability of state $(i, j) \in S$. These probabilities are determined by solving the system of equilibrium equations (SEE) together the normalizing condition. SEE is compiled on the basis of (1) and (2) and due to their evidence, explicit form of this SEE does not given here.

5 Exact Formulas for QoS Metrics

After the solving SEE, characteristics of the studied system are defined as the marginal distributions of the 2-D MC. Thus, indicated above performance measures of the system are defined as follows:

average number of servers in working status N_w is

$$N_w = \sum_{i=1}^{N} i \sum_{j=0}^{R} p(i,j) \tag{3}$$

average number of jobs in system L_s is

$$L_s = \sum_{j=1}^{R} j \sum_{i=0}^{N} p(i,j) \tag{4}$$

probability of loss of jobs P_b is

$$P_b = \sum_{i=0}^{N} p(i,R) \tag{5}$$

Average throughput TH and average response time RT are calculated as follows:

$$\Gamma_{av} = \mu N_w \tag{6}$$

$$RT = L_s/\Gamma_{av} \tag{7}$$

The dimension of indicated SEE determined by the dimension of the state space S, which is defined as $(N+1)$ $(R+1)$. Unfortunately, it is difficult to find the analytical solution of this system of equations. Matrix-geometric [31] and spectral expansion [29] methods can be employed for the numerical solution of the given problem. However implementation of these methods require additional conditions on the transition rates. First, it's mandatory that transitions rates due to impatience of jobs are independent on number of jobs in queue. However from (1) it's obvious that transitions rates due to impatience of jobs are linear function of number of jobs in queue but not constant one. Secondly, in cloud computing systems numbers of servers and buffer size are in orders of hundreds to thousands. So these methods becomes inefficient due to large dimension of the state space S since their applying causes problems related to ill conditionality of high-dimensional matrices used at different stage of appropriate algorithms.

Therefore, to eliminate computational difficulties, we use the space merging method (SMM) [30] for approximate calculation of the stationary distribution of 2-D MC. This method can be applied here correctly, because, in cloud system servers are reliable in terms of hardware and software configuration. Namely, drop of incoming packet are rarely observed due to high sustainability of the servers. Furthermore, the repair time of the server are measured in hours. However arrival intensity and service intensity of packets much higher than drop intensity or repair intensity of the server [28].

6 Approximate Method

In accordance to the above mentioned comments we concluded that the transitions intensity between the states within the columns in the state diagram is

much higher than the lateral transitions intensity between them (see Fig. 2). Then considering the following splitting of state space S

$$S = \cup_{i=0}^{N} S_i \tag{8}$$

where $S_i = \{(i,j) \in S : j = 0,1,...,R\}, i = 0,1,...,N$. In other words, it is investigated splitting transition diagram by column (see Fig. 2).

Merging function given in state space S is determined based on the splitting (8) as follows:

$$U((i,j)) = <i> \tag{9}$$

where $<i>$ is a merging state, which includes all the states of the class S_i. Let $\Omega = \{<i>: i = 0,1,...,N\}$ According to the space merging method (SMM) algorithms [30], we find that the state probability of the initial model is defined as follows:

$$p(i,j) \approx \rho_i(j)\pi(<i>), \tag{10}$$

where $\rho_i(j)$ denotes the state probability of (i,j) within the splitting model with state space S_i, and $\pi(<i>)$ is the probability of the merging state $<i> \in \Omega$

From splitting scheme (8) it is clear that all the splitting models are one-dimensional birth and death processes (1-D BDP), so that in the class of states S_i the first component is constant. Therefore, in the study of the splitting model with state space S_i microstate $(i,j) \in S$ of given model can be represent by scalar $j, j = 0,1,...,R$. From (2) and (3) we get that these parameters for the splitting model with state space S_i are defined as follows (see Fig. 2):

For the case $i = 0$:

$$\rho_0(j) = \frac{\frac{\nu_1^j}{j!}}{\sum_{i=0}^{R} \frac{\nu_1^i}{i!}}, j = 0,1,...,R, \text{where } \nu_1 = \lambda/\alpha; . \tag{11}$$

For the case $i = 1,...,N$

$$\rho_i(j) = \begin{cases} \frac{\nu_2^j}{j!}\rho_i(0), & \text{if } 0 \le j \le i, \\ \frac{\nu_2^j}{j!i^{j-i}}\rho_i(0), & \text{if } i+1 \le j \le R, \end{cases} \tag{12}$$

where $\nu_2 = \lambda/\mu$ and $\rho_i(0)$ is calculated from normalizing condition, i.e. $\sum_{j=0}^{R} \rho_i(j) = 1$.

The transition intensity from the merging state $<i_1>$ to other merging state i_2 is denoted $q(<i_1>,<i_2>), <i_1>, <i_2> \in \Omega$ Then after certain algebras on the bases of (1), (2), (11) and (12) we obtain:

$$q(<i_1>,<i_2>) = \begin{cases} \eta, & \text{if } i_2 = i_1 + 1, \\ i_1\xi, & \text{if } i_2 = i_1 + 1 \\ 0, & \text{in other cases.} \end{cases} \tag{13}$$

From (13) we obtain that the merging state probabilities $\pi(<j>), <j> \in \Omega$ are calculated as the state probabilities of classical Erlang's model $M/M/N/N$

with load $\sigma = \eta/\xi$ erl, i.e.

$$\pi_1(< j >) = \frac{\frac{\sigma^j}{j!}}{\sum_{i=0}^{N} \frac{\sigma^i}{i!}}, j = 0, 1, ..., N \tag{14}$$

Therefore, taking into account the relations (10)-(14) we obtain steady-state probabilities $p(i, j), (i, j) \in S$.

After certain algebras we obtain the following simple approximate formulas for calculating the desired QoS metrics of the cloud system:

$$N_w \approx \sigma(1 - E_B(\sigma, N)) \tag{15}$$

where $E_B = \frac{\frac{\sigma^N}{N!}}{\sum_{i=0}^{N} \frac{\sigma^i}{i!}}$ is Erlang's B-formula

$$L_s = \sum_{j=1}^{R} j \sum_{i=0}^{N} \rho_i(j)\pi_1(<i>) \tag{16}$$

$$P_b = \sum_{i=0}^{N} \rho_i(R)\pi_1(<i>) \tag{17}$$

From (15)–(17) by using (6) and (7) performance measures TH and RT might be calculated.

Note 1. Formula (15) indicated that an average number of servers in working status does not dependent on load parameters of incoming jobs λ and μ and it is determined by failure and repair rates ξ and η. This result was expected one since in Schema I the events failure and repair of servers does not depend on number of jobs in cloud system. In other words, by using approximate approach we have find exact result for one of performance measure. This is undirected proof of high accuracy of proposed approach.

Now consider Schema II. Since the state diagram of the Schema II is very similar to the diagram of Schema I, we didn't included the diagram of Schema II. In this schema elements of Q-matrix for the case $i_1 = 0$ are calculated by the formula (1) as well. However in this schema elements of Q-matrix for the case are calculated as follows:

$$q((i_1, j_1), (i_2, j_2)) = \begin{cases} \lambda, & \text{if } (i_2, j_2) = (i_1, j_1 + 1), \\ \eta, & \text{if } (i_2, j_2) = (i_1 + 1, j_1), \\ (i_1 - j_1)^+\xi, & \text{if } (i_2, j_2) = (i_1 - 1, j_1), \\ \min(i_1, j_1)\mu, & \text{if } (i_2, j_2) = (i_1, j_1 - 1), \\ 0, & \text{in other cases.} \end{cases} \tag{18}$$

From (1) and (18) we can obtain SEE for the Schema II. In order to solve computational difficulties of the steady-state probabilities of this Schema, we can apply proposed above approximate method. No repeating above procedures

note that in this case steady-state probabilities within splitting models are calculated by (11) and (12) also. However in this case merging state probabilities $\pi_2(<j>), <j> \in \Omega$ are calculated as follows:

$$\pi_2(<j>) = \frac{\sigma^j}{\prod_{i=1}^{j} \theta(i)} \pi_2(<0>), j = 1, ..., N \tag{19}$$

where $\pi_2(<0>)$ is calculated from normalizing condition, i.e.

$\pi_2(<0>) = \left(1 + \sum_{j=1}^{N} \frac{\sigma^j}{\prod_{i=1}^{j} \theta(i)}\right)^{-1}$ and $\theta(i) = \sum_{j=1}^{i} j\rho_i(i-j)$.

QoS metrics L_s and P_b are calculated similar to (16) and (17) where $\pi_1(<j>)$ are substituted by $\pi_2(<j>)$ but N_w is calculated as follows:

$$N_w \approx \sum_{i=1}^{N} i\pi_2(<i>) \tag{20}$$

As we seen from formula (20) in this schema an average number of servers in working status depend on load parameters of incoming jobs λ and μ as well as on failure and repair rates ξ and η.

7 Numerical Results

As numerical results we have calculated exact and approximate solution and compared them. From (1) and (2) we get exact values of the system at the given values of the parameters $R = 100$, $N = 50$, $\lambda = 50, \alpha = 0.005, \mu = 1, \xi = 0.008, \eta = 0.0001$. Exact values were driven from Matlab 9.0 running on the personal computer with Intel i7 processor and 16 GB of RAM. Since size of matrix is equal to $(N+1)(R+1)$, it is too difficult to calculate exact values by using SEE at higher value of N and R.

In order to compare an exact and approximate values, first of all, we took comparison of $p(i, j)$ (steady-state probabilities) as an absolute and cosine values (see Table 1). As we mentioned above, for lower values it is possible to calculate exact values of the steady-state probabilities, but for big number it is impossible. Our approximate methods calculate these probabilities at higher values of N and R. In real cloud computing system always involves many server. This means, in real practice we always have to take into consideration a large amount of servers.

Accuracy of the approximate method to calculation of steady-state probabilities is estimated by two norms: (1) Absolute value of maximum of differences between values of steady-state probabilities (Norm 1); (2) cosine similarity (Norm 2) (see Table 1).

High arrival intensity of requests shouldn't dramatically affect the general performance of the cloud system with N servers. For a certain number of server, the higher arrival intensity means that request will wait in the queue, but after all they will be served.

The same behavior can be seen in the nature of the Schema II. However, in comparison with the Schema I, we get more optimal QoS metric in the same

Table 1. Values of various norms for Schema I

R	50	50	50	50	100	100	100	100	200	200	200	200	500	500	500	500
N	10	20	30	40	10	20	30	40	20	40	60	80	20	40	60	80
Norm 1	0.00545	0.00645	0.01246	0.01046	0.09345	0.09011	0.08351	0.09022	0.09055	0.08625	0.08254	0.08952	0.07854	0.0825	0.08625	0.07905
Norm 2	0.9108	0.9276	0.9404	0.9003	0.9035	0.9034	9.062	0.9118	0.9331	0.9099	0.9078	0.9045	0.9339	0.9782	0.9325	0.9225

initial values. Although, there isn't a big difference in QoS metrics between Schema I and II, this differences make huge values in real practice in terms of cost efficiency.

Tables 2 and 3 gives exact and approximate values of QoS metrics for Schema I and Schema II, respectively. Because of multiplication operation in the QoS metrics formulas we have a little bit distinguish, that can be assumed as the calculation error. Probability of blocking for almost the same. The bigger number of server in the cloud system, the lower the loss of the requests. It doesn't mean that we should have a number of server in the cloud system. The point here is that we should utilize the servers as much as optimal in order to reduce the loss of request. The same scenario can be said for the Schema II.

Table 2. Exact and approximate value of QoS metrics for Schema I, $R = 100$, *– Exact Values, **– Approximate Values

N	N_w^*	N_w^{**}	L_s^*	L_s^{**}	Γ_{av}^*	Γ_{av}^{**}	RT^*	RT^{**}	PB^*	PB^{**}
6	5.513083	5.921185	8.523196	9.15412	5.513083	5.921185	1.439441	1.545994	0.00015	1.61E−04
11	10.07285	10.84566	10.36	11.15485	10.07285	10.84566	0.95522	1.028508	3.09E−05	3.32E−05
16	14.59946	15.75917	14.41439	15.55939	14.59946	15.75917	0.914667	0.987323	2.08E−05	2.24E−05
21	19.09073	20.65924	18.61579	20.14527	19.09073	20.65924	0.901088	0.975122	7.21E−06	7.79E−06
26	24.32102	25.54262	23.52669	24.70838	24.32102	25.54262	0.921076	0.96734	4.95E−07	5.36E−07
31	30.87354	30.40499	29.6279	29.17826	30.87354	30.40499	0.974442	0.959653	1.13E−07	1.18E−07
36	35.80708	35.24053	34.02942	33.491	35.80708	35.24053	0.965633	0.950354	6.82E−08	6.71E−08
41	40.71159	40.04118	38.18912	37.56024	40.71159	40.04118	0.953746	0.93804	5.45E−08	5.36E−08

Table 3. Exact and approximate value of QoS metrics for Schema II, $R = 100$,*– Exact Values, **– approximate values

N	N_w^*	N_w^{**}	L_s^*	L_s^{**}	Γ_{av}^*	Γ_{av}^{**}	RT^*	RT^{**}	PB^*	PB^{**}
6	5.576369	5.989155	8.447154	9.072448	5.576369	5.989155	1.41041	1.514813	0.000167335	1.80E−04
11	10.12505	10.90187	10.40127	11.19928	10.12505	10.90187	0.95408	1.02728	6.50718E−05	7.01E−05
16	14.64189	15.80497	14.45358	15.6017	14.64189	15.80497	0.9145	0.987138	2.92429E−05	3.16E−05
21	19.13031	20.70207	18.65292	20.18546	19.13031	20.70207	0.90102	0.975045	1.08088E−06	1.17E−06
26	24.3647	25.58849	23.56736	24.7511	24.3647	25.58849	0.92101	0.967275	1.62671E−06	1.71E−06
31	30.93018	30.46078	29.6796	29.22917	30.93018	30.46078	0.97435	0.959568	7.03302E−07	6.93E−07
36	35.88418	35.31641	34.09742	33.55792	35.88418	35.31641	0.96548	0.950208	6.40072E−08	6.30E−08
41	40.82667	40.15436	38.28501	37.65455	40.82667	40.15436	0.95345	0.937745	1.12406E−08	1.11E−08

High arrival intensity of requests shouldn't dramatically affect the general performance of the cloud system with N servers. For a certain number of server,

the higher arrival intensity means that request will wait in the queue, but after all they will be served. Total response time is acceptable. The same behavior can be seen in the nature of the Schema II. However, in comparison with the Schema I, we get more optimal QoS metric in the same initial values. Although, there isn't a big difference in QoS metrics between Schema I and II, this differences make huge values in real practice in terms of cost efficiency. For example, at the given initial values we have around 10 server difference in the values of Nav. This 10 servers implies extra budget for the cloud system.

8 Conclusion and Future Works

As we mentioned above modern web servers are extremely distinguished from small server to the giant computing system depending on services they're performing and number of users, because of network capabilities some arrival packet may be dropped because of no idle place in the processing queue, even though the server is under fairly light load. Here we considered queuing management in cloud computing centers with large numbers of web servers. Our contribution in this kind research is feasibility of implementation of state merging algorithm because of unit difference in the server repair time and interarrival times of jobs. Namely, server repair time is calculated in hours, where interarrival times of jobs is with seconds. Our result is applicable in the real cloud system in order to calculate the QoS metrics depending of application area.

References

1. Briscoe, G., Marinos, A.: Digital ecosystems in the clouds: towards community cloud computing. In: Proceedings of 3rd IEEE International Conference on Digital Ecosystems and Technologies (DEST 2009), pp. 103–108. IEEE (2009)
2. Marinos, A., Briscoe, G.: Community cloud computing. In: Jaatun, M.G., Zhao, G., Rong, C. (eds.) CloudCom 2009. LNCS, vol. 5931, pp. 472–484. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-10665-1_43
3. Cardellini, V., Casalicchio, E., Colajanni, M.: A performance study of distributed architectures for the quality of web services. In: Proceedings of the 34th Annual Hawaii International Conference on System Sciences (HICSS-34), Washington, DC, USA, vol. 9, p. 9019. IEEE Computer Society (2001)
4. Athula, G., Murugesan, G.: Guest editors' introduction Web engineering: an introduction. IEEE MultiMedia **8**(1), 14–18 (2001)
5. Mei, R.D., Hariharan, R., Reeser, P.K.: Web server performance modeling. Telecommun. Syst. **16**(3–4), 361–378 (2001)
6. Asho, E., Onuttom, N., Walter, W.: Experimental queueing analysis with long-range dependent packet traffic. IEEE/ACM Trans. Netw. **4**(2), 209–223 (1996)
7. Cao, J., Andersson, M., Nyberg, C., Kihl, M.: Web server performance modeling using an M/G/l/K*PS queue. In: 10th International Conference on ICT 2003, vol. 2, pp. 1501–1506, February/March 2003
8. Xiong, K., Perros, H.: Service performance and analysis in cloud computing. In: Proceedings of IEEE World Conference Services, pp. 693–700 (2009)

9. Slothouber, L.: A model of web server performance. In: Proceedings of the Fifth International World Wide Web Conference (1996)
10. Yang, B., Tan, F., Dai, Y.-S., Guo, S.: Performance evaluation of cloud service considering fault recovery. In: Jaatun, M.G., Zhao, G., Rong, C. (eds.) CloudCom 2009. LNCS, vol. 5931, pp. 571–576. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-10665-1_54
11. Ma, N., Mark, J.: Approximation of the mean queue length of an M/G/c queueing system. Oper. Res. **43**, 158–165 (1998)
12. Karlapudi, H., Martin, J.: Web application performance prediction. In: Proceedings of the IASTED International Conference on Communication and Computer Networks, pp. 281–286 (2009)
13. Mei, R.D., Meeuwissen, H.B.: Modelling end-to-end Quality-of-Service for transaction-based services in multidomain environment. In: Proceedings of the 19th International Teletraffic Congress (ITC 19), pp. 1109–1121 (2005)
14. Meisner, D., Gold, B.T., Wenisch, T.F.: PowerNap: eliminating server idle power. SIGPLAN Not. **44**, 205–216 (2009)
15. Meisner, D., Sadler, C.M., Barroso, L.A., Weber, W.-D., Wenisch, T.F.: Power management of online data-intensive services. In: Iyer, R., Yang, Q., González, A. (eds.) ISCA, pp. 319–330. ACM (2011)
16. Calheiros, R.N., Ranjan, R., Beloglazov, A., De Rose, C.A., Buyya, R.: CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. Softw. Pract. **41**(1), 23–50 (2011)
17. Deelman, E., Singh, G., Livny, M., Berriman, B., Good, J.: The cost of doing science on the cloud: the montage example. In: SC-International Conference for High Performance Computing, Networking, Storage and Analysis, pp. 1–12 (2008)
18. Bruneo, D.: A stochastic model to investigate data center performance and QoS in IaaS cloud computing systems. IEEE Trans. Parallel Distrib. Syst. **25**(3), 560–569 (2014)
19. Cao, J., Hwang, K., Li, K., Zomaya, A.Y.: Optimal multiserver configuration for profit maximization in cloud computing. IEEE Trans. Parallel Distrib. Syst. **24**(6), 1087–1096 (2013)
20. Chiang, Y.J., Ouyang, Y.C., Hsu, C.H.: An efficient green control algorithm in cloud computing for cost optimization. IEEE Trans. Cloud Comput. **3**(2), 145–155 (2015)
21. Ghosh, R., Longo, F., Naik, V.K., Trivedi, K.S.: Modeling and performance analysis of large scale IaaS clouds. Future Gen. Comput. Syst. **29**(5), 1216–1234 (2015)
22. Jin, Y., Wen, Y., Zhang, W.: Content routing and lookup schemes using global bloom filter for content-delivery-as-a-service. IEEE Syst. J. **8**(1), 268–278 (2014)
23. Khazaei, H., Misic, J., Misic, V.: Performance analysis of cloud computing centers using M/G/m/m+r queuing systems. IEEE Trans. Parallel Distrib. Syst. **23**(5), 936–943 (2012)
24. Mei, J., Li, K., Ouyang, A., Li, K.: A profit maximization scheme with guaranteed quality of service in cloud computing. IEEE Trans. Comput. **64**(11), 3064–3078 (2015)
25. Vilaplana, J., Solsona, F., Teixidó, I., Mateo, J., Abella, F., Rius, J.: A queuing theory model for cloud computing. J. Supercomput. **69**(1), 492–507 (2014)
26. Yang, B., Tan, F., Dai, Y.S.: Performance evaluation of cloud service considering fault recovery. J. Supercomput. **65**(1), 426–444 (2013)
27. János, S., Bérczes, T.: Tool supported analysis of queueing systems with Future Internet applications. In: Pre-Proceedings of 9th International Conference on Applied Mathematics, Baia Mare, Romania, pp. 114–117 (2013)

28. Enver, E.J.: Performability analysis of cloud computing centers with large number of servers. J. Supercomput. **73**(5), 2130–2156 (2017)
29. Chakka, R.: Spectral expansion solution for some finite capacity queues. Ann. Oper. Res. **79**, 27–44 (1998)
30. Ponomarenko, L., Kim, C.S., Melikov, A.Z.: Performance Analysis and Optimization of Multi-Traffic on Communication Networks. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15458-4
31. Neuts, M.F.: Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach, p. 332. John Hopkins University Press, Baltimore (1981)

The Time-Out Length Influence on the Available Bandwidth of the Selective Failure Mode of Transport Protocol in the Load Data Transmission Path

Denis Bogushevsky, Pavel Mikheev, Pavel Pristupa, and Serguey Suschenko(✉)

Tomsk State University, Lenina 36, Tomsk 634050, Russia
ssp.inf.tsu@gmail.com

Abstract. The model of asynchronous control procedure of virtual connection of the transport Protocol in the mode of selective failure in the form of a two-dimensional Markov chain with discrete time, taking into account the influence of Protocol parameters window size and the length of the timeout of waiting for end-to-end acknowledgement, the probability of packets distortion in individual links of the data path and queue lengths in transit nodes from the "external" flows onto the bandwidth of the virtual connection. The analysis of the dependence of the throughput of the control procedure on the protocol parameters, the error level in the communication channels, the length of the data transmission path, and the distribution of the queue sizes at the transit nodes.

Keywords: Transport protocol · Loaded data path
Mathematical model · Markov chain · Speed of virtual connection
Window size · Duration of end-to-end time-out
Distribution of queue lengths

1 Introduction

In modern networks that transmit computer data and multimedia subscribers traffic through a single network infrastructure, the requirements are significantly increased for the availability of accessible bandwidth and as a consequence to improve the efficiency of its use. The most important operational characteristic of a virtual connection managed by the transport protocol of a computer network is its bandwidth. This indicator is determined not only by the speed and reliability of data transmission in the information channels of the transport connection, but also by the intensity of external flows in relation to this connection, which have a part of the common route with it. A natural model of a multilink transport connection is a network of queuing systems. However, obtaining a meaningful analytical solution is possible only in special cases. At

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 120–131, 2018.
https://doi.org/10.1007/978-3-319-99447-5_11

the same time, the main indicator of the "external" load on the path in which the transport connection is researched is the queue sizes in front of the protocol data blocks of the connection in transit nodes. It is obvious that monitoring of such indicator allows estimating the distribution of queue lengths in transit nodes from external network flows relative to the connection being analyzed and used in calculating the operational characteristics of the connection and selecting protocol parameters for the communication session time between given pair of subscribers. The known models of asynchronous control procedures of a separate data link and transport protocol [4–10] do not allow to take into account the load on shared network resources (bandwidth of separate inter-node channels) provided by the proximity to other virtual connections aggregated on different sections of the path in separate links of the route of given interpersonal connection and manifested in the form of "external" queues in transit nodes. In this paper, we propose a mathematical model of a virtual connection controlled by a transport protocol in the selective failure mode, taking into account, in addition to the distortion factor in forward and reverse data transmission paths and the retransmission mechanisms, due to distortions and the expiration of timeout of unread response from the recipient of the information stream, from "external" interpersonal connections, developing the results obtained in [2,3]. The process of competition for all admissible values of waiting time for end-to-end acknowledgment has been studied.

2 Mathematical Model of Transport Protocol

Let us consider the exchange of data between subscribers connected by a multilink data transmission path. Suppose that the following assumptions are true. The nodes of the path are connected by duplex communication channels having the same bandwidth in both directions. The length of the data transmission path, expressed in the number of retransmission areas, is equal to D_n. The reverse channel, through which the confirmation is sent to the sender about the correct reception of sequence of data segments, has a length D_o. The sum of distance of the forward and reverse paths is interpreted as the round-trip delay time $D = D_n + D_o$. The probabilities of distortion of the segment in the communication channel for the forward $R_n(d)$, $d = \overline{1, D_n}$ and reverse $R_o(d)$, $d = \overline{1, D_o}$ directions of transmission of each retransmission section is given. Then the reliability of the transmission of data segments along the path from the source to the destination and back will be $F_n = \prod_{d=1}^{D_n}(1 - R_n(d))$; $F_o = \prod_{d=1}^{D_o}(1 - R_o(d))$. The processing time of segments in the nodes of the path is the same. Interactive subscribers have an unlimited flow of segments for transmission, and the exchange is performed by segments of the same length. The recipients acknowledgment of the correct reception of the received data is carried in the counter-flow segments. We believe that the retransmission of segments is organized in accordance with the selective failure procedure [1]. We believe that the loss of segments due to the absence of buffer memory in the nodes of the path does not occur. The probability function is specified b_n, $n = \overline{0, N}$ that each segment from the flow of the

analyzed connection in the transit node will meet a queue of size $n \leq N$, where N—is the maximum queue size determined by the capacity of the buffer pools of the transit nodes. The timeout of duration S is triggered before the start of the transmission of the first segment of the sequence and is fixed for all segments within the width of the window. We will assume that the size of the window of the controlled protocol is determined by the value of W, and $S > W$—specifies the duration of the timeout for the confirmation of the correctness of the data delivery. After the next segment is transferred, the protocol copies it to the transferred but not confirmed data queue and starts the timeout. As soon as the queue size becomes equal to the width of window W, the control protocol suspends the transmission pending for acknowledgement or when the timeout is over S for confirmation. Upon receiving of confirmation, the segments that have reached the addressee without distortion are deleted from the queue. When the timeout S is over, the corresponding segment is retransmitted and the timeout is restarted. We will call the clock period the time t, required to output the segment in a line. The clock period is determined by the sum of the segment output time in-line, the propagation time of the signal in the communication channel and the processing time of the segment by the receiving node. Then the time for the sender to receive the pass-through acknowledgement is distributed according to the geometric law with the parameter F_o and the duration of the sampling clock period t.

The dynamics of the queued, but not confirmed, segments on the sending node for different modes of operation of the control protocol due to the Markov-like of the discrete receipt process can be described by a two-dimensional Markov chain with discrete time and number of states in one dimension equal to the duration of the end-to-end time-out S, and by another—increased by one unit maximum length of the queue: $N + 1$. It is obvious that the duration of the timeout must be sufficient to ensure that the packet segment data on the direct channel has reached the recipient and confirmation by the recipient on the reverse channel was accepted by the originator of the stream. It follows that the size of the timeout, expressed in duration of clock periods t, is advisable to choose not less than the sum of the path length and the size of the encountered queue in the transit node $S \geq D + n$. However, the size of the timeout may be less than the duration of the round-trip delay. The acknowledgement for the first segment of the sequence can be received by the sender after the time $s \geq D$ of the intervals of duration t, necessary to achieve the first segment of the addressee and return to the sender confirmation of the correctness of its reception. If a packet with a segment in the forward path (or a confirmation in the reverse direction) encountered a queue of size n, in the transmission of the sequence of by the sender or the transfer of the confirmation, the time for confirmation of acknowledgement is increases by the size of the encountered queue n and amounts to $s \geq D + n$.

The process of transferring the information flow/stream by a transport protocol in a single-link virtual channel is modeled by the Markov chain [4]. A generalization of this model for an empty multi-tier data transmission path was

carried out in [1]. The functioning of a virtual connection managed by a transport protocol in a loaded multi-link data transmission path with segment queues before sending data or confirmations can be described by a Markovized process in which the queue size in front of the forward or reverse data stream/flow of the test connection is an additional variable of the Markovs process. In the state of the Markov chain (i, n), the source sent a sequence of sizes of $i - n$ segments, which during the transfer process in one of the links met a queue of length n segments. The values of the state coordinates $i = 0, W+n$, $n = \overline{0, N}$ of Markov chain correspond to the number of transmitted segments, but they are not acknowledged by the recipient and the time from the beginning of the transmission of the sequence and to the values $i = W + n + 1,\ S - 1$, $n = \overline{0, N}$ the time during which the sender is not active and expects to receive an acknowledgement for the correct reception of the transmitted sequence from W segments. We denote by $P(i, n)$, $i = \overline{0, S - 1}$, $n = \overline{0, N}$—the probabilities of states of Markovs chain. Then the sequence of transmitted, but not confirmed, data segments of the virtual connection under consideration for a zero-length queue grows to the state of Markov chain with coordinates $(D-1, 0)$ with probability b_0. Further growth in the size of this sequence occurs with probability $b_0(1 - F_o)$. In states (i, n), $i = \overline{D - 1 + n, S - 1}$, $n = \overline{0, N}$ it is possible for the sender to receive an acknowledgement and depending on the results of the delivery, the sender transmits new segments (when acknowledgement is positive) or resends again—distorted ones. Since the sent sequence of segments of the studied virtual connection can meet the queue of non-zero length at any time of the transfer process (on the path of the sequence to the addressee or when transferring the confirmation to the sender of the information flow), the transition from state $(i, 0)$, $i = \overline{0, S - 2}$ to state (i, n), $i = \overline{0, S - 2}$, $n = \overline{1, N}$ occurs with probability b_n.

Let us denote π_{in}^{jm} the transition probabilities of Markov chain, where (i, n)—coordinates of the initial source and (j, m)—altered states of the chain. Then the dynamics of the process of information flow in the selective failure mode can be set by the following values of transition probabilities:

$$\pi_{in}^{jm} = \begin{cases} b_0, & j = i + 1, m = 0;\ i = \overline{0, D - 2}, n = 0; \\ b_0(1 - F_o), & j = i + 1, m = 0;\ i = \overline{D - 1, S - 2}, n = 0; \\ b_m, & j = i, m = \overline{1, N};\ i = \overline{0, S - 2}, n = 0; \\ b_0 F_o, & j = D - 1, m = 0;\ i = \overline{D - 1, W - 1}, n = 0; \\ b_0 F_o, & j = W + D - 2 - i, m = 0;\ i = \overline{W, W + D - 2}, n = 0; \\ b_0 F_o, & j = 0, m = 0;\ i = \overline{W + D - 1, S - 2}, n = 0; \\ 1, & j = 0, m = 0;\ i = S - 1, n = \overline{0, N}; \\ 1, & j = i + 1, m = n;\ i = \overline{0, D - 2 + n}, n = \overline{1, N}; \\ 1 - F_o, & j = i + 1, m = n;\ i = \overline{D - 1 + n, S - 2}, n = \overline{1, N}; \\ F_o, & j = D - 1, m = 0;\ i = \overline{D - 1 + n, W - 1 + n}, n = \overline{1, N}; \\ F_o, & j = W + n + D - 2 - i, m = 0; \\ & \qquad i = \overline{W + n, W + n + D - 2}, n = \overline{1, N}; \\ F_o, & j = 0, m = 0;\ i = \overline{W + n + D - 1, S - 2}, n = \overline{1, N}. \end{cases}$$

The bandwidth of a virtual connection managed by a transport protocol is defined as the relation of the average amount of data transmitted between two consecutive confirmations of acknowledgement to the average receipt time [1]. The contribution to the speed of the virtual connection is given by those states of the Markov chain, for which acknowledgement is possible. The normalized bandwidth of a virtual connection in a busy path is determined by the ratio of the average number of data segments transmitted by the sender between the recipients of two consecutive acknowledgements to the average time between confirmation of acknowledgement expressed in the number of intervals of duration t: $Z(W,S) = \overline{V}/\overline{T}$. Since acknowledgements are transferred in each segment independently and arrive to the sender every clock period t provided that they are not distorted in the path of length D from the receiver to the sender of information flow, the average time between confirmation acknowledgement is distributed geometrically with the parameter F_o and will be: $\overline{T} = 1/F_o$. The average volume of data transmitted between confirmation of acknowledgement taking into account the fact that each segment of studied connection with probability b_n, $n = \overline{0,N}$ meets the queue of the size n and contributes to the volume of transmitted information inversely proportional to the value $n+1$, is given by generalization of the relation given in [4]:

$$\overline{V} = \sum_{n=0}^{N} \frac{1}{n+1} \left[\sum_{l=D-1+n}^{W+D-2+n} \bar{l} P(l,n) + \sum_{l=W+D-1+n}^{S-1} \bar{W} P(l,n) \right].$$

The values $\bar{l}$ and $\bar{W}$ are determined by the average number of segments that reached the addressee with a selective procedure for organizing retransmissions of the distorted segments: $\bar{l} = (l-D-n+2)F_n$, $\bar{W} = WF_n$. Finally, the dependence of the bandwidth of the virtual connection on the protocol parameters of the window width W and the duration of end-to-end timeout S will take the form:

$$Z(W,S) = F_n F_o \sum_{n=0}^{N} \frac{1}{n+1} \left[\sum_{l=D-1+n}^{W+D-2+n} (l-D+2-n) P(l,n) + W \sum_{l=W+D-1+n}^{S-1} P(l,n) \right] \quad (1)$$

3 Analysis of the Transmission Process in the Single-Link Path

In general, the system of local equilibrium equations for state probabilities is sensitive to the relationships between the protocol parameters W and S, the duration of the circular delay in the unloaded data transmission path D and the maximum queue size of the competitive protocol data units in the transit nodes $n = \overline{0,N}$. The main parameter that determines the variety of solutions for the probability of states of Markov chain is the duration of the timeout, limited

from below in the general case only by the value of the circular delay D and the window size W. In stationary conditions, the system of equilibrium equations describing the process of data transfer in the virtual channel for the length path $D = 2$, at connections between loading characteristics, parameters of a path of data transmission and a transport protocol of the form

$$W \geq 1, \quad S \geq W + N + 1 \tag{2}$$

according to the transition probabilities has the following form:

$$P(0,0) = F_o \left[b_0 \sum_{i=W}^{S-2} P(i,0) + \sum_{n=1}^{N} \sum_{i=W+n}^{S-2} P(i,n) \right] + \sum_{n=0}^{N} P(S-1,n); \tag{3}$$

$$P(1,0) = b_0 P(0,0) + F_o \left[b_0 \sum_{i=1}^{W-1} P(i,0) + \sum_{n=1}^{N} \sum_{i=n+1}^{W-1+n} P(i,n) \right]; \tag{4}$$

$$P(i,0) = b_0(1-F_o)P(i-1,0), \ i = \overline{2, S-1}; \tag{5}$$

$$P(0,n) = b_n P(0,0), \ n = \overline{1,N}; \tag{6}$$

$$P(i,n) = b_n P(i,0) + P(i-1,n), \ i = \overline{1,n+1}, \ n = \overline{1,N}; \tag{7}$$

$$P(i,n) = b_n P(i,0) + (1-F_o)P(i-1,n), \ i = \overline{n+2, S-2}, \ n = \overline{1,N}; \tag{8}$$

$$P(S-1,n) = (1-F_o)P(S-2,n), \ n = \overline{1,N}. \tag{9}$$

Let us find the solution of this system of equations. From Eq. (5) we obtain:

$$P(i,0) = P(1,0)\big[b_0(1-F_o)\big]^{i-1}, \quad i = \overline{1, S-1}.$$

According to (6), (7) and the obtained expression for we have:

$$P(i,n) = b_n \left[P(0,0) + P(1,0)\frac{1-b_0(1-F_o)^i}{1-b_0(1-F_o)} \right], \ i = \overline{0,n+1}, \ n = \overline{1,N}.$$

Taking into account this relation of (8) and (9) for arbitrary we find:

$$P(i,n) = b_n(1-F_o)^{i-1} \left\{ \frac{P(0,0)}{(1-F_o)^n} + P(1,0) \left[\frac{1}{[1-b_0(1-F_o)](1-F_o)^n} - \frac{b_0^i}{1-b_0} + \frac{F_o b_0^{n+1}}{(1-b_0)[1-b_0(1-F_o)]} \right] \right\}, \ i = \overline{n+1, S-2};$$

$$P(S-1,n) = b_n(1-F_o)^{S-2} \left\{ \frac{P(0,0)}{(1-F_o)^n} + P(1,0) \times \left[\frac{1}{[1-b_0(1-F_o)](1-F_o)^n} - \frac{b_0^{S-2}}{1-b_0} + \frac{F_o b_0^{n+1}}{(1-b_0)[1-b_0(1-F_o)]} \right] \right\}.$$

Substituting the relations found in (4) we obtain an expression for:

$$P(1,0) = \frac{P(0,0)E}{(1-F_o)^{W-1}}, \text{where}$$

$$E = \frac{(1-b_0)[1-b_0(1-F_o)][1-(1-b_0)(1-F_o)^{W-1}]}{(1-b_0)[1-b_0+F_o b_0^W] + b_0 F_o (1-b_0^{W-1}) \sum_{n=1}^{N} b_n (b_0(1-F_o))^n}.$$

From the normalization condition, we find the probability of the initial state $(0,0)$:

$$P(0,0) = (1-F_o)^{W-1}\Bigg\{(1-F_o)^{W-1}\Bigg[\frac{3-2(1+b_0-F_o)+b_0(1-F_o)}{F_o} + \sum_{n=1}^{N} b_n\left(n - \frac{(1-F_o)^{S-n-1}}{F_o}\right)\Bigg] + E\Bigg[\frac{2-(1+b_0-F_o)[1-2b_0(1-F_o)]-b_0(1-F_o)[3+b_0(1-F_o)]}{F_o[1-b_0(1-F_o)]^2} - \frac{[b_0(1-F_o)]^{S-2}(1-b_0)}{1-b_0(1-F_o)} + \sum_{n=1}^{N} b_n\Bigg[\frac{n}{1-b_0(1-F_o)} - \frac{(1-F_o)^{S-n-1}}{F_o[1-b_0(1-F_o)]} + \frac{[b_0(1-F_o)]^{n+1}}{[1-b_0(1-F_o)]^2} - \frac{b_0^{n+1}(1-F_o)^{S-1}}{(1-b_0)[1-b_0(1-F_o)]}\Bigg]\Bigg]\Bigg\}^{-1}.$$

For interval constraints on the transport connection parameters of the form $W \geq 2$, $N \leq W-2$, $W+1 \leq S \leq W+N+1$ Eqs. (3), (4) are transformed to

$$P(0,0) = F_o\left[b_0\sum_{i=W}^{S-2} P(i,0) + \sum_{n=1}^{S-W-1}\sum_{i=W+n}^{S-2} P(i,n)\right] + \sum_{n=0}^{N} P(S-1,n);$$

$$P(1,0) = b_0P(0,0) + F_o\Bigg[b_0\sum_{i=1}^{W-1} P(i,0) + \sum_{n=1}^{S-W-1}\sum_{i=n+1}^{W-1+n} P(i,n) + \sum_{n=S-W}^{N}\sum_{i=n+1}^{S-2} P(i,n)\Bigg].$$

In this case, the form of the solution of the system of equilibrium equations coincides with the previous case to within a coefficient E, which is determined by the relation:

$$E = (1-b_0)[1-b_0(1-F_o)]\Bigg[1-(1-F_o)^{W-1}\sum_{n=1}^{S-W-1} b_n - \sum_{n=1}^{S-W-1} b_n(1-F_o)^{S-n-2}\Bigg] \Bigg/ \Bigg\{(1-b_0)\Bigg[F_ob_0^W + \sum_{n=1}^{S-W-1} b_n + \sum_{n=S-W}^{N} b_n(1-F_o)^{S-W-n-1}\Bigg] + b_0F_o\Bigg[\left(1-b_0^{W-1}\right)\sum_{n=1}^{S-W-1} b_n[b_0(1-F_o)]^n + (1-F_o)^{S-W-1}\Big(\sum_{n=1}^{S-W-1} b_nb_0^n - \sum_{n=S-W}^{N} b_nb_0^{S-2}\Big)\Bigg]\Bigg\}.$$

Restrictions on the duration of the timeout, leading to the possibility of not receiving end-to-end acknowledgments due to too large queue sizes are described as follows:

$$W \geq 2, \quad N \geq W-2, \quad W+1 \leq S \leq N+3. \tag{10}$$

Then Eq. (4) can be rewritten in the form:

$$P(1,0) = b_0 P(0,0) + F_o\left[b_0 \sum_{i=1}^{W-1} P(i,0) + \sum_{n=1}^{S-W-1} \sum_{i=n+1}^{W-1+n} P(i,n) + \sum_{n=S-W}^{S-3} \sum_{i=n+1}^{S-2} P(i,n)\right],$$

and Eqs. (7) and (8) are valid for states $i = \overline{1, n+1}$, $n = \overline{1, S-3}$; $i = \overline{1, S+2}$, $n = \overline{S-2, N}$ and $i = \overline{n+2, S-2}$, $n = \overline{1, S-4}$ respectively. Equation (9) is transformed to the following:

$$P(S-1,n) = (1-F_o)P(S-2,n), \quad n = \overline{1, S-3};$$
$$P(S-1,n) = P(S-2,n), \quad n = \overline{S-2, N}.$$

The probability of the states retains the form of the dependencies corresponding to the constraints (2) with the following relation for the coefficient E:

$$E = (1-b_0)[1-b_0(1-F_o)]\left[\sum_{n=0}^{S-3} b_n - (1-F_o)^{W-1} \sum_{n=1}^{S-W-1} b_n \right.$$
$$\left. - \sum_{n=S-W}^{S-3} b_n (1-F_o)^{S-n-2}\right] \Bigg/ \left\{(1-b_0)\left[(1-F_o)^{1-W}\left(1 - \sum_{n=0}^{S-3} b_n\right) + F_o b_0^W + \sum_{n=1}^{S-W-1} b_n \right.\right.$$
$$+ \sum_{n=S-W}^{S-3} b_n (1-F_o)^{S-W-n-1}\Big] + b_0 F_o \left[\left(1 - b_0^{W-1}\right) \sum_{n=1}^{S-W-1} b_n [b_0(1-F_o)]^n \right.$$
$$\left.\left. + (1-F_o)^{S-W-1}\left(\sum_{n=S-W}^{S-3} b_n b_0^n - b_0^{S-2} \sum_{n=S-W}^{S-3} b_n\right)\right]\right\}.$$

The probability of the initial state $P(0,0)$, found from the normalization condition, is not given here because of its crocktitude nature.

In the case of a start-stop protocol and exceeding the waiting time in queues of transit nodes over the duration of the time-out

$$W = 1, \quad N \geq 0, \quad 2 \leq S \leq N+2 \tag{11}$$

the probability of states $(0,0)$ and $(1,0)$ Markov chain are transformed to $P(1,0) = b_0 P(0,0)$ and

$$P(0,0) = \left\{\frac{1+b_0F_o}{1-b_0(1-F_o)}\left[1 + (S-1)\sum_{n=S-2}^{N} b_n\right] - \frac{b_0[b_0(1-F_o)]^{S-1}}{1-b_0(1-F_o)} \right.$$
$$- \frac{b_0\left[1-[b_0(1-F_o)]^{S-1}\right]}{[1-b_0(1-F_o)]^2} \sum_{n=S-2}^{N} b_n + \sum_{n=1}^{S-3} b_n \left[\frac{(n+2)(1+b_0F_o)}{1-b_0(1-F_o)} - b_0 \frac{1-[b_0(1-F_o)]^{n+2}}{[1-b_0(1-F_o)]^2}\right]$$
$$+ \sum_{n=S-2}^{N} b_n \frac{1+b_0F_o-b_0[b_0(1-F_o)]^{S-2}}{1-b_0(1-F_o)} + \sum_{n=1}^{S-4} b_n \left[\frac{(1+b_0F_o)\left[1-F_o-(1-F_o)^{S-n-2}\right]}{F_o[1-b_0(1-F_o)]}\right.$$
$$\left. + \frac{b_0(1-b_0)[b_0(1-F_o)]^{n+1} - (1-F_o)^{S-2}\left(b_0^{n+2} - b_0^S\right)}{(1-b_0)[1-b_0(1-F_o)]}\right]$$
$$\left. + \sum_{n=1}^{S-3} b_n \left[\frac{(1+b_0F_o)(1-F_o)^{S-n-2}}{1-b_0(1-F_o)} + \frac{F_o b_0^{n+2}(1-F_o)^{S-2}}{(1-b_0)[1-b_0(1-F_o)]} - \frac{b_0[b_0(1-F_o)]^{S-2}}{1-b_0}\right]\right\}^{-1}.$$

Let us analyze the solution obtained in a number of special cases. We note that the solutions of the systems of equilibrium equations are "joined" at the boundaries of the domain of variation of the protocol (W, S) and load (N) parameters. Suppose that constraints (2) are satisfied and there is no external queue at nodes ($b_0 = 1$). Then we obtain the well-known result [1]:

$$P(0,0) = \frac{F_o(1-F_o)^{W-1}}{1+(1-F_o)^{W-1}\left[F_o-(1-F_o)^{S-W}\right]};$$

$$P(i,0) = \frac{F_o(1-F_o)^{i-1}}{1+(1-F_o)^{W-1}\left[F_o-(1-F_o)^{S-W}\right]}, \quad i=\overline{1,S-1};$$

$$P(i,n) = 0, \quad i=\overline{0,S-1}, \quad n=\overline{1,N}.$$

Assume that the flow of studied virtual connections is always met by a queue of non-zero length ($b_0 = 0$). Then the probability of the initial state and the state $(1,0)$ of Markov chain under equity (2) takes the form:

$$P(0,0) = \frac{F_o(1-F_o)^{W-1}}{1+F_o\left[1+\bar{N}\right]+(1-F_o)^{W-1}\left[F_o-\sum_{n=1}^{N} b_n(1-F_o)^{S-W-n}\right]};$$

$$P(1,0) = \frac{P(0,0)\left[(1-F_o)^{W-1}\right]}{(1-F_o)^{W-1}}, \tag{12}$$

where $\bar{N} = \sum_{n=1}^{N} nb_n$. In the case of a queue of deterministic length ($b_n = 1$, $n \geq 1$) for the initial state, we obtain:

$$P(0,0) = \frac{F_o(1-F_o)^{W-1}}{1+F_o(n+1)+(1-F_o)^{W-1}\left[F_o-(1-F_o)^{S-W-n}\right]}.$$

If the window width is unlimited ($W \to \infty$), and therefore the timeout duration S, the initial state is irrevocable ($P(0,0) = 0$), and the probability of the state $(1,0)$ takes the form

$$P(1,0) = F_o[1-b_0(1-F_o)]^2 \Big/ \Big\{2-(1+b_0-F_o)[1-2b_0(1-F_o)]$$

$$-b_0(1-F_o)[3+b_0(1-F_o)]+\bar{N}F_o[1-b_0(1-F_o)]+F_o\sum_{n=1}^{N} b_n[b_0(1-F_o)]^{n+1}\Big\}.$$

If this connection always competes for bandwidth with competitors ($b_0 = 0$), then the probability of this state is converted to

$$P(1,0) = \frac{F_o}{1+F_o(1+\bar{N})}.$$

For absolutely reliable reverse link ($F_o = 1$), the timeout duration exceeding the maximum queue size by twice the path length ($S \geq N+2$), and the start-stop

control procedure ($W = 1$), the set of likely states (i, n) is determined by the index values $i = \overline{0, n+1}$, $n = \overline{0, N}$ and looks like:

$$P(0,0) = \frac{1}{3 - b_0^2 + (1 + b_0)\bar{N}}, \quad P(i,n) = b_n P(0,0).$$

Under the same conditions $W \geq 2$ with and fairness (2), only meaningful states will be (i, n), $i = \overline{1, n+1}$, $n = \overline{0, N}$:

$$P(0,0) = 0,\ P(1,0) = \frac{1}{2 - b_0 + \bar{N}},\ P(i,n) = b_n P(1,0),\ i = \overline{1, n+1},\ n = \overline{1, N}.$$

If (11) is satisfied, the probability of the initial state is

$$P(0,0) = \frac{1}{3 - b_0^2 + (1 + b_0)\left[\sum_{n=1}^{S-3} n b_n + (S-2)\sum_{n=S-2}^{N} b_n\right]}.$$

4 A Study of the Available Bandwidth

The analysis of the information flow transfer process in the virtual channel controlled by the transport protocol shows that the index of the virtual connection performance at absolutely reliable communication channels in separate links of the reverse path ($F_o = 1$) and sufficient time-out duration according to (1) and found probabilities of Markov chain States (11) for a single-link path at $W = 1$ determined by the relation

$$Z(1,S) = \frac{F_n\left[b_0 + (1 + b_0)\sum_{n=1}^{N} \frac{b_n}{n+1}\right]}{3 - b_0^2 + (1 + b_0)\bar{N}},$$

and $W \geq 2$ in accordance with (12) will be:

$$Z(W,S) = \frac{F_n}{2 - b_0 + \bar{N}}\left[1 + \sum_{n=1}^{N} \frac{b_n}{n+1}\right].$$

For the case when the flow of the studied single-link virtual connection always meets the queue of nonzero length ($b_0 = 0$), its speed according to (1) and (8) will be determined by the expression:

$$Z(W,S) = F_n \frac{F_o^2[1 - (1 - F_o)^{W-1}] + \sum\limits_{n=1}^{N} \frac{b_n}{n+1}[1 - (1 - F_o)^W - W F_o (1 - F_o)^{S-n-1}]}{1 + F_o(1 + \bar{N}) + (1 - F_o)^{W-1} - (1 - F_o)^W - \sum\limits_{n=1}^{N} b_n (1 - F_o)^{S-n-1}}.$$

When $b_0 = 0$, $S = W + 1$ and $W \geq N + 2$ in accordance with the found probabilities of states of Markov chain from (1), we obtain:

$$Z(W, W+1) = F_n \Bigg\{ F_o^2 \Bigg[1 - \sum_{n=1}^{N} b_n (1-F_o)^{W-n-1} \Bigg] + \sum_{n=1}^{N} \frac{b_n}{n+1} \Big[1 - (1-F_o)^{W-n} \times \Big(1 - F_o(W-n) \Big) \Big] \Bigg\} \Bigg/ \Bigg\{ 1 + F_o(1+\bar{N}) + (2F_o - 1) \sum_{n=1}^{N} b_n (1-F_o)^{W-n-1} \Bigg\}.$$

With the correctness of (2), $W \geq 2$, $F_o = 1$, and the deterministic queue length ($b_n = 1, n > 0$), the share of the bandwidth of the path (1) available to the virtual connection will be

$$Z(W, S) = \frac{F_n}{n+1}.$$

In the case $W = 1$, $D = 1$, $b_0 = 1$ we arrive at the well-known result [2,3]:

$$Z(1, S) = \frac{F_n}{2}.$$

With an unlimited growth of the window size consequently, the timeout available to the subscriber connection, the share of the bandwidth of the path (1) has the form

$$Z(\infty, \infty) = \frac{F_n[1 + (n+1)F_o^2]}{(n+1)[1 + (n+1)F_o]}.$$

As the competition between subscribers increases for the bandwidth of the data transmission path, the queue size increases, and the information transfer speed decreases rapidly.

5 Conclusions

The analysis is aimed at studying the available bandwidth of the virtual connection in the data path shared by competing subscribers. A mathematical model of the loaded path is proposed for the given distribution of queue lengths of protocol data blocks of competitive subscribers in the form of Markov chain describing the dynamics of the volume of transmitted but not confirmed data of the studied virtual connection. The distributions of Markov chain states are found for different relations between protocol parameters and path characteristics. Analytical dependencies of the speed of the virtual connection for various conditions of its functioning are obtained. Numerical studies of the speed of a virtual channel in a selective retransmission mode have shown that the transmission speed in a virtual connection is largely determined by the intensity of data loss, the distribution of the queue size in transit nodes, and the relation between the path length and the window size. It is obvious that in order to achieve high speed of the transport connection, the duration of the end-to-end timeout should exceed the sum of the circular delay of a separate segment in the unloaded path and the total length of the queues before the information flow of the interacting subscribers of the virtual connection. The obtained results allow to assert that the virtual connection bandwidth index increases with the increase of the end-to-end confirmation timeout duration and practically reaches the theoretical limit at saturation by the protocol parameter W for the values S, exceeding the sum of the window width and the maximum queue size.

References

1. Kokshenev, V.V., Mikheev, P.A., Suschenko, S.P.: Comparative analysis of the performance of selective and group repeat transmission models in a transport protocol. Autom. Remote Control **78**(2), 247–261 (2017)
2. Kokshenev, V.V., Mikheev, P.A., Suschenko, S.P.: Analysis of selective mode of transport Protocol failure in the loaded data path. Vestnik TGU. Ser. Control Comput. Eng. Comput. Sci. (3), 78–94 (2013)
3. Suschenko, S.P.: Analytical models of asynchronous control procedures data link. Avtomatika I vychisleniya technique (2), 32–40 (1988)
4. Kokshenev, V.V., Suschenko, S.P.: Performance analysis of asynchronous data link control procedure. Comput. Technol. **13**(5), 61–65 (2008)
5. Kravets, O.Y.: Mathematical modeling of parameterized TCP protocol. Autom. Remote Control **74**(7), 1218–1224 (2013)
6. Padhye, J., Firoiu, V., Towsley, D.F., Kurose, J.F.: Modeling TCP Reno performance: a simple model and its empirical validation. IEEE/ACM Trans. Netw. **8**(2), 133–145 (2000)
7. Arvidsson, A., Krzesinski, A.: A model of a TCP link. In: Proceedings of the 15th International Teletraffic Congress Specialist Seminar (2002)
8. Altman, E., Avrachenkov, K., Barakat, C., Gelenbe, E., Labetoulle, J., Pujolle, G.: A stochastic model of TCP/IP with stationary random loss. ACM SIGCOMM Comput. Commun. Rev. **30**(4), 231–242 (2000)
9. Olsen, J.: Stochastic modeling and simulation of the TCP protocol. Uppsla Dissertations in mathematics 28, 94 p. (2003)
10. Nikitinskiy, M.A., Chalyy, D.J.: Performance analysis of trickles and TCP transport protocols under high-load network conditions. Autom. Control Comput. Sci. **47**(7), 359–365 (2013)

ICT-Based Beekeeping Using IoT and Machine Learning

Kristina Dineva(✉) and Tatiana Atanasova

Department of Modelling and Optimization, Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Acad. G. Bonchev St., Block 2, 1113 Sofia, Bulgaria
{k.dineva,atanasova}@iit.bas.bg
http://www.iict.bas.bg

Abstract. Integrating information and communication systems into different industries enables storage, management and processing of data and then transforming it into useful information and knowledge. The gradual integration of information technology into beekeeping may provide remarkable results in control over care for bee hives and increase profitability. The aim of this paper is to provide ICT-based means to predict the survival of bee families when the winter season is over. The procedures of collecting and processing of the data from beehives is examined. Several machine learning algorithms are applied and tested. The performance of classification models used for the prediction is estimated. The experimental results are presented.

Keywords: IoT · Sensor data · Beekeeping · Data processing
Machine learning algorithms

1 Introduction

Information and communications technologies (ICT) are the infrastructure and elements that enable modern computing. The term is generally accepted to mean all devices, networking components, applications and systems that combined allow people and organizations to interact in the digital world. ICT also covers the implementation of various components: software, hardware, transactions, communications technology, data, internet access and cloud computing. The combination of these components has been successfully applied in a number of sectors such as education, health, transport, agriculture and many others. Successful implementation of ICT in different industries brings many benefits in the decision-making and managing various processes.

The integration of information technologies in beekeeping can help taking better decisions by the beekeeper in different situations, better adapting to newly occurring conditions and providing better quality of bee products. The ICT components will provide an opportunity to enrich the knowledge of the bees life cycle and the various processes occurring in hives under climate changes.

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 132–143, 2018.
https://doi.org/10.1007/978-3-319-99447-5_12

2 Problem Description

Beekeeping as an agricultural sub-sector needs support through the deployment of information technology in the overall processes management. By building a good architecture and using the right combination of ICT components, serious problems affecting bees can be overcome and occurrence of some of them can be even decreased.

Bees are a very important part of mankind's existence. Bees pollinate 1/6 of the blooming plants and over 400 types of crops. The disappearance of bees will lead to the need for artificial pollination of various crops. Experts in that field estimate that artificial pollination worldwide will worth at about 155 billion dollars a year. On the other hand, there has been a 300% increase in crops that need bee pollination in the world over the last years. This high percentage predicts the occurrence of a "pollination crisis" - a situation where bee pollination services are limited and this can lead to a decline in final crop quantity and quality [1]. Besides this, the number of dying bee colonies grows each year and reaching 44% over the past 10 years [2]. Many factors, independent of each other, lead to the death of the bee families, but a number of more significant reasons standout such as:

1. Varroa mite - a parasite that is attacking bees;
2. The use of pesticides on agricultural crops located close to apiaries;
3. The professional set of skills of the beekeeper, as well as the time and resources invested in growing bee colonies.

The regular inspection of each beehive is of a particular importance for proper bee growing. Inspections are sometimes hampered by bad infrastructure, poor meteorology, etc. During these inspections, beekeepers get to know the status of the bee families. Unfortunately, it can happen that the inspection is late and the bee family is already lost. For this reason, it is necessary to introduce such elements of ICT as IoT devices, wireless connection [3,4] and sophisticated algorithms [5–7] in beekeeping. By building ICT-based system, beekeepers will be able to carry out remote monitoring of the micro climate in the beehive, which will improve the knowledge of the life style of bee families.

Reducing human errors and increasing the work efficiency are key goals for every industry. By introducing the usage of ICT these goals could be successfully achieved. The integration of new technologies must be carried out gradually and should be also considered according the needs and capabilities of the user. It is essential that the technology chosen also does not disturb the natural rhythm of work. The process of transforming collected data into useful information and knowledge is essential in accomplishing this set of goals.

3 Research Approach

Getting knowledge about bee families is a long and difficult process. In order to meet the goals, it is necessary to use an information based communication

system through which all steps are taken to produce useful, accurate and timely information.

The "Smart Bee Hives" [8] IoT-based system combines the main components that ICT includes (data, information, procedures, hardware, software and people). By combining these components, the system performs a complete process (Fig. 1) with collecting, processing, analyzing and visualizing data with the ability to predict events.

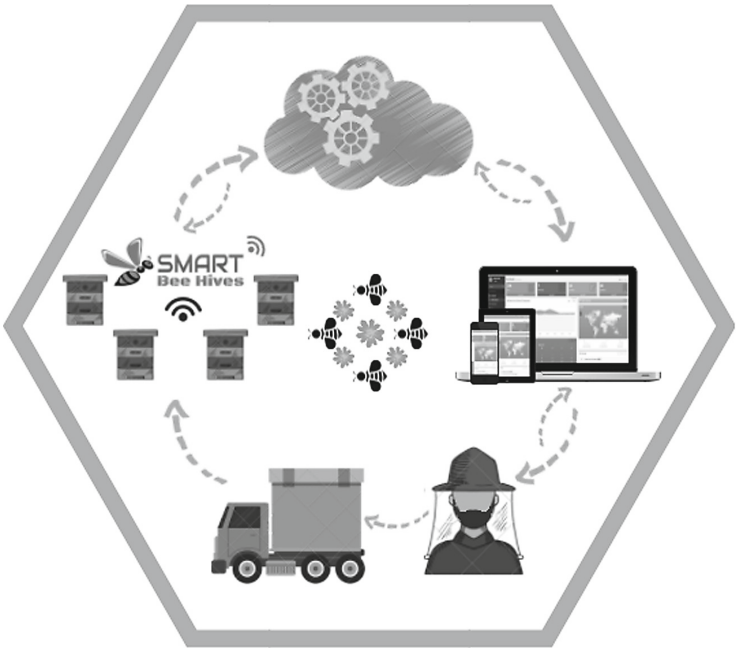


Fig. 1. "Smart Bee Hives" workflow

The gathering, transmission, intelligent processing and visualization of the collected data from the intelligent beehive monitoring system are carried out through several stages:

- collecting data from physical devices;
- organizing and grouping data according to predefined rules and user needs;
- data processing and analysis with appropriate models;
- the results obtained are rendered as visualization logical blocks of data in the different user interfaces.

Data gathering is carried out via IoT sensors devices that collect information about the micro climate in the hive and the outside weather conditions. The devices push the collected data to an intermediate module where primary filtering and data processing is performed. The data is then pushed to the central module.

A standardized process of operation - OSEMN - is applied to the collected data. The use of the OSEMN process provides a clear order of activities - **O**btain, **S**crub, **E**xplore, **M**odel the data and i**N**terpret the data. After complete data processing, machine learning algorithms are applied to the data-driven predictions or decisions, through building classification models from sample inputs. This way, the occurrence of future events in the beehive can be predicted according to the correlation between the analyzed data and situation in the beehive.

Analyzed data is visualized on a dedicated software platform [9] - "Smart Bee Hives": www.smartbeehives.eu.

The platform combines the latest UX concepts using the modern technology. For building the "Smart Bee Hives" platform, technologies such as ASP.NET Core, Angular 5, Bootstrap 4, CSS 3, Python, and others are used. As it can be seen from the site architecture map (Fig. 2), the user interface is divided into two parts - a landing page and a control panel. The presentation part introduces to the user various features, specifics, functionality, innovativeness and objectives of the project. There is a dedicated part with answers to the most frequently asked questions about the implementation and usage of the system.

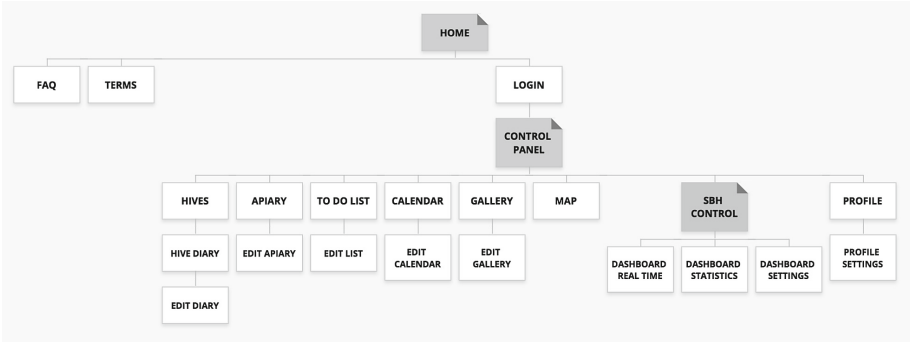


Fig. 2. "Smart Bee Hives" sitemap

The control panel is also divided into two parts - descriptive and control. The different admin/control panel sections allow users to freely create, input, edit, and delete information about their hives and apiaries. In the "Hive Diary" page there is a beekeeping electronic diary. It contains tables filling the data tracking the queen bee changes, date and type of treatment of the hive, diseases if any, frequency of feeding the hive, extracted honey quantity and many more. The state changes of each individual hive over the years can be easily tracked through this log. To-do-list and event-calendar capabilities provide the ability to track/guide and organize various beekeeping activities.

All pages in the "Admin panel" sections are related with each other. The dashboards are available to beekeepers that have installed "Smart Bee Hives" IoT-based sensor systems in their apiaries. These users have the ability to per-

form real-time monitoring of their hives and to have a full set of statistics visualized through charts.

The notable part of the dashboards is the area showing the prediction of possible events in bee hives. Events like hive rooting, sickness, presence or absence of a pile, and many other events typical for bee families can be predicted.

Predictive patterns are being created. When abnormal deviations are detected in the beehive, the beekeeper is immediately being notified with the description of the problem type.

The content is organized in order to improve the usability of the platform and its easy perception. The schema of the platform can be defined as "subjective" [10]. Subjective organization schemes categorize information in a way that may be specific to or defined by the field. This type of categorization helps users understand and draw links between the parts. The subjective schemes of the "Smart Bee Hives" platform include:

- Schemes on topics - Organize content based on the specific subject;
- Task Schemes - Organize content, taking into account the needs, actions, questions and processes users bring from this content.

The organizational schema defines the general characteristics of the content elements and affects the logical grouping of these elements. In addition to building an organizational schema, it is also important to build an organizational structure. The organizational structure defines the relationships between content elements and groups. The organizational structure of "Smart Bee Hives" is hierarchical top-down structure. This is a well-known and easy-to-navigate intuitive structure. A balance is created between the width and depth of the platform relative to its specificity. This approach allows content to be added to one of the two main parts of the platform without significant restructuring, which is important for achieving the goals.

In addition to the well-built platform architecture, it is also important to ensure high security of the data acquisition and data persistence. The web platform security measures include the usage of root certificates issued by Certificate Authorities, HTTPS protocol and authentication for the user access. In the server and in the interface there is a number of measures implemented aiming at SEO optimization with the purpose to increase application easy searchability using various search engines.

4 Applying Data Science Against Bee Set

The workflow of the ICT-based "Smart Bee Hives" system is well-developed and organized, consisting of several logical consecutive activities through which the set goals are achieved. During the process of exploring the data collected by the IoT sensors and the observations that are made on site for checking the survival of bee families after the winter season, a correlation is established between 4 variables: inside hive temperature, outside hive temperature, humidity in the

hive and humidity outside the hive. The classification model in the paper is constructed using these four variables and also the results of the observations.

OSEMN taxonomy of tasks of working with data has been selected. This is standardized and widely accepted process model. It consists of several steps: Obtain, Scrub, Explore, Model the data and iNterpret the data.

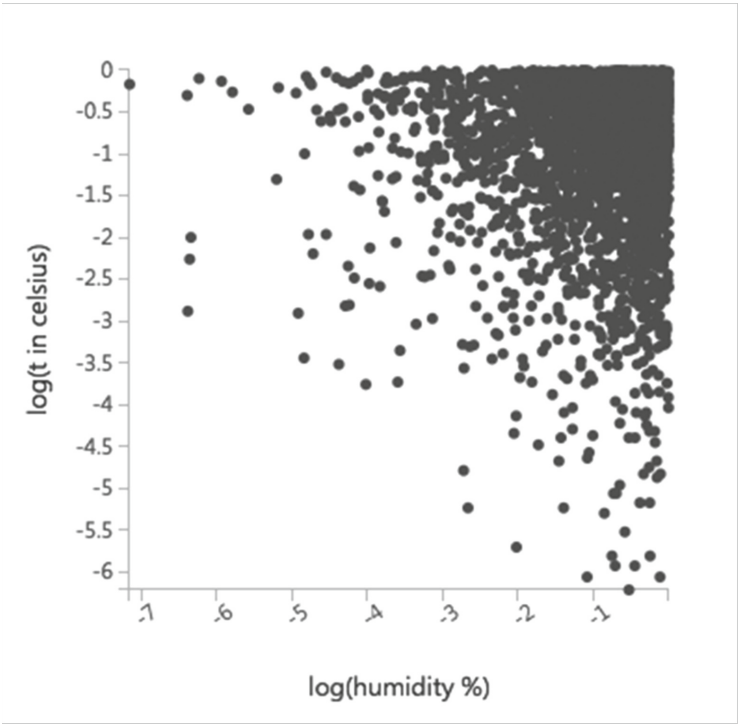


Fig. 3. Data obtained from the "Smart Bee Hives" system

Firstly the data received from IoT sensor devices (Fig. 3) need to be merged into a single table. Each column of this table represents a variable. This kind of data obtained from real observations often contains lots of invalid values - null, NaN or NA. Due to large amount of data and their specific type, such values are removed from the table because they may cause the continuation of data processing to stop and to threat the proper flow of the predicting algorithms.

Extreme values (outliers) are another factor influencing the correctness of the results. A RANSAC algorithm [11] is used to clear the dataset from outliers. It picks up a random number of all records in the database and performs linear regression against them. Input data may include noisy samples. This algorithm provides a statistical estimation of the probability of obtaining reliable forecasts,

i.e., probability within a predetermined number of standard deviations from the true values (1):

$$Y_{measured}(x) = Y_{noise-free}(x) + N, \tag{1}$$

where $Y_{noise-free}(x)$ is the expected activity in a noise-free environment and N is a random internal noise which obeys the homoscedasticity assumption – that it has a constant distribution across all activity values.

For obtaining the best results [12] it is necessary to normalize the data after proper dataset cleaning of missing and extreme values is being performed. The type of normalization used is Min-max. Min-max normalization performs a linear transformation of the initial data, where min_a and max_a are the minimum and maximum values for the attribute a. Min-max normalization maps v values of the range $[min_a, max_a]$ by computing:

$$v' = \frac{(v - min_a)}{(max_a - min_a)}. \tag{2}$$

The Min-max normalize linearly re-scales every feature to the [0, 1] interval. This type of normalization reduces the correlation between column values and leads to improving the performance of the algorithms.

The correct and consistent data preparation (Fig. 4) helps the work of algorithms and significantly increases the reliability of the obtained results.

5 Selecting the Proper Machine Learning Algorithms

The most important step in the entire process is to define the correct algorithm. This is a tricky process requiring the usage of different algorithms and comparing the obtained results.

The machine learning algorithms can be divided into different types of categories [13] according to their purpose:

- Supervised learning - make predictions based on a set of examples. A supervised learning algorithm looks for patterns in the value labels.
- Unsupervised learning - data points have no labels associated with them. The goal of an unsupervised learning algorithm is to organize the data in some way or to describe its structure.
- Semi-Supervised learning - input data is a mixture of labelled and unlabelled examples. There is a desired prediction problem but the model must learn the structures to organize the data as well as make predictions.
- Reinforcement learning - gets to choose an action in response to each data point. The learning algorithm also receives a reward signal a short time later indicating how good the decision was. Based on this, the algorithm modifies its strategy in order to achieve the highest reward.

The purpose of this study is to predict the survival of bee families after the winter season is over. These estimates will be based on a set of examples. In this case, we apply supervised learning.

rows 12000 columns 5

view as

t_in	t_out	humidity_in	humidity_out	observation
0.831667	0.028922	0.830762	0.304206	0
0.613333	0.12989	0.804675	0.053235	0
0.335	0.163373	0.024887	0.322265	0
0.511667	0.146772	0.559689	0.151074	1
0.368333	0.095888	0.771416	0.515259	0
0.066667	0.008521	0.334962	0.849526	0
0.383333	0.023802	0.677524	0.259344	0
0.19	0.180014	0.67841	0.791919	0
0.103333	0.144492	0.054432	0.342782	0
0.465	0.178654	0.588805	0.844296	0
0.841667	0.136411	0.893994	0.359155	1

Fig. 4. Cleaned and normalized dataset

The test data will be classified using two classes: 0 - perished and 1 - survivors. That is why we use a two-class classification model. Four types of algorithms in Machine Learning Azure Studio (Fig. 5) have been examined, applied and tested: Two-class Boosted Decision Tree, Two-Class Bayes Point Machine, Two-Class Logistic Regression and Two-Class Support Vector Machine.

Each of the applied algorithms defines a modeling function that determines how the algorithm works and how the results are predicted. The surest way to properly choose the most appropriate machine learning algorithm for a particular task is to run training and testing on the processed dataset with each of the selected algorithms.

6 Cross Validation Testing

During cross validation testing, the original dataset is randomly distributed in k-sub-units of the same size. The samples k maintain a sub-sample used as validation data for the model testing. The remaining $k-1$ sub-samples are used as training data. The cross validation process is then repeated k times, with each k being used only once as validation data. The result k can be averaged (or otherwise combined) to obtain a single estimate. The advantage of this method is that

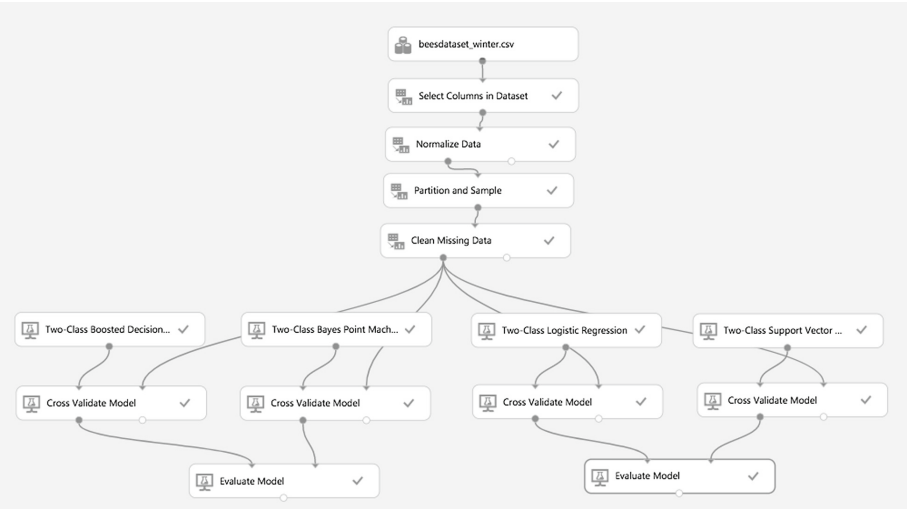


Fig. 5. Machine Learning data and workflow

all observations are used for both training and validation and each observation is used for validation just once.

The purpose of cross validation is to verify the model ability to predict new data that was not used in its assessment to highlight a problem like over fitting and to give an idea of how the model will be adapted to an independent set of data [14].

For classification problems, cross-validation testing is most appropriate. This experiment benefits from using 4-fold cross-validation.

7 Experimental Results

Table 1 presents comparative results obtained after applying the four types of classification algorithms. It can be noticed that the best results after the training were obtained by using the Boosted Decision Tree. The metrics obtained from this algorithm indicate that its precision is above 80%, the sensitivity is nearly 69%, and its precision for predicting values is almost 88%. The harmonic mean of precision and recall is 74%.

Figure 6 shows the comparison between the Boosted decision tree (in blue) and the Bayes Point Machine (in red). The results are visualized as lift curves, which are calculated by the Eq. (3):

$$lift = \frac{a/(a+b)}{a+c/(a+b+c+d)}, \tag{3}$$

where a = True Positive (TP), b = False Negative (FN), c = False Positive (FP) and d = True Negative (TN).

Table 1. Comparative results

Classification algorithms	Boosted Decision Tree	Bayes Point Machine	Logistic Regression	Support Vector Machine
Accuracy	**0.879**	0.762	0.770	0.733
Precision	**0.807**	0.583	0.625	0.440
Recall	**0.688**	0.229	0.245	0.180
*F*1*Score*	**0.743**	0.329	0.352	0.255

The closer the curve is to the upper left corner, the better the classifier's efficiency (this is maximizing the true positive rate while minimizing the false positive rate).

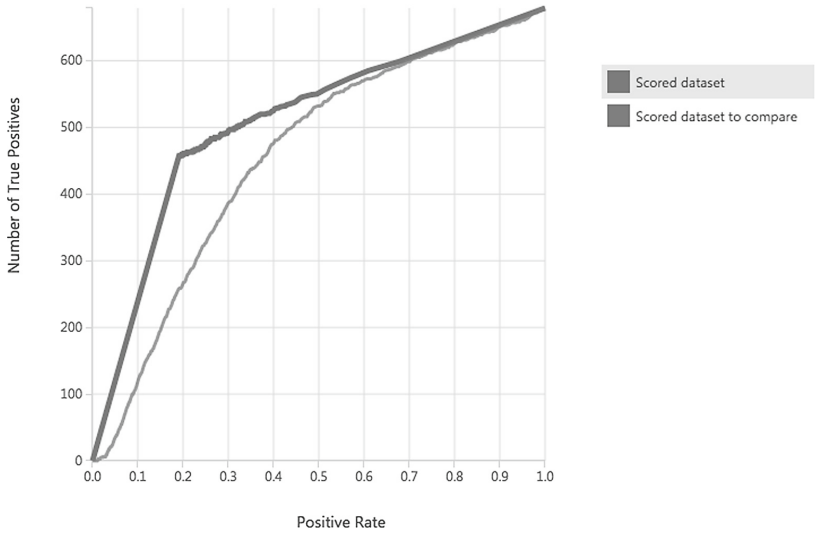


Fig. 6. Lift curves

The accuracy of the algorithm can easily be calculated through the confusion matrix (Table 2).

As it is shown in the metrics of the applied and tested algorithms, it was found that the Boosted decision tree algorithm outperforms significantly other three competitors (Bayes Point Machine, Logistic Regression and Support Vector Machine).

Table 2. Confusion matrix

	Predicted **1**	Predicted **0**
True 1	TP = 467	FN = 212
True 0	FP = 78	TN = 1643

8 Conclusion

The innovative "Smart Bee Hives" information and communication modular system provides collecting of real-time data, as well analyzing, visualizing and predicting the probability of occurrences of future events in beehives.

The most accurate predictive algorithm has been chosen after testing the performance of several different algorithms. Following the experiment in this investigation, the results show the superiority of the Boosted decision tree over the rest algorithms used in the test. From the metrics obtained, Boosted decision tree precision in prediction of events can be clearly seen - 88%. This high percentage of accuracy gives a significant advantage to beekeepers who have implemented "Smart Bee Hive" system into their apiaries. Using it beekeepers increase their knowledge of the processes occurring in the bee families, which helps them to better plan their actions. The gradual integration of information technology into bee-keeping [15] provides remarkable results such as control over honey extraction processes and increased profitability. The use of new technologies in beekeeping allows improvements to be done in the overall state of the industry, which will lead to a number of benefits for agriculture as a whole.

References

1. Allsopp, M., Tirado, R., Johnston, P.: Plan Bee – Living Without Pesticides. Greenpeace International, Amsterdam, The Netherlands, May 2014. https://storage.googleapis.com/p4-production-content/international/wp-content/uploads/2014/05/cac226e7-466-plan-bee.pdf
2. Nosowitz, D.: Honeybee deaths getting worse: we lost 44% of colonies last year. J. Mod. Farmer (2016). https://modernfarmer.com/2016/05/honeybee-colony-loss/
3. Vishnevsky, V.M.: Broadband Wireless Data Transmission Networks. Technosphera, Moscow (2005)
4. Alexandrov, A.: Ad-hoc Kalman filter based fusion algorithm for real-time wireless sensor data integration. In: Andreasen, T., et al. (eds.) Flexible Query Answering Systems 2015. AISC, vol. 400, pp. 151–159. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-26154-6_12
5. Balabanov, T., Zankinski, I., Barova, M.: Strategy for individuals distribution by incident nodes participation in star topology of distributed evolutionary algorithms. Cybern. Inf. Technol. **16**, 80–88 (2016)
6. Tashev, T.D., Hristov, H.R.: Modeling and synthesis of information interactions. Probl. Tech. Cybern. Robot. **52**, 75–80 (2001)

7. Kolchakov, K.: An approach for synthesis performance improvement of non-conflict schedule by decomposition of the connections matrix in the switching nodes. In: Proceedings of the International Workshop DCCN 2010, Moscow, Russia, pp. 168–173 (2010)
8. Dineva, K., Atanasova, T.: Model of modular IoT-based bee-keeping system. In: European Simulation and Modelling Conference ESM 2017, EUROSIS-ETI, pp. 404–406 (2017)
9. Dineva, K., Atanasova, T.: Computer system using internet of things for monitoring of bee hives. In: SGEM, pp. 169–176 (2017)
10. Morville, P., Rosenfeld, L.: Information Architecture for World Wide Web. O'Reilly, Sebastopol (2006)
11. Kaspi, O., Yosipof, A., Senderowitz, H.: RANdom Sample Consensus (RANSAC) algorithm for material informatics: application to photovoltaic solar cells. J. Cheminform. **9**(1), 34–49 (2017). https://doi.org/10.1186/s13321-017-0224-0
12. Yaghini, M.: Data Mining, Part 2 (2010). http://webpages.iust.ac.ir/yaghini/Courses/Data_Mining_882/DM_02_02_Data%20Preparation.pdf. Accessed 2018
13. Dharmendra, S.: Unsupervised, supervised and reinforced learning via spiking computation. DARPA (2016)
14. Vanwinkelen, G., Blockeel, H.: On estimating model accuracy with repeated cross-validation. In: Belgian-Dutch Conference on Machine Learning (2012)
15. Murphy, F., Magno, M.: b+WSN: smart beehive for agriculture, environmental, and honey bee health monitoring - preliminary results and analysis. IEEE (2015)

MAP/PH/1 Retrial Queue with Abandonment, Flush Out and Search of Customers

Dhanya Babu, A. Krishnamoorthy, and V. C. Joshua(✉)

Department of Mathematics, CMS College, Kottayam 686001, Kerala, India
{dhanyababu,krishnamoorthy,vcjoshua}@cmscollege.ac.in
http://www.cmscollege.ac.in

Abstract. This paper considers a single server retrial queueing system with search, abandonment and flush out of customers from the system (system clearance) periodically with exponentially distributed duration. A customer on arrival, enters for service, if the server is found to be idle and enter into an orbit of infinite capacity if the server is busy. Orbital customers receive service either by successful retrials or by an orbital search. At the epoch of completion of a service, sever goes for search with probability p as long as the orbit size is atmost L-1. Search stops the moment there are L or more customers in the orbit. Further orbital customers are assumed to renege with certain probability on an unsuccessful retrial. In addition, clearance of system takes place each time a random duration following exponential distribution, expires. The customers arrive to the system according to Markovian arrival Process, inter-retrial times are exponentially distributed and service time follows phase type distribution. We analyze the resulting GI/M/1 Type queue. Steady-state analysis of the model is performed. Some performance measures are evaluated.

Keywords: Retrial queues · Search mechanism · Abandonment
Flush out

1 Introduction

Queue is a manifestation of congestion in flow of objects through a system or a network consisting of many systems. Queueing theory was developed to provide models to predict the behaviour of systems that attempt to provide service for randomly arising demands. Practically, some important queueing models fails to answer queueing losses and so a new class called retrial queueing systems arise. This class of queues is characterized by the following feature: a customer on arrival, when all servers are busy leaves the service area but after some random time repeat his demand. Also retrial queues can be regarded as networks with re servicing after blocking. This field have many applications in computer and communication networking, aircraft landing and take-off, and in several

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 144–156, 2018.
https://doi.org/10.1007/978-3-319-99447-5_13

other areas. The first mathematical result about retrial queues were published in 1950s and applications in teletraffic theory were presented in the monograph of L. Kosten. About single-server and multi-server queues and their methods and results are described in [9,10,20]. The bibliographical information about retrial queues are in [1–3]. In the retrial queueing system customers arriving to a busy system join a group of blocked customers called orbit and try to capture a free server after a random amount of time. Neuts and Ramalhoto in [16] introduces the idea of search for customers in classical queue. In the retrial set up, each service is preceded and followed by the server(s) idle time because of the ignorance about the status of the server(s) and orbital customers by each other. We are interested in designing retrial queues that reduces the server(s) idle time and achieve this by the introduction of search of orbital customers immediately after a service completion. This is introduced by [6] in retrial queues. This is very much useful in reducing idle time and thereby reducing the waiting time of the customers. Inventory and repair services are some examples of search of orbital customers. Chakravarthy et al. [7] consider a multi-server retrial queue with orbital search in which primary customers arrive according to Markovian Arrival Process (MAP). Artalejo et al. in [5] describes M/G/1 retrial queue with search of customers in the orbit a single server queue with linear retrial policy where the server can go for search of customers immediately after each service completion with probability p_j when there are j customers in the orbit. M/G/1 retrial queue with non persistant customers and orbital search is analyzed in [12]. In [4] there are two objectives one is to introduce retrial queue with orbital search as an appropriate stochastic model for some practical repair models and the other is to provide a link between M/G/1 retrial queue and the standard M/G/1 queue. This is possible by choosing the recovery factor $p_j = \mathrm{r}$ as 0 and 1. Dudin et al. in [8] generalizes the above result by assuming the arrival process to be BMAP. A back and forth movement between classical and retrial queue by taking search probability $\mathrm{p} = 1$ until orbit size reaches a preassigned number and $\mathrm{p} = 0$ when orbit size exceeds that number have been studied in [11]. Extension of this work is discussed in [14] by taking search probability p $(0 < p < 1)$ and with probabilistic abandonment when orbit size is greater than or equal to that preassigned number, since retrial rate (linear) increases when orbit size is atleast that number. In this paper we discussed the same with constant retrial rate by considering flush out of customers by the realization of an exponential clock. Krishnamoorthy et al. in [13] discussed two GI/M/1 Type single server queueing inventory models in which items in the inventory have a common life time. Neuts in [17] developed the theory of phase type (PH)- distributions and related point processes. In stochastic modelling, PH- distributions lend themselves naturally to algorithmic implementation and have nice closure properties with a related matrix formalism that makes them attractive for practical use. Steady state probabilities are computed using Matrix Geometric methods by Neuts. The rate matrix is computed using Ramaswami's Logarithmic reduction Algorithm by [15]. In this paper steady-state analysis of the GI/M/1 Type model is performed by the method in [18]. Markov chains of the GI/M/1 Type is

first considered by Neuts in [19]. Matrix analytic methods have been extensively applied to more elaborate systems than QBD in which one-step transitions are allowed. Markov chains of M/G/1 Type (skip-free to the left) and GI/M/1- Type (skip-free to the right) can be brought under the unifying umbrella of QBD's. In this model we consider a GI/M/1 Type retrial queue with search of customers from the orbit with probabilistic abandonment and derive several system state characteristics.

In Sect. 2 we described the model. Steady-state analysis have done in Sect. 3, system performance measures are evaluated in Sect. 4 .

2 Mathematical Formulation of the Model

We consider a single server retrial queueing model with search of customers from the orbit when orbit size reaches below a pre assigned number say L. An arriving customer enters for service, if the server is found to be idle and enter into an orbit of infinite capacity when the server is found to be busy. Orbital customers receive service either by successful retrials or by an orbital search. On every service completion epoch, the server searches a customer in the orbit with a known probability p, $0 \leq p \leq 1$ until the orbit size exceeds a pre-assigned number L. The search time is assumed to be negligible. The server remains idle with probability $1-p$ until a new customer or a retrial customer gets into service. When the orbit size reaches L, the sever drops search. As a result, customers accumulate in the orbit and retries for service. After unsuccessful retrial customers in the orbit reneged from the system with a probability q. We also assumed that the system vanishes completely on realization of an exponential clock. So the model reduces to a particular case of GI/M/1 Type queue. Customers arrive according to Markovian Arrival Process (MAP). The service time assumed as phase type distributed amount of time with an irreducible representation $PH(\beta, S)$ of order m where the vector S^0 is given by $S^0 = -Se$. The customers in the orbit make retrial attempts which are exponentially distributed with parameter 'μ'. As $L \to \infty$, the server is always busy either with primary customer or with orbital customer through search. As a result, the model partially remains as classical queue when orbit size reaching $L-1$ and remains as retrial queue when orbit size is atleast L. As long as search is sure, this model acts as classical queue and acts as retrial queue when the search is dropped on the orbit size reaching L. But when L becomes larger and larger, i.e. $L \to \infty$ server idle time decreases. So by taking L larger and larger, we can reduce customer abandonment. In this paper we consider flush out of all customers in the system periodically with exponentially distributed duration clock with parameter θ.

The MAP, a special class of tractable Markov renewal process, is a rich class of point processes that includes many well-known processes such as Poisson, PH-renewal processes, and Markov-Modulated Poisson process (MMPP). One of the most significant features of the MAP is the underlying Markovian structure and fits ideally in the context of matrix-analytic solutions to stochastic models. Matrix analytic methods were first introduced and studied by Neuts. Poisson

processes are the simplest and most tractable ones used extensively in stochastic modelling. The idea of the MAP is to significantly generalize the Poisson processes and still keep the tractability for modelling purposes. In many practical applications, mainly in communications engineering, production and manufacturing engineering, the arrivals do not usually form a renewal process. So, MAP is a convenient tool to model both renewal and non-renewal arrivals.

The customers arrive to the system with a stochastic process $\{\nu_t, t \geq 0\}$ with a state space {1,2,...n}. The sojourn time of the chain in the state i is exponentially distributed with the positive finite parameter λ_i. When the sojourn time in the state i expires, with probability $p_0(i,j)$ the process ν_t jumps to the state j without generation of customers where $i,j = \{1,2,...n\}; i \neq j$ and with probability $p_1(i,j)$ the process ν_t jumps to the state j with generation of customers where $i,j = \{1,2,...n\}$.

The MAP process is completely characterized by the matrices D_0 and D_1 defined by $(D_0)_{i,i} = -\lambda_i, i = 1,2,...n$

$(D_0)_{i,j} = \lambda_i p_0(i,j); i,j = 1,2,...n, i \neq j$
$(D_1)_{i,j} = \lambda_i p_1(i,j); i,j = 1,2,...n.$

The point process described by the MAP is a special class of semi-Markov processes with transition probability matrix given by

$$\int_0^x \mathbf{e}^{(D_0 t)} dt D_1$$

By assuming D_0 to be a non-singular matrix, the inter arrival times will be finite with probability one and the arrival process does not terminate. Hence, we see that D_0 is a stable matrix. The matrix $D(1) = D_0 + D_1$ represents the generator of the process $\{\nu_t, t \geq 0\}$. The average arrival rate λ is given by $\lambda = \boldsymbol{\theta} D_1 \mathbf{e}$ where $\boldsymbol{\theta}$ is the invariant vector of the stationary distribution of the Markov chain $\{\nu_t, t \geq 0\}$. The vector $\boldsymbol{\theta}$ is the unique solution to the system

$$\boldsymbol{\theta} D(1)\mathbf{e} = 0, \boldsymbol{\theta}\mathbf{e} = \mathbf{1}. \tag{1}$$

Here $\mathbf{e}$ is a column vector of appropriate size consisting of 1' s and $\mathbf{0}$ is a row vector of appropriate size consisting of zeros. The squared integral coefficient of variation of intervals between successive arrivals is given by $c_{var} = 2\lambda\boldsymbol{\theta}(-D_0)^{-1}\mathbf{e} - 1$.

3 Analysis of the System

In this section we perform the steady-state analysis of the model by establishing the stability condition. The model described in the previous section is a GI/M/1 Type model. Analysis of the model is done by using the method described in [18].

Some Notations

Let

- I_n - an identity matrix of dimension n
- $\otimes$ - Kronecker product of two matrices
- If A is a matrix of order $p \times q$ and if B is a matrix of order $r \times s$, then $A \otimes B$ will denote a matrix of order $pr \times qs$ whose $(i,j)^{th}$ block matrix is given by $a_{ij}B$
- $N(t)$ be the number of customers in the orbit at time t
- $C(t)$ be the status of the server

$$C(t) = \begin{cases} 0, & \text{if the server is idle} \\ 1, & \text{if the server is busy} \end{cases}$$

- $S_1(t)$ be the phase of the service process
- $S_2(t)$ be the phase of the markovian arrival process.

The above model can be represented by the Markov process
$\chi = \{X(t)/t \geq 0\} = \{N(t), C(t), S_1(t), S_2(t); t \geq 0\}$
The state space is $\Omega = l^* \cup l(n)$ where

$l^* = \{(i, 0, 0, s_2) : s_2 = 1, 2, ...n\}$ and for $i \geq 0$,
$l(i) = \{(i, j, s_1, s_2), i \geq 0, 0 \leq j = 0, 1, \leq s_1 \leq m, 1 \leq s_2 \leq n\}$.

Enumerating the states of a continuous time Markov chain in lexicographic order, the infinitesimal generator of the Markov chain is of the form:

$$Q = \begin{pmatrix} A_{10} & A_0 & & & & & & \\ {A_{21}}^1 & A_1^1 & A_0 & & & & & \\ C & A_2^1 & A_1^1 & A_0 & & & & \\ C & & A_2^1 & A_1^1 & A_0 & & & \\ \vdots & & \ddots & \ddots & \ddots & & & \\ \vdots & & & \ddots & \ddots & \ddots & & \\ \vdots & & & & A_1^1 & A_0 & & \\ & & & & A_2^1 & A_1^1 & A_0 & \\ & & & & & A_2 & A_1 & A_0 \\ & & & & & & A_2 & A_1 & A_0 \\ \vdots & & & & & & & \ddots & \ddots \end{pmatrix}$$

where A_{10}, A_0, ${A_{21}}^1$, A_1^1, A_2^1, A_1, A_2 and C are square matrices of order n(m+1).

– A_{10} represents the transition matrix corresponding to level 0 to level 0.
– A_0 represents the transition matrix corresponding to an arrival of primary customer to the orbit.

- ${A_{21}}^1$ represents the level down of transition matrix by one either by realization of an exponential clock, successful retrial of or by corresponding to level 1.
- A_1^1 represents the transition matrix corresponding to level i to level i when $i \leq L-1$.
- A_2^1 represents the transition matrix corresponding to departure of customers either by service completion or abandonment of customers in level i when $i \leq L-1$.
- A_1 represents the transition matrix corresponding to level i to level i.
- A_2 represents the transition matrix corresponding to departure of customers either by service completion or abandonment of customers in level i.
- C represents the transition matrix of flush out of customers in the system.

The transitions can be described by the following matrices

$$A_{10} = \begin{pmatrix} D_0 & \beta \otimes D_1 \\ S_0 \otimes I_n & S \oplus D_0 \end{pmatrix}$$

$$A_0 = \begin{pmatrix} O & O \\ O & I_m \otimes D_1 \end{pmatrix}$$

$${A_{21}}^1 = \begin{pmatrix} \theta I_n & i\mu\beta \otimes I_n \\ \mathbf{e}_m \otimes \theta I_n & pS^0\beta \otimes I_n \end{pmatrix}$$

$$C = \begin{pmatrix} \theta I_n & O \\ \mathbf{e}_m \otimes \theta I_n & O \end{pmatrix}$$

$$A_1^1 = \begin{pmatrix} D_0 - (\mu+\theta)I_n & \beta \otimes D_1 \\ (1-p)S^0 \otimes I_n & (S-\theta I_n) \oplus D_0 \end{pmatrix}$$

$$A_2^1 = \begin{pmatrix} O & \mu\beta \otimes I_n \\ O & pS^0\beta \otimes I_n \end{pmatrix}$$

for $i \geq L$,

$$A_1 = \begin{pmatrix} D_0 - (\mu+\theta)I_n & \beta \otimes D_1 \\ S^0 \otimes I_n & (S-(\mu q+\theta)I_n) \oplus D_0 \end{pmatrix}$$

$$A_2 = \begin{pmatrix} O & \mu\beta \otimes I_n \\ O & \mu q \otimes I_n \end{pmatrix}$$

3.1 Stability Condition

Theorem. The irreducible Markov process Q is positive recurrent if and only if the minimal non negative solution R of the equation

$$\sum_{k=0}^{\infty} R^k A_k = 0$$

has

$$sp(R) < 1$$

and if there exists a positive vector x_0 such that

$$x_0 B[R] = \mathbf{0}$$

The matrix

$$B[R] = \sum_{k=0}^{\infty} R^k B_k$$

is a generator.

The stationary probability vector x, satisfying

$$\boldsymbol{x}Q = 0$$

and

$$\boldsymbol{x}\mathbf{e} = 1$$

is then given by $x_k = x_0 R^k$, for $k \geq 0$, and x_0 is normalized by

$$x_0 (I - R)^{-1} \mathbf{e} = \mathbf{1}$$

.

The matrix R has a positive maximal eigenvalue ζ. If the generator A is irreducible, the left eigen vector u^* of R, corresponding to ζ, is determined up to a multiplicative constant and may be chosen to be positive. The matrix R then satisfies

$$sp(R) < 1$$

if and only if

$$\boldsymbol{\pi} A_0 \mathbf{e} < \sum_{k=2}^{\infty} (k-1) \boldsymbol{\pi} A_k \mathbf{e}.$$

where $\boldsymbol{\pi}$ is given by

$$\boldsymbol{\pi} A = 0$$

and

$$\boldsymbol{\pi}\mathbf{e} = 1$$

whenever

$$\zeta = sp(R) < 1$$

the equality

$$A_0\mathbf{e} = \sum_{k=1}^{\infty} R^k \sum_{\nu=k+1}^{\infty} A_\nu \mathbf{e}.$$

Let π denote the steady-state probability vector of the generator $A = A_0 + A_1 + A_2$. Then A can be written as

$$A = \begin{pmatrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{pmatrix}$$

where the matrices C_{00}, C_{01}, C_{10} and C_{11} are as follows:

$$C_{00} = D_0 - (\mu + \theta) I_n$$

$$C_{01} = \beta \otimes D_1 + \mu\beta \otimes I_n$$

$$C_{10} = S^0 \otimes I_n$$

$$C_{11} = (S - \theta I_n) \oplus D$$

We see that A is an irreducible infinitesimal generator matrix and so there exists the stationary vector $\boldsymbol{\pi}$ of A such that

$$\boldsymbol{\pi} A = 0$$

and

$$\boldsymbol{\pi}\mathbf{e} = 1$$

where

$$\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1)$$

The vector $\boldsymbol{\pi}$, partitioned as $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1)$ is computed by solving the equations

$$\boldsymbol{\pi}_0(D_0 - (\mu + \theta) I_n) + \boldsymbol{\pi}_1(S^0 \otimes I_n) = 0$$

$$\boldsymbol{\pi}_0(\beta \otimes D_1 + \mu\beta \otimes I_n) + \boldsymbol{\pi}_1((S - \theta I_n) \oplus D) = 0$$

subject to

$$\boldsymbol{\pi}_0 + \boldsymbol{\pi}_1 = 1$$

The system X^* is stable if and only if

$$\boldsymbol{\pi} A_0 \mathbf{e} < \boldsymbol{\pi} A_2 \mathbf{e}$$

ie.

$$\boldsymbol{\pi}_1(I_m \otimes D_1) < \boldsymbol{\pi}_0(\mu\beta \otimes I_n) + \boldsymbol{\pi}_1(\mu q \otimes I_n)$$

4 Computation of the Steady-State Vector

Let $\mathbf{x}$ be the steady-state probability vector of Q.
Partition this vector as: $\mathbf{x} = (\mathbf{x_0}, \mathbf{x_1}, \mathbf{x_2}, \ldots\ldots)$ where $\mathbf{x}_i = (\mathbf{x}_{i0}, \mathbf{x}_{i1})$

$$\mathbf{x}_{ij} = (\mathbf{x}_{ij1}, \mathbf{x}_{ij2}, \mathbf{x}_{ij3}, \ldots \mathbf{x}_{ijm})$$

for

$$j = 0, 1$$

whereas for

$$s_1 = 1, 2 \ldots m$$

the vectors

$$\mathbf{x}_{ijs_1} = (\mathbf{x}_{ijs_11}, \mathbf{x}_{ijs12} \ldots\ldots\ldots, \mathbf{x}_{ijs_1k})$$

$\mathbf{x}_{ijs_1s_2}$ is the probability of being in state (ijs_1s_2) for $i \geq 0$; $s_1 = 1, 2, \ldots m; j = 0, 1; s_2 = 1, 2, \ldots\ldots n$.

Under the stability condition the steady-state probability vector is obtained as

$$\mathbf{x}_{i+L-1} = \mathbf{x}_{(L-1)} R^i, i \geq 0$$

where R is the minimal non negative solution to the matrix quadratic equation

$$\sum_{k=0}^{M} R^k A_k = 0 \tag{2}$$

and the vectors $\mathbf{x}_0, \ldots.\mathbf{x}_{L-1}$ are obtained by solving

$$\mathbf{x}_{i-1} A_0 + \mathbf{x}_i A_1^1 + \mathbf{x}_{(i+1)} A_2^1 = 0, 1 \leq i \leq L-2$$

$$\mathbf{x}_{L-2} A_0 + \mathbf{x}_{L-1} \left[A_1^1 + A_2 R\right] = 0$$

$$\mathbf{x}_0 A_{10} + \mathbf{x}_1 A_{21}^1 + (\mathbf{x}_3 + \mathbf{x}_4 + \ldots + \mathbf{x}_{L-2}) C + \mathbf{x}_{L-1} C (I - R)^{-1} \mathbf{e} = 0$$

subject to the normalizing condition

$$\sum_{i=0}^{(L-2)} \mathbf{x}_i \mathbf{e} + \mathbf{x}_{(L-1)} (I - R)^{-1} \mathbf{e} = 1.$$

5 Performance Measures of the System

1. Expected Number of customers in the orbit before realization of clock

$$E[N] = \sum_{i=0}^{\infty} i \mathbf{x}(i) \mathbf{e}$$

2. Probability that the server is idle

$$P_0 = \sum_{i=0}^{\infty} \mathbf{x}_i(0)\mathbf{e}$$

3. Probability that the server is busy

$$P_1 = \sum_{i=0}^{\infty} \mathbf{x}_i(1)\mathbf{e}$$

4. The overall rate of retrials at which the orbiting customers request service

$$\mu^* = \sum_{i=0}^{\infty} \mu \sum_{j=0}^{1} \mathbf{x}_{ij}\mathbf{e}$$

5. The successful rate of retrials

$$\mu^{**} = \sum_{i=0}^{\infty} \mu \mathbf{x}_i(0)\mathbf{e} = \mu E[N]$$

6. Probability that an arriving customer will enter into service immediately

$$P_s = \frac{1}{\lambda} \sum_{i=0}^{\infty} \mathbf{x}_i(0) D_1 \mathbf{e}$$

7. Probability that an arriving customer will enter into service with atleast one customer waiting in the orbit

$$P_{sw} = \frac{1}{\lambda} \sum_{i=1}^{\infty} \mathbf{x}_i(0) D_1 \mathbf{e}$$

8. The rate at which a customer abandoned the system

$$\eta = q\mu \sum_{i=L}^{\infty} \mathbf{x}_i(1)\mathbf{e}$$

9. The rate at which the orbiting customers flushed out from the sytem

$$\eta^* = \theta \sum_{i=0}^{\infty} \sum_{j=0}^{1} \mathbf{x}_i(j)\mathbf{e}$$

6 Cost Function

We construct a cost function in L by assigning a fixed cost for search, switching, abandonment and flush out/system clearance. We define the cost function as

$$C(p, L) = E[SR]C_{sr} + E[SW]C_{sw} + E[AB]C_{ab} + E[SC]C_{sc} + E[N]C_h$$

where

- $E[SR]$ = Expected number of searches during a particular period of time
- $E[SW]$ = Expected number of switchings during a particular period of time
- $E[AB]$ = Expected number of customers abandoned from the orbit during a particular period of time
- $E[SC]$ = Expected number of customers flushed out after realization of an exponential clock
- C_{sr} = cost for one search
- C_{sw} = cost for one switching
- C_{ab} = cost for abandonment of one customer from the system
- C_{sc} = cost for flush out off one customer from the system
- C_h = holding cost of one customer in the orbit.

7 Numerical Results

In this section, we present some numerical examples that describe some performance measures of the system under study. In Fig. 1 we plot the probability of abandonment versus mean number of customers in the orbit before realization of exponential clock for the M/M/1 case with specific parameters N = 10, $\lambda = 2, \theta = .2, \nu = 5, \mu = 4$ and p = .1. From Fig. 1, it is clear that $E[N]$ decreases as q increases.

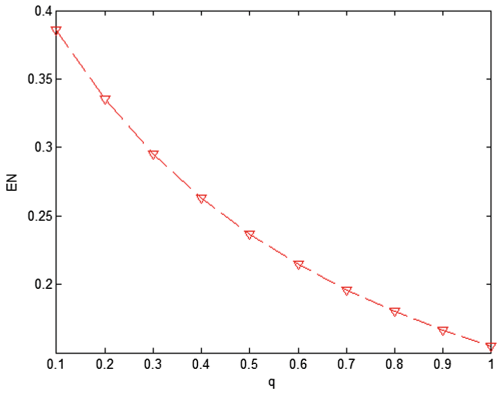


Fig. 1. q vs $E[N]$

By fixing the parameters N = 10, $\lambda = 2, \nu = 5, \mu = 4, p = .1$ and q = .1, Fig. 2 shows that $E[N]$ decreases as θ increases.

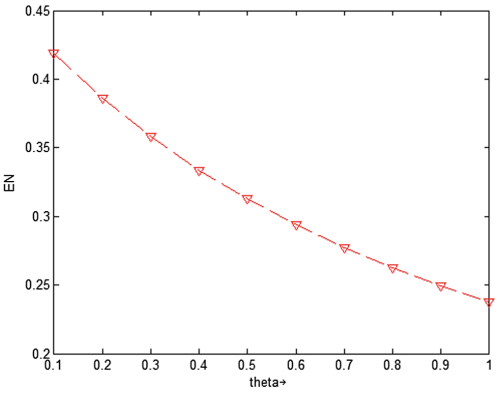


Fig. 2. θ vs $E[N]$

8 Conclusion

In this paper we analyzed a retrial queueing model with search, abandonment and flush out of customers on realization of an exponential clock. The extension of the paper to dependent model with phase type distributed realization clock is possible.

Acknowledgments. The work of the third author is supported by the Maulana Azad National fellowship [$F1-17.1/2015-16/MANF-2015-17-KER-65493$] of University Grants commission, India.

References

1. Artalejo, J.R., Gomez-Corral, A.: Retrial Queueing Systems: A Computational Approach. Springer, Berlin (2008). https://doi.org/10.1007/978-3-540-78725-9
2. Artalejo, J.R.: Accessible bibliography of research on retrial queues. Math. Comput. Model. **30**, 1–6 (1999)
3. Artalejo, J.R.: A classified bibliography of research on retrial queues. Progress in 1990–1999. Top **7**, 187–211 (1999)
4. Artalejo, J.R., Falin, G.I.: Standard and retrial queueing systems: a comparative analysis. Rev. Math. Comput. **15**, 101–129 (2002)
5. Artalejo, J.R., Joshua, V.C., Krishnamoorthy, A.: An M/G/1 retrial queue with orbital search by server. In: Advances in Stochastic Modelling, pp. 41–54. Notable Publications, New Jersey (2002)
6. Artalejo, J.R., Phung-Duc, T.: Single server retrial queues with two way communication. Appl. Math. Model. **37**, 1811–1822 (2013)
7. Chakravarthy, S.R., Krishnamoorthy, A., Joshua, V.C.: Analysis of a multi-server retrial queue with search of customers from the orbit. Perf. Eval. **63**(8), 776–798 (2006)

8. Dudin, A.N., Krishnamoorthy, A., Joshua, V.C., Tsarenkov, G.: Analysis of BMAP/G/1 retrial system with search of customers from the orbit. Euro. J. Oper. Res. **157**, 169–179 (2004)
9. Falin, G.I.: A survey of retrial queues. Queueing Syst. **7**(2), 127–167 (1990)
10. Falin, G.I., Templeton, J.G.C.: Retrial Queues. Chapman and Hall, London (1997)
11. Krishnamoorthy, A., Joshua, V.C.: Excursion between classical and retrial queue
12. Krishnamoorthy, A., Deepak, T.G., Joshua, V.C.: An M/G/1 retrial queue with non persistent customers and orbital search. Stoch. Anal. Appl. **23**, 975–997 (2005)
13. Krishnamoorthy, A., Shajin, D., Lakshmi, B.: GI/M/1 type queueing-inventory systems with postponed work, reservation, cancellation and common life time. Indian J. Pure Appl. Math. **47**(2), 357–388 (2016)
14. Krishnamoorthy, A., Joshua, V.C., Babu, D.: A MAP/PH/1 retrial queue with search and abandonment. In: Accepted for ECQT conference (2018)
15. Latouche, G., Ramaswami, V.: Introduction to Matrix analytic Methods in Stochastic Modelling. American Statistical Association, Siam, Alexandria, Philadelphia (1999)
16. Neuts, M.F., Ramalhoto, M.F.: A service model in which the server is required to search for customers. J. Appl. Probab. **21**, 157–166 (1984)
17. Neuts, M.F.: Probability distributions of phase type. In: Liber Amicorum Professor Emeritus H. Florin, Department of Mathematics, University of Louvain, pp. 173–206 (1975)
18. Neuts, M.F.: Matrix-Geometric Solutions in Stochastic Models - An Algorithmic Approach. The Johns Hopkins University Press, Baltimore and London (1981)
19. Neuts, M.F.: Markov chains with applications in queueing theory which have a matrix- geometric invariant probability vector. Adv. Appl. probab. **10**, 185–212 (1978)
20. Yang, T., Templeton, J.G.: A survey on retrial queues. Queueing Syst. **2**, 201–233 (1987)

Risk Overbounding for a Linear Model

Igor Nikiforov(✉)

Université de Technologie de Troyes, UTT/ICD/LM2S, UMR 6281, CNRS,
12, rue Marie Curie, CS 42060, 10004 Troyes Cedex, France
Igor.Nikiforov@utt.fr
http://www.utt.fr

Abstract. For some safety-critical applications, the risk indicator is represented as a time series of estimation errors in the case of a linear model. The safety of the system is compromised if the probability that this risk indicator leaves a given confidence zone at least once during a certain period becomes too important. Sometimes, we are also interested in the calculation of the instantaneous risk probability. The main difficulty is that the Cumulative Distribution Functions (CDFs) (with infinite support) of measurement noise in the above-mentioned linear model are unknown and only their upper and lower bounds are available. The present paper continues the study of previously developed conservative bounds for the above-mentioned risk probabilities as functions of the bounds for the measurement noise CDFs. The original contribution of the present paper consists in the generalization of the previously obtained results to the case of a linear model.

Keywords: Risk overbounding · Autoregressive process · Linear model

1 Introduction, Motivation and Original Contribution

For some safety-critical applications, it is important to calculate the probability that a discrete time vector autoregressive (AR) process $\{Q_n\}_{n\geq 1}$ leaves the open ball $\|X\|_2 < h$ during a certain period of time T. For example, such AR process can be interpreted as an autocorrelated risk indicator and the ball $\|X\|_2 < h$ as a target zone of the process. The safety of the observed system is compromised if the above-mentioned probability becomes too important. It is assumed that the AR process is represented by the following equation:

$$Q_n = (1-\lambda)Q_{n-1} + \lambda y_n, \quad n = 1, 2, 3, \ldots, \tag{1}$$

where $Q_n \in \mathbb{R}^k$, $0 < \lambda < 1$ is the autoregressive coefficient and the independent random vectors $y_n \in \mathbb{R}^k$ obey a certain distribution F_y, i.e., $y_n \sim F_y$, as well as the initial state Q_0 obeys F_{Q_0}.

The author gratefully acknowledges the partial support of this work from the Thales Alenia Space, France.

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 157–169, 2018.
https://doi.org/10.1007/978-3-319-99447-5_14

The key hypothesis in such safety-critical applications is the assumption that the CDFs F_y of y_n and F_{Q_0} of Q_0 (both with infinite support) are unknown and only their upper and lower bounds are available (sometimes it is called "overbounding").

Let us define the stopping time N:

$$N = \inf\{n \geq 1 : \|Q_n\|_2 \geq h\}. \tag{2}$$

Under assumption that $Q_0 \sim F_{Q_0}$, we are usually interested in the calculation of the following conditional probability:

$$p_r = \mathbb{P}(N \leq T | \|Q_0\|_2 < h). \tag{3}$$

The very first results on conservative bounds for the conditional probability given by (3) as functions of bounds for F_y and F_{Q_0} have been obtained in [1–4] provided that the components of AR process (1) are independent. A conservative bound for the instantaneous risk probability, interpreted as a particular case of (1) and (3), where $\lambda = 1$ and $T = 1$, respectively, has been also proposed in [2–4].

The present paper continues the study of conservative bounds for the above-mentioned probabilities. The original contribution of this paper is a new case where the AR model describes the least squares (LS) estimation errors in a linear regression model.

2 Linear Model and Risk Indicator

Let us consider a linear regression model

$$Z = HX + \Xi, \tag{4}$$

where $Z \in \mathbb{R}^m$ is the vector of observations, $X \in \mathbb{R}^k$ is the vector of unknown (and non-random) regression coefficients, $m > k$, H is a full column rank matrix and $\Xi = (\xi_1, \ldots, \xi_m)^T \in \mathbb{R}^m$ is a random noise with the independent components $\xi_1, \ldots, \xi_m$. If $\mathbb{E}(\Xi) = 0$ and the diagonal variance-covariance matrix $\mathrm{cov}(\Xi) = \Sigma = \mathrm{diag}\left\{\sigma_1^2, \ldots, \sigma_m^2\right\}$ of Ξ is known then the optimal LS estimation of X is given by the following equation

$$\widehat{X} = AZ, \; A = (H^T \Sigma^{-1} H)^{-1} H^T \Sigma^{-1}. \tag{5}$$

Starting from now, the risk indicator is the estimation error $Q = \widehat{X} - X$. It is defined as a function of Ξ by the following equation

$$Q = \widehat{X} - X = A\Xi. \tag{6}$$

If the distribution of Ξ is Gaussian $\mathcal{N}(0, \Sigma)$ and the variance-covariance matrix Σ is known, the estimation $\widehat{X}$ realizes the smallest possible ellipsoid of errors $\widehat{X} - X$. Hence, the risk probability per a given period of time T (defined by (3)) and the instantaneous risk probability $\mathbb{P}(\|Q\|_2 \geq h)$ are minimum.

Unfortunately, the LS estimation given by (5) is optimal only under the above-mentioned conditions. We are interested what happens in practice, if $\mathbb{E}(\Xi) \neq 0$, the variance-covariance matrix Σ is only partially known and, moreover, the CDFs $F_{\xi,i}(x)$ of ξ_i, $i = 1, \ldots, m$, are unknown and only their upper $\overline{F}_{\xi,i}(x)$ and lower $\underline{F}_{\xi,i}(x)$ bounds (overbounds) are available.

Let us assume now that the time series $Z_1, Z_2, \ldots$ of outputs is sequentially generated by regression model (4) with arbitrary vectors of coefficients $X_1, X_2, \ldots$. The estimation $\widehat{X}_n$ of X_n is calculated at each step n. It is also assumed that the autocorrelated noise $\Xi_1, \Xi_2, \ldots$ is defined by the AR model

$$\Xi_n = (1-\lambda)\Xi_{n-1} + \lambda\zeta_n, \quad n = 1, 2, 3, \ldots, \tag{7}$$

where the components of the m dimensional innovation process $\{\zeta_n\}_{n\geq 1}$ are independent. Hence, we get a new AR model for the risk indicator $Q_n = \widehat{X}_n - X_n$ (see (6))

$$Q_n = (1-\lambda)Q_{n-1} + \lambda A\zeta_n, \quad n = 1, 2, 3, \ldots. \tag{8}$$

The question is how to get the conservative bounds for the conditional risk probability p_r defined by (3) if the AR model is given by (8). The difference between the AR models given by (1) and (8) is the presence of the matrix A of size $k \times m$, which induces the dependence between the components of the innovation $A\zeta_n$. A solution to this new problem will be presented in the sequel.

3 Case of AR Model with Independent Components

Let us first consider the AR model defined by (1). The following results have been considered in [1–4] for a special case of independent components of the innovation process $\{y_n\}_{n\geq 1}$. For the sake of simplicity, we consider the case of $k = 2$ in the rest of the paper.

Assumption 1. *Let us assume that the CDF $F_y(X) = F_{y,1}(x_1)F_{y,2}(x_2)$ of the i.i.d. random vectors $\{y_n\}_{n\geq 1}$ and the CDF $F_{Q_0}(X) = F_{Q_{0,1}}(x_1)F_{Q_{0,2}}(x_2)$ of the initial state Q_0 obey the following inequalities $\underline{F}_{y,i}(x) \leq F_{y,i}(x) \leq \overline{F}_{y,i}(x)$ and $\underline{F}_{Q_{0,i}}(x) \leq F_{Q_{0,i}}(x) \leq \overline{F}_{Q_{0,i}}(x)$ for $x \in \mathbb{R}$, where $i = 1, 2$.*

Lemma 1. *Let us consider that Assumption 1 is satisfied. Then the upper bound $\overline{p}_n(U)$ for the probability $p_n(U) = \mathbb{P}(N = n|Q_0 = U)$ is given by*

$$\begin{aligned}\overline{p}_n(U) = &- \int_{-h}^{h} \underline{F}_{y,1}\left(\tfrac{z_1-(1-\lambda)u_1}{\lambda}\right) \mathbb{I}_{\{\overline{I}'(z_1,u_2)\geq 0\}} \overline{I}'(z_1,u_2)dz_1 \\ &- \int_{-h}^{h} \overline{F}_{y,1}\left(\tfrac{z_1-(1-\lambda)u_1}{\lambda}\right) \mathbb{I}_{\{\overline{I}'(z_1,u_2)<0\}} \overline{I}'(z_1,u_2)dz_1,\end{aligned} \tag{9}$$

for $\|U\|_2 \leq h$, *where* $\mathbb{I}_{\{A\}} = \begin{cases} 1 \text{ if } A \text{ is true} \\ 0 \text{ if } A \text{ is false} \end{cases}$ *is the indicator function of* A, $n = 2, 3, \ldots, T$, *and the function* $\overline{I}(z_1, u_2)$ *is defined as follows*

$$\begin{aligned}\overline{I}(z_1,u_2) = {} & \overline{p}_{n-1}\left(z_1, \sqrt{h^2-z_1^2}\right)\overline{F}_{y,2}\left(\frac{\sqrt{h^2-z_1^2}-(1-\lambda)u_2}{\lambda}\right) \\ & -\overline{p}_{n-1}\left(z_1, -\sqrt{h^2-z_1^2}\right)\underline{F}_{y,2}\left(\frac{-\sqrt{h^2-z_1^2}-(1-\lambda)u_2}{\lambda}\right) \\ & -\int_{-\sqrt{h^2-z_1^2}}^{\sqrt{h^2-z_1^2}}\underline{F}_{y,2}\left(\frac{z_2-(1-\lambda)u_2}{\lambda}\right)\mathbb{I}_{\{\overline{p}'_{n-1}(Z)\geq 0\}}\overline{p}'_{n-1}(Z)dz_2 \\ & -\int_{-\sqrt{h^2-z_1^2}}^{\sqrt{h^2-z_1^2}}\overline{F}_{y,2}\left(\frac{z_2-(1-\lambda)u_2}{\lambda}\right)\mathbb{I}_{\{\overline{p}'_{n-1}(Z)<0\}}\overline{p}'_{n-1}(Z)dz_2 \quad (10)\end{aligned}$$

for $-h \leq z_1 \leq h$, $-h \leq u_2 \leq h$, *where* $\overline{I}(-h, u_2) = \overline{I}(h, u_2) = 0$ *and* $\overline{p}'_{n-1}(z_1, z_2) = \dfrac{\partial \overline{p}_{n-1}(z_1, z_2)}{\partial z_2}$. *The upper bound* $\overline{p}_1(U)$ *for the probability* $p_1(U)$ *is given by*

$$\begin{aligned}\overline{p}_1(U) = {} & 1 + \int_{-h}^{h}\overline{F}_{y,1}\left(\frac{z_1-(1-\lambda)u_1}{\lambda}\right)\mathbb{I}_{\{\underline{I}'_1(z_1,u_2)\geq 0\}}\underline{I}'_1(z_1,u_2)dz_1 \\ & + \int_{-h}^{h}\underline{F}_{y,1}\left(\frac{z_1-(1-\lambda)u_1}{\lambda}\right)\mathbb{I}_{\{\underline{I}'_1(z_1,u_2)<0\}}\underline{I}'_1(z_1,u_2)dz_1, \quad (11)\end{aligned}$$

for $\|U\|_2 \leq h$, *where* $\underline{I}'_1(z_1, u_2) = \dfrac{\partial \underline{I}_1(z_1, u_2)}{\partial z_1}$ *and*

$$\underline{I}_1(z_1,u_2) = \max\left\{\underline{F}_{y,2}\left(\frac{\sqrt{h^2-z_1^2}-(1-\lambda)u_2}{\lambda}\right) - \overline{F}_{y,2}\left(\frac{-\sqrt{h^2-z_1^2}-(1-\lambda)u_2}{\lambda}\right), 0\right\}$$

for $-h \leq z_1 \leq h$, $-h \leq u_2 \leq h$.

Proposition 1. *Let us consider that Assumption 1 is satisfied. Then the upper bound* $\overline{p}_r$ *for the conditional probability* p_r *defined in* (3) *is given by*

$$\begin{aligned}p_r \leq \overline{p}_r = \frac{1}{a}\Bigg[& \overline{J}(h)\overline{F}_{Q_{0,1}}(h) - \overline{J}(-h)\underline{F}_{Q_{0,1}}(-h) - \int_{-h}^{h}\underline{F}_{Q_{0,1}}(x_1)\mathbb{I}_{\{\overline{J}'(x_1)\geq 0\}}\overline{J}'(x_1)dx_1 \\ & - \int_{-h}^{h}\overline{F}_{Q_{0,1}}(x_1)\mathbb{I}_{\{\overline{J}'(x_1)<0\}}\overline{J}'(x_1)dx_1\Bigg], \quad (12)\end{aligned}$$

where $\overline{J}'(x_1) = \dfrac{d\overline{J}(x_1)}{dx_1}$,

$$\begin{aligned}\overline{J}(x_1) = {} & \overline{p}_T\left(x_1, \sqrt{h^2 - x_1^2}\right) \overline{F}_{Q_{0,2}}\left(\sqrt{h^2 - x_1^2}\right) \\ & - \overline{p}_T\left(x_1, -\sqrt{h^2 - x_1^2}\right) \underline{F}_{Q_{0,2}}\left(-\sqrt{h^2 - x_1^2}\right) \\ & - \int_{-\sqrt{h^2-x_1^2}}^{\sqrt{h^2-x_1^2}} \underline{F}_{Q_{0,2}}(x_2) \mathbb{I}_{\{\overline{p}'_T(x_1,x_2) \geq 0\}} \overline{p}'_T(x_1, x_2) dx_2 \\ & - \int_{-\sqrt{h^2-x_1^2}}^{\sqrt{h^2-x_1^2}} \overline{F}_{Q_{0,2}}(x_2) \mathbb{I}_{\{\overline{p}'_T(x_1,x_2) < 0\}} \overline{p}'_T(x_1, x_2) dx_2 \end{aligned} \tag{13}$$

for $-h \leq x_1 \leq h$, *where* $\overline{p}_T(X) = \sum_{n=1}^{T} \overline{p}_n(X)$, $\overline{p}'_T(x_1, x_2) = \dfrac{\partial \overline{p}_T(x_1, x_2)}{\partial x_2}$. *The constant* a *is given by*

$$a = -\int_{-h}^{h} \overline{F}_{Q_{0,1}}(x_1) \mathbb{I}_{\{\underline{K}'(x_1) \geq 0\}} \underline{K}'(x_1) dx_1 - \int_{-h}^{h} \underline{F}_{Q_{0,1}}(x_1) \mathbb{I}_{\{\underline{K}'(x_1) < 0\}} \underline{K}'(x_1) dx_1, \tag{14}$$

where $\underline{K}'(x_1) = \dfrac{d\underline{K}(x_1)}{dx_1}$ *and*

$$\underline{K}(x_1) = \max\left\{\underline{F}_{Q_{0,2}}\left(\sqrt{h^2 - x_1^2}\right) - \overline{F}_{Q_{0,2}}\left(-\sqrt{h^2 - x_1^2}\right), 0\right\}.$$

4 AR Model Which Describes the LS Estimation Errors

Let us consider the risk probability p_r per a given period of time T defined in (3) for the AR model describing the LS estimation errors. The instantaneous risk probability is considered as a particular case of (3) and (8), where $T = 1$ and $\lambda = 1$, respectively. As it follows from Assumption 1, the conservative bound given by Lemma 1 and Proposition 1 is applicable if the joint CDF of the innovation $\{y_n\}_{n \geq 1}$ (the initial state Q_0, respectively) is equal to the product of marginal CDFs $F_y(X) = F_{y,1}(x_1)F_{y,2}(x_2)$ ($F_{Q_0}(X) = F_{Q_{0,1}}(x_1)F_{Q_{0,2}}(x_2)$, respectively).

The goal is to find the conservative (upper) bound for the probability $\mathbb{P}(\|Q\|_2 \geq h)$ by using the paired Gaussian CDF overbounding (see [5,6]) for the components $\xi_1, \ldots, \xi_m$ of the regression noise Ξ. Let us recall equation (6) : $Q = (\widehat{x} - x, \widehat{y} - y)^T$, where $\widehat{x} - x = a_1^T \Xi$, $\widehat{y} - y = a_2^T \Xi$, a_1 and a_2 are the first and second rows of the matrix A defined in (5).

Let us define the unit vector $\delta = (\cos\theta, \sin\theta)^T$, $\theta \in [0, 2\pi]$ in the orthogonal coordinate system $(\widehat{x} - x, \widehat{y} - y)$. The projection of the error vector Q on the direction δ is given by the inner product $Q^T\delta$. Hence,

$$\begin{aligned}\mathbb{P}(\|Q\|_2 \geq h) &= 1 - \mathbb{P}(\|Q\|_2 \leq h) = 1 - \mathbb{P}\left(\left|Q^T\delta(\theta)\right| \leq h \ \forall\theta \in [0, 2\pi]\right) \\ &= 1 - \mathbb{P}\left(\left|\sum_{i=1}^{m} (a_{1,i}\cos\theta + a_{2,i}\sin\theta)\, \xi_i\right| \leq h \ \forall\theta \in [0, 2\pi]\right). \end{aligned} \tag{15}$$

In the case of paired Gaussian symmetric CDF overbounding for the components $\xi_1, \ldots, \xi_m$ of the regression noise Ξ

$$\underline{F}_{\xi,i}(x) = \mathcal{N}(\bar{b}_i, \sigma_i^2) \leq F_{\xi,i}(x) \leq \overline{F}_{\xi,i}(x) = \mathcal{N}(-\bar{b}_i, \sigma_i^2), \quad i = 1, \ldots, m, \quad \forall x \in \mathbb{R}, \tag{16}$$

the overbounding of the term $\tau = Q^T \delta(\theta)$ is defined by the mean and variance as functions of $\theta \in [0, 2\pi]$:

$$\underline{F}_{\tau}(x) = \mathcal{N}(m(\theta), \sigma^2(\theta)) \leq F_{\tau}(x) \leq \overline{F}_{\tau}(x) = \mathcal{N}(-m(\theta), \sigma^2(\theta)), \ \forall \theta \in [0, 2\pi], \tag{17}$$

where

$$m(\theta) = \sum_{i=1}^{m} |a_{1,i} \cos\theta + a_{2,i} \sin\theta| \, \bar{b}_i$$

and

$$\sigma^2(\theta) = \sum_{i=1}^{m} \sigma_i^2 (a_{1,i} \cos\theta + a_{2,i} \sin\theta)^2.$$

Let us now define another random vector $\widetilde{Q} = (\zeta, \eta)$, where ζ and η are independent random variables distributed following $F_{\zeta}(x)$ and $F_{\eta}(x)$. This new vector $\widetilde{Q}$ substitutes the vector Q for the calculation of the conservative bound for the risk probabilities (instantaneous and per a given period of time).

Let us re-write Eq. (15) for the random vector $\widetilde{Q}$ with independent components:

$$\begin{aligned} \mathbb{P}\left(\|\widetilde{Q}\|_2 \geq h\right) &= 1 - \mathbb{P}\left(\left|\widetilde{Q}^T \delta(\theta)\right| \leq h \ \forall \theta \in [0, 2\pi]\right) \\ &= 1 - \mathbb{P}\left(|\zeta \cos\theta + \eta \sin\theta| \leq h \ \forall \theta \in [0, 2\pi]\right). \end{aligned} \tag{18}$$

To simplify the notation and without loss of generality, let us assume that the random variables ζ and η are overbounded by using the same Gaussian bounds (a more general case of different Gaussian bounds can be also easily considered)

$$\underline{F}_{\zeta}(x) = \mathcal{N}(m_{\zeta,\eta}, \sigma_{\zeta,\eta}^2) \leq F_{\zeta}(x) \leq \overline{F}_{\zeta}(x) = \mathcal{N}(-m_{\zeta,\eta}, \sigma_{\zeta,\eta}^2) \tag{19}$$

and

$$\underline{F}_{\eta}(x) = \mathcal{N}(m_{\zeta,\eta}, \sigma_{\zeta,\eta}^2) \leq F_{\eta}(x) \leq \overline{F}_{\eta}(x) = \mathcal{N}(-m_{\zeta,\eta}, \sigma_{\zeta,\eta}^2). \tag{20}$$

The overbounding of the term $\widetilde{\tau} = \widetilde{Q}^T \delta(\theta)$ is defined by the mean and variance as functions of $\theta \in [0, 2\pi]$:

$$\underline{F}_{\widetilde{\tau}}(x) = \mathcal{N}(\widetilde{m}(\theta), \sigma_{\zeta,\eta}^2) \leq F_{\tau}(x) \leq \overline{F}_{\tau}(x) = \mathcal{N}(-\widetilde{m}(\theta), \sigma_{\zeta,\eta}^2), \quad \forall \theta \in [0, 2\pi], \tag{21}$$

where $\widetilde{m}(\theta) = (|\cos\theta| + |\sin\theta|) \, m_{\zeta,\eta}$.

Let us define the maximum variance of the random variable τ which coincides with the maximum eigenvalue $\varrho_{\max}$ of the variance-covariance matrix $\Sigma = \mathrm{cov}(Q)$ of the estimation error Q:

$$\varrho_{\max} = \max_{\theta \in [0, 2\pi]} \sigma^2(\theta) = \max_{\theta \in [0, 2\pi]} \sum_{i=1}^{m} \sigma_i^2 (a_{1,i} \cos\theta + a_{2,i} \sin\theta)^2. \tag{22}$$

We are looking for an upper bound for the probability $\mathbb{P}(\|Q\|_2 \geq h)$. This is equivalent to find a lower bound for $\mathbb{P}\left(\left|Q^T\delta(\theta)\right| \leq h\ \forall\theta \in [0,2\pi]\right)$. As it follows from the properties of the Gaussian distribution, to find a lower bound for $\mathbb{P}\left(\left|Q^T\delta(\theta)\right| \leq h\ \forall\theta \in [0,2\pi]\right)$ we need to choose the parameters of the substitute vector $\widetilde{Q}$ overbounding defined in (19)–(20) such that $\sigma^2_{\zeta,\eta} = \varrho_{\max}$ and the value of $m_{\zeta,\eta}$ as follows

$$m_{\zeta,\eta} = \max_{\theta\in[0,2\pi]} \left\{ \frac{\sum_{i=1}^{m} |a_{1,i}\cos\theta + a_{2,i}\sin\theta)|\,\overline{b}_i}{\sigma(\theta)\,(|\cos\theta| + |\sin\theta|)} \right\} \varrho_{\max}. \tag{23}$$

Instantaneous Risk Probability. It follows from (11) (with $\lambda = 1$) that the instantaneous risk probability is upper bounded in the following manner:

$$\begin{aligned}\mathbb{P}(\|Q\|_2 \geq h) \leq \overline{p}_1 = 1 &+ \int_{-h}^{h} \overline{F}_{Q_1}(z_1)\mathbb{I}_{\{\underline{I}'_1(z_1)\geq 0\}}\underline{I}'_1(z_1)dz_1 \\ &+ \int_{-h}^{h} \underline{F}_{Q_1}(z_1)\mathbb{I}_{\{\underline{I}'_1(z_1)<0\}}\underline{I}'_1(z_1)dz_1,\end{aligned} \tag{24}$$

where $\underline{I}_1(z_1) = \max\left\{\underline{F}_{Q_2}\left(\sqrt{h^2-z_1^2}\right) - \overline{F}_{Q_2}\left(-\sqrt{h^2-z_1^2}\right), 0\right\}$ for $-h \leq z_1 \leq h$, $\underline{I}'_1(z_1) = \frac{d\underline{I}_1(z_1)}{dz_1}$, $\underline{F}_{Q_1}(x) = \underline{F}_{Q_2}(x) = \mathcal{N}(m_{\zeta,\eta}, \sigma^2_{\zeta,\eta})$ and $\overline{F}_{Q_1}(x) = \overline{F}_{Q_2}(x) = \mathcal{N}(-m_{\zeta,\eta}, \sigma^2_{\zeta,\eta})$.

Risk Probability Per a Given Period of Time. To calculate the conservative bound for the risk probability p_r per a given period of time (see Proposition 1), it is necessary to define the parameters of the Gaussian marginal bounds for Q_0 from (19)–(23) and the bounds $\underline{F}_{y,i}(x)$ and $\overline{F}_{y,i}(x)$ for the i.i.d. random vectors $\{y_n\}_{n\geq 1}$:

$$\underline{F}_{y,1}(x) = \underline{F}_{y,2}(x) = \mathcal{N}\left(\mu_y, \sigma_y^2\right) \leq F_y(x) \leq \overline{F}_{y,1}(x) = \overline{F}_{y,1}(x) = \mathcal{N}\left(-\mu_y, \sigma_y^2\right),$$

where $\mu_y = m_{\zeta,\eta}$ and $\sigma^2_{y,i} = \dfrac{1-(1-\lambda)^2}{\lambda^2}\varrho_{\max}$.

5 Numerical Examples

Instantaneous Risk Probability. Let us consider the linear regression model (4)–(5). The matrix A is given as follows

$$A = \begin{pmatrix} -0.124 & -0.201 & 0.243 & -0.190 & -0.061 & 0.320 & 0.036 & 0.112 & -0.135 \\ -0.080 & 0.164 & 0.139 & -0.258 & 0.161 & -0.344 & 0.531 & -0.025 & -0.289 \end{pmatrix}. \tag{25}$$

The paired Gaussian symmetric CDF overbounding for the components $\xi_1, \ldots, \xi_m$ of the regression noise Ξ is given by

$$\underline{F}_{\xi,i}(x) = \mathcal{N}(\overline{b}_i, \sigma_i^2) \leq F_{\xi,i}(x) \leq \overline{F}_{\xi,i}(x) = \mathcal{N}(-\overline{b}_i, \sigma_i^2),\ i = 1, \ldots, m,\ \forall x \in \mathbb{R}, \tag{26}$$

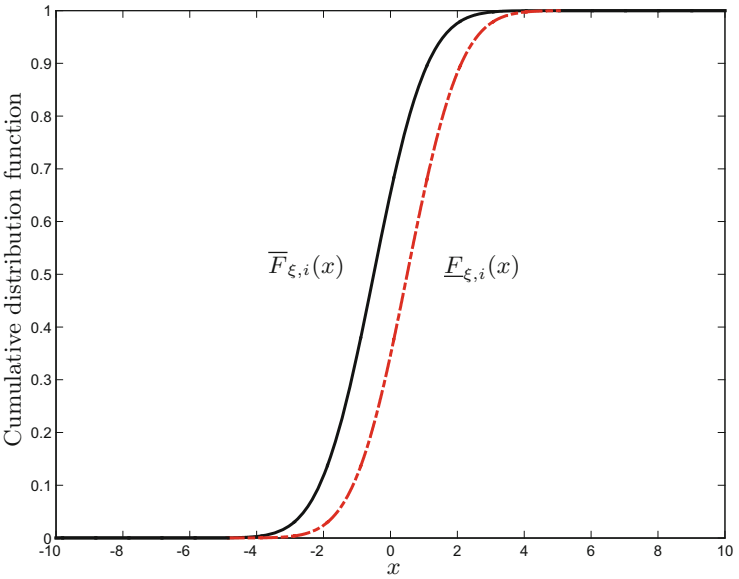


Fig. 1. The paired Gaussian symmetric CDF overbounding for $\xi_1, \ldots, \xi_m$.

where $m = 9$, $\overline{b}_i = 0.477$ and $\sigma_i^2 = 1.6$. The upper and lower bounds for the cumulative distribution function of the components $\xi_1, \ldots, \xi_m$ are shown in Fig. 1. The boundary h of the stopping time (2) is chosen such that $h = 4$.

Let us first compute the expectation and the variance of the substitute vector: $m_{\zeta,\eta} = 1$, $\sigma^2_{\zeta,\eta} = \varrho_{\max} = 1$. Now we can compute the expectations

$$\widetilde{m}(\theta) = (|\cos\theta| + |\sin\theta|)\, m_{\zeta,\eta}$$

and

$$\overline{m}(\theta) = \frac{\varrho_{\max} m(\theta)}{\sigma(\theta)} = \frac{\varrho_{\max} \sum_{i=1}^{m} |a_{1,i}\cos\theta + a_{2,i}\sin\theta)|\, \overline{b}_i}{\sigma(\theta)}$$

as functions of $\theta \in [0, 2\pi]$. These functions are shown in Fig. 2. The expectation $\widetilde{m}(\theta)$ is shown in solid line and the expectation $\overline{m}(\theta)$ in dotted line. To compute the upper bound for the instantaneous risk given by (24), we have to compute the lower bound $\underline{I}_1(z_1)$ as a function of $z_1 \in [-h; h]$ This lower bound is shown in Fig. 3. Finally, we have to compute the derivative $\underline{I}'_1(z_1) = \frac{d\underline{I}_1(z_1)}{dz_1}$ of the lower bound. It is shown in Fig. 4 as a function of $z_1 \in [-h; h]$.

Let us compare this upper bound (overbound) $\overline{p}_1$ with the risk probabilities computed for two particular models of the components $\xi_1, \ldots, \xi_m$.

The first model represents the worst case Gaussian distribution (see details in [2]). First, we replace the covariance matrix $\Sigma = \operatorname{cov} Q$ by $\overline{\Sigma} = \operatorname{diag}\{\varrho_{\max}, \varrho_{\max}\}$, where $\varrho_{\max}$ is the maximum eigenvalue of Σ. Next, we define the following hyperrectangle $\mathbb{B} = \left\{X \in \mathbb{R}^m | x_i \in [-\overline{b}_i, \overline{b}_i],\ i = 1, \ldots, m\right\}$ and a linear mapping (defined by the matrix A) of the set $\mathbb{B}$ onto the set $\mathbb{P}$. The

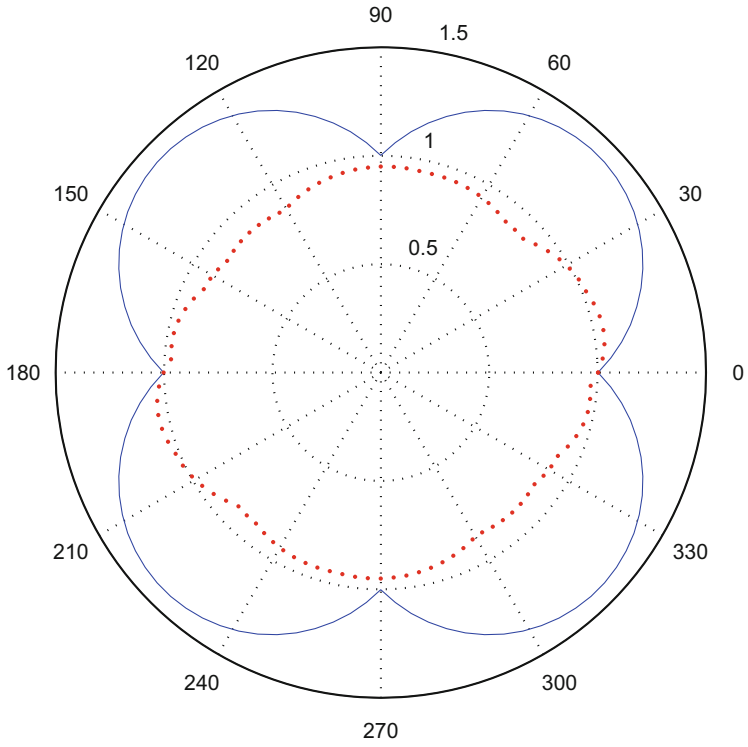


Fig. 2. The expectations $\widetilde{m}(\theta)$ and $\overline{m}(\theta)$ as functions of $\theta \in [0, 2\pi]$.

set $\mathbb{P}$ is a convex polygon. Finally, we compute the worst case mean vector $\overline{B} = AB_j$, $j = \arg \max_{i=1,\ldots,2^m} \{\|AB_i\|_2\}$, where B_i is a vertex of the hyperrectangle $\mathbb{B}$, $i = 1, \ldots, 2^m$, and compute an upper bound for the instantaneous risk probability $\mathbb{P}(\|Q\|_2 \geq h)$ by numerical method (see details in [2]).

The second model represents the following non-Gaussian distribution

$$F_{\xi,i}(x) = \begin{cases} \overline{F}_{\xi,i}(x) \text{ if } & x \leq -\overline{b}_i \\ 1/2 \quad \text{if } -\overline{b}_i < & x < \overline{b}_i \\ \underline{F}_{\xi,i}(x) \text{ if } & x \geq \overline{b}_i \end{cases} . \tag{27}$$

This CDF is shown in Fig. 5. The idea of such a CDF is to redistribute the "probabilistic mass" from the central part to the periphery, close to the bounds $\underline{F}_{\xi_i}(x)$ and $\overline{F}_{\xi,i}(x)$. The instantaneous risk probability $\mathbb{P}(\|Q\|_2 \geq h)$ is estimated by a 10^5-repetition Monte Carlo simulation.

The conservative bound $\overline{p}_1$ and the risk probabilities for the two above-mentioned models are given in the following table:

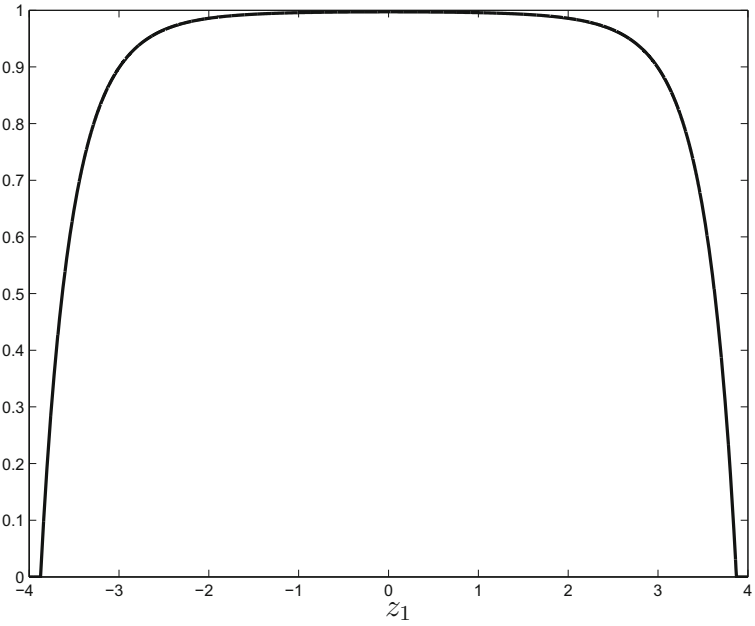


Fig. 3. The lower bound $\underline{I}_1(z_1)$ as a function of $z_1 \in [-h; h]$.

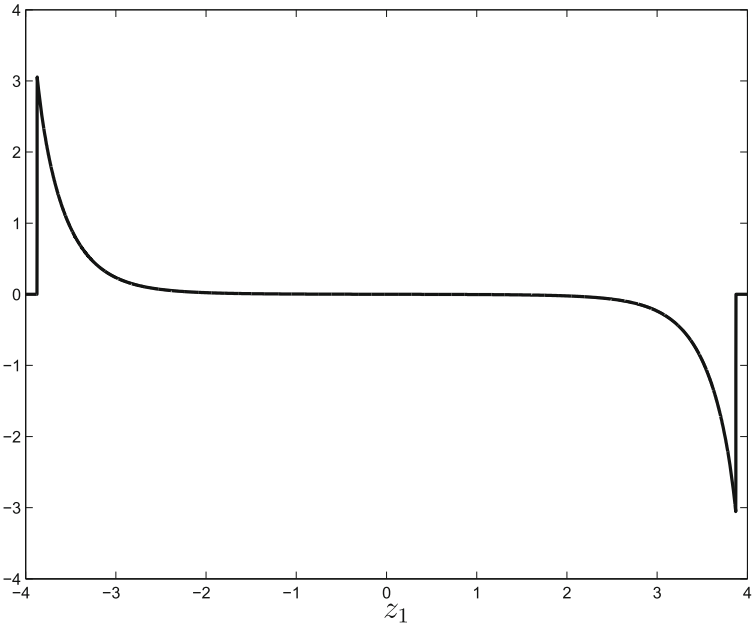


Fig. 4. The derivative of $\underline{I}_1(z_1)$ as a function of $z_1 \in [-h; h]$.

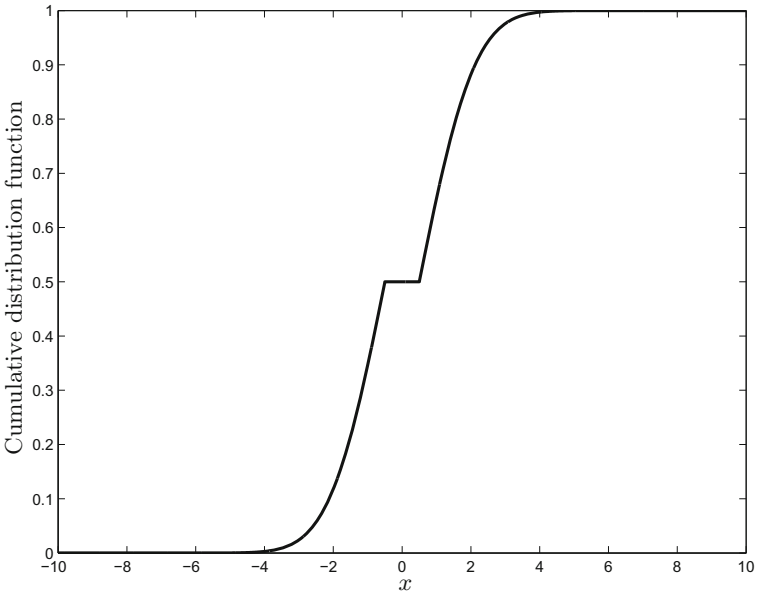


Fig. 5. The non-Gaussian distribution.

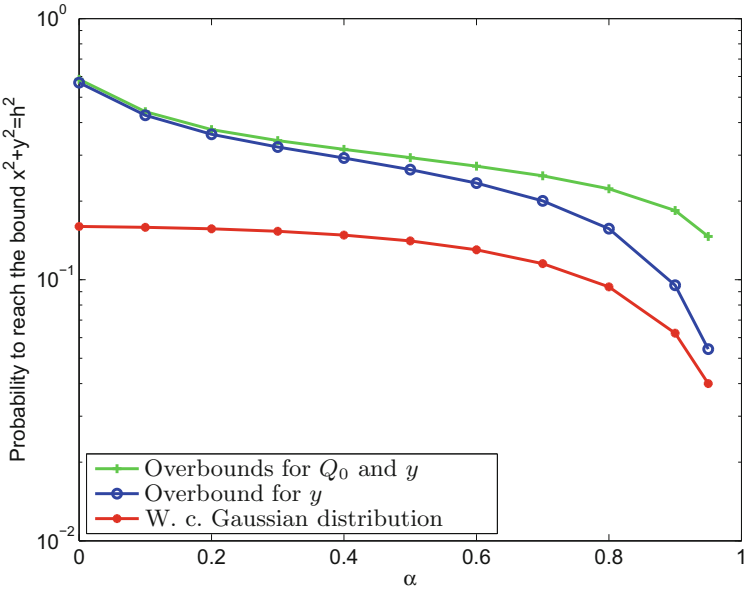


Fig. 6. The risk probability per a given period of time and its conservative bounds.

Overbound $\overline{p}_1$	First model	Second model (27)
$2.943 \cdot 10^{-2}$	$2.624 \cdot 10^{-3}$	$1.750 \cdot 10^{-3}$
		$[1.500 \cdot 10^{-3}, 2.029 \cdot 10^{-3}]$

The 95% confidence interval for the second model risk probability is shown in the third column of the table.

Risk Probability Per a Given Period of Time. Let us consider again the linear regression model (4)–(5). The matrix A is given by (25) and the paired Gaussian symmetric CDF overbounding for the components $\xi_1, \ldots, \xi_m$ of the regression noise Ξ is defined by (26). The previously defined vector $\widetilde{Q}$ which substitutes the vector Q for the computation of the conservative bound for the risk probabilities. Our goal is to compute the conservative bound $\overline{p}_r$ for the risk probability p_r per a given period of time $T = 20$ (see Proposition 1) based on the properties of the substitute vector $\widetilde{Q}$.

This example is devoted to the comparison of the conservative bounds $\overline{p}_r$ with the risk probability p_r per a given period of time induced by the Gaussian AR(1) process with the worst case mean. The risk probability for the AR model and its conservative overbounds as functions of the AR-coefficient $\alpha = 1 - \lambda$ are presented in Fig. 6. Two conservative bounds are calculated by using Proposition 1. The first conservative bound is calculated by using the paired Gaussian CDF overbounding for the i.i.d. random vectors $\{y_n\}_{n\geq 1}$ and the initial condition Q_0. The second one is calculated by using the paired Gaussian CDF overbounding only for the i.i.d. random vectors $\{y_n\}_{n\geq 1}$. The initial condition Q_0 follows the worst case Gaussian distribution $\underline{F}_{Q_0,i}(x)$. It is assumed that the substitute AR(1) vector process $\left\{\widetilde{Q}_n\right\}_{n\geq 1}$ with the CDF $F_y(X) = F_{y,1}(x_1)F_{y,2}(x_2)$ of $\{y_n\}_{n\geq 1}$ and the CDF $F_{Q_0}(X) = F_{Q_0,1}(x_1)F_{Q_0,2}(x_2)$ of Q_0 overbounded in the following manner:

$$\begin{aligned}
\underline{F}_{y,i}(x) = \mathcal{N}\left(m_{\zeta,\eta}, \sigma_y^2\right) \leq F_{y,i}(x) \leq \overline{F}_{y,i}(x) = \mathcal{N}\left(-m_{\zeta,\eta}, \sigma_y^2\right), \\
\underline{F}_{Q_0,i}(x) = \mathcal{N}(m_{\zeta,\eta}, \sigma_{Q_0}^2) \leq F_{Q_0,i}(x) \leq \overline{F}_{Q_0,i}(x) = \mathcal{N}(-m_{\zeta,\eta}, \sigma_{Q_0}^2). \quad (28)
\end{aligned}$$

where $i = 1, 2$, $\sigma_{Q_0}^2 = \sigma_{\zeta,\eta}^2 = 1$, $m_{\zeta,\eta} = 1$, $\sigma_y^2 = \frac{1-(1-\lambda)^2}{\lambda^2}\sigma_{Q_0}^2$ and $\alpha = 1 - \lambda \in [0, 0.95]$.

6 Conclusion

The present paper considers the calculation of conservative bounds for the risk probability per a given period of time and for the instantaneous risk probability in the case where the risk indicator is defined by the estimation errors in a linear model. The original contribution of the paper consists in the generalization of the previously obtained results to the case of a linear model. In this linear model, the measurement noise CDFs are unknown and only upper and lower

bounds for these CDFs are available. The new twist consists in the definition of a specially chosen substitute random vector which allows us to adapt the previously obtained conservative bounds to the new case of a linear model. Finally, the risk probabilities are defined as functions of the upper and lower bounds for the measurement noise CDFs. The risk probabilities are defined as functions of the upper and lower bounds for the measurement noise CDFs.

References

1. Nikiforov, I.: Bounding the risk probability. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds.) DCCN 2017. CCIS, vol. 700, pp. 135–145. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66836-9_12
2. Nikiforov, I.: From the pseudo-range overbounding to the integrity risk overbounding. In: Proceedings of the International Technical Symposium on Navigation and Timing, Toulouse, France, pp. 1–12 (2017)
3. Nikiforov, I.: Pseudo-range errors and the integrity risk overbounding. In: Proceedings of 6th International Colloquium Scientific and Fundamental Aspects of the Galileo Programme, Technical University of Valencia, Valencia, Spain, pp. 1–8 (2017)
4. Nikiforov, I.: From the pseudo-range overbounding to the integrity risk overbounding. J. Inst. Navig. (2018, submitted)
5. Rife, J., Pullen, S., Pervan, B., Enge, P.: Paired overbounding and application to GPS augmentation. In: Proceedings of The Position Location and Navigation Symposium, PLANS, USA, pp. 439–446 (2004)
6. Rife, J., Pullen, S., Enge, P., Pervan, B.: Paired overbounding for nonideal LAAS and WAAS error distributions. IEEE Trans. Aerosp. Electron. Syst. **42**(4), 1386–1395 (2006)

Analysis of Resource Sharing Between MBB and MTC Sessions with Data Aggregation Using Matrix-Analytic Methods and Simulation

Natalia Yarkina[1(✉)], Konstantin Samouylov[1], and Vladimir Vishnevskiy[2]

[1] Department of Applied Probability and Informatics,
RUDN University, Moscow, Russia
{natyarkina,ksam}@sci.pfu.edu.ru
[2] V.A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences,
Moscow, Russia
vishn@inbox.ru

Abstract. Mobile broadband (MBB) and ubiquitous machine type communications (MTC) appear essential in future 5G networks, however their requirements in terms of data rate, expected number of users, latency and reliability are largely contradictory, which constitutes a major challenge for 5G architecture design. MTC data aggregation via device clustering may prevent signaling congestion due to a massive number of MTC devices in the cell. Resource sharing between MBB and MTC sessions with MTC data aggregation can be modeled using a multi-service loss system, however its numerical analysis is impeded due to the considerable difference in scales of MBB and MTC parameters. In the paper, we use simulation to assess the performance measures of the loss system and its more realistic modification. Cases of light, moderate and heavy MTC loading and various parameters of the MMPP of MTC arrivals are considered.

Keywords: Queueing system · MMPP · Simulation · OMNeT++
5G · Internet of Things (IoT) · Machine-type communications (MTC)
Device clustering · Device grouping · Data aggregation

1 Introduction

Fifth generation mobile networks, 5G, aim at increasing further the broadband capabilities already present in existing systems but also at providing extensive support for ubiquitous machine-to-machine (M2M) or machine-type communications (MTC) for both consumer and business needs, thus enabling the full-fledged

N. Yarkina—This work was supported by Russian Science Foundation and the Department of Science and Technology (India) via grant # 16-49-02021 and INT/RUS/RFS/16 for the joint research project at V.A. Trapeznikov Institute of Control Sciences and CMS College Kottayam.

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 170–183, 2018.
https://doi.org/10.1007/978-3-319-99447-5_15

Internet of Things (IoT) [1]. For this reason, the following three types of communication are expected to be predominant in 5G networks: extreme or enhanced mobile broadband (eMBB), massive machine type communication (mMTC or massive IoT) and critical machine type communication (cMTC or critical IoT) [2,3]. eMBB, mMTC and cMTC are also referred to as the generic 5G services, while a 5G system can be defined as "one common network that can provide all generic 5G services and is flexible enough to change the service mix dynamically" [3, p. 34].

eMBB is characterised by extremely high data rates (from at least 50–100 Mbps to several Gbps per user), enhanced coverage, but also low latency to support interactive applications. Massive MTC is intended to provide delay-tolerant connectivity to a very large number of relatively simple, energy-efficient devices transmitting small payloads. Moreover, due to the stringent requirements for low device complexity, the recommended peak data rate for mMTC is limited to about 2 Mbps [4]. Finally, cMTC is destined for ultra-reliable low-latency MTC connectivity, e.g. for such applications as autonomous vehicle or remote health care.

The requirements of eMBB, mMTC and cMTC in terms of data rate, expected number of users, latency and reliability are largely contradictory, which constitutes a major challenge for 5G architecture design [5]. In particular, a serious difficulty when combining MTC and MBB traffic in LTE networks consists in random access channel congestion and overloading due to the massive number of MTC devices in the cell [6–8]. One of the possible solutions to the access congestion is MTC data aggregation via device grouping or clustering, where a few network nodes – aggregators or cluster heads – relay data from numerous MTC devices to the base station [7,9,10].

A Markov model of a network cell providing radio connectivity for a combination of broadband and MTC services with MTC device grouping was proposed in [11] and extended for MAP arrivals in [12], where explicit expressions for important performance measures of the system were obtained using matrix-analytic methods along with an efficient computational algorithm. However, considerable discrepancy in traffic characteristics between MBB and MTC impedes the direct application of the algorithms in certain ranges of parameter values thus hindering numerical analysis. In the present paper, we build on the work presented in [12] and further explore the service system by means of simulation. Furthermore, the system is simulated in two variants: as it was devised in [12], but also with certain simplifying assumptions discarded. We compare the numerical results for the two systems under light, moderate and heavy MTC loading and examine both performance measures' behavior and simulation issues.

The remainder of the paper is structured as follows. In Sect. 2, we briefly outline the service system under study. Section 3 presents the simulation model developed for the analysis. In Sect. 4, selected numerical results are presented and discussed. Finally, Sect. 5 provides some concluding remarks.

2 Service System Description

The network cell is modeled as a multi-service, multi-server loss system with two classes of jobs: streaming jobs corresponding to MBB sessions and elastic jobs representing MTC transmissions. When dealing with wireless data transmission, we work on a number of simplifying assumptions as in [11,13,14], in particular, we do not take into account interference between devices. Interarrival and service times of streaming jobs are exponentially distributed and each streaming job occupies exactly d_S servers during its service time. Elastic jobs arrive according to a Markovian arrival process (MAP) [15–17], have exponentially distributed lengths (service time when served by exactly one server) and are served according to the egalitarian processor sharing (EPS) discipline. In order to reflect MTC device clustering, resources to elastic jobs are allocated in batches of c servers in such a way that each batch can simultaneously serve no more than M jobs. Also, we assume that at each elastic arrival and departure, servers' capacity is reallocated in such a way that all l elastic jobs in service equally share $c(l) = c \cdot \lceil l/M \rceil = c \cdot \min\{y \in \mathbb{N} : y \geq l/M\}$ servers. The system consists of C servers corresponding to the resource units (RU) of the cell, and R of C servers are available to streaming jobs only. There is no queue, so a job that does not find enough free resources upon its arrival is lost (blocked).

Table 1. Model notation.

Notation	Definition
C	Number of servers in the system (cell resource units)
$R < C$	Number of servers available to streaming jobs only
$C_E = C - R$	Number of servers available to both job classes
$c \leq C$	Cluster (batch) size
M	Cluster capacity
d_S	Number of servers taken by one streaming job
K	Number of transient states of the MAP Markov chain
$\mathbf{Q}_0 = (q_{ij}^{(0)})$	Matrix governing MAP MC transitions with no arrivals
$\mathbf{Q}_1 = (q_{ij}^{(1)})$	Matrix governing MAP MC transitions with arrivals
λ	Mean arrival rate of elastic jobs
μ^{-1}	Mean elastic job length (service time on one server)
α	Mean arrival rate of streaming jobs
β^{-1}	Mean service time of streaming jobs
B_E	Blocking probability of elastic jobs
B_S	Blocking probability of streaming jobs
T_E	Mean service time of elastic jobs
U	System utilization
c_E^{avg}	Mean number of servers occupied by elastic jobs
N_E	Mean number of elastic jobs in service

Model parameters are listed in Table 1. We refer the reader to [12] for the detailed description of the system under analysis and for the derivation of the stationary probability distribution of the corresponding Markov process.

3 Simulation Model

We simulate the service system under study using OMNeT++ simulation framework [18]. The resulting model is depicted in Fig. 1. It largely relies on modules from OMNeT++ queueing library org.omnetpp.queueing (modules *Source*, *Sink*, *Allocate*, *Deallocate* and *Delay*), however three additional modules have been developed in C++ in order to implement EPS service discipline (modules *ElasticDelay* and *ElasticResourcePool*) and MAP (module *MAPSource*). It is worth noting that an OMNeT++ module implementing arrivals generation according to MAP along with other stochastic processes was also proposed in [19], where the problem of simulating correlated arrival streams was addressed in a broader context. The algorithm for modeling MAP arrivals is rather straightforward and is based upon simulating transitions of the underlying continuous-time Markov chain (MC) using matrices $\mathbf{Q}_0$ and $\mathbf{Q}_1$ (for details, see e.g. [17]).

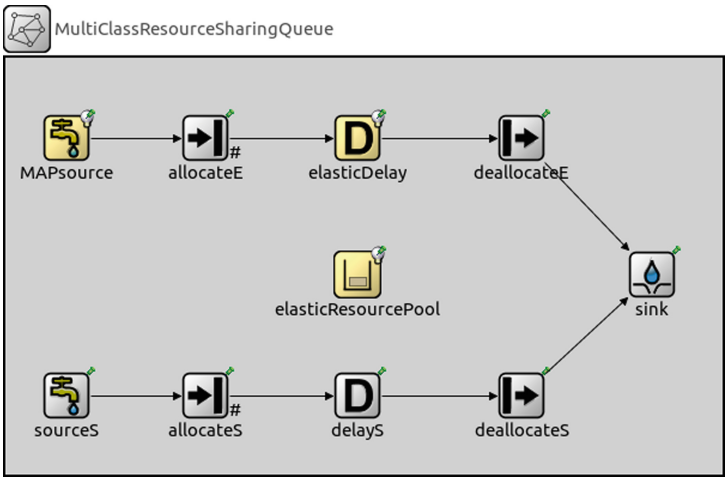


Fig. 1. System model in OMNeT++.

The simulation model can operate in two modes. The first mode (labeled *Resh*) simulates the service discipline for elastic jobs as described in Sect. 2. The second mode (labeled *NonResh*) corresponds to a more realistic service discipline for MTC sessions, where the service rate of elastic jobs in a cluster depends only on the number of jobs in this cluster. Moreover, jobs in clusters are not "reshuffled" upon each elastic departure: if a job departs from a cluster the remaining jobs in this cluster keep being served in it and share equally its

capacity; cluster resources are vacated only when all jobs leave the cluster. If an arriving elastic job finds a cluster with less than M jobs in service it is accepted to this cluster. If all active clusters are fully busy (serve M jobs each), resources for additional cluster are requested.

The modules of the simulation model depicted in Fig. 1 are C++ objects and operate as follows:

- *MAPSource* generates elastic jobs according to a MAP with any given parameters values and sends them out to module *allocateE*.
- *allocateE* requests resources for the job from module *elasticResourcePool* and if the resources are allocated successfully sends the job further to *elasticDelay*. If resources are denied the job is dropped.
- *elasticDelay* sets an initial delay to an arriving job based upon two parameters: the given distribution of elastic jobs lengths and the current service rate for this job. In mode *Resh*, the service rate is determined from the number of elastic jobs currently in service and the number of servers allocated to elastic jobs; in mode *NonResh* it is computed from the number of jobs currently served by the cluster that has been assigned to the arrived job and from the cluster size. Moreover, upon every elastic arrival or departure the delay of all elastic jobs in service (*Resh*) or of the jobs served in the affected cluster (*NonResh*) is reset in accordance to the new service rate. Delay adjustment is implemented in a method invoked from module *elasticResourcePool*. When the service delay of a job is over, the job is send to module *deallocateE*. Total delay time of the job is recorded.
- *deallocateE* requests resource deallocation from *elasticResourcePool* and sends the job to *sink*.
- *elasticResourcePool* manages resource allocation to jobs from multiple sources based upon the preset total amount of resources, the cluster size and the number of jobs in service. If the requested amount of resources is -1 then the job is treated as elastic, otherwise it is considered streaming and the requested amount is allocated if available. Also, in *NonResh* mode, the number of the cluster to which the job has been assigned is returned to be saved in a job object attribute for the use of *elasticDelay* and *deallocateE*.
- *sourceS* generate jobs according to a given distribution and sends them to module *allocateS*.
- *allocateS* requests the preset amount of resources from *elasticResourcePool* and sends the job further if resources are allocated. Otherwise, the job is dropped.
- *delayS* assigns to the incoming job a delay according to the preset distribution. When the delay is over, the job is sent further.
- *deallocateS* requests *elasticResourcePool* to free the preset amount of resources.
- *sink* destroys the jobs.

The simulator gathers data that permits estimating various characteristics of the system. In particular, we consider three estimators for blocking probabilities:

natural, *simple* and *indirect* [20], given respectively by

$$\widehat{B}_E^{ntrl}(t) = \frac{L_E(t)}{A_E(t)}, \tag{1}$$

$$\widehat{B}_E^{smpl}(t) = \frac{L_E(t)}{\lambda t}, \tag{2}$$

$$\widehat{B}_E^{ind}(t) = 1 - \frac{\widehat{n}_E(t)}{\lambda \widehat{T}_E(t)} \tag{3}$$

for elastic jobs, and by

$$\widehat{B}_S^{ntrl}(t) = \frac{L_S(t)}{A_S(t)}, \tag{4}$$

$$\widehat{B}_S^{smpl}(t) = \frac{L_S(t)}{\alpha t}, \tag{5}$$

$$\widehat{B}_S^{ind}(t) = 1 - \frac{\widehat{n}_S(t)\beta}{\alpha} \tag{6}$$

for streaming jobs. Here, $L_{\{E,S\}}(t)$ denotes the number of lost elastic or streaming jobs and $A_{\{E,S\}}(t)$ is the number of total elastic or streaming arrivals in $[0, t]$. $\widehat{n}_{\{E,S\}}(t)$ is a time-average estimator of the steady-state mean number of elastic or streaming jobs in the system, and $\widehat{T}_E(t)$ is an estimator of the average elastic service time based upon data over $[0, t]$.

4 Numerical Results

For our numerical example, we consider a network cell having the peak data rate of 750 Mbps, 100 Mbps of which are reserved for MBB traffic only. Let 50 Mbps be the minimum required data rate for MBB connections. We set the resource unit (RU) equal to 1 Mbps. Now, the structural parameters of the model are $C = 750$, $R = 100$, $C_E = 650$ and $d_S = 50$. Also, we assume $M = c$, which limits the minimum data rate for MTC traffic to 1 RU. Let $\alpha = 0.04$ and $\beta = 0.0055$, which corresponds to the mean MBB session duration of about three minutes. Note that the resulting MBB offered load is $d_S \frac{\alpha}{\beta} = 363.64$.

As for the load parameters of MTC traffic, in order to make the example more intuitively comprehensible we assume that MTC sessions arrive according to a two-state ($K = 2$) Markov-modulated Poisson process (MMPP) [16,17]. Thus, there are alternating low-rate and high-rate regimes with exponential holding times having means $1/q_{1,2}^{(0)}$ and $1/q_{2,1}^{(0)}$, which we set to 200 and 25 respectively (8:1 ratio). The arrival rates in these regimes, $q_{1,1}^{(1)}$ and $q_{2,2}^{(1)}$, will be determined in each case from the ratio $r = q_{2,2}^{(1)}/q_{1,1}^{(1)}$ and the average MTC arrival rate, which equals

$$\lambda = \frac{q_{1,1}^{(1)}/q_{1,2}^{(0)} + q_{2,2}^{(1)}/q_{2,1}^{(0)}}{1/q_{1,2}^{(0)} + 1/q_{2,1}^{(0)}}. \tag{7}$$

In our first example, we assume $\mu = 100$, which corresponds to the average MTC packet length of 1.25 kB, and $\lambda = 200$, thus $\frac{\lambda}{\mu} = 2$. In what follows we refer to this case as *light MTC loading*. For $r = 10$, the model's MAP parameters are $\mathbf{Q}_0 = \begin{bmatrix} -100.005 & 0.005 \\ 0.04 & -1000.04 \end{bmatrix}$ and $\mathbf{Q}_1 = \begin{bmatrix} 100 & 0 \\ 0 & 1000 \end{bmatrix}$.

Figures 2 and 3 show selected performance measures of the system obtained analytically from the stationary distribution derived in [12] and through simulation in mode *NonResh*. Here, the cluster size c is plotted on the x-axis. Simulation results are plotted only for $r = 10$, however, in order to reveal the influence of MMPP structure on the system's behavior, we also present exact values for cases $r = 1000$ and $r = 1$ (Poisson arrivals). Incidentally, the structure of $\mathbf{Q}_0$ did not affect the stationary characteristics under study: a modification of the holding times ratio or their multiplication by the same constant did not affect the performance measures values as long a the MMPP average rate remained unchanged.

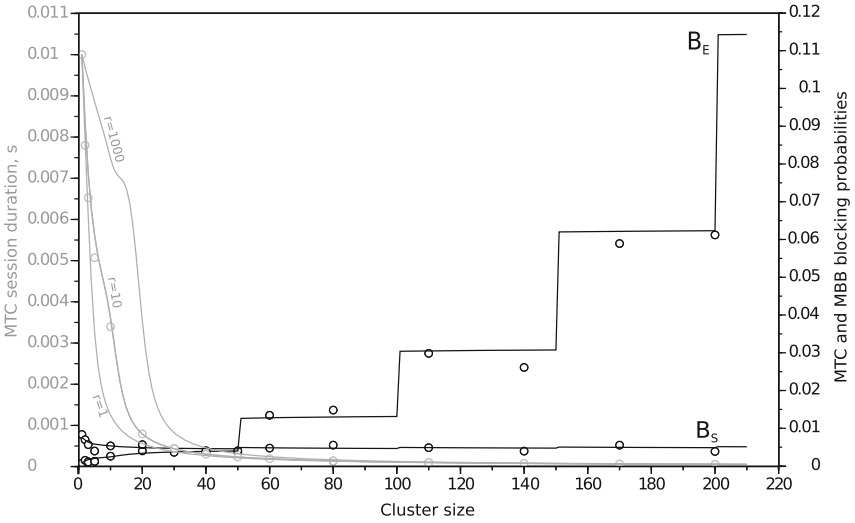


Fig. 2. Blocking probabilities B_E and B_S of MTC and MBB sessions respectively (black, right y-axis) and mean MTC session duration T_E (gray, left y-axis) as functions of the cluster size c. Circles show corresponding *NonResh* simulation results.

The graphs give insight into the order of the values and their behavior with the growth of the cluster size. In Fig. 2, the average MTC session duration T_E is calculated using Little's law and is graphed against the left y-axis, whereas the blocking probabilities B_E and B_S are plotted against the right y-axis. We note that MTC blocking probability increases gradually for small c and in steps of width $d_S = 50$ after approx. $c = 40$. Parameter r affects the shape of the session duration curve for $c < 50$. The influence of r on blocking probabilities is rather insignificant and therefore not shown on the plot (B_E is slightly lower for larger r).

The plotted simulation blocking probability values are obtained using the natural estimator, however the simple estimator yields very close values. The indirect estimator, on the other hand, appears unsuitable for the case under study. For instance, for $c = 10$ we have $\widehat{B}_E^{ntrl} = 0.0027 \pm 0.0013$, $\widehat{B}_E^{smpl} = 0.0027 \pm 0.0014$, $\widehat{B}_E^{ind} = -0.0055 \pm 0.0222$ and $\widehat{B}_S^{ntrl} = 0.0054 \pm 0.0018$, $\widehat{B}_S^{smpl} = 0.0054 \pm 0.0017$, $\widehat{B}_S^{ind} = 0.0071 \pm 0.0180$.

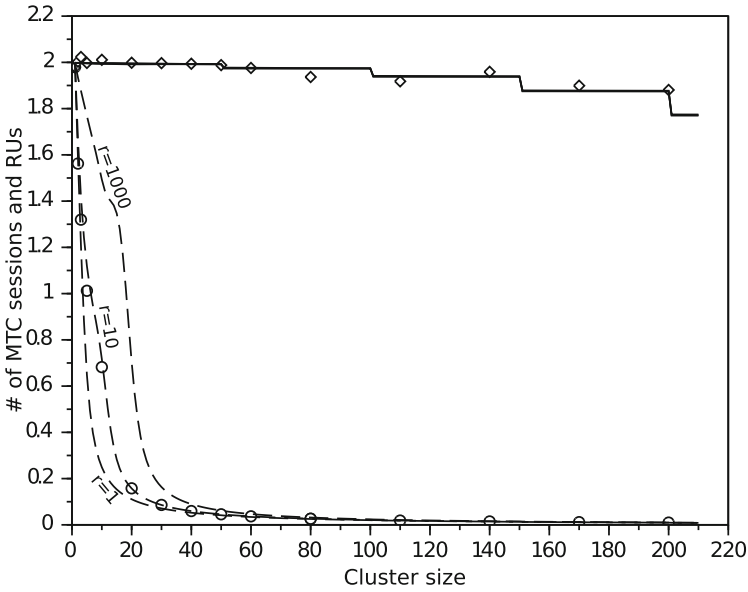


Fig. 3. The mean number N_E of active MTC sessions (dashed line) and the mean number c_E^{avg} of RU allocated to MTC (solid line) as functions of the cluster size c. Circles and diamonds show corresponding *NonResh* simulation results.

Figure 3 shows the mean number of active MTC sessions N_E and the mean number c_E^{avg} of RU allocated to them. We see that the graph of the number of sessions (elastic jobs in service) behaves similarly to the average session duration, whereas the number of allocated RU remains close to λ/μ for smaller c and slowly decreases in steps thereafter. Incidentally, the standard deviation of N_E (not plotted) comes down from over 5 to 3.6 between $c = 1$ and $c = 10$ and then decreases to 0.5 for large c. The standard deviation of c_E^{avg}, on the contrary, remains close to 5 for $c \leq 10$ and equals approx. $c/2$ for larger c. The total system utilization U (not shown) increases slightly (by about 10^{-3}) for small c and then decreases in steps but by no more than 2×10^{-4} in total within the considered range. Parameter r affects only the first portion of the curve, which is flatter for larger r.

Overall, *NonResh* simulation results are within the margin of error from the exact values for a "reshuffled"system. We ran our simulation model dur-

ing 200,000 s simulation time in five replications. The precision vary from one performance measure to another. For instance, for $c = 10$, the values of other characteristics under study with the half widths of the 95% confidence intervals calculated from the replications are $N_E = 0.682 \pm 0.026$, $c_E^{avg} = 5.273 \pm 0.013$, $U = 0.4844 \pm 0.0087$ and $T_E^{avg} = 0.0034 \pm 0.0001$.

In Tables 2, 3 and 4 we present simulation results for *moderate MTC loading* $\lambda/\mu = 20$ (cases (r, μ)=(10,100), (10,1) and (1000,100) respectively) and compare *Resh* and *NonResh* simulation modes. Half widths of the 95% confidence intervals over five repetitions are given along the values. The exact values are not provided for comparison because the chosen range of parameters produces an ill-conditioned generator matrix impeding the direct application of the algorithm. The length of the simulation is 25,000 s for $\mu = 100$ and 200,000 s for $\mu = 1$.

Table 2. Simulation results for moderate MTC loading, case $r = 10$, $\mu = 100$, run length 2.5×10^4 s.

		c = 10	c = 40	c = 70
Resh	$\widehat{B}_E^{ntrl}$	0.0019 ± 0.0018	0.0041 ± 0.0019	0.0061 ± 0.0048
	$\widehat{B}_E^{smpl}$	0.0019 ± 0.0018	0.0042 ± 0.0019	0.0061 ± 0.0049
	$\widehat{B}_E^{ind}$	0.0072 ± 0.0872	-0.0207 ± 0.076	0.0052 ± 0.0679
	$\widehat{B}_S^{ntrl}$	0.0124 ± 0.0049	0.0115 ± 0.0081	0.0080 ± 0.0076
	$\widehat{B}_S^{smpl}$	0.0128 ± 0.0051	0.0118 ± 0.0083	0.0078 ± 0.0075
	$\widehat{B}_S^{ind}$	-0.012 ± 0.021	0.0675 ± 0.017	0.0248 ± 0.0414
	T_E	0.0078 ± 0.0002	0.0048 ± 0.0003	0.0040 ± 0.0003
	N_E	15.44 ± 1.74	9.72 ± 1.35	7.95 ± 1.06
	c_E^{avg}	19.86 ± 1.75	20.41 ± 1.52	19.90 ± 1.36
	U	0.517 ± 0.024	0.523 ± 0.031	0.500 ± 0.022
NonResh	$\widehat{B}_E^{ntrl}$	0.0013 ± 0.0018	0.0090 ± 0.0147	0.0141 ± 0.0046
	$\widehat{B}_E^{smpl}$	0.0013 ± 0.0017	0.0094 ± 0.0156	0.0142 ± 0.0052
	$\widehat{B}_E^{ind}$	0.0349 ± 0.0628	-0.0068 ± 0.0620	0.0111 ± 0.0897
	$\widehat{B}_S^{ntrl}$	0.0083 ± 0.0058	0.0098 ± 0.0084	0.0062 ± 0.0038
	$\widehat{B}_S^{smpl}$	0.0084 ± 0.0057	0.0100 ± 0.0086	0.0062 ± 0.0040
	$\widehat{B}_S^{ind}$	0.0063 ± 0.0387	0.0130 ± 0.0437	0.0131 ± 0.0303
	T_E	0.0072 ± 0.0001	0.0046 ± 0.0003	0.0039 ± 0.0003
	N_E	13.91 ± 1.18	9.27 ± 1.12	7.79 ± 1.38
	c_E^{avg}	19.3 ± 1.26	20.14 ± 1.24	19.78 ± 1.80
	U	0.508 ± 0.019	0.506 ± 0.021	0.506 ± 0.016

We note that the precision of the indirect estimator for blocking probabilities remains insufficient compared to the other two estimators yielding, yet again, very close values. However, if we compare values in Tables 2 and 3, we can see

Table 3. Simulation results for moderate MTC loading, case $r = 10$, $\mu = 1$, run length 2×10^5 s.

		c = 10	c = 40	c = 70
Resh	$\widehat{B}_E^{ntrl}$	0.0018 ± 0.0011	0.0060 ± 0.0023	0.0100 ± 0.0025
	$\widehat{B}_E^{ind}$	0.0091 ± 0.0610	-0.025 ± 0.0295	0.0212 ± 0.0270
	$\widehat{B}_S^{ntrl}$	0.0097 ± 0.0019	0.0099 ± 0.0020	0.0086 ± 0.0023
	$\widehat{B}_S^{ind}$	0.0051 ± 0.0093	0.0036 ± 0.0206	0.0028 ± 0.0189
	T_E	0.7768 ± 0.0139	0.4726 ± 0.0120	0.3843 ± 0.0110
	N_E	15.40 ± 1.22	9.69 ± 0.52	7.53 ± 0.42
	c_E^{avg}	19.81 ± 1.23	20.49 ± 0.59	19.58 ± 0.54
	U	0.509 ± 0.005	0.511 ± 0.010	0.510 ± 0.010
NonResh	$\widehat{B}_E^{ntrl}$	0.0014 ± 0.0005	0.0036 ± 0.0009	0.0089 ± 0.0021
	$\widehat{B}_E^{ind}$	0.0311 ± 0.0251	0.0138 ± 0.0043	0.0292 ± 0.0214
	$\widehat{B}_S^{ntrl}$	0.0097 ± 0.0024	0.0085 ± 0.0020	0.0085 ± 0.0024
	$\widehat{B}_S^{ind}$	0.0138 ± 0.0267	0.0156 ± 0.0218	0.0153 ± 0.0247
	T_E	0.7169 ± 0.0054	0.4424 ± 0.0022	0.3757 ± 0.0084
	N_E	13.89 ± 0.46	8.73 ± 0.08	7.30 ± 0.33
	c_E^{avg}	19.38 ± 0.50	19.72 ± 0.08	19.41 ± 0.41
	U	0.504 ± 0.013	0.504 ± 0.011	0.504 ± 0.012

Table 4. Simulation results for moderate MTC loading, case $r = 1000$, $\mu = 100$, run length 2.5×10^4 s.

		c = 10	c = 40	c = 70
Resh	B_E	0.0086 ± 0.0096	0.0103 ± 0.0117	0.0330 ± 0.0107
	B_S	0.0124 ± 0.0045	0.0138 ± 0.0084	0.0169 ± 0.0027
	T_E	0.0097 ± 0.0000	0.0089 ± 0.0000	0.0080 ± 0.0000
	N_E	18.81 ± 3.04	19.02 ± 3.14	16.54 ± 2.04
	c_E^{avg}	19.44 ± 3.11	21.39 ± 3.52	20.65 ± 2.54
	U	0.512 ± 0.038	0.501 ± 0.021	0.515 ± 0.027
NonResh	B_E	0.0129 ± 0.0138	0.0090 ± 0.0067	0.0253 ± 0.0145
	B_S	0.0143 ± 0.0148	0.0117 ± 0.0047	0.0150 ± 0.0086
	T_E	0.0092 ± 0.0000	0.0087 ± 0.0000	0.0079 ± 0.0001
	N_E	17.58 ± 1.83	18.30 ± 0.52	16.98 ± 2.55
	c_E^{avg}	19.13 ± 2.00	21.10 ± 0.60	21.44 ± 3.01
	U	0.500 ± 0.065	0.506 ± 0.021	0.512 ± 0.039

its precision growing with the simulation length. The difference between *Resh* and *NonResh* values is negligible compared to the statistical precision for most performance measures, although we notice that service times and the number of

Table 5. Simulation results for heavy MTC loading, case $r = 10$, $\mu = 100$, run length 2.5×10^4 s.

		c = 10	c = 40	c = 70
Resh	$\widehat{B}_E^{ntrl}$	0.3034 ± 0.0245	0.3138 ± 0.0334	0.3179 ± 0.0145
	$\widehat{B}_E^{smpl}$	0.2685 ± 0.0796	0.2446 ± 0.0558	0.2025 ± 0.0665
	$\widehat{B}_E^{ind}$	0.3019 ± 0.0422	0.2909 ± 0.0646	0.3302 ± 0.0364
	$\widehat{B}_S^{ntrl}$	0.1219 ± 0.0197	0.1365 ± 0.0282	0.0894 ± 0.0220
	$\widehat{B}_S^{smpl}$	0.1128 ± 0.0209	0.1030 ± 0.0275	0.0505 ± 0.0186
	$\widehat{B}_S^{ind}$	0.0721 ± 0.0513	0.1319 ± 0.0508	0.1079 ± 0.0312
	T_E	0.0097 ± 0.0000	0.0088 ± 0.0001	0.0080 ± 0.0001
	N_E	135.58 ± 8.53	125.05 ± 13.45	107.43 ± 7.64
	c_E^{avg}	139.63 ± 8.45	141.83 ± 12.93	133.95 ± 7.26
	U	0.634 ± 0.024	0.610 ± 0.021	0.612 ± 0.005
NonResh	$\widehat{B}_E^{ntrl}$	0.2861 ± 0.0176	0.2921 ± 0.0121	0.3141 ± 0.0310
	$\widehat{B}_E^{smpl}$	0.2063 ± 0.0245	0.2121 ± 0.0268	0.2406 ± 0.0498
	$\widehat{B}_E^{ind}$	0.3065 ± 0.0290	0.3063 ± 0.0344	0.3107 ± 0.0661
	$\widehat{B}_S^{ntrl}$	0.1092 ± 0.0223	0.1273 ± 0.0184	0.1057 ± 0.0231
	$\widehat{B}_S^{smpl}$	0.0792 ± 0.0193	0.0916 ± 0.0137	0.0834 ± 0.0172
	$\widehat{B}_S^{ind}$	0.1579 ± 0.0391	0.1155 ± 0.0763	0.1012 ± 0.0561
	T_E	0.0093 ± 0.0000	0.0086 ± 0.0001	0.0080 ± 0.0003
	N_E	129.28 ± 5.90	119.96 ± 7.21	110.66 ± 13.95
	c_E^{avg}	138.69 ± 5.80	138.74 ± 6.89	137.86 ± 13.22
	U	0.593 ± 0.013	0.614 ± 0.032	0.620 ± 0.016

elastic jobs are smaller for *NonResh*, as one could expect, the difference being particularly visible for $\mu = 1$, $c = 10$. Also, we notice that the value of μ does not appear to affect considerably the performance measures under study (obviously apart from T_E) as long as λ/μ remains unchanged.

Overall, the nature of the dependence from c of all the performance measures under study is similar to the *light MTC loading* considered previously. For $r = 10$, T_E and N_E decrease rapidly between $c = 10$ and $c = 40$ and slower between $c = 40$ and $c = 70$. For $r = 1000$, the drop seams shifted to the right on the x-axis. U and c_E^{avg} remain roughly constant with $c_E^{avg} \approx \frac{\lambda}{\mu}$, and B_E increases slowly. The precision of blocking probabilities estimation appears generally higher for $r = 10$ compared to $r = 1000$, which could be explained by higher model variability in the latter case [20]. Also, interestingly, as it can be seen from the half widths, the precision is generally slightly higher in mode *NonResh*, especially for MTC-related measures.

Tables 5 and 6 present results under *heavy MTC loading* $\lambda/\mu = 200$ for cases (r, μ)=(10,100) and (10,0.1) respectively. We ran the simulation for 25,000 s for $\mu = 100$ and for 100,000 s for $\mu = 0.1$ in five replications. The simulation in

Table 6. Simulation results for heavy MTC loading, case $r = 10$, $\mu = 0.1$, run length 10^5 s.

		c = 10	c = 40	c = 70
Resh	$\widehat{B}_E^{ntrl}$	0.2477 ± 0.0105	0.2525 ± 0.0068	0.2690 ± 0.0162
	$\widehat{B}_E^{smpl}$	0.2437 ± 0.0196	0.2491 ± 0.0154	0.2669 ± 0.0238
	$\widehat{B}_E^{ind}$	0.2513 ± 0.0207	0.2541 ± 0.0164	0.2621 ± 0.0160
	$\widehat{B}_S^{ntrl}$	0.1117 ± 0.0074	0.1123 ± 0.0050	0.0987 ± 0.0066
	$\widehat{B}_S^{smpl}$	0.1096 ± 0.0084	0.1109 ± 0.0049	0.0955 ± 0.0060
	$\widehat{B}_S^{ind}$	0.1141 ± 0.0201	0.1208 ± 0.0058	0.1036 ± 0.0469
	T_E	9.7303 ± 0.0117	8.8386 ± 0.0393	8.1244 ± 0.0524
	N_E	145.70 ± 4.16	131.85 ± 3.48	119.91 ± 3.33
	c_E^{avg}	149.84 ± 4.14	149.16 ± 3.34	147.50 ± 3.22
	U	0.630 ± 0.011	0.625 ± 0.003	0.632 ± 0.022
NonResh	$\widehat{B}_E^{ntrl}$	0.2493 ± 0.0118	0.2496 ± 0.0064	0.2485 ± 0.0204
	$\widehat{B}_E^{smpl}$	0.2477 ± 0.0152	0.2431 ± 0.0130	0.2380 ± 0.0349
	$\widehat{B}_E^{ind}$	0.2447 ± 0.0089	0.2571 ± 0.0246	0.2730 ± 0.0264
	$\widehat{B}_S^{ntrl}$	0.1289 ± 0.0064	0.1310 ± 0.0127	0.1058 ± 0.0099
	$\widehat{B}_S^{smpl}$	0.1294 ± 0.0078	0.1284 ± 0.0135	0.1035 ± 0.0098
	$\widehat{B}_S^{ind}$	0.1132 ± 0.0221	0.1369 ± 0.0244	0.1164 ± 0.0242
	T_E	8.9720 ± 0.0181	8.3054 ± 0.0354	7.6644 ± 0.0881
	N_E	135.53 ± 1.86	123.41 ± 4.53	111.46 ± 5.31
	c_E^{avg}	151.10 ± 1.92	148.55 ± 4.86	145.36 ± 5.23
	U	0.632 ± 0.009	0.617 ± 0.009	0.622 ± 0.012

case of $\mu = 100$ was particularly time consuming: in average over 10 hours a replication compared to under 2 min for $\mu = 0.1$.

We can observe a substantial decrease in precision for heavy MTC load cases, which corresponds to findings of [20]. For MTC blocking probability, the indirect estimator is more adequate compared to lighter loading and sometimes outperforms the simple estimator; for MBB blocking probability, however, we do not observe any major amelioration.

Yet again, *Resh* and *NonResh* modes yield rather close results, with somewhat higher T_E and N_E in *Resh*. MTC blocking probabilities appear consistently higher for $\mu = 100$, although B_E, N_E and U have relatively close values. As for the relation to the cluster size, the behavior of the performance measures appears similar to the cases considered previously (with $c_E^{avg} \approx (1 - B_E)\frac{\lambda}{\mu}$), although B_E does not grow significantly within the studied range.

5 Concluding Remarks

A major difficulty in analyzing a combination of MBB and MTC communications consists in substantial difference in scales between the two types of traffic.

Indeed, a longer simulation time is needed for system events related to MBB and (in the case of our model) MAP transitions, whereas a very small timescale of events related to MTC considerably lengthen CPU time. Our numerical results show that for lower loading a longer MTC session duration and, accordingly, a smaller arrival rate can be used for a quicker estimation of blocking probabilities and system utilization, however, under heavier loading the effect on blocking probabilities becomes noticeable. Also, according to the simulation results, the loss system with cluster "reshuffling" proposed in [12] appears suitable for analysis of MTC clustering and performs particularly well for small MTC session durations typical for MTC traffic.

References

1. Ericsson: 5G systems: enabling the transformation of industry and society. Ericsson White paper (2017). https://www.ericsson.com/en/white-papers/5g-systems-enabling-the-transformation-of-industry-and-society
2. Tullberg, H.M., et al.: The METIS 5G system concept: meeting the 5G requirements. IEEE Commun. Mag. **54**, 132–139 (2016)
3. Tullberg, H., Fallgren, M., Kusume, K., Höglund, A.: 5G use cases and system concept. In: Osseiran, A., Monserrat, J.F., Marsch, P. (eds.) 5G Mobile and Wireless Communications Technology. Cambridge University Press, Cambridge (2016)
4. Sachs, J., Popovski, P., Höglund, A., Gozalvez-Serrano, D., Fertl, P.: Machine-type communications. In: Osseiran, A., Monserrat, J.F., Marsch, P. (eds.) 5G Mobile and Wireless Communications Technology. Cambridge University Press, Cambridge (2016)
5. Droste, H., Da Silva, I.L., Rost, P., Boldi, M.: The 5G architecture. In: Osseiran, A., Monserrat, J.F., Marsch, P. (eds.) 5G Mobile and Wireless Communications Technology. Cambridge University Press, Cambridge (2016)
6. Laya, A., Alonso, L., Alonso-Zarate, J.: Is the random access channel of LTE and LTE - a suitable for M2M communications? A survey of alternatives. IEEE Commun. Surv. Tutor. **16**(1), 4–16 (2014)
7. Biral, A., Centenaro, M., Zanella, A., Vangelista, L., Zorzi, M.: The challenges of M2M massive access in wireless cellular networks. Digit. Commun. Netw. **1**(1), 1–19 (2015)
8. Dawy, Z., Saad, W., Ghosh, A., Andrews, J.G., Yaacoub, E.: To wards massive machine type cellular communications. IEEE Wirel. Commun. **24**(1), 120–128 (2017)
9. Kim, D.M., Sørensen, R., Mahmood, K., Østerbø, O., Zanella, A., Popovski, P.: Data aggregation and packet bundling of uplink small packets for monitoring applications in LTE. IEEE Netw. **31**(6), 32–38 (2017)
10. Gharbieh, M., Bader, A., ElSawy, H., Alouini, M.-S., Adinoyi, A.: The advents of device-to-device relaying for massively loaded 5G networks. In: GLOBECOM 2017–2017 IEEE Global Communications Conference, pp. 1–7 (2017)
11. Buturlin, I., Gudkova, I., Chukarin, A.: On radio resource allocation scheme model with fixed capacities for machine type communications in LTE network. T-Comm - Telecommunications and Transport, vol. 8, pp. 14–18 (2014). (in Russian)
12. Vishnevsky, V.M., Samouylov, K.E., Naumov, V.A., Krishnamoorthy, A., Yarkina, N.: Multiservice queueing system with MAP arrivals for modelling LTE cell with

H2H and M2M communications and M2M aggregation. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds.) DCCN 2017. CCIS, vol. 700, pp. 63–74. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66836-9_6
13. Samuylov, A., Moltchanov, D., Gaidamaka, Y., Andreev, S., Koucheryavy, Y.: Random triangle: a baseline model for interference analysis in heterogeneous networks. IEEE Trans. Veh. Technol. **65**, 6778–6782 (2016)
14. Begishev, V., et al.: An analytical approach to SINR estimation in adjacent rectangular cells. In: Balandin, S., Andreev, S., Koucheryavy, Y. (eds.) ruSMART 2015. LNCS, vol. 9247, pp. 446–458. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23126-6_39
15. Basharin, G.P., Naumov, V.A.: Simple matrix description of peaked and smooth traffic and its applications. In: Proceedings of the 3 International Seminar on Teletraffic Theory "Fundamentals of Teletraffic Theory", pp. 38–44. VINITI, Moscow (1984)
16. Naumov, V.A.: Markovskie modeli potokov trebovaniy [Markovian Models of Arrival Processes], pp. 67–73. UDN Publisher, Moscow (1987) .(in Russian)
17. Chakravarthy, S.R.: Markovian arrival processes. In: Wiley Encyclopedia of Operations Research and Management Science (2010)
18. OMNeT++ Discrete Event Simulator. https://omnetpp.org
19. Kriege, J., Buchholz, P.: Simulating stochastic processes with OMNeT++. In: Proceedings of the 4th International OMNeT++ Workshop (OMNeT++ 2011), pp. 367–374. ICST (2011)
20. Srikant, R., Whitt, W.: Simulation run lengths to estimate blocking probabilities. ACM Trans. Model. Comput. Simul. **6**, 7–52 (1996)

Efficiency Enhancement of Tethered High Altitude Communication Platforms Based on Their Hardware-Software Unification

V. N. Perelomov[2], L. O. Myrova[2], D. A. Aminev[1], and D. V. Kozyrev[1,3(✉)]

[1] V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, 65, Profsoyuznaya Street, Moscow 117997, Russia
aminev.d.a@ya.ru

[2] JSC "Moscow Order of the Red Banner of Labor Scientific Research Institute of Radio Engineering", Bolshoi Trehsvyatitel'skii lane 2/1, Moscow, Russia
astra@mnirti.ru

[3] Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow 117198, Russia
kozyrev_dv@rudn.university

Abstract. The article presents the results of the analysis of the need for unification of tethered telecommunication high-altitude platforms (HAP). Possible ways of unification and standardization of HAP, as well as criteria and methods of their structural optimization are considered. It is shown that one of the methods for increasing the efficiency of telecommunication networks is the unification of requirements, component parts and their composite systems. The requirements that should be laid at the stage of setting a technical task for product development are given and it is shown that such parameters as the repeatability factor, the applicability factor, the inter-project unification coefficient are used to assess the level of unification. It is shown that to solve the problems of hardware-software unification and ensure information compatibility, it is necessary to develop appropriate models that take into account the specifics of modern information transmission systems in HAP. The rationale for using the apparatus of logical-dynamic models for solving unification problems is proved, and their joint use with network models allows to formalize the information necessary for solving the problems of hardware-software unification of HAP and ensuring their information compatibility.

Keywords: Telecommunication systems
Tethered high-altitude platforms · Unmanned aerial vehicles
Hardware-software unification
Structural optimization · Unification efficiency · Graph apparatus

This work has been financially supported by the Russian Science Foundation and the Department of Science and Technology (India) via grant 16-49-02021 for the joint research project by the V.A. Trapeznikov Institute of Control Sciences and the CMS College Kottayam.

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 184–200, 2018.
https://doi.org/10.1007/978-3-319-99447-5_16

1 Introduction

Despite the rapid development of land-based telecommunications in cities and densely populated areas, and satellite telecommunications, there is a need to use tethered communication high-altitude platforms (TCHAP), since in remote and underdeveloped areas they are one of the main means of providing information interaction with mobile networks and the Internet [1–4].

The tethered telecommunication high-altitude platform consists of the ground terminal's equipment, communication channels, power supply lines, and equipment of the multi-copter that can hang at a height of several hundred meters above the ground [5] (Fig. 1a).

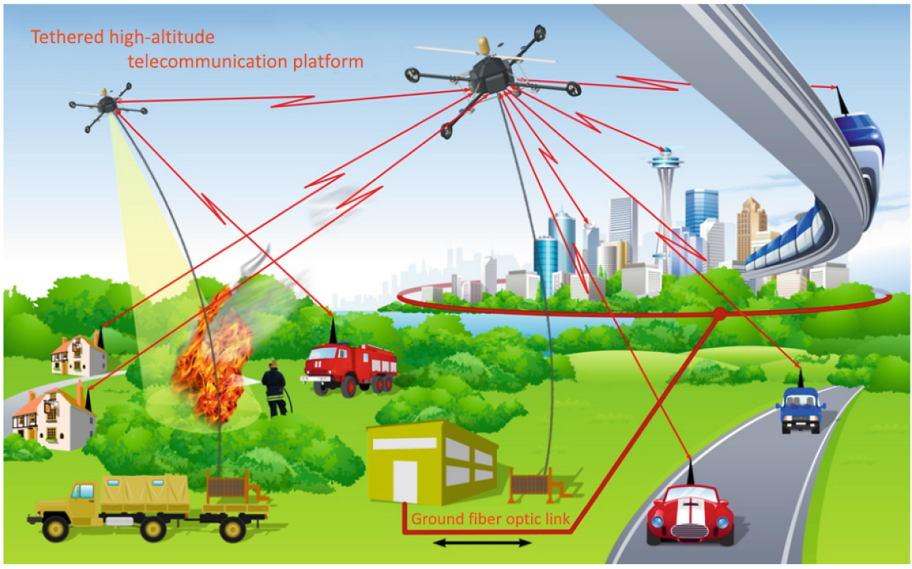
State agencies, first of all, are interested in the development of TCHAP, for example, as a means of providing communication in emergency situations. The development of this field of science and technology will allow the rapid deployment of network coverage of some territory with mobile communication, Internet, radio and television broadcasting. Thus, the increase in the pace of implementation and development of TCHAP leads to the need for enhancement of its utilization efficiency [9–11].

One of the methods to enhance their efficiency is the unification of requirements, components and composite systems of TCHAP, so the issue of bringing to a certain single line of requirements and components is a very urgent task that would simplify and significantly reduce the cost of their development and exploitation, and thereby increase the efficiency. In this regard, there is a need to consider possible approaches and ways of unification or standardization of the TCHAP under development and in operation [12].

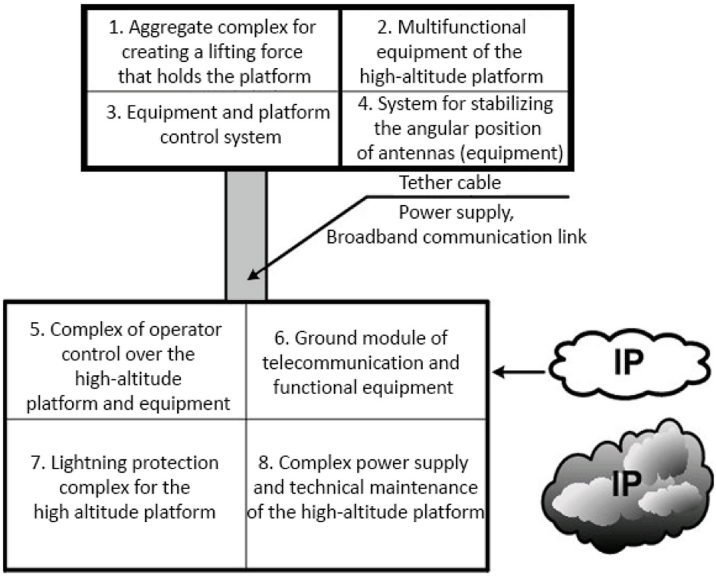
2 TCHAP Enhancement by the Means of Unification

Unification is the establishment of the optimal number of sizes or types of products, processes or services needed to meet basic needs [6]. From the point of view of the standardization theory, we can talk about several methods, which allow for unification. These are: the basic unit method, the method of compounding, the method of modification, the modularity principle. Thus, unification turns out to be beforehand rooted in the process of product development and in the process of its creation.

Considering the possibilities of unification it is necessary to determine, first of all, the goal it is aimed at, because unification can concern both the development processes and the processes of application of the product to its intended purpose, while one can speak both about the unification of a sample or a complex as a whole, and about the unification of the component parts of this product. Since the target function of the effectiveness of TCHAP is still its intended usage, and not its development process, it is necessary to consider unification from the point of view of applying the results obtained during the operation of the product, i.e. the objective function of unification in this case is to simplify the process of using the product for the intended purpose. Thus, it is assumed that,



a)



b)

Fig. 1. TCHAP network field deployment (a) and structure of its equipment (b).

in order to achieve the required level of unification at the operational stage, the requirements should be laid even at the stage of setting the terms of reference for the product development.

Such requirements include the requirements for:

- ensuring the inter-project unification, which will indicate which specific products or units should firstly be unified in the new product being developed;
- application of basic structures and elements of the base;
- the development of the product as a basic one, using the basic unit method, and at that, there should be specified the requirements for the product being developed, the implementation of which provides the possibility of its use as a basic unit in the process of modification as well, which will allow the creation of new modified products based on one and the same basic unit;
- ensuring compatibility conditions;
- the use of restrictive lists, i.e. the composition of hardware components that can be used;
- the application of technological processes, equipment rigging, tools that can be used for making samples;
- use of methods and means of testing and control, necessary for putting products into operation;
- the use of parametric and standard series indicating the nomenclature of the component parts of the product;
- connectivity with elements and characteristics used for operation.

Note, that the effectiveness of unification is characterized by its level. In the classical version, parameters such as the repeatability factor, the applicability factor, the coefficient of inter-project unification are used to estimate the level of unification.

The repeatability coefficient k_{r} of component parts in the total number of constituent parts of this product characterizes the level of unification and interchangeability of component parts of products of a certain type and is calculated by the formula:

$$k_{\text{r}} = \frac{N-n}{N-1} \cdot 100[\%], \tag{1}$$

where N is a total number of components in the product; n is a total number of typical sizes in the product.

The average repeatability of constituent parts in a product is characterized by the coefficient:

$$k_{\text{r}}^{mean} = N/n \,[\text{in fractions of}]. \tag{2}$$

The coefficient k_{r}^{mean} characterizes the level of in-project unification of the product and the interchangeability of the component parts within the product. The numerical value of k_{r}^{mean} is determined on the basis of statistical data on the average repeatability of the component parts in a group of previously developed products with a similar functional purpose. The mean value of k_n is determined by the following formula:

$$k_{\text{r}}^{mean} = \frac{1}{h} \sum_{i=1}^{h} k_{r,i}, \tag{3}$$

where h is a total number of products in the group; $k_{r,i}$ is a repeatability coefficient of an i-th product in the group; $k_{\text{r}} \geq k_{\text{r}}^{mean}$. If N is known, then k_{r} can be

determined based on the correlation dependence of $k_\mathrm{r} = f(N)$. But it is necessary to have statistics (a fairly large group of previously developed products). To build this dependence for each product of the selected group, k_r is determined and the total number of component parts is calculated. These data are ordered by value and approximated by a relationship of the form:

$$k_\mathrm{r} = 1 + dN_a, \tag{4}$$

The values of coefficients d and a are determined using the least squares technique. In the absence of the sufficient volume of statistical data, the k_r for a projected product can be obtained approximately based on the data on the value of k_r and the prototype of the projected product, using the following formula:

$$k_\mathrm{r} = k_r^n \sqrt{\frac{N'}{N_n}}, \tag{5}$$

where k_r is the repeatability coefficient of the prototype; N_n is the total number of component parts in the prototype product; N' is the expected number of parts of the product being developed.

The applicability coefficient k_app, respectively, indicates the level of applicability of the components, i.e. the level of use in the newly developed designs of the nodes and mechanisms of the components, previously used in similar former products. This coefficient is determined by the formula:

$$k_\mathrm{app} = \frac{n - n_0}{n} \cdot 100[\%], \tag{6}$$

$$k_\mathrm{app} = \frac{n - n_0}{n} [\text{in fractions of}], \tag{7}$$

where n is the total number of typical sizes in the sample; n_0 is the number of original typical sizes.

The coefficient k_app can also be calculated by the constituent parts and by their cost by the formulas:

$$k_\mathrm{app} = \frac{N - N_0}{N} [\text{in fractions of}], \tag{8}$$

where N is the total number of component parts in the sample; N_0 is the number of original component parts;

$$k_\mathrm{app} = \frac{C - C_0}{C} \cdot 100[\%], \tag{9}$$

where C is the cost of all components in the sample; C_0 is the cost of the original components.

As can be seen from formulas (6) and (7), k_app can be calculated if n and n_0 or N and N_0 are known. Most commonly, these values are known when it comes to dealing with the modernization of an already existing product. Although at the final stage of development of design documentation for a new product, one

can also find these values and calculate the applicability factor using formulas (8) and (9).

The coefficient of inter-project unification k_{iu} shows the level of unification of the final product, in relation to other similar, already existing ones and used for the intended purpose, and is calculated by the formula:

$$k_{\text{iu}} = \frac{\sum_{i=1}^{H} n_i - Q}{\sum_{i=1}^{H} n_i - n_{max}} \cdot 100, \tag{10}$$

where H is the total number of considered projects (products); n_i is the number of typical sizes (component parts) in the i-th project (product); Q is the total number of original typical sizes (component parts) used in a group of H projects (products); n_{max} is the maximum number of typical sizes (component parts) in a single project (product).

Numerical values of k_{iu} are determined on the basis of the analysis of the state of mutual unification of previously developed products, with regard to the possibility and expediency of further increasing of the level of inter-project unification by reduction of the heterogeneity of the components of jointly operated or manufactured products.

Calculation of k_{iu} can be carried out by a simplified formula:

$$k_{\text{iu}} = n_2'/n', \tag{11}$$

where n' is the expected total number of typical sizes in the product under development at the selected level of structural complexity; n_2' is the number of typical sizes that can be borrowed from the selected product group [7].

However, when developing unification approaches, it is necessary to take into account the complex component of this process, and it is necessary to determine the main goals of unification. Unification is a complex issue which should concern all stages of the life cycle of products. At that, the unification should be carried out in two interconnected ways—the unification of systems and the unification of components and materials.

3 Directions of Unification

Unification of Structural Elements. Implies the creation of standard (modular) structures, the dimensions of the sides of which can vary according to the metric ratio applied to all or only to certain dimensions. Under metric ratio, the value of an n-th dimension is defined as $a_n = a_0 + nm$, where a_0 is the initial value of the dimension (width, height, depth), n is the integer or fractional number underlying the dimension-parametric series, m is the increment in the metric ratio.

Unification of Electronic Equipment and Controllers. A uniform software is created, consisting of separate unified modules. At the same time, a unified platform is being developed and manufactured in the form of some kind

of electronic device. Depending on how the program module is loaded into it, the output is a controller that performs various functions of the systems.

Unification of Interfaces. The task of combining various on-board computational aviation units, storage and display devices, peripheral equipment into one complex is solved with the help of unified connectivity systems—interfaces. The main purpose of an interface is the unification of intrasystem and intersystem connections and devices. Unification of the interaction rules is aimed at providing information, electrical and structural compatibility; it is unification and standardization that underlies the construction of interfaces:

- information compatibility is achieved due to unified requirements imposed on the structure and composition of interface lines, interaction algorithms, coding methods and data formats, control and address information, timings between signals;
- electrical compatibility means the consistency of the parameters of electrical or optical signals transmitted by the interface environment, the correspondence of logical states to signal levels; electrical compatibility determines the requirements for the load capacity of components and the characteristics of the transmission lines used (length, permissible active and reactive load, the order of connection of matching circuits, etc.);
- constructive compatibility means the possibility of mechanical connections of electrical circuits, and sometimes the replacement of some blocks; this kind of compatibility is ensured by the standardization of connectors (connectors, plugs, etc.), cables, board designs, etc.

A uniform standard interface would not be able to provide efficient operation of a variety of devices, were they used at different levels of the hierarchy. This explains the existence of a system of interfaces of different ranks, differing in characteristics and the degree of unification. Depending on the requirements of unification, the following are distinguished:

- physical implementation of the interface, i.e. the composition and characteristics of transmission lines, the design of their connection (for example, connector type), the type and characteristics of the signals;
- logical implementation of the interface, i.e. protocols of interaction, or algorithms for the forming of communications signals.

Unification of communication channels:

- use of a common set of frequency resources;
- compatibility of the signal coding;
- use of compatible modems;
- use of compatible signal coding methods;
- software compatibility of on-board and ground equipment;
- compatibility of communication channels of the on-board airborne complex and the ground communication facilities.

Unification of the display means and the software:

- unification of the on-board data acquisition means and the ground-based data display means assumes the use of systems with compatible software and the use of compatible characteristics of storage devices;
- software unification for systems with functional binding;
- unification of ground-based means of manual control over TCHAP, aerodynamic circuits by replacing on-board and ground-based software;
- unification of the automatic flight management software.

4 Methods of Structural Optimization of TCHAP

When solving the problem of hardware unification and providing information interaction, it should be taken into consideration that the configuration of the system (processes) of information interaction is constantly changing. Therefore,after the analysis of various types of graphs [15,16], the AND-OR graphs seem preferred. With the help of these graphs, the description of variable structures is most natural. Besides, in the process of solving this problem, it becomes necessary to compare the alternatives. Using the AND-OR graphs, this is also possible. At that, graphs need to be supplemented with a mechanism for estimating the cost of decisions.

Thus, to describe the structures of the complex TCHAP and information exchange, taking into account the dynamics of their changes and interactions, we use the apparatus of AND-OR graphs. However, in solving problems of unification, along with a description of the structure, it is necessary to have a number of models that allow to solve these problems in full. In order to select the suitable modeling methods, we consider existing approaches to modeling complex systems. Among the available models of systems, Petri nets and models of logic-dynamic systems are the closest to the problems of describing the information interaction.

Logical-dynamic model in a fairly general form can be described by the following vector equation:

$$\boldsymbol{x} = \sum_{i=1}^{s} P_i^f f_i\Big(t, \boldsymbol{x}, \sum_{j=1}^{g} P_j^u \boldsymbol{u_j}\Big), \tag{12}$$

where P_i^f and P_j^u are the predicates which are written in expanded form as follows:

Predicates P_i^f and P_j^u can be subject to certain additional conditions. For example, the uniqueness and completeness conditions can be imposed on the logical values of the predicates:

$$\forall(t)\Big(P_i^f \wedge P_h^f = 0\,(i \neq k), P_j^u \wedge P_k^u = 0\,(j \neq k), \bigwedge_{i=1}^{s} P_i^f = 1, \bigwedge_{j=1}^{g} P_j^u = 1\Big). \tag{13}$$

In order for these conditions to be fulfilled, the predicate functions must, of course, be matched appropriately. The introduction of the conditions means that one and only one of the predicates P_i^f and one and only one of the predicates P_j^u take values equal to 1. The latter, in turn, means that the considered dynamical system at each moment of time is described by one of the possible variants of its local representations:

$$\boldsymbol{x} = f_i\big(t, \boldsymbol{x}, u_j\big). \tag{14}$$

The considered models are used to solve various problems, for example, problems of ensuring the stability of a multi-copter, the management of electrical equipment, and also in the tasks of controlling the structures of computational complexes.

The apparatus of logic-dynamic models allows to take into account the parameters of information exchange after the information transformation by circuit elements in the chosen configuration, and to choose the most rational way of transformation and transmission from the available.

The Petri nets apparatus allows to adequately describe the processes occurring during data transmission and create a model of information interaction between the on-board equipment and the ground control complex, able to take into account the change in the state of the equipment, the requirements for the transformation parameters of the transmitted information, and allow to carry out the unification of information junctions and interaction protocols between various elements. It should be noted that Petri nets in the form presented above have a number of drawbacks that make it difficult to describe the strictly hierarchical structures to which the TCHAP structure relates. In this regard, various amplifications have been applied to the construction of Petri nets. One of these constructions is called the E-network, which can be used to describe the processes of information exchange of TCHAP.

The E-net is a graph consisting, like the Petri net, of two types of vertices-positions and transitions connected to each other by oriented edges or arcs, each arc connecting only a transition with a position or vice versa. Consequently, the E-net, like the Petri net, is structurally equivalent to a bipartite oriented graph, in which one set of vertices contains positions and the other contains transitions. However, unlike Petri nets, in E-nets there are several types of positions: simple, position-queues and allowing. Dynamics of an E-net is determined by the movement of objects (or chips) from one network position to another as a result of triggering transitions.

From a formal point of view, movement of objects is equivalent to changing the marking of the net. At the same time, it is possible to move objects in many elementary networks. Thus, providing the possibility of representing parallel processes in the modeled system, i.e. a particular process or operation can be associated with each movement of objects. The duration of the state change is determined by the time delay procedure τ.

Thus, the E-nets apparatus allows to fully describe the structure of the TCHAP taking into account the cause-effect changes in the information exchange process.

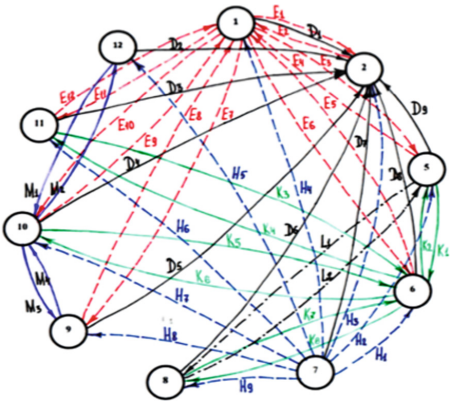


Fig. 2. Graph-signal model of TCHAP (Color figure online)

Table 1. Adjacency matrix of the generalized graph-signal model of TCHAP

	1	2	5	6	7	8	9	10	11	12
1	0	3	2	2	1	0	2	2	2	0
2	3	0	1	1	1	1	1	1	1	1
5	2	1	0	2	1	2	0	0	0	0
6	2	1	2	0	1	2	0	2	2	0
7	1	1	1	1	0	1	1	1	1	1
8	0	1	2	2	1	0	0	0	0	0
9	2	1	0	0	1	0	0	2	0	0
10	2	1	0	2	1	0	2	0	0	2
11	2	1	0	2	1	0	0	0	0	0
12	0	1	0	0	1	0	0	2	0	0

Table 2. Incidence matrix of the generalized graph-signal model of TCHAP

	1	2	5	6	7	8	9	10	11	12
E1	1	-1	0	0	0	0	0	0	0	0
E2	-1	1	0	0	0	0	0	0	0	0
E3	1	0	-1	0	0	0	0	0	0	0
E4	-1	0	1	0	0	0	0	0	0	0
E5	-1	0	0	1	0	0	0	0	0	0
E6	1	0	0	-1	0	0	0	0	0	0
E7	1	0	0	0	0	0	-1	0	0	0
E8	-1	0	0	0	0	0	1	0	0	0
E9	-1	0	0	0	0	0	0	1	0	0
E10	1	0	0	0	0	0	0	-1	0	0
E11	1	0	0	0	0	0	0	0	-1	0
E12	-1	0	0	0	0	0	0	0	1	0
D1	1	-1	0	0	0	0	0	0	0	0
D2	0	-1	0	0	0	0	0	0	0	1
D3	0	-1	0	0	0	0	0	0	1	0
D4	0	-1	0	0	0	0	0	1	0	0
D5	0	-1	0	0	0	0	1	0	0	0
D6	0	-1	0	0	0	1	0	0	0	0
D7	0	-1	0	0	1	0	0	0	0	0
D8	0	-1	0	1	0	0	0	0	0	0
D9	0	-1	1	0	0	0	0	0	0	0
H1	0	0	0	-1	1	0	0	0	0	0
H2	0	0	-1	0	1	0	0	0	0	0
H3	0	-1	0	0	1	0	0	0	0	0
H4	-1	0	0	0	1	0	0	0	0	0
H5	0	0	0	0	1	0	0	0	0	-1
H6	0	0	0	0	1	0	0	0	-1	0
H7	0	0	0	0	1	0	0	-1	0	0
H8	0	0	0	0	1	0	-1	0	0	0
H9	0	0	0	0	1	-1	0	0	0	0
K1	0	0	1	-1	0	0	0	0	0	0
K2	0	0	-1	1	0	0	0	0	0	0
K3	0	0	0	-1	0	0	0	0	1	0
K4	0	0	0	1	0	0	0	0	-1	0
K5	0	0	0	-1	0	0	0	1	0	0
K6	0	0	0	1	0	0	0	-1	0	0
K7	0	0	0	-1	0	1	0	0	0	0
K8	0	0	0	1	0	-1	0	0	0	0
M1	0	0	0	0	0	0	0	1	0	1
M2	0	0	0	0	0	0	0	-1	0	-1
M3	0	0	0	0	0	0	-1	-1	0	0
M4	0	0	0	0	0	0	1	1	0	0
L1	0	0	1	0	0	-1	0	0	0	0
L2	0	0	-1	0	0	1	0	0	0	0

Figure 2 shows the generalized graph-signal model of TCHAP, which reflects the system-level structural composition of the considered complex. In addition, intersystem connections within the complex are reflected. For display purposes, the intersystem links represented by the "edges" of the graph are marked with different colors.

There are several levels of inter-system links:

- level of physical connections of elements,
- level of electrical connections,
- level of information connections.

Information connections are divided into digital and analog ones. The digital ones, in turn, are divided into high-speed and low-speed. On the basis of the obtained graphs, we composed the adjacency matrix (Table 1), and incidence matrix (Table 2) that allow to analyze the mutual influence of systems with their connections.

In particular, the adjacency matrix unambiguously identifies the systems that determine the shape of the entire complex, i.e. those systems that are for sure

subject to unification. At that, the unification of the elements of these systems as basic units will definitely lead to the automatic unification of the systems connected with them.

On the obtained graph, the base systems are defined by the vertices: 1; 2; 6; 7. 1—multi-copter control system; 2—TCHAP stabilizer; 6—data transmission system; 7—power supply system. Analysis of the composition of these systems allows us to identify specific elements that affect the external appearance of the system, i.e. specify which elements should be recognized as basic units.

The incidence matrix characterizes the level of connections that require standardization in order to obtain a unified system. Such connections include: E; D; H; K, where E—command control lines; D—physical connection of the elements of the systems with the stabilizer TCHAP; H—power supply lines; K—information transmission lines.

Let's consider further the basic system of data transmission. Figure 3a shows the graph-signal model of the on-board information transfer system, where: P1—omnidirectional C-band antenna; P2—C-band directional antenna system; P3—UHF antenna; P4—coaxial switch; P5—UHF receiver-controller; P6—power indicator and diplexer; P7—power amplifier; P8—TM/TV transmitter; P9—radio frequency head and communication controller processor; P10—quartz generator; V1—streaming video and digital multiplexer; W1—modular CPU; U1—advanced miniature avionics module; Q1—electro-optical transducer; D1—payload; N—terrestrial data transmission terminal.

We perform the modeling of the structure and interconnection of the elements of the TCHAP's on-board and ground-based data transmission system. When constructing the model, we associate the vertices of a graph with the elements of the structure of the data transmission system and their states. Arcs of the graph will be associated with the directions of information transfer and the form of information transfer. In this case, we unambiguously display the structure of the system and clearly formulate directions and changes in the formats of information.

Since the purpose of the model construction is to determine the main directions for unification and to ensure information compatibility, then we define the rules, as in the previous case. Figure 3b presents a graph-signal model of the information transmission system, which reflects the structural composition of the system at the element level. In addition, inter-element links in the system are reflected. For convenience of perception, the inter-element links represented by the "edges" of the graph are also marked in various colors.

There are several levels of inter-element links:

- level of radio communications;
- level of command information (low-speed);
- level of user information communications.

Based on the obtained graphs, the adjacency and incidence matrices are constructed, which allow to analyze the mutual influence of systems with their connections. Thus, in particular, the adjacency matrix (Table 3) unambiguously

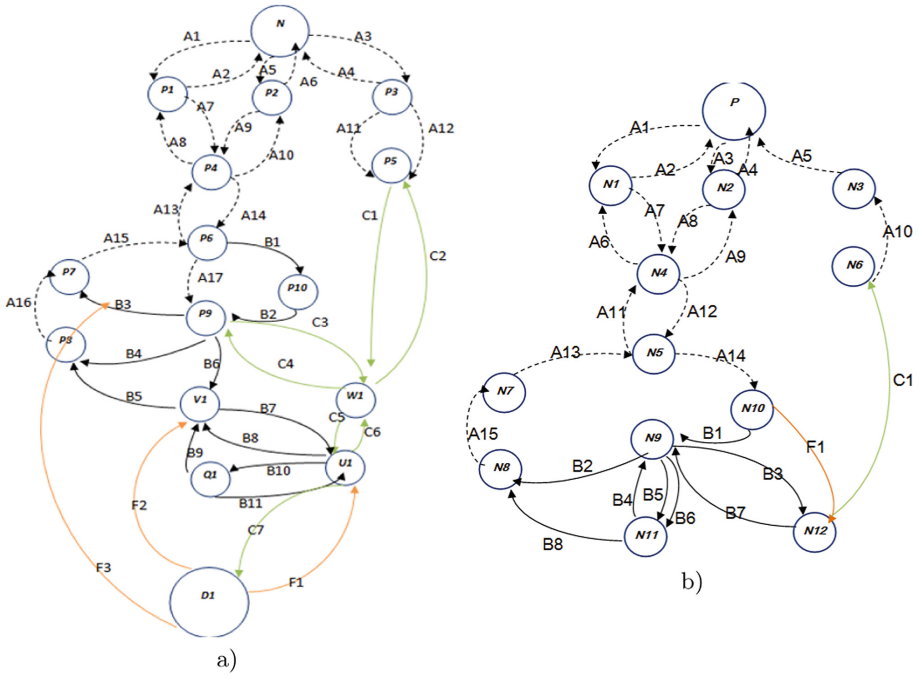


Fig. 3. Graph-signal model of the on-board (a) and ground-based (b) data transmission system (Color figure online)

shows the elements determining the image of the whole complex, i.e. those elements that should be recognized as basic aggregates. Moreover, the unification, or rather the standardization of elements as basic aggregates, will unambiguously lead to automatic unification of the systems associated with them. On the resulting graph, the base systems are the following vertices: P9—processor of the communication controller with a radio-frequency head; V1—digital multiplexer of streaming video; W1—modular CPU; U1—advanced miniature avionics module; N4—power amplifier; N9—ground communication controller hub; N11—ground multiplexer of streaming video and digital data.

Analysis of the composition of these systems allows to identify specific elements that affect the external appearance of the system, i.e. to specify which elements should be recognized as basic units. The incidence matrix (Table 4), characterizes the level of links that require standardization in order to obtain a unified system. These links include: A—radio communications (carrier frequency, signal-code construction); B—high-speed information lines; (protocol, data format); C—low-speed command lines (protocol, data format).

Where: P—on-board data transmission system; N1—omnidirectional C-band antenna; N2—C-band directional antenna system; N3—UHF antenna; N4—power amplifier; N5—radio frequency (RF) board and TM / TV RF head; N6—UHF transmitter-controller; N7—spread spectrum transmitter; N8—

Table 3. Adjacency matrix for the on-board data transmission system

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	V1	W1	U1	Q1	D1	N
P1	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	2
P2	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	2
P3	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	2
P4	2	2	0	0	0	2	0	0	0	0	0	0	0	0	0	0
P5	0	0	2	0	0	0	0	0	0	0	0	2	0	0	0	0
P6	0	0	0	2	0	0	1	0	1	1	0	0	0	0	0	0
P7	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0
P8	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0
P9	0	0	0	0	0	1	1	1	0	1	1	2	0	0	0	0
P10	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0
V1	0	0	0	0	0	0	0	1	1	0	0	0	2	1	2	0
W1	0	0	0	0	2	0	0	0	2	0	0	0	2	0	0	0
U1	0	0	0	0	0	0	0	0	0	0	2	2	0	2	2	0
Q1	0	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0
D1	0	0	0	0	0	0	0	0	0	0	2	0	2	0	0	0
N	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0

intermediate-frequency (IF) modem; N9—ground communication controller hub; N10—receiver of video and telemetry data; N11—ground multiplexer of streaming video and digital data; N12—ground unified control station.

Thus, the analysis of the obtained generalized graph-signal models of the TCHAP, the on-board and ground data transmission systems, and the corresponding adjacency matrices allowed us to determine the image of the whole complex, i.e. subsystems that are definitely subject to unification, and the incidence matrices that characterize the level of links that require standardization in order to obtain a unified system.

Figure 4 shows the results of the efficiency evaluation of the intended use of TCHAP (Tables 5 and 6).

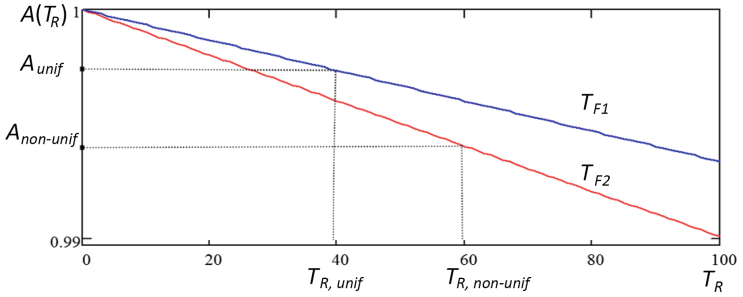


Fig. 4. Efficiency evaluation of the intended use of TCHAP

Analysis of the graph-signal models of the systems of the TCHAP allows to identify specific critical elements that affect the connectivity of the complex as

Table 4. Incidence matrix for the on-board data transmission system

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	V1	W1	U1	Q1	D1	N
A1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
A2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1
A3	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	1
A4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	-1
A5	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
A6	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	-1
A7	1	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0
A8	-1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
A9	0	1	0	-1	0	0	0	0	0	0	0	0	0	0	0	0
A10	0	-1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
A11	0	0	1	0	-1	0	0	0	0	0	0	0	0	0	0	0
A12	0	0	1	0	-1	0	0	0	0	0	0	0	0	0	0	0
A13	0	0	0	-1	0	1	0	0	0	0	0	0	0	0	0	0
A14	0	0	0	1	0	-1	0	0	0	0	0	0	0	0	0	0
A15	0	0	0	0	0	-1	1	0	0	0	0	0	0	0	0	0
A16	0	0	0	0	0	0	-1	1	0	0	0	0	0	0	0	0
A17	0	0	0	0	0	1	0	0	-1	0	0	0	0	0	0	0
B1	0	0	0	0	0	1	0	0	0	-1	0	0	0	0	0	0
B2	0	0	0	0	0	0	0	0	-1	1	0	0	0	0	0	0
B3	0	0	0	0	0	0	-1	0	1	0	0	0	0	0	0	0
B4	0	0	0	0	0	0	0	-1	1	0	0	0	0	0	0	0
B5	0	0	0	0	0	0	0	-1	0	0	1	0	0	0	0	0
B6	0	0	0	0	0	0	0	0	1	0	-1	0	0	0	0	0
B7	0	0	0	0	0	0	0	0	0	0	1	0	-1	0	0	0
B8	0	0	0	0	0	0	0	0	0	0	-1	0	1	0	0	0
B9	0	0	0	0	0	0	0	0	0	0	-1	0	0	1	0	0
B10	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	0	0
B11	0	0	0	0	0	0	0	0	0	0	0	0	-1	1	0	0
C1	0	0	0	0	1	0	0	0	0	0	0	-1	0	0	0	0
C2	0	0	0	0	-1	0	0	0	0	0	0	1	0	0	0	0
C3	0	0	0	0	0	0	0	0	1	0	0	-1	0	0	0	0
C4	0	0	0	0	0	0	0	0	-1	0	0	1	0	0	0	0
C5	0	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0
C6	0	0	0	0	0	0	0	0	0	0	0	-1	1	0	0	0
C7	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
F1	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	1	0
F2	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	1	0
F3	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	1	0

a whole, i.e. specify which elements of the systems must be recognized as basic and are subject to unification. The increase in the effectiveness of the intended use is estimated as follows:

$$A_{\text{eff.incr.}} = f(A), \tag{15}$$

where A is the availability factor of the object, which is determined through the mean time between failures (T_F) of the systems included in its structure and their mean recovery time (T_R) of their failure, i.e.:

$$A = \frac{T_F}{T_F + T_R}, \tag{16}$$

Table 5. Adjacency matrix for the ground data transmission system

	P	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11	N12
P	0	2	2	1	0	0	0	0	0	0	0	0	0
N1	2	0	0	0	2	0	0	0	0	0	0	0	0
N2	2	0	0	0	2	0	0	0	0	0	0	0	0
N3	1	0	0	0	0	0	1	0	0	0	0	0	0
N4	0	2	2	0	0	2	0	0	0	0	0	0	0
N5	0	0	0	0	2	0	0	1	0	0	1	0	0
N6	0	0	0	1	0	0	0	0	0	0	0	0	1
N7	0	0	0	0	0	1	0	0	1	0	0	0	0
N8	0	0	0	0	0	0	0	1	0	1	0	1	0
N9	0	0	0	0	0	0	0	0	1	0	1	3	2
N10	0	0	0	0	0	1	0	0	0	1	0	0	1
N11	0	0	0	0	0	0	0	0	1	3	0	0	0
N12	0	0	0	0	0	0	1	0	0	2	1	0	0

Table 6. Incidence matrix for the ground data transmission system

	P	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11	N12
A1	1	-1	0	0	0	0	0	0	0	0	0	0	0
A2	-1	1	0	0	0	0	0	0	0	0	0	0	0
A3	1	0	-1	0	0	0	0	0	0	0	0	0	0
A4	-1	0	1	0	0	0	0	0	0	0	0	0	0
A5	-1	0	0	0	0	1	0	0	0	0	0	0	0
A6	0	-1	0	0	1	0	0	0	0	0	0	0	0
A7	0	1	0	0	-1	0	0	0	0	0	0	0	0
A8	0	0	1	0	-1	0	0	0	0	0	0	0	0
A9	0	0	-1	0	1	0	0	0	0	0	0	0	0
A10	0	0	0	-1	0	0	1	0	0	0	0	0	0
A11	0	0	0	0	-1	1	0	0	0	0	0	0	0
A12	0	0	0	0	1	-1	0	0	0	0	0	0	0
A13	0	0	0	0	0	-1	0	1	0	0	0	0	0
A14	0	0	0	0	0	1	0	0	0	0	-1	0	0
A15	0	0	0	0	0	0	0	-1	1	0	0	0	0
B1	0	0	0	0	0	0	0	0	0	-1	1	0	0
B2	0	0	0	0	0	0	0	0	-1	1	0	0	0
B3	0	0	0	0	0	0	0	0	0	1	0	0	-1
B4	0	0	0	0	0	0	0	0	0	-1	0	1	0
B5	0	0	0	0	0	0	0	0	0	1	0	-1	0
B6	0	0	0	0	0	0	0	0	0	1	0	-1	0
B7	0	0	0	0	0	0	0	0	0	-1	0	0	1
B8	0	0	0	0	0	0	0	0	-1	0	0	1	0
C1	0	0	0	0	0	0	-1	0	0	0	0	0	1
F1	0	0	0	0	0	0	0	0	0	0	1	0	-1

where $T_F = f(F_{\text{unif}})$, $T_R = f(F_{\text{unif}})$. The unification factor is defined as $F_{\text{unif}} = f(n_{\text{unif}})$, where n_{unif} is the number of unified elements.

The obtained numerical results are: $T_{F1} = 15000\,\text{h}$ (unified case), $T_{F2} = 10000\,\text{h}$ (non-unified case).

5 Conclusions

The problems of unification of TCHAP are of high priority in the process of creating the promising means of telecommunications. Analysis of the process of hardware unification of airborne and ground equipment of TCHAP and the process of establishing their information compatibility, allowed to establish that the tasks of determining the current state of an object and choosing the control action are the main ones and are the decision-finding problems whose properties allow to treat them as difficult problems, requiring automation.

To obtain the automated solution of the problems of unification and information interaction, it is necessary to have tools for modeling the structure of the TCHAP, the data transmission system and the dynamics of information interaction. The selected modeling tools allow to describe changes in the network structure. The system of products (the coordinated system of rules) can be used as a tool that combines the capabilities of network and logical-dynamic models for modeling the structure and data transmission processes. Analysis of the known modeling approaches allowed us to determine that the apparatus of Petri nets should be used to describe the structure of the TCHAP and the information transmission system, and the E-network apparatus - should be used to describe the structure of information interaction. When solving problems of unification, along with the structure description, it is necessary to have means of describing its changes. For this purpose, the apparatus of logical-dynamic models is effective. The joint use of logical-dynamic models with network models allows to formalize the information necessary for solving the problems of hardware unification of the components of TCHAP and ensuring their information compatibility.

References

1. Vishnevsky, V.M., Tereshchenko, B.N.: Principles of construction and implementation of tethered high-altitude telecommunication platforms using small rotorcrafts (In Russian). In: Proceedings of Distributed Computer and Communication Networks. Theory and Applications (DCCN-2009, Moscow), pp. 102–116. R&D Company "Information and Networking Technologies" (2009 in Russian)
2. Vishnevsky, V.M., Tereshchenko, B.N., Shabayev, V.I.: Patent for utility model No. 70067: high-altitude rotorcraft for wireless data transmission networks/registered in the state register of utility models of the Russian Federation on 10 January 2008
3. Aminev, D., Zhurkov, A., Polesskiy, S., Kulygin, V., Kozyrev, D.: Comparative analysis of reliability prediction models for a distributed radio direction finding telecommunication system. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds.) DCCN 2016. CCIS, vol. 678, pp. 194–209. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-51917-3_18
4. Aminev, D.A., Kozyrev, D.V., Zhurkov, A.P., Romanov, A.Y., Romanova, I.I.: Method of automated control of distributed radio direction finding system. In: 2017 Dynamics of Systems, Mechanisms and Machines (Dynamics), Omsk, Russia, pp. 1–9 (2017). https://doi.org/10.1109/Dynamics.2017.8239426

5. Vishnevsky, V.M., Tereshchenko, B.N., Shabaev, V.I.: Patent for invention No. 2319319: method of formation of wireless information transmission networks and a high-altitude rotorcraft for its implementation/Registered in the state register of inventions of the Russian Federation on 10 March 2008
6. Smolyakov, A.V., Fedorovich, O.E.: System simulation of the main characteristics of unmanned aerial systems. Aviat. Space Equip. Technol. **5**(31), 39–42 (2006). (in Russian)
7. Majnika, E.: Optimization Algorithms on Networks and Graphs. Mir, Moscow (1981). 324 p. (in Russian)
8. Reznikov, B.A.: System analysis and methods of system engineering, Part 1. Methodology of system studies. In: Modeling of Complex Systems. USSR Ministry of Defense (1990 in Russian)
9. Rostopchin, V.V., Fedin, S.I.: The use of unmanned aerial vehicles in the fight against the spread of narcotic substances (2006 in Russian)
10. Rostopchin, V.V.: Modern classification of unmanned military aviation systems. Electron.: Sci. Technol. Bus. (4) (2009, in Russian)
11. Foreign military unmanned aerial vehicles and prospects for their development [Electronic resource]. http://aviation.gb7.ru/UAVs.htm
12. Mobile multi-purpose unmanned systems for aerial reconnaissance, security and video surveillance. Ultra-light aviation. Hovercrafts [Electronic resource]. http://www.kbvzlet.com/
13. Unmanned Aerial Vehicles and Targets. Issue Twenty-Three, November 2004
14. Zykov, A.A.: Fundamentals of Graph Theory. Nauka, Gl.red.fizmat. lit., Moscow (1987)
15. GOST 1.1-2002. Interstate system of standardization. Terms and Definitions
16. Slyusar, V.: Data transfer from the UAV: NATO standards. ELECTR.: Sci. Technol. Bus. **3** (2010, in Russian)
17. STANAG 7023/AEDP-9 NATO Primary Image Format. www.nato.int/-structur/AC/224/standard/7023/-7023.htm
18. STANAG 4586 (EDITION 3) - Standard interfaces of UAV control system (UCS) for NATO UAV interoperability

On Cyber-Security of Information Systems

Manfred Sneps-Sneppe[1(✉)], Vladimir Sukhomlin[2], and Dmitry Namiot[2]

[1] Ventspils International Radio Astronomy Centre, Ventspils University College, Inzenieru 101a, Ventspils 3601, Latvia
manfreds.sneps@gmail.com
[2] Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, GSP-1, 1-52, Leninskiye Gory, Moscow 119991, Russia
sukhomlin@mail.ru, dnamiot@gmail.com

Abstract. In this paper, we discuss cyber-security issues for a digital economy. In particular, this paper targets the complexity of software development for information systems and related problems. This article provides the review of development-related problems for the biggest information systems in the world. We discuss Global Services Management-Operations, DISN, MFSS, and JRSS. This article concentrates on the creation (description) of the system architecture. In our opinion, this is the main issue for large systems. In this connection, the paper discusses the concept of "enterprise architecture" and the framework (model) for Information Systems Architecture proposed by A. Zachman. The main idea of the model is to provide the possibility of the sequential description of each individual aspect of the system in coordination with all the others. Also, we discuss DoDAF metamodel. Any model of enterprise architectures should cover, of course, cyber-security aspects.

Keywords: Internet of Things · Search · Services

1 Introduction

This work is a continuation of our proposals for the digital economy of Russia [1] and is related to the discussion of the complexities of software development for information systems.

Let's start with an example that is very typical for complex information systems that have been built in the world for more than 30 years, mainly based on the Zachman model [2], which, in turn, goes back to IBM. According to Zachman's model, many attempts have been made in the world to create complex information systems - from the largest corporations to the so-called electronic governments.

As an example, we have chosen a project to modernize the management of Cyber-Security of the DISN network (Defense Information System Network) of

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 201–211, 2018.
https://doi.org/10.1007/978-3-319-99447-5_17

the US Department of Defense (DOD). Chosen due to the fact, that this project is built at the expense of taxpayers and therefore has a lot of open information. In addition, it is under control by the US Government Accountability Office (GAO), whose reports are also publicly available.

We give a brief history of the question.

June 2012. Lockheed Martin won the largest tender for the development of IT services for managing the DISN (Global Services Management-Operations, GSM-O). The essence of the GSM-O contract is the modernization of the DISN management system for cyber security requirements. The cost of work is a huge amount - 4.6 billion dollars for 7 years.

In 2013, the GSM-O team began to study the status of the four GIG network management centers, which are responsible for the maintenance and uninterrupted operation of all Pentagon computer networks - 8,100 computer systems in more than 460 locations in the world, which in turn are connected by 46,000 cables. The first step was to upgrade the computer network management. It was decided to consolidate the operating centers - from four to two. The centers are expanding at the air bases Scott (Illinois) and Hickam in Hawaii, but the centers in Bahrain and Germany are being closed.

2015. Two years later, the telecommunications world shook the news: Lockheed Martin does not cope with the upgrade of DISN management, that is, with the implementation of a multi-billion dollar GSM-O contract, and sells its LM Information and Global Solutions division to the competing Leidos company. The failure of the work was, most likely, due to the inability to recruit developers.

July 2016. A report GAO-16-593 [4] has appeared, which requires more control overspending of funds for the creation of the Joint Information Environment (JIE) of Pentagon.

November 2016. Defense Department Chief Information Officer Terry Halvorsen is attacking the GAO [5], they do not understand what the Joint Information Environment is: “What we tried to say to the GAO is that in this case do not measure JIE - you should measure its components.” JIE, he said, is the conceptual term, used to describe the modernized infrastructure of the DOD IT infrastructure, rather than one program. JIE consists of various programs, such as JRSS security stacks and partner country environments.

February 2017. Nonetheless, DOD’s response in the GAO audit indicates it will take steps to address all GAO’s recommendations regarding JIE.

January 2018. In January, the Pentagon’s chief weapons tester said the Department of Defense should stop deploying its new network security platform, known as Joint Regional Security Stacks until the JRSS demonstrates that it is capable of helping network defenders to detect and respond to operationally realistic cyber-attacks.

February 2018. Pentagon leadership does not give up [6]. The Defense Department’s Joint Regional Security Stacks initiative to increase information network security is more than halfway complete.

Once again, the tremendous scope of JIE's activities was emphasized [7]. The huge scale of the JIE undertaking incorporates the U.S. military's 65,000 servers, and 7 million endpoints – all connected to 15,000 different networks – used by DOD's 1.3 million military active duty and 742,000 civilian personnel, who are based at more than 555,000 facilities scattered across the globe.
So, the Pentagon continues to spend its multi-billion dollar budget on the development of IT infrastructure.

Further, in Sect. 2 we explain the essence of the Zachman model, and in Sect. 3 - what is Joint Information Environment (JIE). Sections 4 and 5 describe the DISN network: the target architecture and the MFSS softswitch, as the basis for switching to the packet switched network. The JRSS equipment and its implementation difficulties are described in Sect. 6.

2 What Is Joint Information Environment (JIE)

In 1987 there was an article by J.A. Zachman "A Framework for Information Systems Architecture" and for the first time introduced the concept of "enterprise architecture" [1]. John Zachman for many years engaged in the implementation of IBM information systems and rethought their architecture. He became the "father" of the enterprise architecture.

The main idea of the model is to provide the possibility of a sequential description of each individual aspect of the system in coordination with all the others [8]. For any sufficiently complex system, the total number of links, conditions, and rules usually exceeds the possibilities for simultaneous consideration. At the same time, separate, in isolation from others, consideration of each aspect of the system often leads to suboptimal decisions - in terms of productivity and the cost of implementation.

J. Zachman defined the architecture of the enterprise as a set of descriptive ideas (models) that are applicable to the description of the Enterprise in accordance with the requirements of management personnel (quality) and which can develop over a certain period (dynamism). The term "architecture" is not accidental here, it emphasizes the existing analogy between the internal structure of an abstract object - an enterprise and a complex artificial object, such as a building or an orbiting international space station (ISS).

Actually, the model is represented in the form of a table having five rows and six columns. Note that there were five lines in the original Zachman model. The sixth line appeared later, it goes beyond the description of the architecture, but describes the operating system or the enterprise as a whole.

So, Zachman's model contains six levels. At each of these levels, participants look at the same questions, corresponding to the columns in the table, but with different levels of abstraction and detail.

The first line corresponds to the level of business planning in whole (business model). At this level, fairly general basic concepts that define the very essence of business (products, services, customers) are introduced, and a business strategy is formulated. In fact, this line defines the context of all subsequent lines.

The second line (the conceptual model) is designed to define (in terms of managers) the structure of the organization, key and auxiliary business processes.

The third level (logical model) corresponds to consideration from the point of view of the system architect. Here, business processes are described already in terms of information systems, including different types of data, the rules for their transformation and processing to perform certain at the level of business functions.

Levels from the fourth and further describe details that are of interest to IT managers and designers, but the leading role is played by developers.

At the fourth level - the technological (physical) model, the data and operations over them are linked to the selected implementation technologies. For example, here you can define the choice of a relational DBMS, or tools for working with unstructured data, or an object-oriented environment.

The fifth level corresponds to the detailed implementation of the system, including specific hardware models, network topology, the manufacturer and version of the DBMS, development tools and the actual program code itself. Many of the works at this level are often performed by subcontractors.

The sixth level describes the operating system. In Zachman's original work, the content of this level is not detailed. At this level, you can enter objects such as instructions for working with the system, the actual databases, the operation of the HelpDesk service, and so on.

Following Zachman's model, the Joint Information Environment (JIE) of DISN is represented as a seven-level model (Fig. 1) [9]. To date, it is only partially standardized. A decision has been made regarding unified communication networks (lower level): it is a wide area IP backbone network and MPLS protocol (multiprotocol label switching protocol). The Unified Capabilities Reference Architecture is fixed, which refers to the level of the Information Distribution Service. The most important and laborious work in creating the Single Information Environment is the standardization of the application layer (Enterprise and

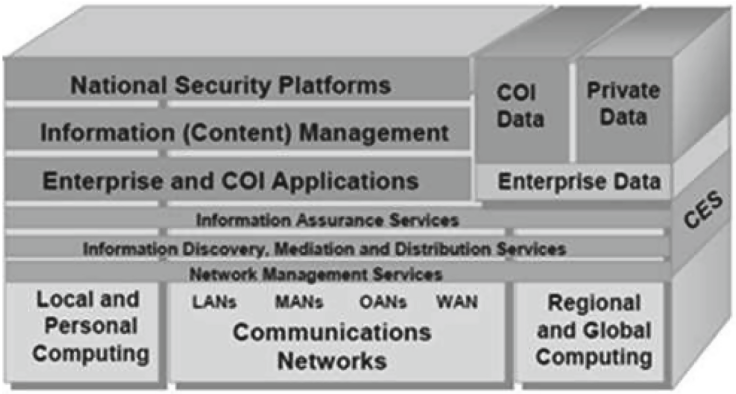


Fig. 1. The target architecture of the JIE.

COI applications, Community-of-Interest). Levels 2 through 4 are provided by CES (Core Enterprise Services).

Based on the Zachman model, the NIST Institute has developed a FEAF (Federal Enterprise Architecture Framework) e-government model for the US federal government. From the Zachman model, only the first three columns were taken and the developers focused on the first three lines of the model.

The architecture of the federal organization is an attempt to bring countless agencies (ministries) of the US federal government to a single and widely used architecture. From the point of view of FEA, the enterprise architecture consists of separate segments. Segment is one of the main aspects of business, for example, labor resources. Segments are divided into two types: basic and service.

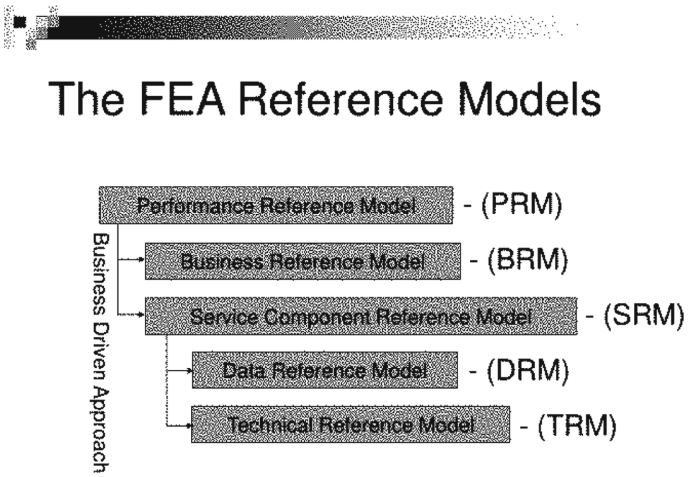


Fig. 2. The FEA reference model.

The core segment is a key aspect of the enterprise's activities within the boundaries of political-administrative division. For example, for the US Department of Health and Human Services, the basic segment is health.

The service segment is a segment that is fundamental, if not for everyone, then for most political organizations. For example, financial management is an official segment, mandatory for all federal agencies.

Another type of assets in the enterprise architecture are the services of the enterprise. The enterprise service is a clearly defined function within the boundaries of a political-administrative division. As an example of the service of an enterprise, we can see the management of security. This is a service that is uniformly implemented throughout the enterprise.

The difference between services and business segments, especially service segments, is not obvious. Both services and segments cover the entire enterprise. The difference lies in the fact that the scope of service segments extends only

to one political organization. The scope of the enterprise services is extended to the whole enterprise.

For example, both in the Ministry of Health and Social Services and in the Environmental Protection Agency of the US Federal Government, the workforce is used in the service segment. However, the labor resources for the Ministry of Health and Social Services differ from the workforce for the Environmental Protection Agency. The same goes for the security management service: it is essentially the same and Ministry of Health and Social Services, and the Agency for Environmental Protection uses this. Effective account management for secure access is provided only if it is carried out at the enterprise level.

Unfortunately, the results of the development of the FEA program were not encouraging. In the official report of the Federal Accounts Chamber for the US Congress on the status of the FEA program in 2002, it was concluded that "in general, the FEA system is not sufficiently developed to make informed investment decisions in the IT field". In addition, the FEA program was extremely expensive. For example, "by the end of 2010, the federal government spent more than a billion dollars on corporate architecture, and much, if not most, of it was wasted".

Following Zachman's model, the DoDAF metamodel (Department of Defense Meta-Model) [10] is being built. It has been developed since 1990. In Fig. 2 is given version 2.0 (is being developed since 2009). The Fig. 3 illustrates the relationship between the basic concepts of the DoDAF metamodel.

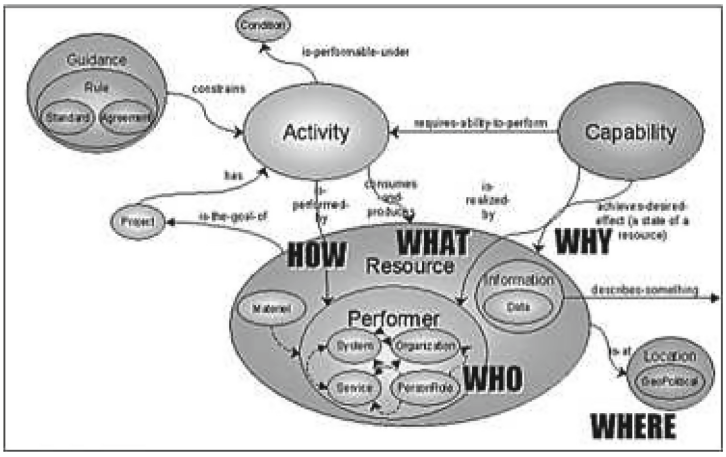


Fig. 3. The DoDAF metamodel

The model contains six descriptions, which are united by the key concept Activity:

1. Description of data (Data Description) - answers the question What;

2. Function Description - answers the question How (also contains the description of Performer);
3. Network Description – Where;
4. Description of participants (People Description) - Who (which includes Organizations);
5. Time Description – When;
6. Motivation Description - Why (with an extension that includes a description of Capability requirements).

The JIE documentation consists of 52 volumes and is presented from eight Viewpoints [11]:

- All Viewpoints – 2 volumes,
- Capability Viewpoint – 7 volumes,
- Data and Information Viewpoint – 3,
- Operational Viewpoint – 9,
- Project Viewpoint – 3,
- Services Viewpoint – 13,
- System Viewpoint – 13,
- Standard Viewpoint – 2.

The development of DoDAF has been going on for more than 25 years, but it can not be completed. The conclusion is that the very idea of creating a single information system for such a complex enterprise as the US Department of Defense is today an impossible task or the Zachman's model development method itself is erroneous.

3 The New DISN Architecture

Currently, a new DISN network is under construction. This is a packet switching network (over IP protocols). Its target architecture contains two levels: Tier 0 and Tier 1. The Tier 0 cluster is responsible for the invulnerability of the entire DISN network. Each Tier 0 cluster contains three softswitches (routers). Tier 0 level communicates by ICCS Protocol (Intra-Cluster Communication Signaling), which is automatically updated databases. A cluster is essential as one distributed softswitch. It is required that the latency in the exchange of database contents does not exceed 40 ms. Since the signal transmission takes 6 ms per 1 km, the distance between softswiches cannot exceed 6,600 km. On the below level, two types of local network are: the protected ASLAN (protocol AS-SIP based) and traditional LAN for H.323 protocol. Thus, the protected hybrid network DISN provides voice and video transmission (Voice and Video over Internet protocol, CVVoIP).

The peculiar "birthmark" of the DISN network, built on a single AS-SIP protocol, is the top-secret government link DRSN (Defense RED Switch Network), which preserves the technology of switching channels, more precisely, ISDN channels and ISDN PRI and CAS signaling protocols (Channel Associated Signaling). The DISN training materials [12] do not provide for the transfer of the DRSN network to packet switching.

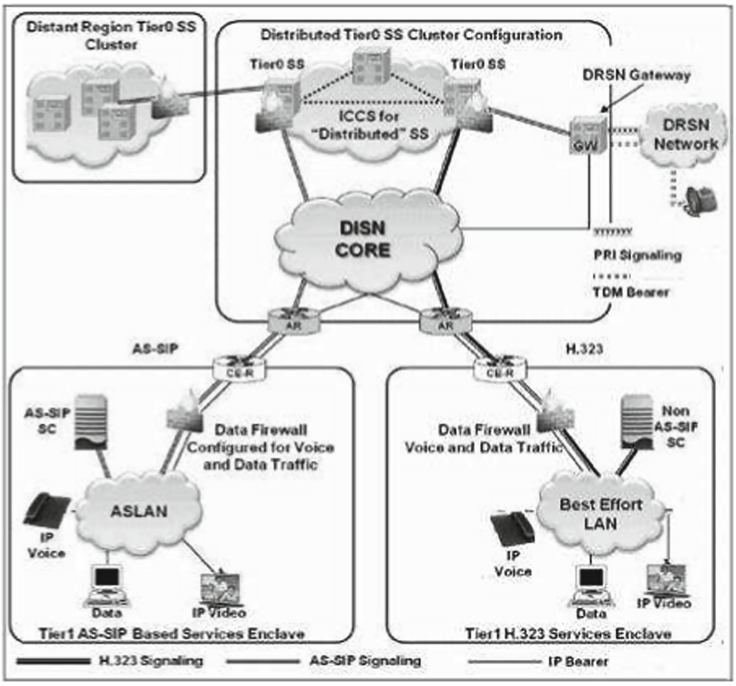


Fig. 4. The target architecture of the DISN communication network.

4 MFSS - The Basis for Transition to the Packet Switching Network

The switch from the circuit switched network, where the SS7 protocol now rules, to packet switching and SIP (or to its secured version AS-SIP) requires the installation of SoftSwitch gateways [13]. Software switches (routers) become the main element of the architecture of the DISN.

The company CISCO - the largest contractor of the Pentagon - installed 22 large Softswitches at military bases around the world. There are two types of top-level softswitches: WAN SS = Wide Area Network SoftSwitch and MFSS = MultiFunction SoftSwitch. Figure 4 also shows the four Global Network Support Center (GNSC) - two in the US (at the Scott and Hawaii airbases), in Germany and Bahrain, which was discussed at the beginning of the article.

5 The Equipment of Cyber Defense JRSS

The main task of the Pentagon's Cyber Command is to ensure the cybersecurity of the JIE, and the regional security stacks (JRSS) play a key role in this. JRSS equipment, in fact, are IP-routers with a complex set of cyberprotection programs.

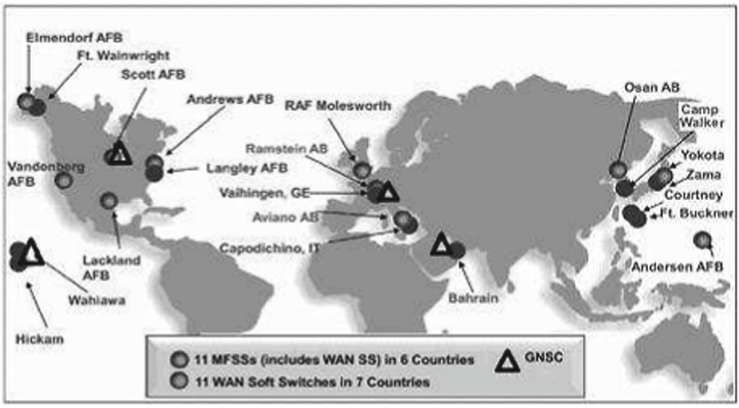


Fig. 5. Plans to install 22 large programmable switches (Softswitch) [14].

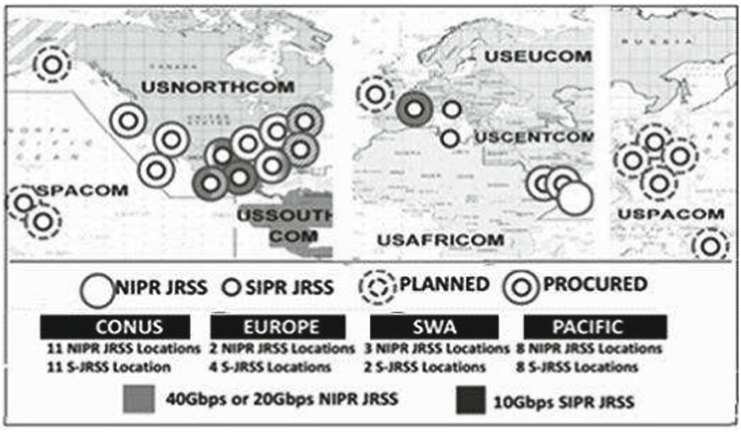


Fig. 6. JRSS stack installation map [17].

Currently, JRSS stacks are installed for the NIPRNet (Non-classified Internet Protocol Router Network) network. This is a network used to exchange unclassified but important service information between "internal" users. It is planned also to install the stacks for the SIPRNet (Secret Internet Protocol Router Network) network. This is a system of interconnected computer networks used by DOD to transmit sensitive information over TCP/IP protocols. The first JRSS stack was installed and successfully operated at the military base of San Antonio, Texas. In 2014, 11 JRSS stacks were installed in the United States, 3 stacks in the Middle East and one in Germany. The state of affairs under the GSM-O contract for 2014 is well described in the article [15].

The total amount of works includes the installation of 24 JRSS stacks on the NIPRNet service network and 25 JRSS stacks on the secret SIPRNet network

(Fig. 5). By 2019, it is planned to transfer to these stacks cybersecurity programs, which are now deployed in more than 400 locations.

Several versions of the JRSS software have already been developed. The complexity of cyberprotection software is shown in Figs. 6 and 7.

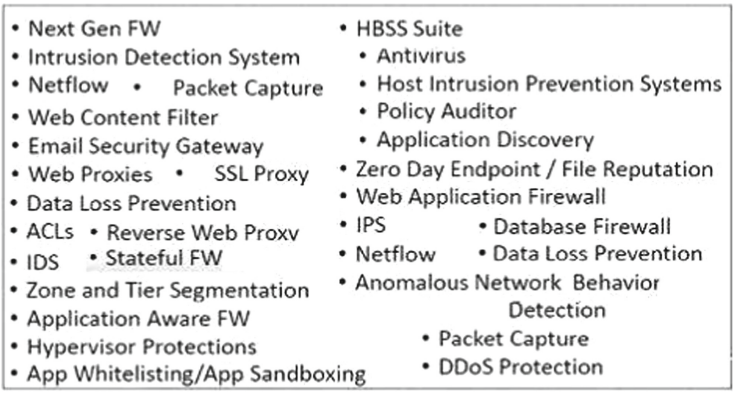


Fig. 7. Cyber defense in the SIPRNet network in accordance with a unified security architecture [18].

Will the Pentagon's grandiose plans be fulfilled? The complexity of the task, in particular, characterizes the set of requirements for potential JRSS developers, named in the invitations to work for Leidos [19]. Requires work experience of 12–14 years and knowledge of at least two or more products from ArcSight, TippingPoint, Sourcefire, Argus, Bro, Fidelis XPS, Niksun FPCAP, Lancope, NetCool, InfoVista, and Riverbed. We note that each of these companies provides its complex software of cyber defense. How to combine them?

6 Summary

American communications technology for the needs of the military passed three generations of transformation: from signaling SS7 and intelligent networks (Joint Vision 2010) to IP protocol (Joint Vision 2020) and, finally, to the extremely ambitious plans of cybersecurity of networks. Cybersecurity targets are the Pentagon's top priority, but the lack of necessary standards hampers the implementation of the cybersecurity program. The failure was most likely also due to the inability to recruit developers capable of combining the "old" circuit switching equipment with the latest packet switching systems.

We recall the GAO claims to the state of affairs for JIE and JRSS (from the report GAO-16-593 mentioned above). Without a strategy that identifies the resources needed to implement the JIE strategy and the timetable for completing the evaluation, the GAO does not have the confidence that the necessary assessments will be completed.

References

1. Sneps-Sneppe, M., Sukhomlin, V., Namiot, D.: On the program "digital economy of the russian federation": how to create an information infrastructure. Int. J. Open Inf. Technol. **6**(3), 37–48 (2018)
2. Zachman, J.A.: A framework for information systems architecture. IBM Syst. J. **26**(3), 276–292 (1987)
3. Leidos-Lockheed merger changes the face of federal IT. https://www.federaltimes.com/it-networks/2016/02/05/leidos-lockheed-merger-changes-the-face-of-federal-it/. Accessed Apr 2018
4. Joint Information Environment: DOD Needs to Strengthen Governance and Management. GAO-16-593, 14 July 2016
5. Pentagon Tech Chief Says He'll 'Take the Hit' for GAO Criticism of JIE, 02 November 2016. http://www.nextgov.com/cio-briefing/2016/11/pentagon-tech-chief-says-hell-take-hit-gao-criticism-jie/132882/
6. Dod CIO: JRSS set for 2019 completion, 05 March 2018. https://fcw.com/articles/2018/03/05/jrss-completion-miller.aspx
7. The Department of Defense. Strategy for Implementing the Joint Information Environment, 18 September 2013
8. Primenenie modeli Zahmana dlja proektirovanija IT-arhitektury predprijatija. http://www.management.com.ua/ims/ims177.html. Accessed Apr 2018
9. DeVries, D.: DoD joint information enterprise. http://c4i.gmu.edu/eventsInfo/reviews/2013/pdfs/AFCEA2013-DeVries.pdf. Accessed Apr 2018
10. Department of Defense. Information Enterprise Architecture (DoD IEA). Version 2.0. vol. II - IEA Description, July 2012
11. DoDAF. http://dodcio.defense.gov/Portals/0/Documents/DODAF/DoDAF-v2-02/web.pdf. Accessed Apr 2018
12. Department of Defense. Unified Capabilities Framework 2013, January 2013
13. Department of Defense. Unified Capabilities Master Plan, October 2011
14. https://www.cisco.com/web/strategy/docs/gov/Cisco-LSC-Overview-Jan2011.pdf . Accessed Apr 2018
15. Meloni, S.: The future of the joint information environment (JIE), 24 September 2014. http://blog.immixgroup.com/2014/09/24/the-futureof-the-joint-information-environment-jie
16. The JRSS program is underway, 01 October 2014. http://archive.c4isrnet.com/article/20141001/C4ISRNET12/310010005/The-JRSS-program-underway. Accessed Apr 2018
17. Welsh, W.: New tools ahead for DOD's global grid, 14 September 2015. https://gcn.com/articles/2015/09/14/dod-global-information-grid.aspx. Accessed Apr 2018
18. Metz, D.: Joint Information Environment Single Security Architecture (JIE SSA), 12 May 2014
19. Cyber Systems Training Support Engineer - JRSS. https://www.energyjobline.com/job/571137/cyber-systems-training-support-engineer-jrss/. Accessed Apr 2018

Inventory Management System with Two-Switch Synchronous Control

Anatoly Nazarov, Valentina Broner(✉), and Alexander Moiseev

Tomsk State University, Lenina ave., 36, Tomsk 634050, Russia
nazarov.tsu@gmail.com, valsubbotina@mail.ru, moiseev.tsu@gmail.com

Abstract. The paper presents mathematical model of multi-period inventory management system with stochastic demands and two-switch synchronous control with two thresholds. The control is realized as changing the intensities of the input and output product flows depending on the amount of the accumulated resource in the system. An explicit expression for probability density function of inventory level is found for this system. Also, an example of applying the inventory management models to analysis of cloud computer service is suggested in the paper. Finally, some numerical examples are given to illustrate the obtained results.

Keywords: Inventory management · Switch control
Cloud computing

1 Introduction

The inventory management theory is usually used to solve various problems of economics, management, business, logistics, etc. But mathematical methods of the theory may be also applied to solve many problems in computer science, e.g., in cloud computing. The problems of optimization of task queues and execution, virtual machines allocation, power saving are very important for cloud computing [1–4]. The solutions to this problems are mainly presented in the form of heuristic recommendations or are based on the results of simulations [5–7]. There are also few works devoted to analytical research in this field [8,9]. A detailed overview of the issues can be found in [10].

The main difficulty of analyzing such systems is the fact that the cloud service is a complex network of servers with a complicate internal organization including both topology and software. If we consider a single server of a cloud service, then the problem may be simplified and analytical methods of analysis, such as inventory management theory, may be applied to its solution. This is discussed in details in Sect. 2 of the paper. Using inventory management approach we may not only analyze the system evolution but obtain appropriate control parameters for handling of the system workload.

In this paper we consider multi-period inventory management model with Poisson input and output product flows. Multi-period systems were considered

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 212–223, 2018.
https://doi.org/10.1007/978-3-319-99447-5_18

in [11–17]. Models with On/Off control and constant or piecewise-constant rate of input flow were discussed in [12–16]. Models of queuing system with hysteretic overload control were discussed in [18,19]. The main difference from these works is the input and output flows that occur according to Poisson process with piecewise-constant intensity. Switching of intensity is performed synchronously both for input and output flows. Using two thresholds for regulating intensity of input and output we can reach the desired stock level of resource.

The goal of the study is to obtain explicit expression for probability density function of stock level process in inventory management system with two-switch synchronous control.

2 Mathematical Model

Consider the following problem. Tasks arrive at a single server of a cloud service (resource arrivals in the mathematical model). In the server, they are queued and have to be executed. The completion of task execution is corresponds to a demand occurring in the mathematical model. Each arriving resource brings some volume into the system. This volume may be interpreted as a job which should be done or as a number of parallel processes that should be executed on cloud service processor devices.

When the critical load level of the server is reached, the cloud service redistributes the workload of its servers in such a way as to unload this server. So, the arrival rate of incoming tasks became lower for this server. On the contrary, when there are few executing tasks in this server, the cloud service redistributes the workload of its servers in such a way as to load this server more. We call the load level of the server as a stock level in the mathematical model.

So, let us consider the following mathematical model in the form of inventory management system with two-switch synchronous of input and output flows (Fig. 1).

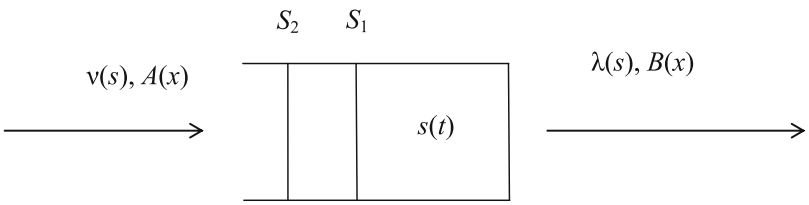


Fig. 1. Resource management model

We denote the stock level accumulated in the system at the time t as $s(t)$. S_1 and S_2 are fixed threshold values of resource. Lets resource arrive in the system according to a Poisson process with piecewise-constant intensity $\nu(s)$, where s is a value of the process $s(t)$. We assume that the process $s(t)$ can take negative

values, i.e. the customer waits for the arrival the required amount of resource. If the stock level $s(t)$ accumulated in the system is below threshold value S_1, then intensity of input flow is equal to ν_1. If the inventory level is above threshold value S_1 but below S_2, then we switch intensity of input flow to ν_2. In the other case, when $s > S_2$, intensity is equal to ν_3:

$$\nu(s) = \begin{cases} \nu_1, s < S_1, \\ \nu_2, S_1 < s < S_2, \\ \nu_3, s > S_2. \end{cases} \tag{1}$$

Demands occur according to a Poisson process with intensity $\lambda(s)$ which also has different values determined by the thresholds S_1 and S_2

$$\lambda(s) = \begin{cases} \lambda_1, -\infty < s < S_1, \\ \lambda_2, S_1 < s < S_2, \\ \lambda_3, s > S_2. \end{cases} \tag{2}$$

The batch sizes of input and output flows are independent, identically distributed random variables. In this paper we consider exponential distribution $A(x)$ and $B(x)$ of batch sizes

$$A(x) = 1 - e^{-\alpha x}, B(x) = 1 - e^{-\beta x}. \tag{3}$$

The condition of existence of steady-state regime is determined by inequalities

$$\beta\nu_1 > \alpha\lambda_1, \beta\nu_3 < \alpha\lambda_3. \tag{4}$$

In the paper we allow some simplifications to the model of real cloud service. For example, we do not consider changes in the intensity of total task flow to the service, features of its topology and software. In addition, for theoretical derivations we assume that process $s(t)$ may have negative values. This allows us to obtain analytical results. In practice, we can apply derived expressions as an approximation for workload parameters by setting such values of ν_1, S_1, and λ_1 that make the probability of states $s(t) < 0$ be nearby zero (e.g., $\lambda_1 = 0$).

3 Main Results

According to the mathematical model, process $s(t)$ is Markovian in continuous time t and has a continuous set of values $-\infty < s < \infty$. Denoting

$$P(s) = \frac{\partial P\{s(t) < s\}}{\partial s},$$

we can write the following Kolmogorov equation:

$$\begin{aligned} (\nu(s) + \lambda(s))\, P(s) = &\int_0^\infty \nu(s-x)P(s-x)dA(x) \\ &+ \int_0^\infty \lambda(s+x)P(s+x)dB(x), \end{aligned} \tag{5}$$

which solution $P(s)$ satisfies boundary conditions

$$P(-\infty) = P(\infty) = 0. \tag{6}$$

Let us prove the following statement.

Theorem 1. *If $A(x)$ and $B(x)$ are exponential distributions (3) and $\beta\nu_2 \neq \alpha\lambda_2$, then solution $P(s)$ of Eq. (5) has the form*

$$P(s) = \begin{cases} P_1(s) = C_1 e^{\gamma_1 s}, -\infty < s < S_1, \\ P_2(s) = C_2 e^{\gamma_2 s}, \ S_1 < s < S_2, \\ P_3(s) = C_3 e^{\gamma_3 s}, \ S_2 < s < \infty, \end{cases} \tag{7}$$

where

$$\gamma_i = \frac{\beta\nu_i - \alpha\lambda_i}{\lambda_i + \nu_i}, i = 1, 2, 3, \tag{8}$$

$$C_2 = \left[e^{\gamma_2 S_1} \left\{ \frac{\nu_2 + \lambda_2}{\gamma_1 (\nu_1 + \lambda_1)} - \frac{1}{\gamma_2} \right\} + e^{\gamma_2 S_2} \left\{ \frac{1}{\gamma_2} - \frac{\nu_2 + \lambda_2}{\gamma_3 (\nu_3 + \lambda_3)} \right\} \right]^{-1}, \tag{9}$$
$$C_1 = C_2 \frac{\nu_2 + \lambda_2}{\nu_1 + \lambda_1} e^{(\gamma_2 - \gamma_1) S_1}, C_3 = C_2 \frac{\nu_2 + \lambda_2}{\nu_3 + \lambda_3} e^{(\gamma_2 - \gamma_3) S_2}.$$

Proof. Considering case of exponential distributions $A(x)$ and $B(x)$, Eq. (5) can be transformed into the form

$$(\nu(s) + \lambda(s)) P(s) = \alpha e^{-\alpha s} \int_{-\infty}^{s} \nu(z) P(z) e^{\alpha z} dz + \beta e^{\beta s} \int_{s}^{\infty} \lambda(z) P(z) e^{-\beta z} dz. \tag{10}$$

Multiplying Eq. (10) by $e^{\alpha s}$ and differentiating, we derive

$$(\nu(s) + \lambda(s)) P'(s) + (\alpha + \beta)\lambda(s) P(s) = \beta(\alpha + \beta) e^{\beta s} \int_{s}^{\infty} \lambda(z) P(z) e^{-\beta z} dz,$$

Similarly, multiplying equation by $e^{-\beta s}$ and differentiating, we obtain

$$(\nu(s) + \lambda(s)) P''(s) + (\alpha\lambda(s) - \beta\nu(s)) P'(s) = 0.$$

One may obtain a solution of this equation as follows

$$P(s) = \begin{cases} P_1(s) = C_1 e^{\gamma_1 s}, & -\infty < s < S_1, \\ P_2(s) = C_0 + C_2 e^{\gamma_2 s}, & S_1 < s < S_2, \\ P_3(s) = C_3 e^{\gamma_3 s}, & S_2 < s < \infty, \end{cases} \tag{11}$$

where

$$\gamma_i = \frac{\beta\nu_i - \alpha\lambda_i}{\lambda_i + \nu_i},$$

$i = 1, 2, 3$, and $P_1(-\infty) = P_3(\infty) = 0$.

To determine the constants C_0, C_1, C_2, C_3, we substitute expression (11) into Eq. (5) and consider obtained equation on the intervals $s < S_1$, $S_1 < s < S_2$, and $s > S_2$:

– Case $S_1 < s < S_2$:
We obtain

$$(\nu_2+\lambda_2)\,P_2(s) = \alpha e^{-\alpha s}\left[\int\limits_{-\infty}^{S_1}\nu_1 P_1(z)e^{\alpha z}dz + \int\limits_{S_1}^{s}\nu_2 P_2(z)e^{\alpha z}dz\right] + \beta e^{\beta s}\left[\int\limits_{s}^{S_2}\lambda_2 P_2(z)e^{-\beta z}dz + \int\limits_{S_2}^{\infty}\lambda_3 P_3(z)e^{-\beta z}dz\right].$$

Substituting (11) into this equation, we derive

$$\begin{aligned} &C_2 e^{\gamma_2 s}\left[\nu_2+\lambda_2-\frac{\nu_2\alpha}{\gamma_2+\alpha}+\frac{\lambda_2\beta}{\gamma_2-\beta}\right] \\ &= \alpha e^{-\alpha s}\left[\frac{\nu_1 C_1}{\gamma_1+\alpha}e^{\gamma_1 S_1}-\frac{\nu_2 C_0}{\alpha}-\frac{\nu_2 C_2}{\gamma_2+\alpha}e^{\gamma_1 S_1}\right]e^{\alpha S_1} \\ &+\beta e^{\beta s}\left[-\frac{\lambda_2 C_0}{\beta}+\frac{\lambda_2 C_2}{\gamma_2-\beta}e^{\gamma_2 S_2}-\frac{\lambda_3 C_3}{\gamma_3-\beta}e^{\gamma_3 S_2}\right]e^{-\beta S_2}. \end{aligned}$$

Equating coefficients of linear combinations of exponents $e^{\gamma_2 s}, e^{-\alpha s}$ and $e^{\beta s}$ to zero, we obtain system of equations for C_0, C_1, C_2, C_3:

$$\begin{cases} \nu_2+\lambda_2-\dfrac{\nu_2\alpha}{\gamma_2+\alpha}+\dfrac{\lambda_2\beta}{\gamma_2-\beta}=0, \\ \dfrac{\nu_1 C_1}{\gamma_1+\alpha}e^{\gamma_1 S_1}-\dfrac{\nu_2 C_0}{\alpha}-\dfrac{\nu_2 C_2}{\gamma_2+\alpha}e^{\gamma_1 S_1}=0, \\ -\dfrac{\lambda_2 C_0}{\beta}+\dfrac{\lambda_2 C_2}{\gamma_2-\beta}e^{\gamma_2 S_2}-\dfrac{\lambda_3 C_3}{\gamma_3-\beta}e^{\gamma_3 S_2}=0. \end{cases} \tag{12}$$

Taking into account expression (8) for γ_2, we conclude that the first expression of the system (12) is true for all values of C_0, C_1, C_2, C_3.

– Case $s < S_1$:
From the first equation of system (12)

$$\frac{\lambda_1 C_1}{\gamma_1-\beta}e^{\gamma_1 S_1}+\frac{\lambda_2 C_0}{\beta}-\frac{\lambda_2 C_2}{\gamma_2-\beta}e^{\gamma_2 S_1}=0.$$

– Case $s > S_2$:
It can be shown that

$$\frac{\nu_2 C_0}{\alpha}+\frac{\nu_2 C_2}{\gamma_2+\alpha}e^{\gamma_2 S_2}+\frac{\nu_3 C_3}{\gamma_3+\alpha}e^{\gamma_3 S_2}=0.$$

Thereby we have following system of four equations

$$\begin{cases} \dfrac{\nu_1 C_1}{\gamma_1+\alpha}e^{\gamma_1 S_1}-\dfrac{\nu_2 C_0}{\alpha}-\dfrac{\nu_2 C_2}{\gamma_2+\alpha}e^{\gamma_2 S_1}=0, \\ -\dfrac{\lambda_2 C_0}{\beta}+\dfrac{\lambda_2 C_2}{\gamma_2-\beta}e^{\gamma_2 S_2}-\dfrac{\lambda_3 C_3}{\gamma_3-\beta}e^{\gamma_3 S_2}=0, \\ \dfrac{\lambda_1 C_1}{\gamma_1-\beta}e^{\gamma_1 S_1}+\dfrac{\lambda_2 C_0}{\beta}-\dfrac{\lambda_2 C_2}{\gamma_2-\beta}e^{\gamma_2 S_1}=0, \\ \dfrac{\nu_2 C_0}{\alpha}+\dfrac{\nu_2 C_2}{\gamma_2+\alpha}e^{\gamma_2 S_2}+\dfrac{\nu_3 C_3}{\gamma_3+\alpha}e^{\gamma_3 S_2}=0. \end{cases} \tag{13}$$

In order to simplify the system of equations we use expressions (8), then system (13) have form

$$\begin{cases} \dfrac{C_1(\nu_1+\lambda_1)}{\alpha+\beta}e^{\gamma_1 S_1} - \dfrac{\nu_2 C_0}{\alpha} - \dfrac{C_2(\nu_2+\lambda_2)}{\alpha+\beta}e^{\gamma_2 S_1} = 0, \\ -\dfrac{\lambda_2 C_0}{\beta} + \dfrac{C_2(\nu_2+\lambda_2)}{\alpha+\beta}e^{\gamma_2 S_2} - \dfrac{C_3(\nu_3+\lambda_3)}{\alpha+\beta}e^{\gamma_3 S_2} = 0, \\ -\dfrac{C_1(\nu_1+\lambda_1)}{\alpha+\beta}e^{\gamma_1 S_1} + \dfrac{\lambda_2 C_0}{\beta} + \dfrac{C_2(\nu_2+\lambda_2)}{\alpha+\beta}e^{\gamma_2 S_1} = 0, \\ \dfrac{\nu_2 C_0}{\alpha} + \dfrac{C_2(\nu_2+\lambda_2)}{\alpha+\beta}e^{\gamma_2 S_2} - \dfrac{C_3(\nu_3+\lambda_3)}{\alpha+\beta}e^{\gamma_3 S_2} = 0. \end{cases} \tag{14}$$

Summing up the first and the third equations of the system (14), we obtain

$$C_0\left(\frac{\lambda_2}{\beta} - \frac{\nu_2}{\alpha}\right) = 0. \tag{15}$$

On other hand, summing up the second and the fourth equations, we also obtain Eq. (15). So, analyzing it, we may conclude that the following two cases are possible:

1. $\beta\nu_2 - \alpha\lambda_2 \neq 0$ and $C_0 = 0$, or
2. $\beta\nu_2 - \alpha\lambda_2 = 0$ and $\gamma_2 = 0$.

In first case $P_2(s) = C_2 e^{\gamma_2 s}$, and $P_2(s) = C_0 + C_2$ in the second one.

Considering $\beta\nu_2 \neq \alpha\lambda_2$ and $C_0 = 0$, we obtain

$$\begin{cases} C_1(\nu_1+\lambda_1)e^{\gamma_1 S_1} - C_2(\nu_2+\lambda_2)e^{\gamma_2 S_1} = 0, \\ C_2(\nu_2+\lambda_2)e^{\gamma_2 S_2} - C_3(\nu_3+\lambda_3)e^{\gamma_3 S_2} = 0. \end{cases} \tag{16}$$

Now constants C_1 and C_2 can be expressed as follows

$$C_1 = C_2\frac{\nu_2+\lambda_2}{\nu_1+\lambda_1}e^{(\gamma_2-\gamma_1)S_1},$$

$$C_3 = C_2\frac{\nu_2+\lambda_2}{\nu_3+\lambda_3}e^{(\gamma_2-\gamma_3)S_2}.$$

From $\int\limits_{-\infty}^{\infty} P(s)ds = 1$ follows that:

$$1 = \frac{C_1 e^{\gamma_1 S_1}}{\gamma_1} + \frac{C_2\left(e^{\gamma_2 S_2} - e^{\gamma_2 S_1}\right)}{\gamma_2} - \frac{C_3 e^{\gamma_3 S_2}}{\gamma_3},$$

hence, we obtain

$$C_2 = \left[e^{\gamma_2 S_1}\left\{\frac{\nu_2+\lambda_2}{\gamma_1(\nu_1+\lambda_1)} - \frac{1}{\gamma_2}\right\} + e^{\gamma_2 S_2}\left\{\frac{1}{\gamma_2} - \frac{\nu_2+\lambda_2}{\gamma_3(\nu_3+\lambda_3)}\right\}\right]^{-1}.$$

Thus, we have proved that in the case of parameters $\beta\nu_2 \neq \alpha\lambda_2$ the solution of Eq. (5) has the form (7), which parameters are determined by expressions (8) and (9).

The theorem is proved.

The following theorem can be proved by a similar way.

Theorem 2. *If $A(x)$ and $B(x)$ are exponential distributions (3) and $\beta\nu_2 = \alpha\lambda_2$, then solution $P(s)$ of Eq. (5) has the form*

$$P(s) = \begin{cases} P_1(s) = C_1 e^{\gamma_1 s}, -\infty < s < S_1, \\ P_2(s) = C_2, \quad S_1 < s < S_2, \\ P_3(s) = C_3 e^{\gamma_3 s}, \quad S_2 < s < \infty, \end{cases} \tag{17}$$

where

$$\gamma_i = \frac{\beta\nu_i - \alpha\lambda_i}{\lambda_i + \nu_i}, i = 1, 3, \tag{18}$$

$$C_2 = \left[\frac{\nu_2 + \lambda_2}{\beta\nu_1 + \alpha\lambda_1} + S_2 - S_1 + \frac{\nu_2 + \lambda_2}{\beta\nu_3 + \alpha\lambda_3} \right]^{-1},$$

$$C_1 = C_2 \frac{\nu_2 + \lambda_2}{\nu_1 + \lambda_1} e^{-\gamma_1 S_1}, C_3 = C_2 \frac{\nu_2 + \lambda_2}{\nu_3 + \lambda_3} e^{-\gamma_3 S_2}. \tag{19}$$

Proof. In the proof of Theorem 1 it was shown that the solution of Eq. (5) has the form (11). Taking into account $\beta\nu_2 - \alpha\lambda_2 = 0$, we conclude that $\gamma_2 = 0$ and the solution of Eq. (5) has the form (17), where γ_1 and γ_2 are expressed as (18).

To determine the constants C_1, C_2, C_3, we solve system (14) under the conditions $\gamma_2 = 0$ and $C_0 + C_2 = C_2$. We obtain

$$C_1 = C_2 \frac{\nu_2 + \lambda_2}{\nu_1 + \lambda_1} e^{-\gamma_1 S_1},$$

$$C_3 = C_2 \frac{\nu_2 + \lambda_2}{\nu_3 + \lambda_3} e^{-\gamma_3 S_2}.$$

From $\int_{-\infty}^{\infty} P(s)ds = 1$ it follows that:

$$1 = \frac{C_1 e^{\gamma_1 S_1}}{\gamma_1} + \frac{C_2 (S_2 - S_1)}{\gamma_2} - \frac{C_3 e^{\gamma_3 S_2}}{\gamma_3} = C_2 \left[\frac{1}{\gamma_1} \frac{\nu_2 + \lambda_2}{\nu_1 + \lambda_1} + S_2 + S_1 - \frac{1}{\gamma_3} \frac{\nu_2 + \lambda_2}{\nu_3 + \lambda_3} \right]$$

$$= \left[\frac{\nu_2 + \lambda_2}{\beta\nu_1 + \alpha\lambda_1} + S_2 - S_1 + \frac{\nu_2 + \lambda_2}{\beta\nu_3 + \alpha\lambda_3} \right].$$

Hence, we derive that

$$C_2 = \left[\frac{\nu_2 + \lambda_2}{\beta\nu_1 + \alpha\lambda_1} + S_2 - S_1 + \frac{\nu_2 + \lambda_2}{\beta\nu_3 + \alpha\lambda_3} \right]^{-1}.$$

Thus, in the case of $\beta\nu_2 = \alpha\lambda_2$ we obtain parameters γ_1, γ_2 and constants C_1, C_2, C_3 of the solution $P(s)$ of Eq. (5) in the form (18)–(19).

The theorem is proved.

Theorems proved in this section allow us to find the probabilistic characteristics of the process $s(t)$ such as its mean, variance, probability of stockout and others.

4 Numerical Results

In this section, we discuss the numerical results about considering model.

Let us use the following values of the system parameters:

intensities of input product flow $\nu_1 = 5, \nu_2 = 10, \nu_3 = 3$;
intensities of output flow $\lambda_1 = 3, \lambda_2 = 9, \lambda_3 = 5$;
threshold values $S_1 = 5, S_2 = 7; \alpha = 1, \beta = 1$.

So, we obtain the following values of function $P(s)$ parameters from expressions (7) and (8):

$$\gamma_1 = 0.25, C_1 = 0.024,$$

$$\gamma_2 = 0.53, C_0 = 0, C_2 = 0.035,$$

$$\gamma_3 = -0.25, C_3 = 0.474.$$

The numerical results are illustrated by the left part of Fig. 2.

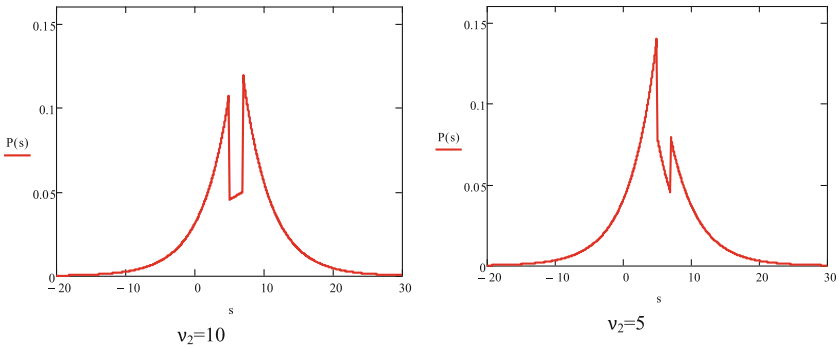


Fig. 2. Inventory management model

It is worth noting, that there are no conditions on the parameters ν_2 and λ_2, so it is interesting to consider the different relation between these parameters. If we change the value of the parameter ν_2 to be equal to 5, then the slope of the function $P_2(s)$ graph will change. Results are illustrated by the right part of Fig. 2.

Although only one parameter of the system has changed, this may affect the behavior of the whole inventory management system. The same effect is presented at Fig. 3. Here we can see changes both in range of values and in the shape of probability density curve.

Let us consider an example close to a cloud computing server functioning when values of parameters affect the probability density function. We consider the system with the following characteristics:

$$\nu_2 = 1, \ \nu_3 = 0.1,$$

$$\lambda_1 = 0,\ \lambda_2 = \lambda_3,$$

$$\alpha = \beta = 1,\ S_1 = 2\nu_2,\ S_2 = 5\nu_2.$$

Values of parameters ν_1 and $\lambda_2 = \lambda_3$ will be varied to obtain a common picture of their influence on the workload of the system. Using such values we control input intensity both to avoid overloading of the system ($\nu_3 = 0.1$) and to start accept additional tasks when the system is close to be empty ($\nu_1 = 5$ or 20). Also using $\lambda_1 = 0$ we try to avoid mock completions of task executing when we have no tasks to be executed ($s(t) \leq 0$).

Results of numerical experiments for the considering model is presented on Fig. 4. We may see big changes in the workload distribution when we vary intensity ν_1 from 1 to 5 but the changes are rather small when we vary this parameter from 5 to 20 (left graphs of the figure). From the right graphs of this figure, we can see that the system workload is very sensitive to its processors performance (parameters λ_2 and λ_3). When we increase these parameters, workload of the system became more controllable - we avoid both an overload of the system and its weak load.

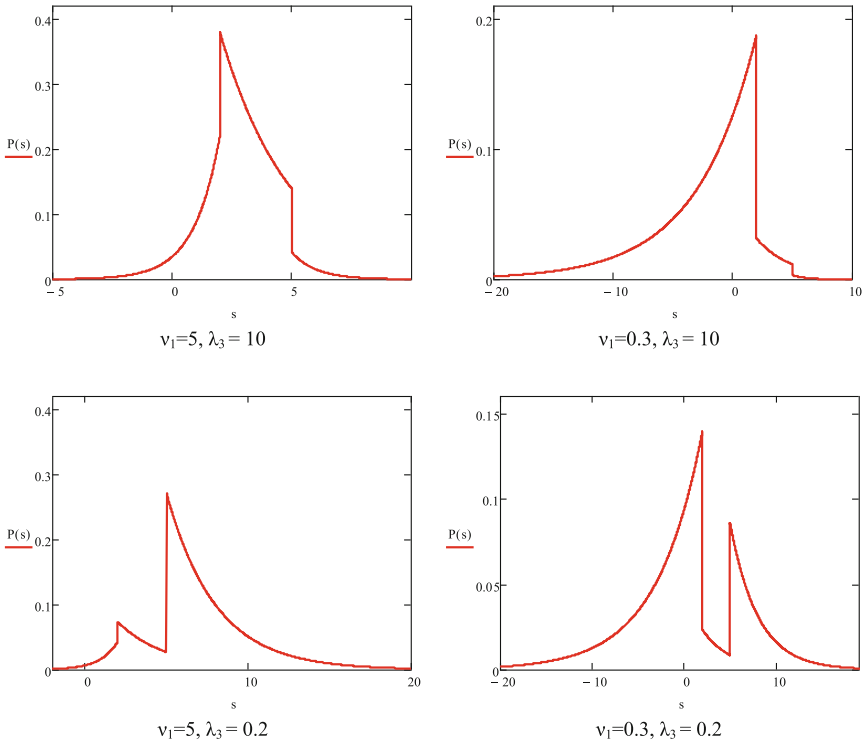


Fig. 3. Changes of the probability density function of process $s(t)$ for various values of parameters ν_1 and λ_3 (here $\nu_2 = 1, \nu_3 = 0.1, \lambda_1 = 0.2, \lambda_2 = 2, \alpha = \beta = 1, S_1 = 2\nu_2, S_2 = 5\nu_2$)

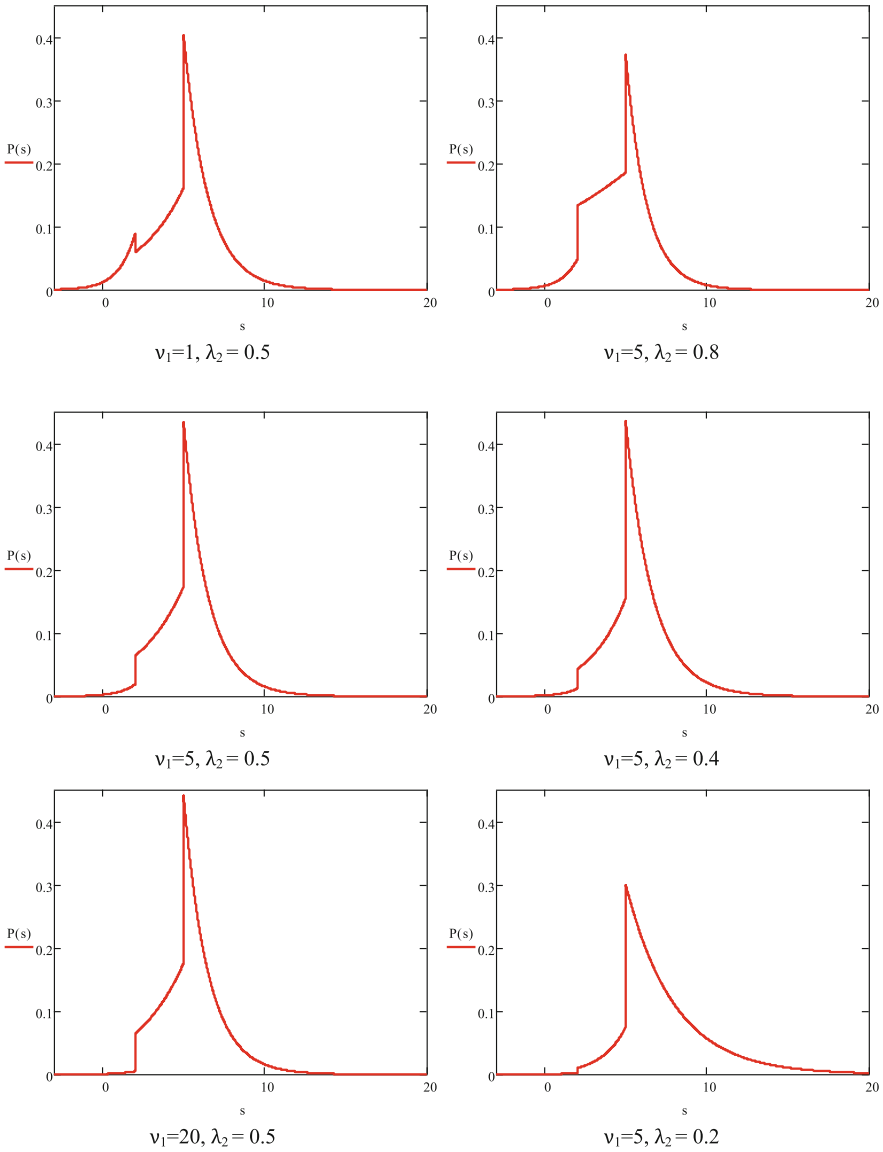


Fig. 4. Changes of the probability density function of process $s(t)$ for the cloud server example

Using such results, we may choose values of the system parameters that provide us an appropriate regime of the server workload.

5 Conclusions

The paper presents a study of the inventory management system with two switch synchronous control. We have found explicit expression for the stationary probability distribution of the stock level accumulated in the system in the case of exponential distributions of batch sizes in the input and output flows. Numerical analysis of the solution shows that small changes in even one parameter of the model can lead us to serious changes in the characteristics of the entire system functioning.

An example of applying the inventory management models to analysis of cloud computer service is suggested in the paper. At the current stage of the research we may only analyze workload of a single server in the service but the approach may be used as a base for future studies.

In the future research, we also would like to consider the inventory management models with two switch synchronous control for the cases of arbitrary probability functions of resource/demand batches and for the multi-threshold switching.

References

1. Patil, S.D., Mehrotra, S.C.: Resource allocation and scheduling in the cloud. Int. J. Emerg. Trends Technol. Comput. Sci. (IJETTCS) **1**(1), 47–52 (2012)
2. Vignesh, V., Sendhil Kumar, K.S., Jaisankar, N.: Resource management and scheduling in cloud environment. Int. J. Sci. Res. Publ. **3**(6), 1–6 (2013)
3. Tian, W., Zhao, Y.: Optimized Cloud Resource Management and Scheduling: Theory and Practice. Elsevier, New York City (2015)
4. da Fonseca, N.L.S., Boutaba, R. (eds.): Cloud Services, Networking, and Management. Wiley Online Library, Hoboken (2015)
5. Abdelsalam, H., Maly, K., Kaminsky, D.: Analysis of energy efficiency in clouds. In: Proceedings of Computation World: Future Computing, Service Computation, Cognitive, Adaptive, Content, Patterns 2009, pp. 416–421. IEEE, Athens (2009)
6. Van, H., Tran, F., Menaud J.-M.: Performance and power management for cloud infrastructures. In: Proceedings of 3rd International Conference on Cloud Computing 2010, pp. 329–336. IEEE, Miami (2010)
7. Mohsenian-Rad, A.-H., Leon-Garcia, A.: Energy-information transmission tradeoff in green cloud computing. In: Proceedings of GLOBECOM 2010. IEEE, Miami (2010)
8. Gelenbe, E., Lent, R., Douratsos, M.: Choosing a local or remote cloud. In: Proceedings of 2nd Symposium on Network Cloud Computing and Applications 2012, London, UK, pp. 25–30 (2012)
9. Kumar, N., Agarwal, S.: An analytical model for dynamic resource allocation framework in cloud environment. Res. J. Recent Sci. **3**(IVC–2014), 1–6 (2014)

10. Sakellari, G., Loukas, G.: A survey of mathematical models, simulation approaches and testbeds used for research in cloud computing. Simul. Model. Pract. Theory **39**, 92–103 (2013)
11. Mousavia, S.M., Hajipoura, V., Niakib, S.T.A., Alikar, N.: Optimizing multi-item multi-period inventory control system with discounted cash flow and inflation: two calibrated meta-heuristic algorithms. Appl. Math. Model. **37**(4), 2241–2256 (2013)
12. Nazarov, A., Broner, V.: Inventory management system with erlang distribution of batch sizes. In: Dudin, A., Gortsev, A., Nazarov, A., Yakupov, R. (eds.) ITMM 2016. CCIS, vol. 638, pp. 273–280. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44615-8_24
13. Nazarov, A.A., Broner, V.I.: Resource control for physical experiments in the cramer-lundberg model. Russ. Phys. J. **59**(7), 1024–1036 (2016)
14. Nazarov, A., Broner, V.: Inventory management system with On/Off control of output product flow. In: Rykov, V.V., Singpurwalla, N.D., Zubkov, A.M. (eds.) ACMPT 2017. LNCS, vol. 10684, pp. 132–144. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-71504-9_13
15. Nazarov, A., Broner, V.: Inventory management system with On/Off control of input product flow. In: Dudin, A., Nazarov, A., Kirpichnikov, A. (eds.) ITMM 2017. CCIS, vol. 800, pp. 370–381. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68069-9_30
16. Kitaeva, A., Subbotina, V., Zmeev, O.: Diffusion appoximation in inventory management with examples of application. In: Dudin, A., Nazarov, A., Yakupov, R., Gortsev, A. (eds.) ITMM 2014. CCIS, vol. 487, pp. 189–196. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13671-4_23
17. Zhang, D., Xu, H., Wu, Y.: Single and multi-period optimal inventory control models with risk-averse constraints. Eur. J. Oper. Res. **199**, 420–434 (2009)
18. Gaidamaka, Y., Pechinkin, A., Razumchik, R., Samouylov, K., Sopin, E.: Analysis of an MG1R queue with batch arrivals and two hysteretic overload control policies. Int. J. Appl. Math. Comput. Sci. **24**(3), 519–534 (2014)
19. Samouylov, K.E., Abaev, P.O., Gaidamaka, Y.V., Pechinkin, A.V., Razumchik, R.V.: Analytical modelling and simulation for performance evaluation of sip server with hysteretic overload control. In: Proceedings - 28th European Conference on Modelling and Simulation, ECMS 2014, pp. 603–609 (2014)

A Retrial Queueing System with Multiple Hierarchial Orbits and Orbital Search

A. Krishnamoorthy, V. C. Joshua(✉), and Ambily P. Mathew

Department of Mathematics, CMS College, Kottayam 686001, Kerala, India
{krishnamoorthy,vcjoshua,ambilypm}@cmscollege.ac.in
http://www.cmscollege.ac.in

Abstract. We consider a $MAP/PH/1$ retrial queueing model with orbital search, consisting of a finite number of orbits which are hierarchially ordered. The model consists of an initial orbit of infinite capacity and a finite number, say M finite capacity orbits, each of which is hierarchially numbered according to the number of unsuccessful retrials made by the customers present in them. Each of the M orbits can hold both individually and collectively a maximum of N customers where $N \geq M$ and this results in the loss of customers from the system after each unsuccessful retrial. The server searches for those customers who failed to get service even after making a maximum number, say N retrials. At the end of each service completion epoch, the server searches for customers in orbit$_M$ with probability p where $0 \leq p \leq 1$ and with its complementary probability $(1 - p)$ the server remains idle. Search time is assumed to be negligible. Steady state analysis of the system is performed. Some performance measures of the system are evaluated.

Keywords: Retrial queue · Hierarchial orbits · Orbital search

1 Introduction

The theory of retrial queues serve as an effective tool in the modelling of queueing networks which are more realistic as well as practically important. An introduction to the theory of retrial queues can be found in the monograph by Falin and Templeton [7]. Detailed and classified bibliography of articles on retrial queues can be found in [1,2]. In most of the retrial queueing models the retrying customers are unaware of the status of the server and only the retrials made at the epoch when the server is idle will result in a success. So there is a possibility that the server may become idle even if there are customers in the orbit waiting for being served. While dealing with such situations, orbital search by the server helps to reduce the idle time of the server. The concept of search of customers in the classical queueing models was introduced by Neuts and Ramalhoto [14]. In [3] Artalejo et al. introduced the idea of search of orbital customers in a retrial set up. Krishnamoorthy et al. [8] extended the search mechanism to a queueing model in which the blocked customers leave the system for ever. More literatures on retrials queue with orbital search can be found in [4,5,9].

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 224–233, 2018.
https://doi.org/10.1007/978-3-319-99447-5_19

In most of the retrial queueing models all retrying customers are treated alike in the process of getting in to service and no preference shall be given for the number of unsuccessful attempts made by them while in the system. In such queueing models usually a retrying customer is either taken for service if the server is free or returns to the same orbit if the server is busy. In such a situation no preference shall be given for the number of retrials made by such a customer. In the present paper we investigate a single server retrial queueing model which takes in to account the number of retrials made by each retrying customer and hierarchially allot separate orbits for them based on the number of retrials made by them. So the model consists of the primary orbit and a finite number say N of hierarchial orbits. Any Customer who has completed i retrials is placed in orbit - i where $i = 1, 2, ..., N$. Each of the N orbits can hold both individually and collectively a maximum of M customers where $M \geq N$ and this results in the loss of customers in the system. In order to reduce loss of customers due to the finiteness of the holding capacity, to reduce the idle time of the server and to give a preference to those customers who have made N unsuccessful retrials we introduce search for those customers in $orbitN$. At the end of each service completion epoch the server searches for customers in $orbitN$ with probability p and it remains idle with probability $(1-p)$. Customers arrive to the system according to a Markovian Arrival Process (MAP) with representation (D_0, D_1). MAP is an important modelling tool in the theory of point processes. It was introduced by Neuts [12] as versatile markovian process. But later it was redefined as MAP [13]. The service times are assumed to follow Phase type distribution.

In the next section a description of the model under consideration is given. In Sect. 3 stability condition is established and the steady state analysis of the model is performed using matrix analytic Methods. In Sect. 4 some performance measures are evaluated.

2 Model Description

We consider a single server retrial queueing system with a primary orbit and a finite number, say M hierarchial orbits. Customers arrive directly to the server according to a Markovian Arrival Process (MAP) and those customers are named as primary customers. If the server is free at the time of arrival of a primary customer, they are immediately taken for service and they leave the service area after being served. If the server is busy at the time of arrival of a primary customer, they enter an orbit named as the primary orbit from where they can make retrials for entering in to service. We assume that the primary orbit can hold an infinite number of customers. Customers in the primary orbit retries for getting in to the service facility at time intervals which are exponentially distributed with rate $n_p\mu$ where n_p is the number of customers present in the primary orbit at the time of the retrial. If the retrial attempt was successful, a customer from the primary orbit enters the service facility and leaves the system after being served. If it was not successful then instead of returning to the same

orbit as in the case of ordinary retrial queues, a customer moves to another orbit, say $orbit-1$ of finite capacity and retries for service from that orbit. If a retrial from $orbit-1$ is unsuccessful, a customer is placed in another orbit, say $orbit-2$ and retries for service from there and if retrials from $orbit-2$ is not successful it is placed in $orbit-3$ and so on. In such a situation there multiple orbits each of which is numbered hierarchially based on the number of unsuccessful retrials made by the customers present in them and all of them are of finite capacity. Suppose that there are M hierarchial orbits. In this case, despite from the ordinary retrial queue we can identify the number of customers in the system who have made i unsuccessful retrials. A competition for getting in to the service facility occurs among a customer who arrives directly in to the system, retrying customer from the primary orbit and retrying customers from each of the hierarchial orbits.

Let n_i denote the number of customers present in $orbit-i$ where $i = 1, 2, \ldots, M$ and time between successive retrials from $orbit-i$ is assumed to follow exponential distribution with rate $n_i\mu_i$. We assume that each of the finite capacity hierarchial orbits can hold both individually and collectively a finite number, say N customers where $N \geq M$. So loss of customers in this system can occur from any of the M hierarchial orbits and the maximum capacity of the $orbit-i$ depends on the total number of customers present in the preceding orbits in this hierarchy. That is, the maximum number of customers in $orbit-i$ is restricted to $N - \sum_{r=1}^{i-1} n_r$. At each service completion epoch the server searches for customers in $orbit-M$ and this provides some sort of preference to those customers who have made a maximum of M retrials which in turn helps to reduce the idle time of the server as well the number of customers lost from the system due to the capacity restriction. Let p be the probability with which the server searches for customers from $orbit-M$ where $0 \leq p \leq 1$.

We assume that the arrival process in the model under consideration is Markovian Arrival Process (MAP) which takes in to account the correlation between the inter arrival times. MAP may be considered as a generalized version of the Poisson process. For the construction of a MAP, let $\{(A(t), P(t)\}$ be a Markov process on the state space (i, j) where $i \geq 0; 1 \leq j \leq m$ with an infinitesimal generator

$$\begin{pmatrix} D_0 & D_1 & & \\ & D_0 & D_1 & \\ & & \ddots & \ddots \\ & & & \ddots \end{pmatrix}$$

where $D_0 = (D_{ij}(0))$ has negative diagonal components and non negative off diagonal components. The matrix $D_1 = (D_{ij}(1))$ has non negative components.

If $A(t)$ represents the number of arrivals during $[0, t]$ and $P(t)$ indicate the auxiliary state representing the phase of the arrival, then the above Markov process defines a MAP where $(D_{ij}(0)), i \neq j$ is the transition rate from state i to state j in the underlying Markov chain without an arrival and $(D_{ij}(1))$ is the state transition rate with an arrival. $D = D_0 + D_1$ is an irreducible infinitesimal

generator for underlying Markov chain $\{A(t)\}$ with $1 \times m$ stationary probability vector θ such that $\theta D\mathbf{e} = 0; \theta\mathbf{e} = 1$ where $\mathbf{e}$ is a column vector of ones having dimension m. The fundamental arrival rate λ is given by $\lambda = \theta D\mathbf{e}$. The squared integral coefficient of variation of intervals between successive arrivals is given by $C_{var} = 2\lambda\boldsymbol{\theta}(-D_0)^{-1}\mathbf{e} - 1$.

The service times in this model is assumed to follow Phase Type distribution with irreducible representation $PH(\beta, S)$ with n phases. This means that the service of each customer continues for a time untill the Continuous time Markov Chain $\{\psi(t), t \geq 0\}$ with state space $\{0, 1, 2, \ldots, n\}$ reaches the single absorbing phase 0. Transitions of the chain within the state space $\{1, 2, \ldots, n\}$ is given by the matrix S and the intensities of transitions in to the absorbing state S^0 is given by $S^0 = -S\mathbf{e}$ where $\mathbf{e}$ is a column vector of dimension n consisting of all 1's.

In the sequel we use the following notations:
Let

- $\mathbf{e}$ be a column vector all one's of appropriate order
- $\mathbf{O}$ be a zero matrix of appropriate order
- I_l be an identity matrix of dimension l
- $\otimes$ be the Kronecker product of two matrices.

The Kronecker product of two matrices is defined as follows: If A is a matrix of order $m \times n$ and if B is a matrix of order $p \times q$, then $A \otimes B$ will denote a matrix of order $mp \times nq$ whose $(i, j)^{th}$ block matrix is given by $a_{ij}B$.

3 Matrix Analytic Solution

This above queueing model can be analyzed by considering it as a Level dependent Quasi-Birth-Death (LDQBD) process and a steady state solution is obtained by Matrix Analytic Method.

3.1 Process of the System States

We introduce the necessary random variables as follows:
Let

- $N_p(t)$ be the number of customers in the primary orbit at time t
- $N_i(t)$ be the number of customers in the orbit i at time t for $i = 1, 2, \ldots, M$
- $C(t)$ be the status of the server

$$C(t) = \begin{cases} 0, & \text{if the server is idle} \\ 1, & \text{if the server is busy} \end{cases}$$

- $S(t)$ be the phase of the service process at time t
- $A(t)$ be the phase of the arrival process at time t.

Then $\phi(t) = \{N_p(t), N_M(t), N_{(M-1)}(t), \ldots, N_1(t), C(t), S(t), A(t)\}$ is a an irreducible Continuous time Markov chain and it describes the process under consideration. This model can be considered as a Level dependent Quasi Birth Death (LDQBD) process.

We define the state space of the QBD under consideration and analyze the structure of its infinitesimal generator.

The state space Ω consists of all elements of the form $(i, j_M, j_{(M-1)}, \ldots, j_1, r, s, t)$ where i is designated as the level and $i \geq 0$. We also have

$$0 \leq \sum_{k=1}^{M} j_k \leq N$$

$$r = \begin{cases} 0, & \text{if the server is idle} \\ 1, & \text{if the server is busy} \end{cases}$$

$$s = 1, 2, \ldots, n$$

$$t = 1, 2, 3 \ldots, m$$

Let the ordering of the elements of Ω be lexicographical. We form the so called macro states $(i, j_M, j_{(M-1)}, \ldots, j_1, r)$ from the corresponding states of the Markov chain $\phi(t)$.

We analyze the transitions of the Markov chain $\phi(t)$ during an interval having an infinitesimal duration we can define the matrices defining the transition rates of this chain. The infinitesimal generator Q of the LDQBD describing the model under consideration is of the form

$$Q = \begin{pmatrix} A_1 & A_0 & & & \\ A_2^1 & A_1 & A & & \\ & A_2^2 & A_1 & A_0 & \\ & & A_2^3 & A_1 & A_0 \\ & & & \cdots\cdots\cdots & \\ & & & & \cdots\cdots\cdots \end{pmatrix}$$

where A_0, A_1, A_2^i are all square matrices whose entries are block matrices of appropriate dimensions.

A_0 represents the rate matrix corresponding to the arrival of a customer to the primary orbit, that is transition from level $i \to i+1$ where $i \geq 0$ and it is independent of i.

A_2^i represents the rate matrix corresponding to the loss of a customer from the primary orbit as a result of the retrial, that is transitions from level $i \to i-1$; for $i = 1, 2, \ldots,$ and

A_1 describes all transitions in which the level does not change (transitions within levels i) and it is independent of i.

The structure of the A_1, A_0, A_2^i for $i \geq 0$ can all be defined in terms of transition matrices corresponding to the transitions of the macro states given in the following Table 1.

The first column defines the macro state from which a transition can occur, the second column defines a macro state to which a transition can occur, the third column describes the condition when this transition occurs and the last column contains the block matrix defining the rate of the corresponding transition.

Table 1. Intensities of transitions.

From	To	Description	Transition rate
$(i, j_M, .., j_1, 0)$	$(i, j_M, .., j_1, 1)$	Arrival of a primary customer directly in to an idle server	$\beta \otimes D_1$
$(i, j_M, .., j_1, 1)$	$(i+1, j_M, .., j_1, 1)$	Arrival of a customer to the primary orbit	$I_n \otimes D_1$
$(i, j_M, .., j_1, 0)$	$(i-1, j_M, .., j_1, 1)$	Successful retrial from the primary orbit	$\beta \otimes n_p \mu_p I_m$
$(i, j_M, .., j_i, .., j_1, 0)$	$(i-1, j_M, .., j_i - 1, .., j_1, 1)$	Successful retrial from $orbit - i$	$\beta \otimes n_i \mu_i I_m$
$(i, j_M, .., j_i, .., j_1, 1)$	$(i, j_M, .., j_i, .., j_1 + 1, 1)$	Entry in to the $orbit - 1$ from the primary orbit	$n_p \mu_p I_{mn}$
$(i, .., j_i, .., j_1, 1)$	$(i, .., j_{i+1} + 1, .., j_1, 1)$	Entry in to the $orbit - (i+1)$ from $orbit - i$ if $i \neq M$[a]	$n_i \mu_i I_{mn}$
$(i, j_M, .., j_1, 1)$	$(i, j_M - 1, .., j_1, 1)$	Search of a customer from $orbit - M$ if $N_M \neq 0$	$pS^0 \otimes \beta \otimes I_m$
$(i, j_M, .., j_1, 1)$	$(i, j_M, .., j_1, 0)$	Service completion of a customer if $N_M \neq 0$	$(1-p)S^0 \otimes I_m$
$(i, j_M, .., j_1, 1)$	$(i, j_M, .., j_1, 0)$	Service completion of a customer if $N_M = 0$	$S^0 \otimes I_m$
$(i, j_M, .., j_1, 1)$	$(i, j_M, .., j_1, 1)$	No changes in levels and phases of busy server	$S \oplus [D_0 - (n_p + \sum_{k=1}^{M} n_k) I_m]$
$(i, j_M, .., j_1, 0)$	$(i, j_M, .., j_1, 0)$	No changes in levels and phases of idle server	$[D_0 - (n_p + \sum_{k=1}^{M} n_k) I_m]$

[a] Subject to capacity restriction

The matrices A_2^i, representing the rates at which customers leave the primary orbit is

$$A_2^i = i\mu_p I$$

where I is an identity matrix of appropriate order.

The matrix A_0 is given by $I \otimes D_1$.

3.2 Stability Condition

The present model is a level dependent QBD and we apply Neuts-Rao truncation for the analysis of the model. We assume that when the number of customers in the primary orbit exceeds a certain limit, say K, retrial from primary orbit occurs at constant rates $K\mu_p$. In that situation the matrices A_2^i becomes A_2^K for $i \geq K$. The infinitesimal generator Q^1 of the modified model becomes

$$Q^1 = \begin{pmatrix} A_1^0 & A_0 & & & & & \\ A_2^1 & A_1 & A_0 & & & & \\ & A_2^2 & A_1 & A_0 & & & \\ & & \cdots & \cdots & & & \\ & & & A_2 & A_1 & A_0 & \\ & & & & A_2 & A_1 & A_0 \\ & & & & & \cdots & \cdots \\ & & & & & & \cdots \end{pmatrix}$$

where $A_2 = A_2^i$ where $i \geq K$

Let the matrix A be defined as $A = A_0 + A_1 + A_2$. We can see that A is an irreducible infinitesimal generator matrix of the underlying process and so there exists the stationary vector $\boldsymbol{\pi}$ of A such that

$$\boldsymbol{\pi} A = 0$$

and

$$\boldsymbol{\pi}\mathbf{e} = 1.$$

The Markov chain with generator Q^1 is positive recurrent if and only if

$$\boldsymbol{\pi} A_0 \mathbf{e} < \boldsymbol{\pi} A_2 \mathbf{e}$$

that is, the Markov chain is positive recurrent if and only if

$$\lambda < K\mu_p$$

where μ_p is the rate of retrials per customer from the primary orbit and λ is the fundamental arrival rate.

3.3 Steady State Distribution

The stationary distribution of the Markov process under consideration is obtained by solving the set of equations

$$\mathbf{xQ^1} = \mathbf{0}, \mathbf{xe} = \mathbf{1}.$$

Let $\mathbf{x}$ be the steady-state probability vector of Q^1.

Partition this vector in conformity with Q^1 as follows:

$$\mathbf{x} = (\mathbf{x(0)}, \mathbf{x(1)}, \mathbf{x(2)}, \ldots,)$$

$\mathbf{x}(i) = \mathbf{x}(i, j_M)$ and so on. Continuing like this finally we get

$$\mathbf{x}(i) = \mathbf{x}(i, j_M, j_{(M-1)}, \ldots, j_1, r, s, t); i \geq 0$$

Here $\mathbf{x}(i, j_M, j_{(M-1)}, \ldots, j_1, r, s, t)$ is the probability of being in state $(i, j_M, j_{(M-1)}, \ldots, j_1, r, s, t)$ We have

$$i \geq 0$$

$$0 \leq \sum_{k=1}^{M} j_k \leq N$$

$$r = \begin{cases} 0, & \text{if the server is idle} \\ 1, & \text{if the server is busy} \end{cases}$$

$$s = 1, 2, \ldots, n$$

$$t = 1, 2, 3 \ldots, m$$

Under the stability condition the steady-state probability vector is obtained as

$$\mathbf{x}(K-1) + i = \mathbf{x}(K-1)R^i, i \geq 0$$

where R is the minimal non negative solution to the matrix quadratic equation

$$R^2 A_2 + RA_1 + A_0 = 0$$

and the vectors $\mathbf{x}0, \ldots, \mathbf{x}(K-1)$ are obtained by solving

$$\mathbf{x}(\mathbf{0})A_1 + \mathbf{x}(\mathbf{1})_1 A_2^1 = 0$$

$$\mathbf{x}(i-1)A_0 + \mathbf{x}(i)iA_1 + \mathbf{x}(i+1)A_2^{(i+1)} = 0; 1 \leq i \leq (K-2)$$

$$\mathbf{x}(K-2)A_0 + \mathbf{x}(K-1)\left[A_1 + A_2 R\right] = 0$$

subject to the normalizing condition

$$\sum_{i=0}^{(K-2)} \mathbf{x}(i) + \mathbf{x}(K-1)(I-R)^{-1}\mathbf{e} = 1.$$

4 Some Performance Measures of the System

Some measures of performance, which helps the operators of the system to make decisions concerning the optimal values of the number M of hierarchial orbits to be maintained and the total number N of customers to be allowed in those hierarchial orbits are evaluated. Loss of customers can happen due to these capacity restrictions. The effect of this loss on the system can be can be minimized by means of orbital search.

Following are some performance measures which helps us to make a detailed study about the problem under consideration.

1. Expected number of customers in the primary orbit

$$E[PO] = \sum_{i=0}^{\infty} ix_i\mathbf{e}$$

where $\mathbf{e}$ is a column vector of appropriate order consisting of all ones.

2. Expected number of customers in the i^{th} hierarchial orbit

$$E[HO(i)] = \sum_{i_p, J_M, .., J_{i+1}} \sum_{j_i=0}^{L_i} n_i \mathbf{x}(i_p, j_M, j_{(M-1)}, \ldots, j_i)\mathbf{e}$$

where $L_i = min[K, K - (\sum_{k=1}^{(i-1)} n_k)]$

3. Expected number of customers in the system

$$E[sys] = E[PO] + \sum_{i=1}^{M} E[HO(i)]$$

4. Probability that a customer is lost from the system

$$P[Loss] = \sum_{i=0}^{\infty} x_i \mathbf{e}$$

where the summation is restricted by the condition that the number of customers in the hierarchial orbits are individually and collectively should not exceed K.

5. Probability that a customer enters the service facility as a result of orbital search

$$P[OS] = \sum_{i=0}^{\infty} \sum_{J_M=1}^{L_M} p\mathbf{x}(i, j_M)\mathbf{e}$$

where L_M is defined as above

6. Expected number of customers enter the service facility as a result of orbital search

$$E[OS] = P[OS] * p * n_M * \mu_M.$$

5 Conclusion

The results in this paper may be extended by proposing an optimization problem which determines the optimum value of K, the number of customers that the hierarchial orbits can hold individually and collectively. To construct an objective function we may assume that the service of each customer provides an additional revenue if it enters the system by means of orbital search. A holding cost is associated with each of the hierarchial orbits. There is a search cost associated with each customer entering the main station by means of orbital search. The search cost is an expenditure encountered by the system. So we can find an optimal value for the total capacity of the orbits. We plan to analyze this problem numerically for finding the optimal value of search probability P.

Acknowledgement. A. Krishnamoorthy and V. C. Joshua thanks the Department of Science and Technology, Government of India, for the support given under the Indo-Russian Project $INT/RUS/RSF/P - 15$. A. Krishnamoorthy also thanks the UGC

India for the Award of Emeritus Fellowship $No.F6-6/2017/-18/EMERITUS-2017-18-GEN-10822/(SA-II)$. Ambily P. Mathew thanks the UGC-India for the teacher fellowship sanctioned under the Faculty Development Programme [$F.No.FIP/12thplan/KLMG002TF06$].

References

1. Artalejo, J.R.: Accessible bibliography on retrial queues: progress in 2000–2009. Math. Comput. Model. **51**, 1071–1081 (2010)
2. Artalejo, J.R.: A classified bibliography of research on retrial queues: progress in 1990–1999. Top **7**, 187–211 (1999)
3. Artalejo, J.R., Joshua, V.C., Krishnamoorthy, A.: An M/G/1 retrial queue with orbital search by server. In: Advances in Stochastic Modelling, pp. 41–54. Notable Publications, New Jersey (2002)
4. Dudin, A.N., Krishnamoorthy, A., Joshua, V.C., Tsarenkov, G.: Analysis of BMAP/G/1 retrial system with search of customers from the orbit. Eur. J. Oper. Res. **157**, 169–179 (2004)
5. Dudin, A., Deepak, T.G., Joshua, V.C., Krishnamoorthy, A., Vishnevsky, V.: On a $BMAP/G/1$ retrial system with two types of search of customers from the orbit. In: Dudin, A., Nazarov, A., Kirpichnikov, A. (eds.) ITMM 2017. CCIS, vol. 800, pp. 1–12. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68069-9_1
6. Falin, G.I.: A survey of retrial queues. Queueing Syst. **7**, 127–167 (1990)
7. Falin, G.I., Templeton, J.G.C.: Retrial Queues. Chapman and Hall, London (1997)
8. Krishnamoorthy, A., Deepak, T.G., Joshua, V.C.: An M/G/1 retrial queue with non persistent customers and orbital search. Stoch. Anal. Appl. **23**, 975–997 (2005)
9. Krishnamoorthy, A., Joshua, V.C., Mathew, A.P.: A retrial queueing system with abandonment and search for priority customers. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds.) DCCN 2017. CCIS, vol. 700, pp. 98–107. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66836-9_9
10. Latouche, G., Neuts, M.F.: Efficient algorithmic solutions to exponential tandem queues with blocking. SIAM J. Algebr. Discret. Meth. **1**, 93–106 (1980)
11. Latouche, G., Ramaswami, V.: Introduction to Matrix Analytic Methods in Stochastic Modeling, vol. 5. Siam, Philadelphia (1999)
12. Neuts, M.F., Lucantoni, D.M.: A Markovian queue with n servers subject to breakdowns and repairs. Manag. Sci. **25**, 849–861 (1979)
13. Lucantoni, D.M., Hellstern, K.S., Neuts, M.F.: A single server queue with server vacation and a class of non renewal arrival process. Adv. Appl. Probab. **22**, 676–705 (1990)
14. Neuts, M.F., Ramalhoto, M.F.: A service model in which the server is required to search for customers. J. Appl. Probab. **21**, 57–166 (1984)
15. Neuts, M.F.: Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach. Courier Corporation, North Chelmsford (1981)
16. Neuts, M.F., Rao, B.M.: Numerical investigation of a multiserver retrial model. Queueing syst. **7**, 169–189 (1990)

On Sensitivity Analysis of Steady State Probabilities of Double Redundant Renewable System with Marshall-Olkin Failure Model

Vladimir Rykov[1,2], Elvira Zaripova[1(✉)], Nika Ivanova[1], and Sergey Shorgin[3]

[1] Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St., Moscow 117198, Russian Federation
vladimir_rykov@mail.ru, zaripova_er@pfur.ru, ivanovanika1995@gmail.com
[2] Gubkin Russian State University of Oil and Gas, 65 Leninsky Prospekt, Moscow 119991, Russian Federation
[3] Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation
sshorgin@ipiran.ru

Abstract. A heterogeneous double redundant hot standby renewable system with Marshall-Olkin failure model is considered. In one of previous papers the stationary characteristics for such system for the case of exponential lifetime distributions has been found and its asymptotic insensitivity to the shape of its components repair time distributions has been studied. In this paper the problem of asymptotic insensitivity of such system for both life- and repair time general distributions with the help of simulation method is studied.

Keywords: Heterogeneous standby renewable system
Marshall-Olkin failure model · Stationary probabilities
Sensitivity analysis

1 Introduction and Motivation

Stability of different systems' characteristics to the changes in initial states or exterior factors are the key problems in all natural sciences. For stochastic systems stability often means insensitivity or low sensitivity of their output characteristics to the shapes of some input distributions.

One of the earliest results concerning insensitivity of systems' characteristics to the shape of service time distribution has been obtained in 1957 by Sevast'yanov [1], who proved the insensitivity of Erlang formulas to the shape of

V. Rykov—The publication has been prepared with the support of the "RUDN University Program 5-100" and funded by RFBR according to the research projects No. 17-01-00633, No. 17-07-00142, and No. 18-07-00576.

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 234–245, 2018.
https://doi.org/10.1007/978-3-319-99447-5_20

service time distribution with fixed mean value for loss queueing systems with Poisson input flow. In 1976 Kovalenko [2] found the necessary and sufficient conditions for insensitivity of stationary reliability characteristics of redundant renewable system with exponential life time and general repair time distributions of its components to the shape of the latter. These conditions consist in sufficient amount of repairing facilities, i.e. in possibility of immediate start to repair any of failed element. The sufficiency of this condition for the case of general life and repair time distributions has been found in 2013 by Rykov [3] with the help of multi-dimensional alternative processes theory. However, in the case of limited possibilities for restoration these results do not hold, as it was shown, for example, in [4] with the help of additional variable method.

On the other hand in series of work of Gnedenko, Solov'ev [5–7] it was shown that under "quick" restoration the reliability function of a cold standby double redundant heterogeneous system tends to the exponential one for any life and repair time distributions of its elements. This result also means the asymptotic insensitivity of the reliability characteristics of such system to the shapes of their elements life and repair times distributions. The problem of the convergence rate does not enough investigated. In the paper of Kalashnikov [8] the evaluation of the convergence rate has been done in terms of moments of appropriate distributions. At that the numerical investigations, proposed in [9–11] demonstrate enough quick appearance of practical insensitivity of the time dependent as well as stationary reliability characteristics to the shapes of life and repair time distributions with fixed their mean values.

In the papers [12,13,15,16] the problem of systems' steady state reliability characteristics sensitivity to the shape of life and repair time distributions of its components for the same type of systems has been considered, for the case, when one of the input distributions (either of life or repair time lengths) is exponential. For these models explicit expressions for stationary probabilities have been obtained which show their evident dependence on the non-exponential distributions in the form of their Laplace-Stiltjes transforms.

Most of these investigations deal with system which components fail independently. In 1967 Marshall and Olkin [14] proposed bivariate distribution with depend components governing by independent Poisson processes that can be used as a failure model for two-components reliability system.

Many textbooks give a special attention to the bivariate lack of memory property of this distribution and related bivariate exponential distribution exhibiting singularity along the main diagonal in R_+^2, see Barlow and Proschan [17], Singpurwalla [18], Balakrishnan and Lai [19], Gupta et al. [20], McNeil et al. [21] among others. Many articles complement and extend Marshall-Olkin's bivariate exponential distribution, justifying advantages in analysis of various data sets from engineering, medicine, insurance, finance, biology, etc. For example, Li and Pellerey [22] launched the generalized MO model considering non-exponential independent components. In 2014 the model is extended to the multidimensional case by Lin and Li [23]. As a further step, in 2015 Kolev and Pinto [24] introduced the extended MO model assuming dependence between components.

Most of these investigations deal with bivariate distributions and their properties, use the MO model only for the first failure and does not include it into the reliability model.

In the present paper MO model will be used for renewable heterogeneous double redundant standby renewable systems, which elements failure satisfies to the MO model. The reliability function in terms of its Laplace transforms for this model in [26] has been considered. The stationary probabilities for the case of components' lifetime exponential distributions in the paper [25] has been proposed. In this paper the simulation method will be used for the same model study with general distributions both the life- and the repair time distributions.

In this case the renovation procedure after the system components failures is very important and it will be included into the model.

The paper is organized as follows. In the next section the problem setting and some notations will be introduced. In Sect. 3 analytical results for stationary distribution and their sensitivity analysis of the model for the case of exponential life time distributions will be done without proofs from [25], and the next 4-th section devoted to the sensitivity analysis of the model in general case with the help of simulation method. The paper ends with conclusion and some problems description.

2 The Problem Setting and Notations

Consider a heterogeneous hot double redundant repairable system [26], the failure of which components satisfies to the MO model. This means that there exists three type of shocks, which leads to the system failure. The first shock act only to the first component, the second one act only to the second one, while the third one act to both components and leads to the system failure (Fig. 1).

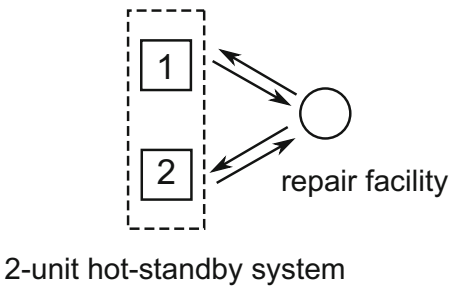


Fig. 1. 2-unit hot-standby repairable system with one repair facility.

Thus accordingly to the MO failure model the vector random variable (r.v.) of the system and its components failure has a form

$$(T_1, T_2) = (\min(A_1, A_3), \min(A_2, A_3)), \tag{1}$$

where A_1, A_2, A_3 are representatives of the sequences of independent r.v.'s, which represent times to the different shocks. Its cumulative distribution functions (c.d.f.) are denoted by $A_i(x)$ $(i = 1, 2, 3)$. Dealing with the renewable model we need to consider some procedure of the system restoration after the partial and/or full failure. In this paper it is supposed that after partial failure (when only one, say i-th component, fails) the partial repair of random time B_i $(i = 1, 2)$ begins, which means that the system prolong to work and the failed component begins to repair. But after the system failure the renewal of whole system begins that demand some random time, say B_3, and after this time the system became as a new one and goes to the state 0. In any case the repair times B_k $(k = 1, 2, 3)$ that are representatives of sequences independent repair times of components and the whole system has absolute continuous distributions with c.d.f. $B_k(x)$ $(k = 1, 2, 3)$ and probability density functions (p.d.f.) $b_k(x)$ $(k = 1, 2, 3)$ correspondingly. All repair times are independent.

The system state space can be represented as $E = \{0, 1, 2, 3\}$, which means:

0 — both components are working,
1 — the first component is repaired, and the second one is working,
2 — the second component is repaired, and the first one is working,
3 — both components are in down states, system is failed and repaired.

For the system behavior description introduce a random process $J = \{J(t),\ t \geq 0\}$ with values into system set of states E : $J(t) = j$, if in the time t the system is in the state $j \in E$.

This paper deals with s.s.p.'s of the system, $\pi_j = \lim\limits_{t\to\infty} \mathbf{P}\{J(t) = j\}$, and properties of their asymptotic insensitivity to the shapes of life- and repair times of its components.

3 Steady State Probabilities for Poisson Shocks

In the case of Poisson shocks the r.v's. A_i $(i = 1, 2, 3)$ have exponential distributions with parameters, say α_i $(i = 1, 2, 3)$. For s.s.p. calculation in this case the method of additional variables introduction is used in the paper [25]. For our case as additional variables we use elapsed time of the failed component. Thus, consider two-dimensional process $Z = \{Z(t),\ t \geq 0\}$, with $Z(t) = (J(t), X(t))$ where $J(t)$ is the system state in the time t, and $X(t)$ represents elapsed time of the failed component. Due to additional components the process Z is a Markov one with states space is $\mathcal{E} = \{0, (1, x), (2, x), (3, x)\}$, and transition graph represented in the Fig. 2.

Here and further we will use the following notations.

- $\alpha = \alpha_1 + \alpha_2 + \alpha_3$ – the summary intensity of the system failure;
- $\bar{\alpha}_k = \alpha_k + \alpha_3$ – the intensity of joint k-th and third shock;
- $b_k = \int_0^\infty (1 - B_k(x))dx$ $(k = 1, 2, 3)$ – k-th component repair time expectations;
- $\rho_k = b_k\alpha_k$;

– $\beta_k(x) = (1 - B_k(x))^{-1} b_k(x)$ $(k = 1, 2, 3)$ – k-th component conditional repair intensity given elapsed repair time is x;
– $\tilde{b}_k(s) = \int_0^\infty e^{-sx} b_k(x) dx$ $(k = 1, 2, 3)$ – Laplace transform (LT) of the k-th component repair time distribution.

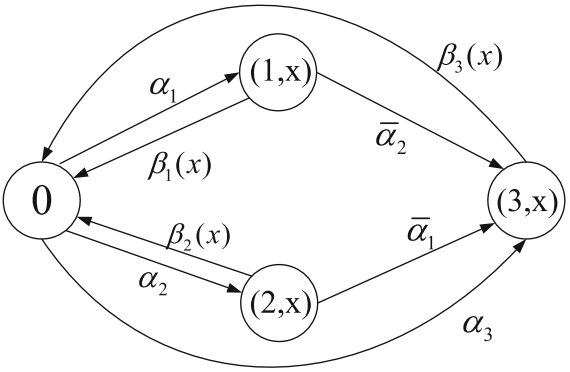


Fig. 2. Transition graph of the system with full repair.

Using the Markov process Z for its first component s.s.p's $\pi_i = \lim\limits_{t\to\infty} \mathbf{P}\{J(t) = i\}$, $i \in \{0, 1, 2, 3\}$ the following theorem has been proved.

Theorem 1. The stationary micro-state probabilities of the system under consideration has the form

$$\begin{aligned}
\pi_1(x) &= \alpha_1 e^{-\bar{\alpha}_2 x}(1 - B_1(x))\pi_0, \\
\pi_2(x) &= \alpha_2 e^{-\bar{\alpha}_1 x}(1 - B_2(x))\pi_0, \\
\pi_3(x) &= [\alpha_1(1 - \tilde{b}_1(\bar{\alpha}_2)) + \alpha_2(1 - \tilde{b}_2(\bar{\alpha}_1)) + \alpha_3](1 - B_3(x))\pi_0
\end{aligned} \tag{2}$$

with appropriate macro-states probabilities

$$\begin{aligned}
\pi_1 &= \frac{\alpha_1}{\bar{\alpha}_2}(1 - \tilde{b}_1(\bar{\alpha}_2))\pi_0, \\
\pi_2 &= \frac{\alpha_2}{\bar{\alpha}_1}(1 - \tilde{b}_2(\bar{\alpha}_1))\pi_0, \\
\pi_3 &= [\alpha_1(1 - \tilde{b}_1(\bar{\alpha}_2)) + \alpha_2(1 - \tilde{b}_2(\bar{\alpha}_1)) + \alpha_3]b_3\pi_0
\end{aligned} \tag{3}$$

where π_0 is given by

$$\pi_0 = \left[1 + \alpha_1(1 - \tilde{b}_1(\bar{\alpha}_2))\left(b_3 + \frac{1}{\bar{\alpha}_2}\right) + \alpha_2(1 - \tilde{b}_2(\bar{\alpha}_1))\left(b_3 + \frac{1}{\bar{\alpha}_1}\right) + \alpha_3 b_3\right]^{-1}. \tag{4}$$

Proof. Using the transition graph, represented at the Fig. 2, analogously to the case of Markov processes with discrete states space one can write down the following system of balance equations for s.s.p. of the process

$$\begin{aligned}
\alpha\pi_0 &= \int_0^\infty \pi_1(x)\beta_1(x)dx + \int_0^\infty \pi_2(x)\beta_1(x)dx + \int_0^\infty \pi_3(x)\beta_3(x)dx,\\
\dot\pi_1(x) &= -(\bar\alpha_2 + \beta_1(x))\pi_1(x),\\
\dot\pi_2(x) &= -(\bar\alpha_1 + \beta_2(x))\pi_2(x),\\
\dot\pi_3(x) &= -\beta_3(x)\pi_3(x).
\end{aligned} \tag{5}$$

with boundary conditions

$$\begin{aligned}
\pi_1(0) &= \alpha_1\pi_0,\\
\pi_2(0) &= \alpha_2\pi_0,\\
\pi_3(0) &= \alpha_3\pi_0 + \bar\alpha_1\int_0^\infty \pi_2(x)dx + \bar\alpha_2\int_0^\infty \pi_1(x)dx.
\end{aligned} \tag{6}$$

Solutions of the last three of equations (5) are

$$\begin{aligned}
\pi_1(x) &= C_1 e^{-\bar\alpha_2 x}(1 - B_1(x)),\\
\pi_2(x) &= C_2 e^{-\bar\alpha_1 x}(1 - B_2(x)),\\
\pi_3(x) &= C_3(1 - B_3(x)).
\end{aligned} \tag{7}$$

Using boundary conditions (6) to find unknown constants C_i gives

$$\begin{aligned}
C_1 &= \pi_1(0) = \alpha_1\pi_0,\\
C_2 &= \pi_2(0) = \alpha_2\pi_0,\\
C_3 &= \pi_3(0) = [\alpha_1(1 - \tilde b_1(\bar\alpha_2)) + \alpha_2(1 - \tilde b_2(\bar\alpha_1)) + \alpha_3]\pi_0.
\end{aligned}$$

The simple integration of the formulas (7) with respect to x allows to find appropriate stationary macro-state probabilities

$$\begin{aligned}
\pi_1 &= \frac{\alpha_1}{\bar\alpha_2}(1 - \tilde b_1(\bar\alpha_2))\pi_0,\\
\pi_2 &= \frac{\alpha_2}{\bar\alpha_1}(1 - \tilde b_2(\bar\alpha_1))\pi_0,\\
\pi_3 &= [\alpha_1(1 - \tilde b_1(\bar\alpha_2)) + \alpha_2(1 - \tilde b_1(\bar\alpha_2)) + \alpha_3]b_3\pi_0,
\end{aligned} \tag{8}$$

The normalizing conditions gives

$$\begin{aligned}
1 &= \pi_0 + \pi_1 + \pi_2 + \pi_3 =\\
&= \Big[1 + \frac{\alpha_1}{\bar\alpha_2}(1 - \tilde b_1(\bar\alpha_2)) + \frac{\alpha_2}{\bar\alpha_1}(1 - \tilde b_2(\bar\alpha_1))\\
&\quad + (\alpha_1(1 - \tilde b_1(\bar\alpha_2)) + \alpha_2(1 - \tilde b_2(\bar\alpha_1)) + \alpha_3)b_3\Big]\pi_0,
\end{aligned}$$

from which the formula (4) follows that proves the theorem.

The above results show the evident sensitivity of the considered systems s.s.p's. to the shape of their components repair time distribution.

Theorem 2. Under the rare components' failures, when $q = \max[\alpha_1, \alpha_2, \alpha_3] \to 0$ the s.s.p. of the considered system take the form

$$\begin{aligned}
&\pi_0 \approx [1 + \rho_1 + \rho_2 + \rho_3 + b_3(\bar{\alpha}_2\rho_1 + \bar{\alpha}_1\rho_2)]^{-1}, \\
&\pi_i \approx \rho_i\pi_0 \quad (i = 1, 2), \\
&\pi_3 \approx [\rho_3 + b_3(\bar{\alpha}_2\rho_1 + \bar{\alpha}_1\rho_2)]\pi_0.
\end{aligned} \tag{9}$$

Proof. Applying Teilor expansion up to the second order of q, and tasking into account that when $q \to 0$

$$1 - \tilde{b}_i(\bar{\alpha}_{i^*}) \approx b_i\alpha_{i^*}$$

one can find from (8)

$$\begin{aligned}
\pi_i &= \frac{\alpha_i}{\bar{\alpha}_{i*}}(1 - \tilde{b}_i(\bar{\alpha}_{i^*}))\pi_0 = \frac{\alpha_i}{\bar{\alpha}_{i\,*}}b_i\alpha_{i^*}\pi_0 = \rho_i\pi_0 \qquad (i = 1, 2), \\
\pi_3 &= [\alpha_1(1 - \tilde{b}_1(\bar{\alpha}_2)) + \alpha_2(1 - \tilde{b}_2(\bar{\alpha}_1)) + \alpha_3]b_3\pi_0 \\
&= [(\rho_1\bar{\alpha}_2 + \rho_2\bar{\alpha}_1)b_3 + \rho_3]\pi_0,
\end{aligned}$$

while for π_0 it follows

$$\begin{aligned}
\pi_0 &= \left[1 + \alpha_1 b_1\bar{\alpha}_2\left(b_3 + \frac{1}{\bar{\alpha}_2}\right) + \alpha_2 b_2\bar{\alpha}_1\left(b_3 + \frac{1}{\bar{\alpha}_1}\right) + \alpha_3 b_3\right]^{-1} \\
&= [1 + \rho_1 + \rho_2 + \rho_3 + (\rho_1\bar{\alpha}_2 + \rho_2\bar{\alpha}_1)b_3]^{-1}.
\end{aligned}$$

Thus from here the theorem result follows.

Remark 1. Taking into account that the expression $b_3(\bar{\alpha}_1\rho_1 + \bar{\alpha}_2\rho_2)$ in probabilities π_0 and π_3 has the second order with respect to q, and therefore using only the first order of this value the above formulas can be rewrite as follows

$$\begin{aligned}
&\pi_0 \approx [1 + \rho_1 + \rho_2 + \rho_3]^{-1}, \\
&\pi_i \approx \rho_i\pi_0 \quad (i = 1, 2, 3).
\end{aligned}$$

The results of the theorem show asymptotic insensitivity of the s.s.p's. to the shapes of their components' repair time distributions, but only on their mean values and the components failure intensities.

4 Simulation Results

The above results gives the closed form representation for s.s.p's. of the considered system for the case of Poisson failures flows, and shows their asymptotic for rare failures insensitivity to the shape of repair time distributions. In this section the asymptotic insensitivity of the s.s.p.'s is shown with the help of simulation method also for general life- and repair time distributions. For system simulation we used multi-method modeling environment AnyLogic.

In the first numerical example we compare the availability of the system $1 - \pi_3$ for two cases:

(1) Exponential distribution for the lifetime and Gamma distribution for the repair time;
(2) both life- and repair time have Gamma distribution.

For the next numerical example we used also Pareto, Gnedenko-Weibull and Uniform distributions. The following notations are used for r.v's. and their distributions.

- A r.v. X that has Exponential distribution with parameter α_0 is denoted as $X \sim \text{Exp}(\alpha_0)$ and its p.d.f. is $f_X(x) = \alpha_0 e^{-\alpha_0 x}$, $x \geq 0$.
- A r.v. X with gamma-distribution of a shape parameter k and a scale parameter θ is denoted $X \sim \Gamma\ (k,\theta)$ with the corresponding p.d.f. $f_X(x) = \frac{x^{k-1}e^{-x/\theta}}{\theta^k \Gamma(k)}$ for $x > 0$ and $k, \theta > 0$.
- A r.v. X with a Pareto distribution of a scale parameter x_m and a shape parameter k denoted $X \sim \text{P}\ (k, x_m)$, its p.d.f. is $f_X(x) = \frac{k x_m^k}{x^{k+1}}$, $x \geq x_m$, and $k, x_m > 0$.
- A r.v. X that has Gnedenko-Weibull distribution with a scale parameter λ and a shape parameter k, $k, \lambda > 0$, (denoted by $X \sim \text{GW}(k, \lambda)$), its p.d.f. is $f_X(x) = \frac{k}{\lambda}(\frac{x}{\lambda})^{k-1} e^{-(\frac{x}{\lambda})^k}$, $x \geq 0$.
- A r.v. X with Uniform distribution on interval $[a, b]$ has p.d.f. $f_X(x) = \frac{1}{b-a}$ for $a \leq x \leq b$, and $0 \leq a < b$ (denoted by $X \sim \text{Unif}(a, b)$).

The following data were chosen for simulation:

- $b_i = b = 1$ $(i = 1, 2, 3)$ is the mean repair time of all components and the whole system;
- $\alpha_i = \alpha_0$ $(i = 1, 2, 3)$ is the intensity of all types of shocks;
- $T = 1000000$ b is the total simulation time;
- the parameters of all distributions are chosen in such a way that the coefficient of variation (the ratio of the standard deviation σ to the mean μ) $c = \frac{\sigma}{\mu}$ takes fixed value and mean life time $a = \frac{1}{\alpha_0}$ increases.

The system availability $1 - \pi_3$ with exponential lifetime can be find by analytical approach, whereas for the second case (when both life- and repair time have gamma distribution with coefficient of variation $c = 10$) we used simulation. As you see on Fig. 3 the system availability for these cases are very close to each other, and from $a = 6$ the difference between values are less than 1%, from $a = 20$ – less than 0.1%. Our example shows asymptotic insensitivity with rare failures.

The second numerical experiment represented on Fig. 4. Here system availability $1 - \pi_3$ versus to mean elements lifetime a in interval from 1 up to 100 are represented for lifetime distributed as gamma, and different repair time distributions: (1) $\Gamma(k,\theta)$, (2) $\text{P}(k, x_m)$, (3) $\text{GW}(k, \lambda)$. The coefficient of variation for all distributions is equal to $c = 10$.

The next numerical experiments with Uniform distribution is represented on Figs. 5 and 6. We compare the system availability $1 - \pi_3$ for (1) Uniform and (2) exponential lifetime a in interval from 1 up to 100 on Fig. 5 and from 100 to

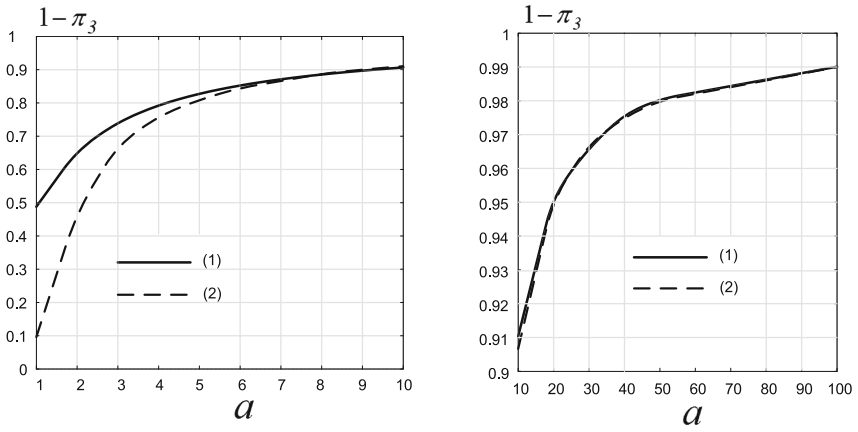


Fig. 3. The system availability $1 - \pi_3$: (1) lifetime ∼Exp, repair time $\sim \Gamma$; (2) lifetime $\sim \Gamma$, repair time $\sim \Gamma$.

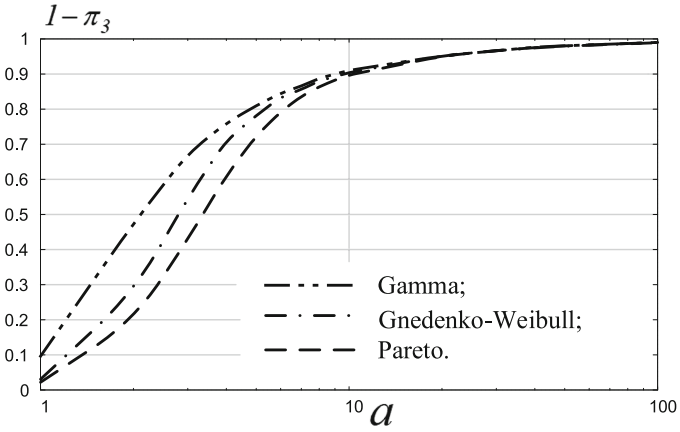


Fig. 4. The system availability $1 - \pi_3$ for lifetime $X \sim \Gamma\ (k, \theta)$.

1000 on Fig. 6. In this cases a repair time has Uniform distibution with a mean value 1. As you see on Figs. 5 and 6 the system availability for these cases are also very close to each other, and from $a = 6$ the difference between values are less than 5%, from $a = 50$ – less than 1%. The asymptotic insensitivity with rare failures are shown on our examples.

The analysis of the system availability $1 - \pi_3$ shows that the type of distribution does not matter when $a \to \infty$, and it approaches to the values given by analytical formulas for different distributions with different parameters.

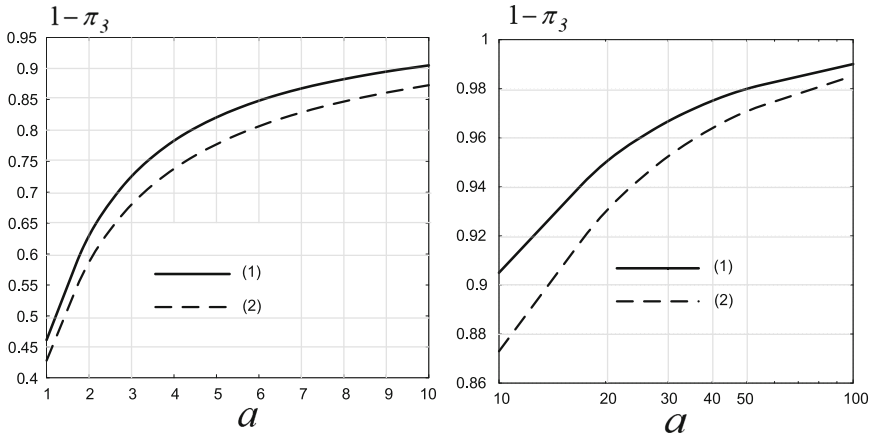


Fig. 5. The system availability $1-\pi_3$ for repair time $X \sim Unif$, (1) lifetime $X \sim Unif$, (2) lifetime $X \sim$ Exp.

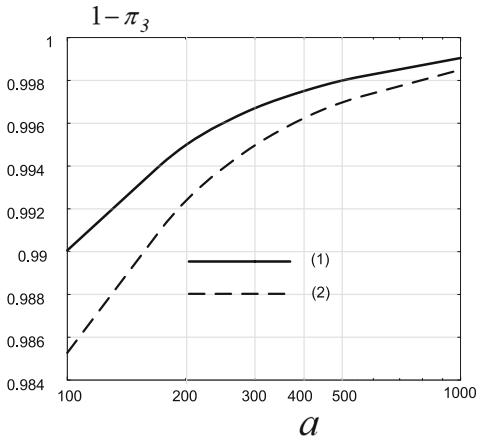


Fig. 6. The system availability $1-\pi_3$ for repair time $X \sim Unif$, (1) lifetime $X \sim Unif$, (2) lifetime $X \sim$ Exp.

5 Conclusion

Simulation method is used for heterogeneous double redundant hot standby renewable system s.s.p. analysis. It was shown that under rare failures the s.s.p. asymptotically insensitive to the shape of the components life- and repair time distributions. Investigation of more complex systems with dependent failures of their components is the subject of our further research.

References

1. Sevast'yanov, B.A.: An ergodic theorem for markov processes and its application to telephone systems with refusals. Theory Probab. Appl. **2**(1), 104–112 (1957)
2. Kovalenko, I.N.: Investigations on Analysis of Complex Systems Reliability, p. 210. Naukova Dumka, Kiev (1976). (in Russian)
3. Rykov, V.: Multidimensional alternative processes reliability models. In: Dudin, A., Klimenok, V., Tsarenkov, G., Dudin, S. (eds.) BWWQT 2013. CCIS, vol. 356, pp. 147–156. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-35980-4_17
4. Koenig, D., Rykov, V., Schtoyn, D.: Queueing Theory, p. 115. Gubkin University Press, Moskva (1979). (in Russian)
5. Gnedenko, B.V.: On cold double redundant system. Izv. AN SSSR. Texn. Cybern. **4**, 3–12 (1964). (in Russian)
6. Gnedenko, B.V.: On cold double redundant system with restoration. Izv. AN SSSR. Texn. Cybern. **5**, 111–118 (1964). (in Russian)
7. Solov'ev, A.D.: On reservation with quick restoration. Izv. AN SSSR. Texn. Cybern. **1**, 56–71 (1970). (in Russian)
8. Kalashnikov, V.V.: Geometric Sums: Bounds for Rare Events with Applications: Risk Analysis, Reliability, Queueing, p. 256. Kluwer Academic Publishers, Dordrecht (1997)
9. Kozyrev, D.V.: Analysis of asymptotic behavior of reliability properties of redundant systems under the fast recovery. Math. Inf. Sci. Phys. **3**, 49–57 (2011). (in Russian). Bulletin of Peoples' Friendship University of Russia
10. Rykov, V.V., Kozyrev, D.V.: Analysis of renewable reliability systems by markovization method. In: Rykov, V.V., Singpurwalla, N.D., Zubkov, A.M. (eds.) ACMPT 2017. LNCS, vol. 10684, pp. 210–220. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-71504-9_19
11. Rykov, V., Kozyrev, D., Zaripova, E.: Modeling and simulation of reliability function of a homogeneous hot double redundant repairable system. In: Paprika, Z.Z., Horák, P., Váradi, K., Zwierczyk, P.T., Vidovics-Dancs, Á., Rádics, J.P. (eds.) ECMS 2017 Proceedings of European Council for Modeling and Simulation, pp. 701–705 (2017). https://doi.org/10.7148/2017-0701
12. Rykov, V., Ngia, T.A.: On sensitivity of systems reliability characteristics to the shape of their elements life and repair time distributions. Vestnik PFUR. Ser. Math. Inform. Phys. **3**, 65–77 (2014). (in Russian)
13. Efrosinin, D., Rykov, V.: Sensitivity analysis of reliability characteristics to the shape of the life and repair time distributions. In: Dudin, A., Nazarov, A., Yakupov, R., Gortsev, A. (eds.) ITMM 2014. CCIS, vol. 487, pp. 101–112. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13671-4_13
14. Marshall, A., Olkin, I.: A multivariate exponential distribution. J. Am. Stat. Assoc. **62**, 30–44 (1967)
15. Rykov, V., Kozyrev, D.: On sensitivity of steady-state probabilities of a cold redundant system to the shapes of life and repair time distributions of its elements. In: Pilz, J., Rasch, D., Melas, V., Moder, K. (eds.) IWS 2015. PROMS, vol. 231, pp. 391–402. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76035-3_28
16. Efrosinin, D., Rykov, V., Vishnevskiy, V.: Sensitivity of reliability models to the shape of life and repair time distributions. In: 9th International Conference on Availability, Reliability and Security (ARES 2014), pp. 430–437. IEEE (2014). https://doi.org/10.1109/ARES.2014.65. Published in CD: 978-I-4799-4223-7/14

17. Barlow, R., Proschan, F.: Statistical Theory of Reliability and Life Testing. Silver Spring, Redwood City (1981)
18. Singpurwalla, N.: Reliability and Risk: A Bayesian Perspective. Wiley, Chichester (2006)
19. Lai, C.D., Balakrishnan, N.: Continuous Bivariate Distributions, 2nd edn. Springer, New York (2009). https://doi.org/10.1007/b101765
20. Gupta, A., Zeung, W., Hu, Y.: Probability and Statistical Models: Foundations for Problems in Reliability and Financial Mathematics. Birkhauser, Basel (2010)
21. McNeil, A., Frey, L., Embrechts, P.: Quantitative Risk Management, 2nd edn. Princeton University Press, Princeton (2015)
22. Li, X., Pellerey, F.: Generalized Marshall-Olkin distributions and related bivariate aging properties. J. Multivar. Anal. **102**, 1399–1409 (2011)
23. Lin, J., Li, X.: Multivariate generalized Marshall-Olkin distributions and copulas. Methodol. Comput. Appl. Probab. **16**, 53–78 (2014)
24. Kolev, N., Pinto, J.: A weak version of bivariate lack of memory property. Braz. J. Probab. Stat. (2017, accepted)
25. Rykov, V.: On steady state probabilities of double redundant renewable system with Marshal-Olkin failure model. In: Accepted for publication in Proceedings of IWS-2018 (2018)
26. Kozyrev, D., Rykov, V., Kolev, N.: Reliability function of renewable system with Marshal-Olkin failure model. Reliab.: Theory Appl. **13**(1(48)), 39–46 (2018)

Data Reclassification of Multidimensional Information System Designed Using Cluster Method of Metadata Description

M. B. Fomin(✉)

Department of Information Technologies, RUDN University,
117198 Miklukho-Maklaya st. 6, Moscow, Russia
fomin_mb@rudn.university

Abstract. The structure of a multidimensional data cube is determined by the aspects of analysis that are used. It is considered as static. If the observed phenomenon changes, the structure of the cube must also be changed. The structure also changes if it is necessary to analyze the data in accordance with the new requirements. In this paper we consider the problem of data reclassification of a multidimensional information system, the results of which should be presented in a cluster form. Cluster description of the structure of a multidimensional data cube is based on identification of groups of members which are connected with groups of members of other dimensions. The use of cluster description allows to identify the semantics of a multidimensional data cube.

Keywords: Multidimensional data model · OLAP
Data reclassification · Sparse data cube
Set of possible members combinations
Cluster of members combinations

1 Introduction

The appearance of new information about the observed phenomenon or a change in the method of data analysis necessitates a change in the way the data is presented. In the case of multidimensional information systems, the data are represented by the values of measures and members that are the values of the multidimensional cube dimensions [1]. The transformations of this data change the structure of the multidimensional cube. The corresponding procedure is called "reclassification". If we use a large amount of semantically heterogeneous data for the description of the observed phenomenon the multidimensional cube is characterized by high sparseness [2–6]. A way to describe the structure of a sparse multidimensional cube is provided by the cluster method. It is based on the detection of semantically related groups of members. This paper considers a

The publication has been prepared with the support of the "RUDN University Program 5-100".

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 246–256, 2018.
https://doi.org/10.1007/978-3-319-99447-5_21

reclassification method in which the resulting multidimensional cube is described in the form of a set of clusters. Such description allows to reveal semantics of the data received as a result of reclassification.

2 Multidimensional Data Model

The structure of multidimensional data model should reflect the aspects of subject domain which are used in the data analysis process. Each aspect corresponds to one dimension of a multidimensional cube H. A full set of dimensions forms a set $D(H) = \{D^1, D^2, .., D^n\}$, there D^i is i–dimension, and $n = dim(H)$—dimensionality of multidimensional cube [7]. Each dimension is characterized by a set of members $D^i = \{d_1^i, d_2^i, .., d_{k_i}^i\}$, there i is a number of dimension, k_i—the quantity of members. Members of D^i are drawn from a set of positions of the basic classifier which corresponds to an aspect of the observed phenomenon associated with D^i [8,9].

The multidimensional data cube is a structured set of cells. Each cell c is defined by a combination of members $c = (d_{i_1}^1, d_{i_2}^2, .., d_{i_n}^n)$. The combination includes one member for each of the dimensions. If the analysis of the observed phenomenon is performed using a large set of diverse aspects, not all member combinations define the possible cells of multidimensional cube, i.e. the cells corresponding to a certain fact. This effect occurs due to semantic inconsistencies of some members from different dimensions to each other and generates a sparseness in the cube.

The complex structure of the compatibility of members may lead to a situation where a certain dimension becomes semantically uncertain if combined with a set of members from other dimensions. In this situation, while describing the possible cell of multidimensional cube the special value “Not in use” can be used to set the member of semantically unspecified dimension. The structure of the multidimensional data cube in the information system can be described as the set of possible member combinations. Different values from the classifiers, which comply with the dimensions, and the special value “Not in use” can be applied in the combinations of this set. To refer to the set of possible member combinations we will use the abbreviation “SPMC”.

The subject domain is characterized by the measure values defined in possible cells of the multidimensional cube. The full set of measures composes the set $V(H) = \{v_1, v_2, .., v_p\}$, where v_j is j-measure, p—the quantity of measures in the hypercube. Not all the measures from the $V(H)$ can be defined in the possible cell. This situation can appear in case of semantic inconsistency between the members defining the cell and some measures. While describing multidimensional data cube structure for every possible sell it is necessary to define its own set $V(c) = \{v_1, v, .., v_{p_c}\}$, which consists of certain measures for this cell, $1 \leq p_c \leq p$. We can use the special value “Not in use” for the description of c measures, which are not included in the set $V(c)$.

The description of the SPMC can be obtained with the help of the cluster method based on the analysis of links between members [14]. The cluster method

allows identifying the groups of members. The group $G_j^i = \left\{d_1^i, d_2^i, ..., d_{m_j}^i\right\}$ of members in i-dimension includes m_j members ($1 \leq m_j \leq k_j$), where j is a group number and contains members, which equally coincide in the SPMC with the members from some groups of members of other dimensions.

It is possible to define connected groups of member in different dimensions with the help of the semantic analysis. The cluster of member combinations K is the set of member combinations, which can be obtained with the help of Cartesian product where operands are groups of members or special value "Not in use"; one operand stands for every dimension used in the cluster $\text{SPMC}(K) = G_1 \times G_2 \times ... \times G_n$. Clusters of member combinations can be used for the description of the SPMC.

3 Data Reclassification of Multidimensional Information System

Metadata of a multidimensional information system is usually a fairly stable structure. However, if the way the data is analyzed changes, or if changes occur in the structure of the data domain, the corresponding changes should be made to the metadata. Changes can occur in a set of dimensions of a multidimensional data cube, in the structure of one or more dimensions, in a set of members, in a set of valid members combinations. In accordance with these changes, the data of the multidimensional cube should be reclassified.

Reclassification of a multidimensional data cube is a transformation, in which information about the facts, the description of which is presented in the cells of the cube, transformed and transferred to the cells of other multidimensional cube. In this case, the transformed data of the facts become presented in the aspects of new characteristics of the observed phenomenon related to the dimensions of the new cube.

There are two options for reclassification. In the first case, the set of multidimensional cube dimensions does not change. Transformations change the set of members and the structure of hierarchy of members [10–13]. In the second case, changes occur in the set of dimensions. We will consider the second option. In this case, the following transformations should be specified in the information system:

1. The set of dimensions must be converted to a new set of dimensions in the resulting cube: $D(H) = \left\{D^1, D^2, .., D^n\right\} \rightarrow D'(H) = \left\{D'^1, D'^2, .., D'^l\right\}$, there $l = dim(H')$ is the dimension of the resulting cube;
2. The set of members must be converted to a new set of members for each reclassified dimension: $D^i = \left\{d_1^i, d_2^i, .., d_{k_i}^i\right\} \rightarrow D'^i = \left\{d'^i_1, d'^i_2, .., d'^i_{k'_i}\right\}$, there i is a number of dimension, k_i—the quantity of members;
3. The set of measures must be converted to a new set of measures: $V(H) = \{v_1, v_2, .., v_p\} \rightarrow V'(H) = \left\{v'_1, v'_2, .., v'_{p'}\right\}$, there p is a number of measures;
4. The set of possible members combinations must be converted to a new set of possible members combinations in reclassified multidimensional data cube.

In order to transform the original multidimensional data cube into a reclassified data cube, the following transformation rules must be formulated:

1. The correspondence of members $d'^i \in D'^i$ of the reclassified dimension D'^i in the resulting multidimensional cube to the members combinations $\left(d^1_{i_1}, d^2_{i_2}, .., d^n_{i_n}\right)$, which describe the cell c in the original multidimensional cube, must be set;
2. The correspondence of the measure values v'_i, specified in cells c' of the resulting multidimensional data cube to measure values v_i, located in the cells c, which are the prototypes of cells c' upon the reclassification, must be set. Or the formula calculation v'_i depending on v_i must be specified. The correspondence can be established by specifying a formula of calculation v'_i. Often the a calculation formula for the measure value v'_i in the cells c' is the aggregation function of this measure.

The described rules of reclassification allow to build the resulting multidimensional data cube, to identify possible cells in it and to form a set of possible members combinations, to calculate the measures values in possible cells. To build members combinations in the SPMC of the reclassified data cube, you must specify the member for each dimension individually. This method does not allow the cluster structure to be transferred from the original multidimensional data cube to the reclassified data cube. To identify a cluster structure in the reclassified data cube, you must perform additional actions.

As an illustrative example, let us consider the reclassification of data for the observed phenomenon of "Granting of loans". The following aspects of analysis can be selected as dimensions of the multidimensional data cube:"Type of loan", "Term of the loan","Loan repayment method" and "Commission type". Members for the dimensions of the original data cube are presented in Table 1.

Table 1. Set of members for the dimensions of the original data cube for the observed phenomenon of "Granting of loans"

Type of loan	Term of the loan	Loan repayment method	Commission type
Operating	Short	At a time	Account opening fee
Consumer	Medium-term	Partially	For maintaining the account
Mortgage	Long		For the provision of services
Interbank			

As the measures for the analysis of the observed phenomenon "Granting of loans" indicators "Lending" and "Interest on the loan" can be taken.

Consider the case when there is a need to analyze the observed phenomenon in terms of the characteristics of the recipients of the loan. It is necessary to

reclassify the data of a multidimensional cube and to establish the dependence of the measures on new characteristics. Members for the dimensions of the reclassified data cube are presented in Table 2.

Table 2. Set of members for the dimensions of the reclassified data cube for the observed phenomenon of "Granting of loans"

Type of loan	Debtor type	Occupation	Debtor gender
Operating	Legal entity	Manufacturing	Male
Consumer	Natural person	Trade	Female
Mortgage		Banking	
Interbank			

In order to reclassify the multidimensional data cube in accordance with the new structure of dimensions, it is required to establish a correspondence between the members' combinations of the original cube and the members of the dimensions of reclassified cube. In case of observed phenomenon of "Granting of loans", the values of measures "Lending" and "Interest on the loan" in the cells of the reclassified data cube can be calculated by linking the members "Debtor type", "Occupation" and "Debtor gender" to the members combinations of the original data cube. For the values of measures, formulas should be obtained that take into account the type of relationship between the values of the original and the values of reclassified dimensions. The dependence on the values of the dimension "Debtor gender" should be taken into account in the formulas for the values of measures using additional data that are not available in the original data cube.

4 Description of the Reclassified Data Cube Using the Cluster Method

To identify semantics in the data specified in the multidimensional form, it is convenient to present possible members combinations of a multidimensional cube in the form of the set of clusters [14]. This method gives good results if the members are characterized by a complex compatibility and the multidimensional cube has a significant sparsity.

The main task in identifying the cluster structure of the multidimensional data cube is to build groups of members. To solve this problem, you need to analyze the properties of the set of possible members combinations. The group $G_j^i = \left\{d_1^i, d_2^i, .., d_{m_j}^i\right\}$ of members of the i-th dimension includes m_j values $(1 \leq m_j \leq k_i)$, there j is the group number. Members belonging to the group are combined "equally" in SPMC with members from some other groups of other dimensions. The cluster of possible members combinations can be defined

by linked groups of members from different dimensions. Cluster of members combinations is a set of combinations of members which can be obtained by means of the Cartesian product operation in which the operands are the groups of members or the special member "Not in use", one operand for each of the dimensions.

The task is to classify objects through possible combinations of these objects with other objects. You should use logical methods to solve this problem. The combinations of members should be represented as logical structures: the member must be represented as a pair "DimensionName.Value", and the members combination—in the form of a conjunction of pairs "DimensionName.Value", taken for all dimension. Described thus, all member combination from SPMC for reclassified multidimensional data cube must be connected with the disjunction.

In the constructed expression, we will combine elements that differ only in one conjunct for one of the dimensions, and continue the process of combining, considering other dimensions.

As a result of this process, similar to the process of bringing to a perfect disjunctive normal form, SPMC can be presented in a logical form as a conjunction of clusters, each of which contains a disjunctive description of groups of members.

In the illustrative example that is considered, the SPMC for reclassified multidimensional data cube has the structure that is presented in Table 3.

Table 3. Set of possible members combinations of reclassified multidimensional data cube

Type of loan	Debtor type	Occupation	Debtor gender
Operating	Legal entity	Manufacturing	Not in use
Operating	Legal entity	Trade	Not in use
Consumer	Natural person	Not in use	Male
Consumer	Natural person	Not in use	Female
Mortgage	Legal entity	Manufacturing	Not in use
Mortgage	Legal entity	Trade	Not in use
Mortgage	Legal entity	Banking	Not in use
Mortgage	Natural person	Not in use	Male
Mortgage	Natural person	Not in use	Female
Interbank	Legal entity	Banking	Not in use

The structure of SPMC presented in Table 1 is determined by the fact that no gender is defined for legal entities, the occupation is not used in the description of natural person, debtors of different types can take different loans. The complex structure of the observed phenomenon is the cause of the sparsity of the multidimensional data cube, which describes the facts of this phenomenon.

The problem of cluster description of the structure of the multidimensional data cube is divided into two subtasks:

1. Construction of groups of members using information about the compatibility of these members in the SPMC;
2. Construction of subsets of combinations of members—clusters of members combinations in which these groups of members play the same role.

The features of the problem statement allow to classify it as a problem of intellectual analysis of the logical type. An important argument in favor of the choice of logical methods is the need to interpret the results of the decision. The following conclusions can be drawn: logical methods can be used to solve the problem. Combinations of members must be interpreted as logical structures; the problem to be solved is a clustering problem. To build a cluster description of the structure of the multidimensional data cube, you need to perform the following steps.

Step 1. Description of the combinations of members in a logical form. Combinations of members, the description of which is presented in Table 1, should be represented as a conjunction of pairs "DimensionName.Value". Logical expressions describing SPMC have the following form:

TypeOfLoan.Operating & DebtorType.LegalEntity & Occupation.Manufacturing & DebtorGender.NotInUse;
TypeOfLoan.Operating & DebtorType.LegalEntity & Occupation.Trade & DebtorGender.NotInUse;
TypeOfLoan.Consumer & DebtorType.NaturalPerson & Occupation.NotInUse & DebtorGender.Male;
TypeOfLoan.Consumer & DebtorType.NaturalPerson & Occupation.NotInUse & DebtorGender.Female;
TypeOfLoan.Mortgage & DebtorType.LegalEntity & Occupation.Manufacturing & DebtorGender.NotInUse;
TypeOfLoan.Mortgage & DebtorType.LegalEntity & Occupation.Trade & DebtorGender.NotInUse;
TypeOfLoan.Mortgage & DebtorType.LegalEntity & Occupation.Banking & DebtorGender.NotInUse;
TypeOfLoan.Mortgage & DebtorType.NaturalPerson & Occupation.NotInUse & DebtorGender.Male;
TypeOfLoan.Mortgage & DebtorType.NaturalPerson & Occupation.NotInUse & DebtorGender.Female;
TypeOfLoan.Interbank & DebtorType.LegalEntity & Occupation.Banking & DebtorGender.NotInUse.

Step 2. Connection of members combinations. Logical expressions for members combinations that differ only by one conjunct must be sequentially connected by disjunctions. As a result, the following expression can be obtained:

TypeOfLoan.{Consumer—Mortgage} & DebtorType.NaturalPerson & Occupation.NotInUse & DebtorGender.{Male—Female};

TypeOfLoan.{Operating—Mortgage} & DebtorType.LegalEntity & Occupation.{Manufacturing—Trade} & DebtorGender.NotInUse;
TypeOfLoan.{Mortgage—Interbank} & DebtorType.LegalEntity & Occupation.Banking & DebtorGender.NotInUse.

Step 3. Formation of groups of members. As a result of connecting of members combinations, groups of members are formed. These groups are characterized by the same compatibility of members that belong to them in SPMC. Each expression in the resulting description of SPMC defines a separate cluster of members combinations.

The groups of members identified in the illustrative example are as follows:

- for dimension "Type of loan"—{Consumer, Mortgage}, {Operating, Mortgage}, {Mortgage, — Interbank};
- for dimension "Debtor type"—{Legal entity}, {Natural person}, {Not in use};
- for dimension "Occupation"—{Manufacturing, Trade}, {Banking}, {Not in use};
- for dimension "Debtor gender"—{Male, Female}, {Not in use}.

Figure 1 presents clusters of members combination that are described using groups of members, which are described above.

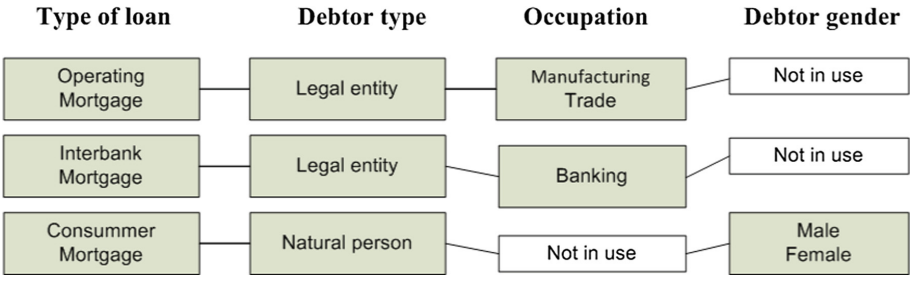


Fig. 1. Clusters of members combinations for SPMC

The use of cluster description of a multidimensional data cube allows significantly simplify the description of its structure and to identify its semantics.

The cluster description allows you to identify relationships between the dimensions in the multidimensional data cube. Figure 2 presents the diagrams containing the designations of the pairwise relations between the dimensions of the reclassified data cube.

The cluster description allows you to identify the following relationships between the dimensions in the reclassified data cube:

1. Association. There is an association in a pair of dimensions D^1 and D^2 if n groups, $n \geq 2$, can be singled out of a set of members of each of them, and a bijection can be established between these groups which manifests as follows: if a combination of SPMC includes the members D^1 and D^2, they come in pairs, taken from the corresponding groups of members;

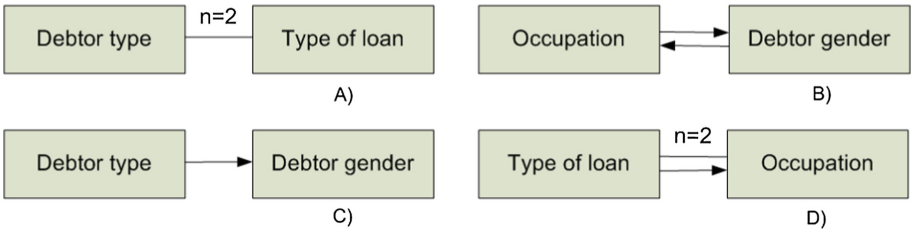


Fig. 2. Pairwise relations between the dimensions of the reclassified data cube: association (A), two-sided dependence (B), dependence (C), association and dependence (D)

2. Dependence. There is a dependence between dimensions D^1 and D^2 (D^2 depends on D^1) if the members of D^1 can be divided in two groups of members in such way that if a certain combination from SPMC includes the member from the first group of members D^1, the member of D^2 in this combination is possible, and if the member of the second group of members D^1 is included into the combination, the D^2 in such combination is set to the "Not in use" member.
3. Association and dependence. There is an association and dependence between D^1 and D^2 if n groups can be singled out of D^2, $n \geq 1$, and $(n+1)$—out of D^1 in such way that there is an association between first n groups from D^1 and D^2, and if the combination of SPMC includes the member from $(n+1)$ group of D^1 members, D^2 in this combination is set to the "Not in use" member. Besides, the members from $(n+1)$ group of D^1 members can not be met in other groups of this dimension;
4. Two-sided dependence. There is a two-sided dependence between D^1 and D^2 dimensions if the following rule holds: in case of SPMC combination includes the member from D^1, the D^2 in this combination is set to the "Not in use" member, and when the combination includes the member from D^2, the D^1 in this combination is set to the "Not in use" member.

In order to reclassify the multidimensional data cube in accordance with the new structure of dimensions, it is required to establish a correspondence between the members combinations of the original cube and the members of the dimensions of reclassified cube. In case of observed phenomenon of "Granting of loans", the values of measures "Lending" and "Interest on the loan" in the cells of the reclassified data cube can be calculated by linking the members "Debtor type", "Occupation" and "Debtor gender" to the members combinations of the original data cube. The formulas for values of measures "Lending" and "Interest on the loan" should be obtained in accordance with relationship between the values of the dimensions "Type of loan","Term of the loan","Loan repayment method" and "Commission type" of the original dimensions and the reclassified dimensions.

5 Conclusions

In this paper the method of data reclassification of a multi-dimensional information system was considered. Reclassification allows you to present the data of a multidimensional cube depending on new characteristics, added as a result of reclassification, and allowing you to use new aspects of analysis. The need for reclassification can arise for two reasons. The first is a change in the properties of the observed phenomenon, the description of which is presented in the information system. The second is the need to analyze the indicators of the observed phenomenon using new aspects.

The method of representation of the reclassified data in the cluster form is offered. A clustered description of the data of a multidimensional information system makes it possible to identify the semantics of data. The method is based on the construction of groups of members that are semantically related to groups of members of other dimensions. This allows you to combine several facts that have similar properties to the analysis aspects. The method of data representation in the cluster form is based on the description of set of facts in the form of logical expressions and transformation of these expressions into a union of clusters, each of which contains a disjunctive description of groups of values.

References

1. Thomsen, E.: OLAP Solution: Building Multidimensional Information System. Willey, New York (2002)
2. Hirata, C.M., Lima, J.C.: Multidimensional cyclic graph approach: representing a data cube without common sub-graphs. Inf. Sci. **181**, 2626–2655 (2011)
3. Luo, Z.W., Ling, T.W., Ang, C.H., Lee, S.Y., Cui, B.: Range top/bottom k queries in OLAP sparse data cubes. In: Mayr, H.C., Lazansky, J., Quirchmayr, G., Vogel, P. (eds.) DEXA 2001. LNCS, vol. 2113, pp. 678–687. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44759-8_66
4. Vitter, J.S., Wang, M.: Approximate computation of multidimensional aggregates of sparse data using wavelets. In: Proceedings of the 1999 International Conference on Management of Data – SIGMOD 1999, pp. 193–204. ACM, New York (1999)
5. Messaoud, R.B., Boussaid, O., Rabaseda, S.L.: A multiple correspondence analysis to organize data cube. In: Databases and Information Systems IV – DB&IS 2006, pp. 133–146. IOS Press, Vilnius (2007)
6. Fu, L.: Efficient evaluation of sparse data cubes. In: Li, Q., Wang, G., Feng, L. (eds.) WAIM 2004. LNCS, vol. 3129, pp. 336–345. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-27772-9_34
7. Chen, C., Feng, J., Xiang, L.: Computation of sparse data cubes with constraints. In: Kambayashi, Y., Mohania, M., Wöß, W. (eds.) DaWaK 2003. LNCS, vol. 2737, pp. 14–23. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-45228-7_3
8. Salmam, F.Z., Fakir, M., Errattahi, R.: Prediction in OLAP data cubes. J. Inf. Knowl. Manag. **15**, 449–458 (2016)

9. Gomez, L.I., Gomez, S.A., Vaisman, A.: A generic data model and query language for spatiotemporal OLAP cube analysis. In: Rundensteiner, E., Markl, V., Manolescu, I., Amer-Yahia, S., Naumann, F., Ari, I. (eds.) Proceedings of the 15-th International Conference on Extending Database Technology – EDBT 2012, pp. 300–311. ACM, New York (2012)
10. Karayannidis, N., Sellis, T., Kouvaras, Y.: CUBE file: a file structure for hierarchically clustered OLAP cubes. In: Bertino, E., Christodoulakis, S., Plexousakis, D., Christophides, V., Koubarakis, M., Böhm, K., Ferrari, E. (eds.) EDBT 2004. LNCS, vol. 2992, pp. 621–638. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24741-8_36
11. Jin, R., Vaidyanathan, J.K., Yang, G., Agrawal, G.: Communication and memory optimal parallel data cube construction. IEEE Trans. Parallel Distrib. Syst. **16**, 1105–1119 (2005)
12. Wang, W., Lu, H., Feng, J., Yu, J.X.: Condensed cube: an effective approach to reducing data cube size. In: Proceedings of the 18th International Conference on Data Engineering — ICDE 2002, pp. 155–165. IEEE Computer Society, Washington (2002)
13. Goil, S., Choudhary, A.: Design and implementation of a scalable parallel system for multidimensional analysis and OLAP. In: Rolim, J. (ed.) IPPS/SPDP 1999, pp. 576–581. Springer, Heidelberg (1999)
14. Fomin, M.B.: Cluster method of description of information system data model based on multidimensional approach. In: Vishnevskiy, V., Samouylov, K., Kozyrev, D. (eds.) DCCN 2016. CCIS, vol. 678, pp. 657–668. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-51917-3_56

Method for Adaptive Node Clustering in AD HOC Wireless Sensor Networks

Alexander Alexandrov(✉) and Vladimir Monov

Institute of Information and Communication Technologies,
Bulgarian Academy of Sciences, Acad. Georgi Bonchev Str, Block 2,
Office 302, 1113 Sofia, Bulgaria
akalexandrov@iit.bas.bg
http://www.iikt.bas.bg

Abstract. The research propose a new approach related to WCA (Weighted Clustering Algorithm) based on a modified method for ad-hoc clustering in wireless sensor networks (WSN). The proposed method include a new functionality for sensor node clustering by including a method for link quality (LQ) prioritization. The method reduce sensitively the risk of crash of the WSN cluster coordinators and help for the energy optimization of the implemented routing protocols.

Keywords: WSN · Ad-hoc · Sensor · Node · Clustering

1 Related Work

The proposed research is based on the WCA (Weighted Clustering Algorithm) for neighbor sensor nodes analysis and ad-hoc clustering. The WCA is developed on ideas proposed by Chatterjee et al. [1] Zabian et al. [2] and Lehsaini et al. [3]. In the current research, the original WCA version is modified and upgraded by adding new parameters and related equations referencing the sensor nodes for quality link communications. The proposed new approach allows sensitively faster and energy efficient ad hock nodes clustering process. The new method is based on the following preliminary assumptions: - all the sensor nodes communicate in only one radio channel during the ad-hock clustering process. - every sensor node has fixed his coordinates; - every sensor node know his energy level; - during the clustering process only one Cluster Head (CH) exist; - during the CH choice every sensor node can communicate directly with his neighbor sensor nodes.

2 Problem Statement - Method for Dynamical Ad-Hock Cluster Generation

The proposed method defines 3 main phases—discovery phase, assigning phase and monitoring phase. The discovery phase starts a process of neighbor nodes

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 257–263, 2018.
https://doi.org/10.1007/978-3-319-99447-5_22

discovery and mapping. The task of the assigning phase is to assign special weight coefficients to every node depend on their specific parameters and to choose the CH. The main role of the monitoring phase is to ensure an adaptive dynamical reconfiguration of the sensor network.

2.1 Discovery Phase

The discovery phase starts a process for neighbor nodes discovery by transmitting a broadcast message.

Based on the received answers with node coordinates and the signal level, calculated by integrated into the sensor node RSSI (Received Signal Strength Indicator) starts an analysis of the node parameters and calculation of the weight coefficients. The weight coefficients depends on the number of the directly answered neighbor nodes and the received signal strength.

In this phase, we propose a new additional key approach based on a dynamical link quality measurement LQI (Link Quality Indicator). During the latest research in real sensor network environment, we found that very often the short distances between sensor nodes cannot guarantee an acceptable link quality. In more than 30% of the cases in real networks, the link quality between sensor nodes situated on a relatively large distance is much better and consume a less energy for packages transmitting than a link quality between sensor nodes situated in radio noise environment or behind radio non-permissible barriers or shields.

The parameters based on which we calculate the weight coefficient K_{weight} for every sensor node N_i are as follows: - Ci represents the number of neighbor sensor nodes in the communication distance of node N_i

$$C_i = |N(i)| = \sum_{j \in N(i), i \neq j} \{dist(i,j) < S_{range}\} \tag{1}$$

- $dist(i,j)$ represent the distance between a couple nodes in communication distance.
- S_{range} is a coefficient who represent the maximum communication distance between two nodes and depend on the current hardware implementation of the network.
- D_i represent the average distance between node Ni and the neighbor nodes j.

For every node, we calculate D_i by the Eq. (2)

$$D_i = \frac{1}{C_i} \sum_{j \in N(i)} \{dist(i,j)\} \tag{2}$$

- M_i is a coefficient representing the mobility of the sensor node e.g. the probability this node to change his coordinates during the period

$$M_i = \frac{1}{T}\sum_{t=1}^{T}\sqrt{(X_t - X_{t-1})^2 + (Y_t - Y_{t-1})^2} \tag{3}$$

where X_t, Y_t and X_{t-1}, Y_{t-1} are the coordinates of the node I in time t and $t-1$.

– $E_{current}$ represents the current amount of energy of node N_i Generally

$$E_{current} = E_i - (E_{rx}.t1 + E_{tx}.t2 + E_{comp}.t3 + E_{sensor}.t4) \tag{4}$$

where E_i - is the initial amount of energy of the node i;

E_{rx} - the amount of energy needed for the receiving signals mode;

E_{tx} - the amount of energy needed for the transmitting mode of the sensor node;

E_{comp} - the amount of energy needed for the work of the microcontroller block for the clusters generation operations;

E_{sensor} - the amount of energy needed for the sensor node coordinates calculation;

t_1–t_4 - time intervals related to the sensor blocks energy consumption E_{rx}, E_{tx}, E_{comp} and E_{sensor};

- $K_{i_l ink}$ represent the link quality (LQ) of node i. $K_{i_l ink}$ is calculated by the equation:

$$K_{i_l ink} = K_{tx_j} + \frac{1}{C_j}\sum_{i}^{j=0} K_{i\_RSSI} \tag{5}$$

where K_{tx} is a parameter representing the power of the transmitted by N_i signal and KRSSI is a parameter representing the power of the received signal transmitted by the neighbor node N_j.

The $K_{i_l ink}$ coefficient is a key parameter for our research because he represents the quality of the communication environment and allow the optimization of the clustering process. The weight coefficient N_i is calculated by the equation:

$$P_i = \omega_1.C_i + \omega_2.D_i + \omega_3.M_i + \omega_4.K_{i_l ink} \tag{6}$$

where $\omega_1 + \omega_2 + \omega_3 + \omega_4 + \omega_5 = 1$.

The correction coefficients ω_1 till ω_5 are chosen in compliance with the specific requirements related to the ad-hoc clusters generation. For example in sensor networks with reduced mobility $\omega_3 = 0$ and the ω_5 is between 0.5 and 0.7.

The discovery phase finished with the calculation of the weight coefficients of all the nodes in the sensor network.

2.2 Assigning Phase

In the assigning phase we start the process of CH choise. On the basis of the already calculated weight coefficients we choise the sensor node with the maximal weight coefficient as a cluster head (CH) and the connected to him nodes are excluded form the process of other CH choise.

After the first CH choise the procedure is repeated with the rest nodes till the final clustering of the sensor network.

The process of the communication between two sensor nodes is illustrated on Fig. 1.

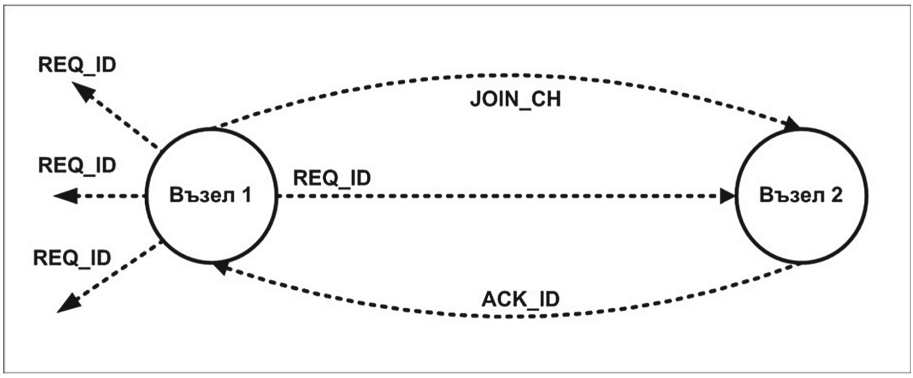


Fig. 1. Process of the communication between two sensor nodes during the clustering

On this stage are used basically 2 types of packages - JOIN_CH and ACK_ID $JOIN_CH$ is a message who is transmitted individually to every node in the communication range from the node with calculated maximal weight coefficient who act as a CH.

The message include ID generated by the CH cluster by which the receiving sensor node is assigning to the cluster as a cluster member.

ACK_ID is a message sent as a confirmation answer of the JOIN_CH request. After the ACK_ID transmission, the sensor node stops to answer of the JOIN_CH or REQ_ID messages from other nodes.

The block diagram of Fig. 2 illustrate the process of clustering.

2.3 Monitoring Phase

The monitoring phase in the ad-hock based WSNs is the most important phase during the process of the network adaptation.

On the basis of the monitoring results is executed a process of adaptive reconfiguration and redistribution of the communication node roles.

In the proposed method in the monitoring phase, we analyze the following 5 situations leading to adaptive reconfiguration:

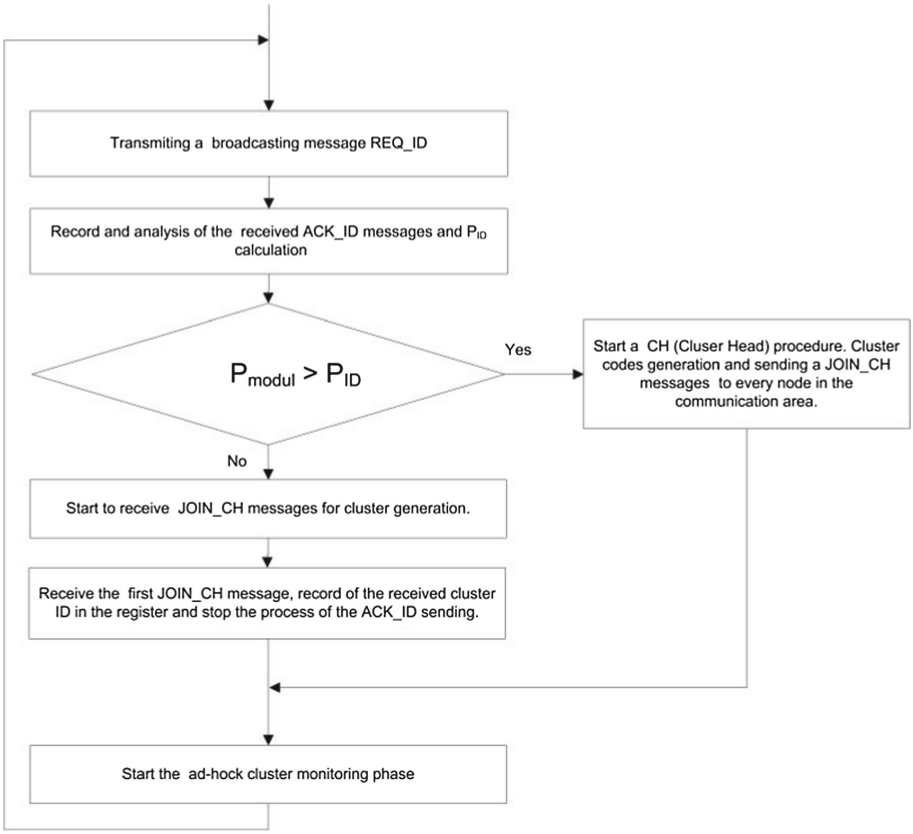


Fig. 2. Diagram of the proposed clustering workflow process.

- sensor nodes battery capacity near to critical minimum;
- add a new sensor node to the cluster (network);
- movement of the sensor node (typical for mobile sensor networks);
- the crash of the sensor node;
- critical change of the link quality (LQ) between existing cluster head and cluster nodes.

Every cluster head CHi periodically starts a monitoring procedure to the nodes assigned to him by sending a broadcasting messages START_MON. The START_MON message include the packages START_MON1 and START_MION2. Every node n_i $(i \neq j)$ from the cluster i when receive a START_MON1 package starts an own procedure for link calculation as follows:

- number of packages sent by CH_i to n_i and the time period between them $\triangle t = [t_0, t] : Nbp\_Send(n_i, \triangle t)$
- the delay calculation between two sequental packages:

$$Delay(n_i, t) = Arrival_{pTi} - Arrival_{pTi-1} \tag{7}$$

– calculation of the energy consumed by node n_i during the process of receiving the packages and sending a confirmation of the received package.

$$E_c(n_i, \triangle t) = E_r n_i, t_0 - E_r n_i, t_1 \tag{8}$$

– $\triangle t$ represents the time period $[t_0, t_1]$;

3 Conclusion

In the proposed research is developed a new conception for WSN ad-hock nodes cluster generation. On the base of WCA defined in previous publications is developed a new clustering method with following key benefits:

– sensor node clusters generation based on the priority of the quality of the links (Ki_link parameter) between sensor nodes not by the distance between sensor nodes. This approach reduces sensitively the risk of network crashes generated by randomly increased radio noise level;
– improved CH criteria approach. In our approach, the CH is the node with the biggest weight coefficient which generates a totally different network topology compared to the original WCA algorithm;
– relatively small number cluster heads in combination with a bigger number of cluster members.

These new key changes allow sensitively optimization of the package routing process and energy efficient communication between sensor nodes, cluster heads, and the PAN coordinator.

References

1. Bianchi, G.: Performance analysis of the IEEE 802.11 distributed coordination function. IEEE J. Sel. Areas Commun. **18**, 535–547 (2000)
2. Lee, D.L., Chen, Q.: A model-based WiFi localization method. In: Proceedings of the 2nd International Conference on Scalable Information Systems, p. 40. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering) (2007)
3. Mautz, R.: Indoor positioning technologies. Doctoral dissertation, Habilitationsschrift ETH Zurich (2012)
4. Bulusu, N., Heidemann, J., Estrin, D.: GPS-less low cost outdoor localization for very small devices. IEEE Pers. Commun. Mag. **7**(5), 28–34 (2000). Special Issue on Smart Spaces and Environments
5. Sikora, A., Groza, V.F.: Coexistence of IEEE802.15.4 with other systems in the 2.4 GHz-ISM-Band. In: Proceedings of the IEEE Instrumentation and Measurement Technology Conference, IMTC 2005, vol. 3, pp. 1786–1791. IEEE (2005)
6. Voute, R.: CGIs indoor positioning for hospitals - operational control of assets and clients (2015)

7. Hall, D.L., Llinas, J.: An introduction to multisensor data fusion. Proc. IEEE **85**(1), 6–23 (1997)
8. Bishop, G., Welch, G.: An Introduction to the Kalman Filter. Department of Computer Science, University of North Carolina, UNC-Chapel Hill, TR 95–041, 24 July 2006
9. Haghighat, M.B.A., Aghagolzadeh, A., Seyedarabi, H.: Multi-focus image fusion for visual sensor networks in DCT domain. Comput. Electr. Eng. **37**(5), 789–797 (2011)
10. Julier, S.J., Uhlmann, J.K.: A new extension of the Kalman filter to nonlinear systems. In: Proceedings of the International Symposium on Aerospace/Defense Sensing, Simulation and Controls, vol. 3 (1997)
11. Luo, R.C., Yih, C.-C., Su, K.L.: Multisensor fusion and integration: approaches, applications, and future research directions. IEEE Sens. J. **2**(2), 107–119 (2002)
12. Llinas, J., Bowman, C., Rogova, G., Steinberg, A., Waltz, E., White, F.: Revisitingthe JDL data fusion model II. Technical report, DTIC Document (2004)
13. Blasch, E.P., Plano, S.: JDL level 5 fusion model user refinement issues and applications in group tracking. In: Proceedings of the Signal Processing, Sensor Fusion, and Target Recognition XI, pp. 270–279, April 2002
14. Durrant-Whyte, H.F., Stevens, M.: Data fusion in decentralized sensing networks. In: Proceedings of the 4th International Conference on Information Fusion, Montreal, Canada, pp. 302–307 (2001)
15. Chen, L., Wainwright, M.J., Cetin, M., Willsky, A.S.: Data association based on optimization in graphical models with application to sensor networks. Math. Comput. Model. **43**(9–10), 1114–1135 (2006)
16. Weiss, Y., Freeman, W.T.: On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. IEEE Trans. Inf. Theory **47**(2), 736–744 (2001)
17. Brown, C., Durrant-Whyte, H., Leonard, J., Rao, B., Steer, B.: Distributed data fusion using Kalman filtering: a robotics application. In: Abidi, M.A., Gonzalez, R.C. (eds.) Data, Fusion in Robotics and Machine Intelligence, pp. 267–309. Academic Press, New York (1992)
18. Atanasova, T.: Modelling of complex objects in distance learning systems. In: Proceedings of the First International Conference - "Innovative Teaching Methodology", Tbilisi, Georgia, 25–26 October 2014, pp. 180–190 (2014). ISBN 978-9941-9348-7-2
19. Tashev, T.D., Hristov, H.R.: Modeling of synthesis of information processes with generalized nets. J. Cybern. Inf. Technol. **2**, 92–104 (2003)
20. Balabanov, T., Zankinski, I., Barova, M.: Strategy for individuals distribution by incident nodes participation in star topology of distributed evolutionary algorithms. Cybern. Inf. Technol. **16**(1), 80–88 (2016). Print ISSN 1311-9702, Online ISSN 1314-4081

The Model and Algorithms for Estimation the Performance Measures of Access Node Serving the Mixture of Real Time and Elastic Data

Sergey N. Stepanov(✉) and Mikhail S. Stepanov

Department of communication networks and commutation systems,
Moscow Technical University of Communication and Informatics,
8A, Aviamotornaya str., Moscow 111024, Russia
stpnvsrg@gmail.com, mihstep@yandex.ru

Abstract. The model of joint servicing by access node the real time and elastic traffics is proposed. Flows of requests for real time servicing is described by Poisson model (narrowband traffic) or Engset model (broadband traffic). The requests for data transmission are coming by groups according to the Poisson model. Requests from the group occupy free transmission resources or free waiting positions if all transmission resources are occupied. Real time traffic has advantage in occupying the link transmission capacity. It exhibits itself in decreasing if necessary the speed of data transmission to some minimum value. When link has free capacity the speed of data transmission is increasing to some maximum value. The exact and approximate algorithms of estimation the model's performance measures are suggested. The obtained results can be used for estimation the required transmission capacity of access node for given volumes of offered real time and elastic traffics.

Keywords: Access node · Finite number of sources
Dynamic resource distribution · Performance measures
System of state equations · Procedure of decomposition

1 Introduction

New services provided to subscribers by operators require much more transmission resources than the traditional speech communication. For fixed networks the need in transmission capacity normally is solved by implementation of high-speed optical lines. For mobile networks the situation is much more harder because the presence of competition and existence of physical constraints on the transmission of radio signals. The formulated problems can be partly solved by using the procedures of the dynamic allocation of the transmission resources between

The publication has been prepared with the support of the Russian Foundation for Basic Research, project No. 16-29- 09497ofi-m.

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 264–275, 2018.
https://doi.org/10.1007/978-3-319-99447-5_23

subscribers being on servicing. Procedures of such kind are implemented by the so-called packet scheduler representing a software analytical complex intended to enhance the efficiency of traffic transmission based on the information of channel state, number of requests on servicing, noise level, and so on. The theoretical study of scheduler is very important from the practical point of view and considered in numerous publications [1–9].

Such models are used not only for the purpose of modern mobile networks optimization. Now 4G networks evolving to 5G and Internet of Things (IoT) becomes new reality requiring to study conjoint servicing of several categories of traffic—multimedia, sensory and so on. Each category has its own performance measures that makes the problem of bandwidth sharing more complex. Such situation takes place, for example, in Narrow Band IoT (NB-IoT) technology. Multimedia traffic include video data from surveillance cameras located in public and private places. Sensory traffic is collected from numerous distributed sensors (machine-to-machine communication) of thousands IoT devices like intelligent road signs, traffic lights etc. Conjoint servicing of different traffic streams is main feature of future networks and needs theoretical study for the purpose of efficient usage of limited radio resource.

The analytical results are mostly restricted to the cases of monoservice models with processor sharing service discipline [2,3]. Multiservice models with elastic traffic streams can be analyzed by recursive algorithms for some choices of resource sharing procedures [2,4]. The estimation of performance measures of the mix of real time and elastic traffic streams presents greater difficulties which is due to the complexity of the random processes describing the changes in the model states [5–9]. In this paper the model of joint servicing by access node the real time and elastic traffic streams will be constructed and the exact and approximate algorithms of estimation the model's performance measures are elaborated. Section 2 describes the model functioning. Section 3 constructs a system of equilibrium equations and suggests an algorithm to solve it numerically. Section 4 defines the characteristics. Section 5 considers the procedure of decomposition that can be used for approximate calculation of the elastic data servicing.

2 Model Description

Let us denote in the model by C the speed of information transmission in bit/s provided by the equipment of the access node of mobile network. In the model the process of conjoint servicing of requests for real time and elastic traffic transmission is considered. The arriving of requests for real time transmission is modelled by two flows. The requests of the first flow follow Poisson process with intensity λ_1. In order to serve one request of the first flow it is necessary to reserve the node transmission capacity of r_1 bit/s for narrowband traffic transmission. The service time has exponential distribution with parameter μ_1. The requests of the second flow are arriving from the group of s subscribers. Each active subscriber sends a request after random time having exponential distribution with parameter γ. In order to serve one requests of the second flow it is necessary to reserve

the node transmission capacity of r_2 bit/s for broadband communication. The service time has exponential distribution with parameter μ_2. The model of the second flow allow to consider the segment of so called heavy subscribers consuming large amount of link transmission capacity. Let us denote as a_1 and a_2 the offered loads in erlangs of the first and the second flows of requests for real time traffic transmission. It is clear that the following relations are valid: $a_1 = \lambda_1/\mu_1$; $a_2 = \gamma s/(\gamma + \mu_2)$.

The requests for data (files) transmission arrive in groups (batches). We assume that arrival of the groups follows the Poisson model with intensity λ_d. With the probability f_k, $k = 1, \ldots, z$, the arriving group has k requests for files transmission. Let us denote by w the number of waiting positions. Depending on the volume of free transmission capacity and the number of free waiting positions some requests of the arriving group are taken for servicing or waiting other are lost without resuming.

The requests for elastic data transmission get the link transmission capacity of r_d bit/s for file downloading satisfying the inequality $r_{d,1} \le r_d \le r_{d,2}$. Here $r_{d,1}$ and $r_{d,2}$ are correspondingly the minimum and maximum link transmission capacity for elastic data transmission. Let us suppose that file volume has exponential distribution with mean value F expressed in bits. It is clear that the time of downloading the file with only minimum or maximum link transmission capacity has exponential distribution with parameters α_1 and α_2. The values of α_1 and α_2 can be found from relations $\alpha_1 = r_{d,1}/F$ and $\alpha_2 = r_{d,2}/F$ correspondingly.

Let us consider in more detail the procedure of requests admission control. Let i_1, i_2 are correspondingly the number of requests of the first and the second flows for real time traffic transmission being on servicing at the moment of request arrival and $\ell = i_1 r_1 + i_2 r_2$ is the node transmission capacity occupied by real time traffic transmission at the moment considered. Let d is the number of requests for elastic data transmission being on servicing or waiting at the moment considered. Let us analyze the arrival of request for real time traffic transmission. If $\ell + dr_{d,2} + r_n \le C$ then a request of n-th flow, $n = 1, 2$, is accepted for servicing. If $\ell + dr_{d,2} + r_n > C$ and $\ell + dr_{d,1} + r_n \le C$ then a request of n-th flow, $n = 1, 2$, is accepted for servicing and the speed of files transmission being on servicing is diminishing from $(C - \ell)/d$ to $(C - \ell - r_n)/d$. If $\ell + dr_{d,1} + r_n > C$ then a request of n-th flow, $n = 1, 2$, gets refusal and does not resumed. In all situation the service time of a request of n-th flow, $n = 1, 2$, has exponential distribution with parameter μ_1 and μ_2 correspondingly.

Now let us consider the admission to service of one request from the group for elastic data transmission. If $\ell + dr_{d,2} + r_{d,2} \le C$ then a request is accepted for servicing and gets the maximum allowed link transmission capacity $r_{d,2}$. In this situation the service time of each request for data transmission has exponential distribution with parameter α_2. If $\ell + dr_{d,2} + r_{d,2} > C$ and $\ell + dr_{d,1} + r_{d,1} \le C$ then a request is accepted for servicing and the speed of files transmission being on servicing is diminishing from $(C - \ell)/d$ to $(C - \ell)/(d + 1)$. In this situation the service time of each request for data transmission has exponential distribution

with parameter $\alpha_d = (C - \ell)/((d+1)F)$. If $\ell + dr_{d,1} + r_{d,1} > C$ then a request for file transmission gets refusal in immediate servicing and occupies free waiting position. If in the situation considered all waiting positions are occupied then coming request is lost and does not resumed. The maximum allowed time of waiting is restricted by random variable having exponential distribution with parameter equals σ.

The process of usage the transmission capacity of of access node serving the mixture of real time and elastic data is shown on the Fig. 1.

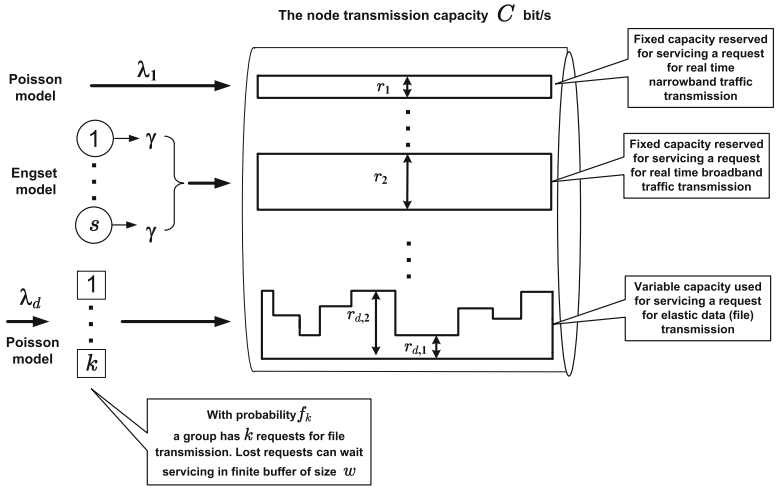


Fig. 1. The model of distribution the transmission capacity of access node serving the mixture of real time and elastic data

3 System of State Equations

Let us denote by symbols $i_1(t)$ and $i_2(t)$ correspondingly the number of requests of the first and the second flows for real time traffic transmission being on servicing at time t and by $d(t)$ denote the number of requests for elastic data transmission being on servicing or waiting at time t. The dynamic of the model states changing is described by multidimensional Markov process with components $r(t) = (i_1(t), i_2(t), d(t))$, defined on the finite set of model's states S with components i_1, i_2, d taking values

$$i_1 = 0, 1, \ldots, \left\lfloor \frac{C}{r_1} \right\rfloor; \quad i_2 = 0, 1, \ldots, \min\left(s, \left\lfloor \frac{C - i_1 r_1}{r_2} \right\rfloor\right); \tag{1}$$
$$d = 0, 1, \ldots, \left(w + \left\lfloor \frac{C - i_1\, r_1 - i_2\, r_2}{r_{d,1}} \right\rfloor\right).$$

Let $p(i_1, i_2, d)$ be the values of stationary probabilities of states $(i_1, i_2, d) \in S$. The model performance measures can be expressed through values of $p(i_1, i_2, d)$ that can be found from the solution of the system of state equations. Let us denote for arbitrary model state $(i_1, i_2, d) \in S$ as ℓ the node transmission capacity occupied by real time traffic transmission $\ell = i_1 r_1 + i_2 r_2$. After equating the intensity of leaving arbitrary model state $(i_1, i_2, d) \in S$ to the intensity of entering the state (i_1, i_2, d) we obtain the following system of linear equations:

$$
\begin{aligned}
P(i_1,i_2,d)\Big\{ & \lambda_1 I(\ell + dr_{d,1} + r_1 \le C) \qquad (2)\\
& + (s-i_2)\gamma I(\ell + dr_{d,1} + r_2 \le C) + \lambda_d I(\ell + dr_{d,1} + r_{d,1} \le C)\\
& + \lambda_d I\left(\ell + dr_{d,1} + r_{d,1} > C, d - \left\lfloor \frac{C-\ell}{r_{d,1}} \right\rfloor < w\right) + i_1\mu_1 I(i_1 > 0)\\
& + i_2\mu_2 I(i_2 > 0) + \min\left(d\alpha_2, \frac{C-\ell}{F}\right) I(\ell + dr_{d,1} \le C, d > 0)\\
& + \left(\left\lfloor \frac{C-\ell}{r_{d,1}} \right\rfloor \frac{r_{d,1}}{F} + \left(d - \left\lfloor \frac{C-\ell}{r_{d,1}} \right\rfloor\right)\sigma\right) I(\ell + dr_{d,1} > C)\Big\}\\
= {} & P(i_1 - 1, i_2, d)\lambda_1 I(i_1 > 0, \ell + dr_{d,1} \le C)\\
& + P(i_1, i_2 - 1, d)(s - i_2 + 1)\gamma I(i_2 > 0, \ell + dr_{d,1} \le C)\\
& + \sum_{k=1}^{d} P(i_1, i_2, d-k)\lambda_d\\
& \times \left(f_k + I\left(\ell + dr_{d,1} \ge C, d - \left\lfloor \frac{C-\ell}{r_{d,1}} \right\rfloor = w\right) \sum_{j=k+1}^{z} f_j\right)\\
& + P(i_1 + 1, i_2, d)(i_1 + 1)\mu_1\Big(I(\ell + dr_{d,1} + r_1 \le C)\\
& + I(\ell + dr_{d,1} + r_1 > C, \ell + r_1 \le C, d - \left\lfloor \frac{C-\ell-r_1}{r_{d,1}} \right\rfloor \le w)\Big)\\
& + P(i_1, i_2 + 1, d)(i_2 + 1)\mu_2\Big(I(\ell + dr_{d,1} + r_2 \le C, i_2 + 1 \le s)\\
& + I(\ell + dr_{d,1} + r_2 > C, i_2 + 1 \le s, \ell + r_2 \le C, d - \left\lfloor \frac{C-\ell-r_2}{r_{d,1}} \right\rfloor \le w)\Big)\\
& + P(i_1, i_2, d+1)\min\left\{(d+1)\alpha_2, \frac{C-\ell}{F}\right\} I\Big(\ell + dr_{d,1} + r_{d,1} \le C\Big)
\end{aligned}
$$

$$+P(i_1,i_2,d+1)\left(\left\lfloor\frac{C-\ell}{r_{d,1}}\right\rfloor\frac{r_{d,1}}{F}+\left(d+1-\left\lfloor\frac{C-\ell}{r_{d,1}}\right\rfloor\right)\sigma\right)$$

$$\times I\left(\ell+dr_{d,1}+r_{d,1}>C, d+1-\left\lfloor\frac{C-\ell}{r_{d,1}}\right\rfloor\le w\right),$$

$$(i_1,i_2,d)\in S.$$

By $I(\cdot)$ in (2) the indicator function is defined

$$I(\cdot)=\begin{cases}1, \text{ if condition formulated in brackets is fulfilled,}\\ 0, \text{ if this condition isn't fulfilled.}\end{cases}$$

The solution of the system (2) gives unnormalized values of probabilities $P(i_1,i_2,d)$. They need to be normalized:

$$p(i_1,i_2,d)=\frac{P(i_1,i_2,d)}{\sum\limits_{(i_1,i_2,d)\in S}P(i_1,i_2,d)}.$$

The matrix of the system of state equations (2) is quite big. For moderate values of C the number of unknowns can reach the order 10^5–10^6. Almost all elements of the matrix are zeros. In this case the effective way to solve (2) and find $p(i_1,i_2,d)\in S$ is to use Gauss-Seidel iterative algorithm [1,2]. Relations (2) can be easily rewritten into the Gauss–Seidel recursions for estimation of the probabilities $P(i_1,i_2,d)$. Let us denote by $P^{(q)}(i_1,i_2,d)$ the q-th approximation for $P(i_1,i_2,d)$ obtained by Gauss-Seidel iterations. The initial approximation can be chosen from relations $P^{(0)}(i_1,i_2,d)=1$, $(i_1,i_2,d)\in S$. The successive approximations are obtained from the following expressions:

$$P^{(q+1)}(i_1,i_2,d)=\frac{1}{L(i_1,i_2,d)} \tag{3}$$

$$\times\left\{P^{(q,q+1)}(i_1-1,i_2,d)\lambda_1 I(i_1>0,\ell+dr_{d,1}\le C)\right.$$

$$+P^{(q,q+1)}(i_1,i_2-1,d)(s-i_2+1)\gamma I(i_2>0,\ell+dr_{d,1}\le C)$$

$$+\sum_{k=1}^{d}P^{(q,q+1)}(i_1,i_2,d-k)\lambda_d$$

$$\times\left(f_k+I\left(\ell+dr_{d,1}\ge C, d-\left\lfloor\frac{C-\ell}{r_{d,1}}\right\rfloor=w\right)\sum_{j=k+1}^{z}f_j\right)$$

$$+P^{(q,q+1)}(i_1+1,i_2,d)(i_1+1)\mu_1\Big(I(\ell+dr_{d,1}+r_1\le C)$$

$$+ I(\ell + dr_{d,1} + r_1 > C, \ell + r_1 \le C, d - \left\lfloor \frac{C - \ell - r_1}{r_{d,1}} \right\rfloor \le w)\Big)$$

$$+ P^{(q,q+1)}(i_1, i_2 + 1, d)(i_2 + 1)\mu_2 \Bigg(I(\ell + dr_{d,1} + r_2 \le C, i_2 + 1 \le s)$$

$$+ I(\ell + dr_{d,1} + r_2 > C, i_2 + 1 \le s, \ell + r_2 \le C, d - \left\lfloor \frac{C - \ell - r_2}{r_{d,1}} \right\rfloor \le w)\Bigg)$$

$$+ P^{(q,q+1)}(i_1, i_2, d + 1) \min\left\{(d + 1)\alpha_2, \frac{C - \ell}{F}\right\} I\Bigg(\ell + dr_{d,1} + r_{d,1} \le C\Bigg)$$

$$+ P^{(q,q+1)}(i_1, i_2, d + 1) \left(\left\lfloor \frac{C - \ell}{r_{d,1}} \right\rfloor \frac{r_{d,1}}{F} + \left(d + 1 - \left\lfloor \frac{C - \ell}{r_{d,1}} \right\rfloor \right) \sigma \right)$$

$$\times I\left(\ell + dr_{d,1} + r_{d,1} > C, d + 1 - \left\lfloor \frac{C - \ell}{r_{d,1}} \right\rfloor \le w\right)\Bigg\}.$$

By $L(i_1, i_2, d)$ in (3) the auxiliary expression is defined

$$L(i_1, i_2, d) = \Bigg\{ \lambda_1 I(\ell + dr_{d,1} + r_1 \le C)$$
$$+ (s - i_2)\gamma I(\ell + dr_{d,1} + r_2 \le C) + \lambda_d I(\ell + dr_{d,1} + r_{d,1} \le C)$$
$$+ \lambda_d I\left(\ell + dr_{d,1} + r_{d,1} > C, d - \left\lfloor \frac{C - \ell}{r_{d,1}} \right\rfloor < w\right) + i_1\mu_1 I(i_1 > 0)$$
$$+ i_2\mu_2 I(i_2 > 0) + \min\left(d\alpha_2, \frac{C - \ell}{F}\right) I(\ell + dr_{d,1} \le C, d > 0)$$
$$+ \left(\left\lfloor \frac{C - \ell}{r_{d,1}} \right\rfloor \frac{r_{d,1}}{F} + \left(d - \left\lfloor \frac{C - \ell}{r_{d,1}} \right\rfloor \right) \sigma \right) I(\ell + dr_{d,1} > C) \Bigg\}.$$

The upper index $(q, q + 1)$ for successive approximations in relations (3) means calculation of the components of the $(q + 1)$-st approximation with usage of the already found components of the $(q + 1)$-st approximation. In case when no such components are available, then with usage of the known components of the q-th approximation. Convergence of the iterative algorithm is estimated by reaching smallness of the normalized difference between two successive approximations to the vector of unknown probabilities.

The found probabilities $p(i_1, i_2, d) \in S$ have interpretation as a portion of time the model stays in the state (i_1, i_2, d). This interpretation gives the possibility to use the values of $p(i_1, i_2, d)$ for calculation the main performance measures of requests servicing.

4 Performance Measures

In the analyzed model we have two types of performance measures that characterized the process of servicing of arriving traffic flows. The first group describes the servicing of requests for real time traffic transmission. Let us denote for n-th flow, $n = 1, 2$, of this type of traffic by π_n the portion of lost calls, by y_n denote the mean number of requests being on servicing and by m_n denote the mean amount of occupied link transmission capacity. The formal definitions of introduced characteristics are looking as follows:

$$\pi_1 = \sum_{\left\{(i_1,i_2,d)\in S \mid \ell+dr_{d,1}+r_1>C\right\}} p(i_1, i_2, d);$$

$$\pi_2 = \frac{\sum\limits_{\left\{(i_1,i_2,d)\in S \mid \ell+dr_{d,1}+r_2>C\right\}} p(i_1, i_2, d)(s - i_2)}{\sum\limits_{\left\{(i_1,i_2,d)\in S\right\}} p(i_1, i_2, d)(s - i_2)};$$

$$y_n = \sum_{\left\{(i_1,i_2,d)\in S\right\}} p(i_1, i_2, d) i_n, \quad n = 1, 2;$$

$$m_n = y_n r_n, \quad n = 1, 2.$$

The second group of characteristics describes the servicing of requests for elastic data transmission. Let us denote for this type of traffic by d_m the mean number of requests for elastic traffic transmission in one group; by π_3 denote the portion of lost requests; by y_3 denote the mean number of requests being in the system on serving or waiting; by $y_{3,s}$ denote the mean number of requests being on servicing; by $y_{3,w}$ denote the mean number of requests being on waiting; by m_3 denote the mean amount of link transmission capacity occupied by serving this type of traffic; by b_d denote the mean amount of link transmission capacity occupied by serving one request for file transmission; by T_d denote the mean time of file transmission. The formal definition of introduced characteristics are looking as follows:

$$d_m = \sum_{k=1}^{z} f_k\, k;$$

$$\pi_3 = \frac{1}{\lambda_d d_m}$$

$$\times \left(\sum_{\left\{(i_1,i_2,d)\in S \mid \ell+dr_{d,1}\geq C, d-\left\lfloor \frac{C-\ell}{r_{d,1}} \right\rfloor = w\right\}} \sum_{k=0}^{d} p(i_1, i_2, d-k)\lambda_d \sum_{j=k+1}^{z} f_j\,(j-k) \right.$$

$$\left. + \sum_{\left\{(i_1,i_2,d)\in S \mid \ell+dr_{d,1}>C\right\}} p(i_1, i_2, d) \left(d - \left\lfloor \frac{C-\ell}{r_{d,1}} \right\rfloor \right) \sigma \right);$$

$$y_3 = \sum_{\{(i_1,i_2,d)\in S\}} p(i_1,i_2,d)d;$$

$$y_{3,s} = \sum_{\{(i_1,i_2,d)\in S \mid \ell+dr_{d,1}\le C, d>0\}} p(i_1,i_2,d)d + \sum_{\{(i_1,i_2,d)\in S \mid \ell+dr_{d,1}>C\}} p(i_1,i_2,d)\left\lfloor \frac{C-\ell}{r_{d,1}} \right\rfloor;$$

$$y_{3,w} = y_3 - y_{3,s};$$

$$m_3 = \sum_{\{(i_1,i_2,d)\in S \mid \ell+dr_{d,1}\le C, d>0\}} p(i_1,i_2,d)\min(dr_{d,2}, C-\ell) + \sum_{\{(i_1,i_2,d)\in S \mid \ell+dr_{d,1}>C\}} p(i_1,i_2,d)\left\lfloor \frac{C-\ell}{r_{d,1}} \right\rfloor r_{d,1};$$

$$b_d = \frac{m_3}{y_{3,s}}; \quad T_d = \frac{y_3}{\lambda_d b_d(1-\pi_3)}.$$

5 The Algorithm of Decomposition

The exact estimation of the model's performance measures is based on the solution of the system of states equation (2) by the Gauss–Seidel iterative algorithm. The usage of this approach is restricted by the number of unknowns in (2). Normally this number should be less than several millions. If this is not true the approximate methods can be used. One of such approaches is based on the decomposition technique that allows to determine the performance measures of servicing the elastic data [5]. The quality of transmission the elastic data traffic is estimated by the mean time of file downloading. The volume of transmission resources found in the process of access node planning has to provide low losses (round one percent) of accepting the requests for real time and elastic data transmission. For such values of input parameters the model under consideration has some properties that can be used for construction of the approximate algorithm of estimation the mean time of file downloading and the mean number of requests for elastic data transmission being on servicing.

Let us represent the node transmission capacity C in terms of its smallest granularity which we call resource unit. Let us denote by v the total amount of resource units and consider the moment of arrival of a request for the elastic data transmission. Let us denote by i the number of resource units occupied at this moment by servicing the real time traffic. Correspondingly $v-i$ is the number of resource units that can be used for servicing the arriving request. Let us denote by $p(i)$ the probability of occupation of i resource units by servicing the real time traffic in the case when no data traffic is considered in the system,

$i = 0, 1, \ldots, v$. The values of $p(i)$ can be easily found from recursions established for BPP traffic streams [1,2].

Let us denote, respectively, by $y_3(v-i)$ the mean number of requests for data transmission being in the system on serving or waiting; by $y_{3,s}(v-i)$ denote the mean number of such requests being on servicing; by $m_3(v-i)$ denote the mean amount of link transmission capacity occupied by serving this type of traffic; by $T_d(v-i)$ denote the mean time of file transmission. The introduced characteristics are calculated with help of the access node model servicing only data traffic and having in total $v-i$ resource units. The characteristics can be calculated by simple recursive algorithm. Denote the resulting estimates of servicing the data traffic by the same symbols as the characteristics themselves, only with the superscript (a). The values of the estimates are determined from relations

$$y_3^{(a)} = \sum_{i=0}^{v} p(i)\, y_3(v-i), \quad y_{3,s}^{(a)} = \sum_{i=0}^{v} p(i)\, y_{3,s}(v-i), \tag{4}$$

$$m_3^{(a)} = \sum_{i=0}^{v} p(i)\, m_3(v-i), \quad T_d^{(a)} = \sum_{i=0}^{v} p(i)\, T_d(v-i).$$

For illustration purposes let us consider the process of joint servicing of the real time and elastic traffic by the access node with the following values of input parameters: $C = 60$; $w = 5$; $r_1 = 1$; $\sigma = 0{,}5$; $r_2 = 5$; $a_1 = 5$; $\mu_1 = 1$; $a_2 = 0{,}5$; $\mu_2 = 1$; $s = \lfloor a_2 \rfloor + 10$; $\gamma = \frac{a_2\mu_2}{s-a_2}$; $F = 10$; $r_{d,1} = 1$; $r_{d,2} = 5$; $z = 5$, $f_k = 1/z$, $k = 1, 2, \ldots, 5$. The values of $c, r_1, r_2, r_{d,1}, r_{d,2}$ are expressed in Mbit/s. The value of F is expressed in Mbit. As a time unit was chosen the mean time of servicing a request for real time traffic transmission. The values of a_1, a_2 are expressed in Erlangs. The values $\lambda_d, \sigma, \gamma$ represent the mean number of corresponding events

Table 1. The results of exact and approximate calculation of characteristics of data transmission

λ_d	π	y_3		T_d		b_d	
		Exact	Approx.	Exact	Approx.	Exact	Approx.
1,0	0,000063	6,8438	6,9291	2,2813	2,3099	4,3835	4,3293
1,1	0,000252	7,9527	8,1085	2,4100	2,4578	4,1495	4,0690
1,2	0,000930	9,3420	9,6060	2,5952	2,6704	3,8534	3,7453
1,3	0,003142	11,1746	11,5718	2,8662	2,9726	3,4894	3,3645
1,4	0,009623	13,7012	14,1837	3,2657	3,3903	3,0631	2,9484
1,5	0,026311	17,2446	17,5982	3,8448	3,9399	2,6029	2,5321
1,6	0,062855	22,0512	21,8567	4,6350	4,6120	2,1611	2,1525
1,7	0,128491	27,9757	26,7829	5,5985	5,3576	1,7916	1,8353
1,8	0,223114	34,2941	31,9531	6,6080	6,0910	1,5207	1,5894
1,9	0,333873	40,0654	36,8192	7,5118	6,7178	1,3405	1,4105
2,0	0,443905	44,7266	40,9416	8,2248	7,1740	1,2271	1,2865

in chosen time unit. The exact and approximate values of y_3, T_d, b_d depending on the value of λ_d are given in Table 1. The exact values of the characteristic are found after solving the system of state equations (2) by the iterative Gauss–Seidel algorithm (3). The approximate values of the characteristics are found by using expressions (4). The level of requests losses was estimated by the value of $\pi = \max_{1 \le k \le 3} \pi_k$. It is clearly seen that the accuracy (4) increases with decrease of the offered traffic.

6 Conclusion

In this paper the model of joint servicing by access node the real time and elastic data traffic is proposed. Flow of requests for real time servicing is described by Poisson model (narrowband traffic) or Engset model (broadband traffic). The requests for data (file) transmission are coming by groups (batches) according to the Poisson model. Requests from the group occupy free transmission resources or free waiting positions if all transmission resources are occupied. The excess of the group is lost when all transmission resources or waiting positions are occupied. The number of requests in the group is varying from one to the some value and defined by the some probability. The sum of these probabilities is equal to one. The volume of the file has exponential distribution with mean value represented in bits. Real time traffic has advantage in occupying the node transmission resource. It exhibits itself in decreasing if necessary the speed of data transmission to some minimum value. When the node has free capacity the speed of data transmission is increasing to some maximum value (elastic property). The duration of servicing of requests for real time traffic transmission has exponential distribution and doesn't depend on the model's state. The duration of servicing of requests for elastic data transmission also has exponential distribution but with parameter depending on the available free node transmission capacity and the allowed values of capacity that can be used for servicing by one request.

In the framework of the proposed model the definitions of main performance measures of traffic transmission are formulated through values of probabilities of model's stationary states. The exact and approximate algorithms of estimation the introduced performance measures are suggested. Exact algorithm based on the solving the system of state equations and approximate algorithm is based on the procedure of decomposition. The proposed model and the results of its analysis can be used for estimation the required transmission capacity of access node for given volumes of offered real time and elastic traffics.

References

1. Stepanov, S.N.: Osnovy teletraffika multiservisnykh setei (Fundamentals of Multiservice Networks). Eqo-Trends, Moscow (2010). (in Russian)
2. Stepanov, S.N.: Teoriya teletraffika: kontseptsii, modeli, prilozheniya (Theory of Teletraffic: Concepts, Models, Applications). Goryachaya Liniya–Telekom, Moscow (2015). (in Russian)

3. Yashkov, S.F., Yashkova, A.S.: Processor sharing: a survey of the mathematical theory. Autom. Remote Control **68**, 1662–1731 (2007)
4. Bonald, T., Virtamo, J.: A recursive formula for multirate systems with elastic traffic. IEEE Commun. Lett. **9**, 753–755 (2005)
5. Stepanov, S.N., Stepanov, M.S.: Planning transmission resource at joint servicing of the multiservice real time and elastic data traffics. Autom. Remote Control **78**, 2004–2015 (2017)
6. Borodakiy, V.Y., Samouylov, K.E., Gudkova, I.A., Markova, E.V.: Analyzing mean bit rate of multicast video conference in LTE network with adaptive radio admission control scheme. J. Math. Sci. **218**, 257–268 (2016). (United States)
7. Gudkova, I., et al.: Analyzing impacts of coexistence between M2M and H2H communication on 3GPP LTE system. In: Mellouk, A., Fowler, S., Hoceini, S., Daachi, B. (eds.) WWIC 2014. LNCS, vol. 8458, pp. 162–174. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13174-0_13
8. Vasiliev, A.P., Stepanov, S.N.: The construction and analysis of mathematical model of ddynamic distribution of channel resources for group arrival of requests for data transfer. T-Comm. **10**, 55–59 (2016)
9. Stepanov, S.N., Romanov, A.M.: The mathematical model of access line serving real time traffic and elastic data. T-Comm. **17**, 74–79 (2017)

Unreliable Single-Server Queue with Two-Way Communication and Retrials of Blocked and Interrupted Calls for Cognitive Radio Networks

Anatoly Nazarov[1(✉)], Tuan Phung-Duc[2], and Svetlana Paul[1]

[1] National Research Tomsk State University, Tomsk, Russia
nazarov.tsu@gmail.com, paulsv82@mail.ru
[2] Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba, Japan
tuan@sk.tsukuba.ac.jp

Abstract. In this paper, we consider an $M/GI/GI/1/1$ retrial queue where incoming fresh calls arrive at the server according to a Poisson process. Upon arrival, an incoming call either occupies the server if it is idle or joins an orbit if the server is busy. From the orbit, an incoming call retries to occupy the server and behaves the same as a fresh incoming call. After some idle time, the server makes an outgoing call to outside. We consider the system with an unreliable server. In a free state and while servicing outgoing calls the server is reliable and unable to crash. If while servicing an incoming call the server crashes, the incoming call goes into the orbit. The service time of such an interrupted call follows the same distribution as that of an incoming call. For that system we obtained probability distribution of the states of the server, the condition for the existence of a stationary mode and probability distribution of a number of calls in the system.

Keywords: Retrial queueing system
Incoming and outgoing calls · Unreliable server · Cognitive networks

1 Introduction

Retrial queues are characterized by the feature that a blocked customer does not lost or queue before the server but leave the server and retry for service after some random time. The retrial phenomenon are ubiquitous in random access systems such as cellular mobile networks or local area networks etc. [1,2,6]. An important application of these models can be found in call centers where customes who cannot connect to the operator will make the phone call again. In modern call centers, the operators not only serve incoming calls but in idle time also initiate outgoing calls to outside [3,4]. Motivated by this situation, a retrial queue with distinct distributions of incoming calls and outgoing calls

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 276–287, 2018.
https://doi.org/10.1007/978-3-319-99447-5_24

is considered in [7,12] while the model with the same distribution of incoming and outgoing calls was considered in [9]. In these works, the server is assumed to be reliable, i.e., there is no interruption during the service. In this paper, we consider an extension of [12], where the server is unreliable while serving incoming calls. A closely related model is investigated in [17] where the authors consider unreliable server when serving both incoming and outgoing calls. In [17], when the server is broken down the ongoing customer stays at the server and the service is resumed upon the completion of the repairment of the server.

In this paper, we consider a new extension of [12] where the interrupted call leaves the server and retries again instead of waiting at the server. Mathematically, the interrupted customer will be served again as other incoming calls.

This feature reflects various real world situations in service systems. For example, in call centers, if the customer does not have enough information, the operator cannot complete the service and the customer must call again in a later time when he has enough information. Also in various real life service systems, the service cannot be completed at the first arrival and the customer must come again in the future. Our model fits very well the situation in cognitive networks where secondary users use the licensed channel when primary users are not present at the system. The behavior primary users could be modeled by the breakdown and repair mechanism presented in this paper. Furthermore, secondary users in cognitive radio networks correspond to incoming calls in our model. When a primary user uses the channel, from the view point of secondary users, the server is broken down. The time to repair could be interpreted as the service time of the primary user. For related works, we mention some papers with either unreliable server or two-way communication in [5,7–12,14–16].

The rest of this paper is organized as follows. In Sect. 2, we describe the model in details. Section 3 represents the set of Kolmogorov equations which describe the dynamic of the system. Section 4 is devoted to the analysis of the distribution of the state of the server and Sect. 5 is devoted to derivation of characteristic functions of the stationary queue length distribution. Finally, concluding remarks are presented in Sect. 6.

2 Model Description and Problem Definition

We consider a single server queueing model with two types of calls: incoming calls and outgoing calls. Incoming calls arrive at the system according to a Poisson process with rate λ. Upon arrival, if the server is free the incoming call enters the system and goes into service for a time duration, distributed with probability function $B_1(x)$. If at the moment of entering system, the server is busy, the call instantly goes to the orbit and stays there for an exponentially distributed duration of time with rate σ, after which the call retries to go into service.

If the server is idle (empty) it starts making outgoing calls to the outside (not from the orbit) with rate α and the service time of outgoing calls has distribution function $B_2(x)$.

We consider the system with an unreliable server [13,17], which crashes with intensity γ and recovers with intensity μ while serving incoming calls. In the free state and while servicing outgoing calls the server is reliable and unable to crash.

In case the server crashes while servicing an incoming call, the incoming call goes into the orbit. The service time of an interrupted call follows the same distribution as other incoming calls. In the other words, the service time distribution of the interrupted call is also $B_1(x)$. When the server is serving an incoming call or recovering, incoming calls enter the orbit.

Let $i(t)$ denote the number of incoming calls in the system at time t. The main purposes of this papers are:

(1) Find the condition for the existence of a stationary mode in the retrial queue described above,
(2) Find the characteristic function of the stationary queue length distribution:

$$P(i) = P\left\{i(t) = i\right\}. \tag{1}$$

3 Kolmogorov System of Equations

We define some notations: the state of the server at time t as $k(t)$:

0 if the server is free,
1 if the server is busy serving an incoming call,
2 if the server is busy serving an outgoing call,
3 if the server is in a state of recovery;

$z(t)$ - remaining time of service, when $k = 1, 2$.

We further define the probabilities:

$$\begin{gathered} P\left\{k(t) = k, i(t) = i, z(t) < z\right\} = P_k(i, z, t), \quad k = 1, 2, \\ P\left\{k(t) = k, i(t) = i\right\} = P_k(i, t), \quad k = 0, 3. \end{gathered} \tag{2}$$

Since the random process

$$\left\{k(t), i(t), z(t)\right\}, k = 1, 2; \left\{k(t), i(t)\right\}, k = 0, 3$$

with a variable number of components is a Markov process, then we have to compose a system of Kolmogorov equations for the probability distribution (2).

We denote $P_k(i, \infty, t) = P_k(i, t)$, for $k = 1, 2$. The stationary probability distribution is the unique solution of the Kolmogorov system of equations:

$$-(\lambda + \alpha + i\sigma)P_0(i,t) + \frac{\partial P_1(i+1,0,t)}{\partial z} + \frac{\partial P_2(i,0,t)}{\partial z} + \mu P_3(i,t) = \\ = \frac{\partial P_0(i,t)}{\partial t},$$

$$\frac{\partial P_1(i,z,t)}{\partial z} - \frac{\partial P_1(i,0,t)}{\partial z} - (\lambda + \gamma)P_1(i,z,t) + \lambda P_1(i-1,z,t) +$$

$$\lambda B_1(z)P_0(i-1,t) + i\sigma B_1(z)P_0(i,t) = \frac{\partial P_1(i,z,t)}{\partial t},$$

$$\frac{\partial P_2(i,z,t)}{\partial z} - \frac{\partial P_2(i,0,t)}{\partial z} - \lambda P_2(i,z,t) + \lambda P_2(i-1,z,t) + \alpha B_2(z)P_0(i,t) = \\ = \frac{\partial P_2(i,z,t)}{\partial t},$$

$$-(\lambda+\mu)P_3(i,t) + \lambda P_3(i-1,t) + \gamma P_1(i,t) = \frac{\partial P_3(i,t)}{\partial t}.$$

Let's write down the last system in stationary mode:

$$\begin{gathered} -(\lambda+\alpha+i\sigma)P_0(i) + \frac{\partial P_1(i+1,0)}{\partial z} + \frac{\partial P_2(i,0)}{\partial z} + \mu P_3(i) = 0, \\ \frac{\partial P_1(i,z)}{\partial z} - \frac{\partial P_1(i,0)}{\partial z} - (\lambda+\gamma)P_1(i,z) + \lambda P_1(i-1,z) + \\ \lambda B_1(z)P_0(i-1) + i\sigma B_1(z)P_0(i) = 0, \\ \frac{\partial P_2(i,z)}{\partial z} - \frac{\partial P_2(i,0)}{\partial z} - \lambda P_2(i,z) + \lambda P_2(i-1,z) + \alpha B_2(z)P_0(i) = 0, \\ -(\lambda+\mu)P_3(i) + \lambda P_3(i-1) + \gamma P_1(i) = 0. \end{gathered} \tag{3}$$

We introduce partial characteristic functions, denoting $j = \sqrt{-1}$:

$$\begin{gathered} H_0(u) = \sum_{i=0}^{\infty} e^{jui} P_0(i), \\ H_k(u,z) = \sum_{i=1}^{\infty} e^{jui} P_k(i,z), \quad k = 1,2, \\ H_3(u) = \sum_{i=0}^{\infty} e^{jui} P_3(i). \end{gathered} \tag{4}$$

We rewrite (3) in the following form:

$$\begin{gathered} -(\lambda+\alpha)H_0(u) + \mu H_3(u) + j\sigma H_0'(u) + e^{-ju}\frac{\partial H_1(u,0)}{\partial z} + \frac{\partial H_2(u,0)}{\partial z} = 0, \\ \frac{\partial H_1(u,z)}{\partial z} - \frac{\partial H_1(u,0)}{\partial z} + \\ (\lambda(e^{ju}-1)-\gamma)H_1(u,z) + \lambda B_1(z)e^{ju}H_0(u) - j\sigma B_1(z)H_0'(u) = 0, \\ \frac{\partial H_2(u,z)}{\partial z} - \frac{\partial H_2(u,0)}{\partial z} + \lambda(e^{ju}-1)H_2(u,z) + \alpha B_2(z)H_0(u) = 0, \\ (\lambda(e^{ju}-1)-\mu)H_3(u) + \gamma H_1(u) = 0. \end{gathered} \tag{5}$$

In the system (5), we take the limit as $z \to \infty$. Denote

$$H_k(u,\infty) = H_k(u), \quad k = 1,2.$$

After summing the resulting equations, we will get

$$\frac{\partial H_1(u,0)}{\partial z} - \lambda e^{ju} H(u) = 0, \tag{6}$$

where

$$H(u) = H_0(u) + H_1(u) + H_2(u) + H_3(u).$$

In the system (5) the first equation is replaced by the equation (6). Then the system of equations for partial characteristic functions (5) could be rewritten in the following form

$$\begin{gathered}
\frac{\partial H_1(u,z)}{\partial z} - \frac{\partial H_1(u,0)}{\partial z} + \\
(\lambda(e^{ju}-1)-\gamma)H_1(u,z) + \lambda B_1(z)e^{ju}H_0(u) - j\sigma B_1(z)H_0'(u) = 0, \\
\frac{\partial H_2(u,z)}{\partial z} - \frac{\partial H_2(u,0)}{\partial z} + \lambda(e^{ju}-1)H_2(u,z) + \alpha B_2(z)H_0(u) = 0, \\
(\lambda(e^{ju}-1)-\mu)H_3(u) + \gamma H_1(u) = 0, \\
\frac{\partial H_1(u,0)}{\partial z} - \lambda e^{ju} H(u) = 0.
\end{gathered} \tag{7}$$

We define the Laplace-Stieltjes transform

$$B_k^*(s) = \int_0^\infty e^{-sx} dB_k(s),$$

$$H_k^*(u,s) = \int_0^\infty e^{-sz} dH_k(u,z), k = 1,2.$$

We rewrite system (7) in the form

$$\begin{gathered}
-\frac{\partial H_1(u,0)}{\partial z} + (\lambda(e^{ju}-1)-\gamma+s)H_1^*(u,s) + \\
B_1^*(s)(\lambda e^{ju}H_0(u) - j\sigma H_0'(u)) = 0, \\
-\frac{\partial H_2(u,0)}{\partial z} + (\lambda(e^{ju}-1)+s)H_2^*(u,s) + \alpha B_2^*(s)H_0(u) = 0, \\
(\lambda(e^{ju}-1)-\mu)H_3(u) + \gamma H_1(u) = 0, \\
\frac{\partial H_1(u,0)}{\partial z} - \lambda e^{ju} H(u) = 0.
\end{gathered} \tag{8}$$

This system will be the main in further research.

4 Probabilities Distribution of the Server States and Condition of Existence of a Stationary Mode

We will prove the following statement.

Theorem 1. *For our retrial queue, denoting* $b_2 = \int_0^\infty x dB_2(x)$*, the probabilities*

$$r_k = P\{k(t) = k\}$$

of the server states have the form

$$\begin{aligned} r_0 &= \frac{1}{1+\alpha b_2}\left(1 - \lambda\frac{\mu+\gamma}{\mu}\cdot\frac{1-B_1^*(\gamma)}{\gamma B_1^*(\gamma)}\right), \\ r_1 &= \lambda\frac{1-B_1^*(\gamma)}{\gamma B_1^*(\gamma)}, \\ r_2 &= \alpha b_2 r_0, \\ r_3 &= \frac{\gamma}{\mu}r_1. \end{aligned} \tag{9}$$

Proof. We denote

$$\begin{aligned} b_2 &= \int_0^\infty x dB_2(x), \\ H_k^*(u,s) &= r_k^*(s), \quad H_k(0) = r_k, k = 0,3, \\ \left.\frac{\partial H_k(u,0)}{\partial z}\right|_{u=0} &= r_k'(0), \quad k = 1,2, \quad H_0'(u)|_{u=0} = jm_0. \end{aligned} \tag{10}$$

Then substituting $u = 0$ to (8), we get the following system of equations:

$$\begin{aligned} -r_1'(0) + (s-\gamma)r_1^*(s) + B_1^*(s)(\lambda r_0 + \sigma m_0) &= 0, \\ -r_2'(0) + sr_2^*(s) + \alpha B_2^*(s) r_0 &= 0, \\ -\mu r_3 + \gamma r_1 &= 0, \\ \lambda - r_1'(0) &= 0. \end{aligned} \tag{11}$$

Substituting $s = \gamma$ into the first and using the fourth equations of the system (11), we obtain

$$\begin{aligned} r_1'(0) &= B_1^*(\gamma)(\lambda r_0 + \sigma m_0), \\ r_1'(0) &= \lambda. \end{aligned} \tag{12}$$

We differentiate the second equation in (11) with respect to s . Substitute the expression $s = 0$ in the resulting equation. From systems (11) and (12) we get the system

$$
\begin{aligned}
-\lambda - \gamma r_1 + \frac{\lambda}{B_1^*(\gamma)} &= 0,\\
r_2 - \alpha b_2 r_0 &= 0,\\
-\mu r_3 + \gamma r_1 &= 0.
\end{aligned} \tag{13}
$$

We'll find the probability value r_0 from the normalization condition in the form of a first equation in (9). Theorem 1 is proved.

Corollary. The condition for the existence of the stationary mode for the model in this paper is the following inequality:

$$
\lambda < \frac{\mu}{\mu+\gamma} \cdot \frac{\gamma B_1^*(\gamma)}{1 - B_1^*(\gamma)}. \tag{14}
$$

Proof. The condition (14) follows from the positivity of the probability r_0 in (9). The corollary is proved.

Let's define the system flow capacity S as a maximum average number of incoming calls that could be served in the system per unit time. By inequality (14) the value S for our system with outgoing calls and unreliable server is defined by

$$
S = \frac{\mu}{\mu+\gamma} \cdot \frac{\gamma B_1^*(\gamma)}{1 - B_1^*(\gamma)}. \tag{15}
$$

If the value of the parameter λ of the incoming flow is defined by the equality $\lambda = \rho S$, then at any values of the parameter $0 < \rho < 1$ the stationary mode exists for our system, and the probabilities r_k from (9) of the server states could be written in the following form

$$
\begin{aligned}
r_0 &= \frac{1-\rho}{1+\alpha b_2},\\
r_1 &= \rho \frac{\mu}{\mu+\gamma},\\
r_2 &= \alpha b_2 \frac{1-\rho}{1+\alpha b_2},\\
r_3 &= \rho \frac{\gamma}{\mu+\gamma},
\end{aligned} \tag{16}
$$

which do not depend on the form of distribution functions $B_1(x)$ and $B_2(x)$ of the service time of both incoming and outgoing calls. When this happens the intensity λ of the incoming flow linearly depends on S, which by (15) depends on the form of distribution function $B_1(x)$.

For the model in this paper, the system flow capacity S from (15) has a non-standard property. Let $\mu = \nu\gamma$, where the parameter ν can take any positive value $\nu > 0$, then $S = S(\gamma)$. We consider the system flow capacity $S(\gamma)$ while increasing the parameter γ. We have

$$
\lim_{\gamma\to\infty} S(\gamma) = \lim_{\gamma\to\infty} \frac{\nu}{\nu+1} \cdot \frac{\gamma B_1^*(\gamma)}{1 - B_1^*(\gamma)} =
$$

$$\frac{\nu}{\nu+1}\lim_{\gamma\to\infty}\gamma B_1^*(\gamma)=\frac{\nu}{\nu+1}\cdot B_1'(0). \tag{17}$$

The distribution density at zero can take any non-negative value that depends on the form of the distribution function $B_1(x)$.

If $B_1'(0)=\infty$, then for any arbitrarily large value of the rate λ, by (17) we can find the value of the rate γ for which inequality (14) holds. Consequently, there is a stationary mode in the retrial queue considered in this paper. This is evident and intuitive because in this case, the distribution of the service time concentrates on zero.

If $B_1'(0)=0$, then for any arbitrarily small value of the rate λ, by (17) we can find the value of the rate γ for which inequality (14) does not hold. Consequently, in the system, there is no stationary mode for any arbitrarily small fraction $\frac{\gamma}{\mu+\gamma}=\frac{1}{\nu+1}$ of the time that is spent on the repair of the server. Let $B_1(x)$ be a function of the gamma distribution with the same values of the form parameter α_1 and the scale parameter β_1, that is, $\alpha_1=\beta_1=\beta$. Then the average value $\frac{\alpha_1}{\beta_1}$ of the service time of the outgoing calls is equal to one, and $B_1^*(\gamma)=\left(\frac{\beta}{\beta+\gamma}\right)^{\beta}$. For $\mu=\nu\gamma$ the system flow capacity S depends on γ and has the form

$$S=S(\gamma)=\frac{\nu}{\nu+1}\cdot\frac{\gamma B_1^*(\gamma)}{1-B_1^*(\gamma)}.$$

We set $\nu=4$. Table 1 shows the system flow capacity $S_1(\gamma)$ for $\beta_1=\beta=5$ and $S_2(\gamma)$ for $\beta_1=\beta=0.2$ and the indicated values of the parameter γ.

Table 1. The system flow capacity S.

γ	0.01	0.1	1	10	100
S_1	0.797	0.769	0.538	0.033	$2\cdot10^{-5}$
S_2	0.816	0.947	1.856	6.692	32.427

Further we find probability distribution $P(i)$ of the number $i(t)$ of calls in our retrial queueing system considered in this paper.

5 Probability Distribution of the Number of Calls in the System

Let's denote

$$\begin{aligned} g_1(u)&=\lambda(1-e^{ju})+\gamma,\\ g_2(u)&=\lambda(1-e^{ju}), \end{aligned} \tag{18}$$

and rewrite system (8) in the following form

$$\begin{gathered} -\frac{\partial H_1(u,0)}{\partial z} + \\ (s-g_1(u))H_1^*(u,s) + B_1^*(s)(\lambda e^{ju}H_0(u) - j\sigma H_0'(u)) = 0, \\ -\frac{\partial H_2(u,0)}{\partial z} + (s-g_2(u))H_2^*(u,s) + \alpha B_2^*(s)H_0(u) = 0, \\ (\lambda(1-e^{ju}) - \mu)H_3(u) + \gamma H_1(u) = 0, \\ \frac{\partial H_1(u,0)}{\partial z} - \lambda e^{ju}H(u) = 0. \end{gathered} \tag{19}$$

Substituting $s = g_1(u)$ in the first equation and $s = g_2(u)$ in the second equation of the system (19) we'll get the following equations:

$$\begin{gathered} \frac{\partial H_1(u,0)}{\partial z} = B_1^*(g_1(u))(\lambda e^{ju}H_0(u) - j\sigma H_0'(u)), \\ \frac{\partial H_2(u,0)}{\partial z} = \alpha B_2^*(g_2(u))H_0(u). \end{gathered} \tag{20}$$

Let's rewrite system (19) in the following form

$$\begin{gathered} (s-g_1(u))H_1^*(u,s) + (B_1^*(s) - B_1^*(g_1(u)))(\lambda e^{ju}H_0(u) - j\sigma H_0'(u)) = 0, \\ (s-g_2(u))H_2^*(u,s) + \alpha(B_2^*(s) - B_2^*(g_2(u)))H_0(u) = 0, \\ (\lambda(e^{ju}-1))H_3(u) + \gamma H_1(u) = 0, \\ B_1^*(g_1(u))(\lambda e^{ju}H_0(u) - j\sigma H_0'(u)) - \lambda e^{ju}H(u) = 0. \end{gathered} \tag{21}$$

Let's denote $H_k^*(u,0) = H_k(u), k = 1,2$. Substituting $s = 0$ in (21), we rewrite system (21) in the following form:

$$\begin{gathered} -g_1(u)H_1(u) + \{1 - B_1^*(g_1(u))\}\left\{\lambda e^{ju}H_0(u) - j\sigma H_0'(u)\right\} = 0, \\ -g_2(u)H_2(u) + \alpha(1 - B_2^*(g_2(u)))H_0(u) = 0, \\ (\lambda(e^{ju}-1) - \mu)H_3(u) + \gamma H_1(u) = 0, \\ B_1^*(g_1(u))(\lambda e^{ju}H_0(u) - j\sigma H_0'(u)) - \lambda e^{ju}H(u) = 0. \end{gathered} \tag{22}$$

We'll rewrite the first three equations of system (22) in the following form:

$$\begin{gathered} H_1(u) = \frac{1 - B_1^*(g_1(u))}{g_1(u)}\left\{\lambda e^{ju}H_0(u) - j\sigma H_0'(u)\right\}, \\ H_2(u) = \alpha\frac{1 - B_2^*(g_2(u))}{g_2(u)}H_0(u), \\ H_3(u) = \frac{\gamma}{\mu - \lambda(e^{ju}-1)}H_1(u). \end{gathered} \tag{23}$$

By substituting these expressions into the fourth equation of system (22) we'll get the following equality

$$0 = \lambda e^{ju} H(u) - B_1^*(g_1(u)) \left(\lambda e^{ju} H_0(u) - juH_0^{'}(u) \right) =$$

$$\lambda e^{ju} \left[H_0(u) \left(1 + \alpha \frac{1 - B_2^*(g_2(u))}{g_2(u)} \right) + H_1(u) \left(1 + \frac{\gamma}{\mu - \lambda(e^{ju} - 1)} \right) \right] -$$

$$B_1^*(g_1(u)) \left\{ \lambda e^{ju} H_0(u) - j\sigma H_0'(u) \right\} =$$

$$\lambda e^{ju} \left[H_0(u) \left(1 + \alpha \frac{1 - B_2^*(g_2(u))}{g_2(u)} \right) + \right.$$

$$\left. \frac{1 - B_1^*(g_1(u))}{g_1(u)} \cdot \frac{\mu - \lambda(e^{ju} - 1) + \gamma}{\mu - \lambda(e^{ju} - 1)} \left(\lambda e^{ju} H_0(u) - j\sigma H_0'(u) \right) \right] -$$

$$B_1^*(g_1(u)) \left\{ \lambda e^{ju} H_0(u) - j\sigma H_0'(u) \right\},$$

which we'll then rewrite in the following form

$$\lambda e^{ju} H_0(u) \left(1 + \alpha \frac{1 - B_2^*(g_2(u))}{g_2(u)} \right) =$$

$$\left(\lambda e^{ju} H_0(u) - j\sigma H_0'(u) \right) \times$$

$$\left(B_1^*(g_1(u)) - \lambda e^{ju} \frac{1 - B_1^*(g_1(u))}{g_1(u)} \cdot \frac{\mu - \lambda(e^{ju} - 1) + \gamma}{\mu - \lambda(e^{ju} - 1)} \right). \tag{24}$$

Let's denote

$$f(u) = \left(1 + \alpha \frac{1 - B_2^*(g_2(u))}{g_2(u)} \right) \times$$

$$\left(B_1^*(g_1(u)) - \lambda e^{ju} \frac{1 - B_1^*(g_1(u))}{g_1(u)} \cdot \frac{\mu - \lambda(e^{ju} - 1) + \gamma}{\mu - \lambda(e^{ju} - 1)} \right)^{-1}, \tag{25}$$

and rewrite equality (24) in the following form

$$\lambda e^{ju} H_0(u) f(u) = \lambda e^{ju} H_0(u) - j\sigma H_0'(u), \tag{26}$$

i.e., in the form of an ordinary differential equation

$$H_0^{'}(u) = j \frac{\lambda}{\sigma} e^{ju} H_0(u)(f(u) - 1),$$

with respect to the function $H_0(u)$, satisfying the condition $H_0(0) = r_0$. The solution $H_0(u)$ of this equation will have the following form

$$H_0(u) = r_0 exp \left\{ j \frac{\lambda}{\sigma} \int_0^u e^{jx} (f(x) - 1) dx \right\}. \tag{27}$$

By substituting last equation into (23) we'll write

$$H_1(u) = \frac{1 - B_1^*(g_1(u))}{g_1(u)} \lambda e^{ju} H_0(u) f_(u),$$
$$H_2(u) = \alpha \frac{1 - B_2^*(g_2(u))}{g_2(u)} H_0(u),$$
$$H_3(u) = \frac{\gamma}{\mu - \lambda(e^{ju} - 1)} H_1(u). \tag{28}$$

Thus the following statement is proved.

Theorem 2. *Using $g_1(u)$ and $g_2(u)$ from (18) and also $f(u)$ from (25) then the characteristic function of the number $i(t)$ of calls in our retrial queueing system in this paper has the following form*

$$H(u) = Me^{ju} = H_0(u) + H_1(u) + H_2(u) + H_3(u),$$

in which the partial characteristic functions $H_k(u), k = \overline{0,3}$ are defined by equalities (27), (28).

Stationary probabilities distribution $P(i) = P\{i(t) = i\}$ of the number of calls in our retrial queue is obtained by the reverse Fourier transform and has the following form

$$P(i) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-jui} H(u) du, \tag{29}$$

in which the expression for a characteristic function $H(u)$ is defined in Theorem 2. The numerical realization of probabilities distribution $P(i)$ from (29) is easily obtained for any values of initial parameters $\alpha, \gamma, \mu, \sigma, \lambda$ and of distribution functions $B_1(x)$ and $B_2(x)$ that satisfy the condition (14).

6 Conclusion

In this paper, we have considered retrial queue *M/GI/GI/1/1* with two-way communication, unreliable server and retrials of interrupted calls. We have found the probability distribution of server states, the condition for the existence of a stationary mode and probability distribution of a number of calls in the system.

References

1. Artalejo, J.R., Gomez-Corral, A.: Retrial Queueing Systems: A Computational Approach. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78725-9
2. Falin, G.I., Templeton, J.G.C.: Retrial Queues. Chapman and Hall, London (1997)
3. Bhulai, S., Koole, G.: A queueing model for call blending in call centers. IEEE Trans. Autom. Control **48**, 1434–1438 (2003)

4. Deslauriers, A., L'Ecuyer, P., Pichitlamken, J., Ingolfsson, A., Avramidis, A.N.: Markov chain models of a telephone call center with call blending. Comput. Oper. Res. **34**, 1616–1645 (2007)
5. Choi, B.D., Choi, K.B., Lee, Y.W.: M/G/1 retrial queueing systems with two types of calls and finite capacity. Queueing Syst. **19**, 215–229 (1995)
6. Tran-Gia, P., Mandjes, M.: Modeling of customer retrial phenomenon in cellular mobile networks. IEEE J. Sel. Areas Commun. **15**, 1406–1414 (1997)
7. Artalejo, J.R., Phung-Duc, T.: Markovian retrial queues with two way communication. J. Ind. Manag. Optim. **8**, 781–806 (2012)
8. Nazarov, A., Phung-Duc, T., Paul, S.: Heavy outgoing call asymptotics for MMPP/M/1/1 retrial queue with two-way communication. In: Dudin, A., Nazarov, A., Kirpichnikov, A. (eds.) ITMM 2017. CCIS, vol. 800, pp. 28–41. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68069-9_3
9. Falin, G.I.: Model of coupled switching in presence of recurrent calls. Eng. Cybern. Rev. **17**, 53–59 (1979)
10. Artalejo, J.R., Resing, J.A.C.: Mean value analysis of single server retrial queues. Asia-Pac. J. Oper. Res. **27**, 335–345 (2010)
11. Falin, G.I., Artalejo, J.R., Martin, M.: On the single server retrial queue with priority customers. Queueing Syst. **14**, 439–455 (1993)
12. Artalejo, J.R., Phung-Duc, T.: Single server retrial queues with two way communication. Appl. Math. Model. **37**(4), 1811–1822 (2003)
13. Senthil Kumar, M., Dadlani, A., Kim, K.: Performance analysis of an unreliable M/G/1 retrial queue with two-way communication. arXiv:1512.08609v3
14. Nazarov, A., Paul, S., Gudkova, I.: Asymptotic analysis of Markovian retrial queue with two-way communication under low rate of retrials condition. In: Proceedings - 31st European Conference on Modelling and Simulation, ECMS, Budapest, pp. 687–693 (2017)
15. Djellab, N.V.: On the M/G/1 retrial queue subjected to breakdowns. RAIRO - Oper. Res. **36**(4), 299–310 (2002)
16. Sherman, N., Kharoufeh, J., Abramson, M.: An M/G/1 retrial queue with unreliable server for streaming multimedia applications. Probab. Eng. Inf. Sci. **23**, 281–304 (2009)
17. Samouylov, K., Naumov, V., Sopin, E., Gudkova, I., Shorgin, S.: Sojourn time analysis for processor sharing loss system with unreliable server. In: Wittevrongel, S., Phung-Duc, T. (eds.) ASMTA 2016. LNCS, vol. 9845, pp. 284–297. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-43904-4_20

Some Aspects of the Discrete Geo/G/1 Type Cyclic Waiting Systems

Laszlo Lakatos(✉)

Eotvos Lorand University, Budapest, Hungary
lakatos1948@freemail.hu

Abstract. Earlier we have investigated the discrete-time cyclic-waiting system in the case of geometrically distributed interarrival time and general service time distribution. We obtained the generating functions of ergodic distributions both for the queue length and the waiting time, we got the stability condition in different forms. In this paper we show their coincidence, and find a relation between the zero probabilities for the two models. We also compute the mean values for the queue length and the waiting time.

Keywords: Discrete cyclic waiting system · Queue length
Waiting time · Stability condition

1 Introduction

Earlier we have considered a single-server queueing system where an entering customer might be accepted for service either at the moment of arrival or at moments differing from it by the multiples of a given cycle time T. The problem was motivated by the transmission of optical signals: optical signals enter a node and they should be transmitted according to the FCFS rule. This information cannot be stored, if it cannot be served at once is sent to a delay line and returns to the node after having passed it. So, the signal can be transmitted from the node at the moment of its arrival or at moments that differ from it by the multiples of time necessary to pass the delay line.

A general description of results about such system can be found in [3] in the case of exponential interarrival and service time distributions, some aspects were investigated in [1,2]. In [5,6] such system was considered from the viewpoints of queue length and waiting time in the discrete time case on condition the interarrival time had geometrical, and the service time had general distributions. Our results were formulated in the following theorems.

Theorem 1. *Let us consider a discrete queueing system in which the interarrival time has geometrical distribution with parameter r, the service time has general distribution with probabilities* q_i $(i = 1, 2, \ldots)$. *The service of a customer may start upon arrival or (in case of busy server or waiting queue) at moments differing from it by the multiples of a given cycle time T (equal to n time units)*

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 288–301, 2018.
https://doi.org/10.1007/978-3-319-99447-5_25

according to the FCFS discipline. Let us define an embedded Markov chain whose states correspond to the number of customers in the system at moments $t_k - 0$, where t_k is the moment of beginning of service of the k-th one. The matrix of transition probabilities has the form

$$\begin{bmatrix} a_0 & a_1 & a_2 & a_3 & \dots \\ a_0 & a_1 & a_2 & a_3 & \dots \\ 0 & b_0 & b_1 & b_2 & \dots \\ 0 & 0 & b_0 & b_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

its elements are determined by the generating functions

$$A(z) = \sum_{i=0}^{\infty} a_i z^i = Q_1 + z\frac{r}{1-r}Q_1 + z\sum_{k=1}^{\infty}(1-r+rz)^{kn} \tag{1}$$
$$\times \left\{ \sum_{i=(k-1)n+2}^{kn+1} q_i + \sum_{i=kn+2}^{\infty} q_i(1-r)^{i-kn-1} - \sum_{i=(k-1)n+2}^{\infty} q_i(1-r)^{i-(k-1)n-1} \right\},$$
$$Q_k = \sum_{i=k}^{\infty} q_i(1-r)^i;$$

$$B(z) = \sum_{i=0}^{\infty} b_i z^i = \sum_{k=0}^{\infty}\sum_{j=1}^{n} q_{kn+j}(1-r+rz)^{kn+j} \tag{2}$$
$$\times \left\{ \frac{r}{1-(1-r)^n} \frac{1-(1-r)^{j-1}(1-r+rz)^{j-1}}{1-(1-r)(1-r+rz)}(1-r+rz)^{n-j+1} \right.$$
$$\left. + \frac{r(1-r)^{j-1}}{1-(1-r)^n} \frac{1-(1-r)^{n-j+1}(1-r+rz)^{n-j+1}}{1-(1-r)(1-r+rz)} \right\}.$$

The generating function of ergodic distribution $P(z) = \sum_{i=0}^{\infty} p_i z^i$ has the form

$$P(z) = \frac{p_0[zA(z) - B(z)] + p_1 z[A(z) - B(z)]}{z - B(z)}, \tag{3}$$

where

$$p_1 = \frac{1-a_0}{a_0}p_0,$$
$$p_0 = \frac{a_0[1-B'(1)]}{a_0 + A'(1) - B'(1)}. \tag{4}$$

The ergodicity condition is

$$\sum_{i=1}^{\infty} q_i \left\lceil \frac{i}{n} \right\rceil < \frac{1}{1-(1-r)^n} \sum_{i=1}^{\infty} q_i (1-r)^{i-1(\mathrm{mod}\ n)}. \tag{5}$$

If the service time distribution is is geometrical [i.e., $q_i = q^{i-1}(1-q)$], the condition (5) means

$$\frac{rq[1-(1-r)^n]}{(1-q)(1-q^n)(1-r)^n} < 1.$$

Let us consider the waiting time. Let t_n denote the time of arrival of the nth customer; its service will begin at the moment $t_n + T \cdot X_n$, where T is the cycle time and X_n is a nonnegative integer. Let $\xi_n = t_{n+1} - t_n$ and η_n the service time of nth customer. Furthermore, let $X_n = i$, if

$$(k-1)T < iT + \eta_n - \xi_n \le kT \qquad (k \ge 1),$$

then $X_{n+1} = k$, and if $iT + \eta_n - \xi_n \le 0$, then $X_{n+1} = 0$. Hence, X_n is a homogeneous Markov chain with transition probabilities p_{ik}, where

$$p_{ik} = P\{(k-i-1)T < \eta_n - \xi_n \le (k-i)T\}$$

if $k \ge 1$, and

$$p_{i0} = P\{\eta_n - \xi_n \le -iT\}.$$

Introduce the notations

$$f_j = P\{(j-1)T < \eta_n - \xi_n \le jT\}, \tag{6}$$

$$p_{ik} = f_{k-i} \quad \text{if} \quad k \ge 1, \quad p_{i0} = \sum_{j=-\infty}^{-i} f_j = \hat{f}_i. \tag{7}$$

Under the conditions of the previous theorem we have

Theorem 2. *Let us consider the above described system and introduce a Markov chain whose states correspond to the waiting time (in the sense the waiting time is the number of actual state multiplied by T) at the arrival time of customers. The matrix of transition probabilities for this chain is*

$$\begin{bmatrix}
\sum\limits_{j=-\infty}^{0} f_j & f_1 & f_2 & f_3 & f_4 & \dots \\
\sum\limits_{j=-\infty}^{-1} f_j & f_0 & f_1 & f_2 & f_3 & \dots \\
\sum\limits_{j=-\infty}^{-2} f_j & f_{-1} & f_0 & f_1 & f_2 & \dots \\
\sum\limits_{j=-\infty}^{-3} f_j & f_{-2} & f_{-1} & f_0 & f_1 & \dots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}$$

its elements are defined by (10) and (13). The generating function of the ergodic distribution is

$$P(z) = \left[1 - \frac{\sum_{i=1}^{\infty} q_i \left\{ \left\lceil \frac{i}{n} \right\rceil - (1-r)^{i-1(\mathrm{mod}\ n)} \frac{1-(1-r)^{\lceil \frac{i}{n} \rceil n}}{1-(1-r)^n} \right\}}{\frac{Q_1(1-r)^n}{(1-r)[1-(1-r)^n]}}\right] \times \frac{\frac{Q_1}{1-r} - \frac{Q_1[1-(1-r)^n]}{1-r} \frac{z}{z-(1-r)^n}}{1 - F_+(z) - \frac{Q_1[1-(1-r)^n]}{1-r} \frac{z}{z-(1-r)^n}}, \tag{8}$$

where

$$F_+(z) = \sum_{j=1}^{\infty} f_j z^j, \qquad Q_1 = \sum_{j=1}^{\infty} q_j (1-r)^j;$$

the condition of existence of ergodic distribution is

$$\frac{\sum_{i=1}^{\infty} q_i \left\{ \left\lceil \frac{i}{n} \right\rceil - (1-r)^{i-1(\mathrm{mod}\ n)} \frac{1-(1-r)^{\lceil \frac{i}{n} \rceil n}}{1-(1-r)^n} \right\}}{\frac{Q_1(1-r)^n}{(1-r)[1-(1-r)^n]}} < 1. \tag{9}$$

2 Coincidence of Ergodicity Conditions

In the case of queue length the ergodicity condition can be obtained in the form (see [6])

$$\sum_{k=0}^{\infty} \sum_{j=1}^{n} q_{kn+j}[(k+1)n+1] - \frac{nr}{1-(1-r)^n} \sum_{k=0}^{\infty} \sum_{j=1}^{n} q_{kn+j}(1-r)^{j-1} < 1,$$

which can be written as

$$\sum_{k=0}^{\infty} \sum_{j=1}^{n} q_{kn+j}(k+1) < \frac{1}{1-(1-r)^n} \sum_{k=0}^{\infty} \sum_{j=1}^{n} q_{kn+j}(1-r)^{j-1}$$

or

$$\sum_{i=1}^{\infty} q_i \left\lceil \frac{i}{n} \right\rceil < \frac{1}{1-(1-r)^n} \sum_{i=1}^{\infty} q_i (1-r)^{i-1(\mathrm{mod}\ n)}.$$

In the case of waiting time we had the ergodicity condition in the form (see [5])

$$\frac{\sum_{i=1}^{\infty} q_i \left\{ \lceil \frac{i}{n} \rceil - (1-r)^{i-1(\mathrm{mod}\ n)} \frac{1-(1-r)^{\lceil \frac{i}{n} \rceil n}}{1-(1-r)^n} \right\}}{\frac{Q_1(1-r)^n}{(1-r)[1-(1-r)^n]}} < 1,$$

or

$$\sum_{i=1}^{\infty} q_i \left\{ \left\lceil \frac{i}{n} \right\rceil - (1-r)^{i-1(\mathrm{mod}\ n)} \frac{1-(1-r)^{\lceil \frac{i}{n} \rceil n}}{1-(1-r)^n} \right\} < \frac{Q_1(1-r)^n}{(1-r)[1-(1-r)^n]}. \tag{10}$$

One has

$$\frac{Q_1(1-r)^n}{(1-r)[1-(1-r)^n]} = \frac{(1-r)^{n-1}}{1-(1-r)^n} \sum_{i=1}^{\infty} q_i (1-r)^i = \frac{1}{1-(1-r)^n} \sum_{i=1}^{\infty} q_i (1-r)^{i+n-1}$$
$$= \frac{1}{1-(1-r)^n} \sum_{i=1}^{\infty} q_i (1-r)^{i-1(\mathrm{mod}\ n)} (1-r)^{\lceil \frac{i}{n} \rceil n}.$$

Putting this value into (10) one gets

$$\sum_{i=1}^{\infty} q_i \left\lceil \frac{i}{n} \right\rceil < \frac{1}{1-(1-r)^n} \sum_{i=1}^{\infty} q_i (1-r)^{i-1(\mathrm{mod}\ n)},$$

i.e., we come to the same inequality.

The Case of Geometrical Distribution. In the general case for the queue length we had the inequality

$$\sum_{k=0}^{\infty} \sum_{j=1}^{n} q_{kn+j}(k+1) < \frac{1}{1-(1-r)^n} \sum_{k=0}^{\infty} q_{kn+j}(1-r)^{j-1}.$$

The corresponding values are

$$\sum_{k=0}^{\infty} \sum_{j=1}^{n} (1-q) q^{kn+j-1}(k+1) = 1 + (1-q) \sum_{k=0}^{\infty} k q^{kn} \sum_{j=1}^{n} q^{j-1}$$
$$= 1 + (1-q) \sum_{k=0}^{\infty} k q^{kn} \frac{1-q^n}{1-q} = 1 + (1-q^n) q^n \sum_{k=1}^{\infty} k q^{(k-1)n}$$
$$= 1 + (1-q^n) q^n \frac{1}{(1-q^n)^2} = 1 + \frac{q^n}{1-q^n} = \frac{1}{1-q^n};$$

$$\sum_{k=0}^{\infty} \sum_{j=1}^{n} q^{kn+j-1}(1-q)(1-r)^{j-1} = (1-q) \sum_{k=0}^{\infty} q^{kn} \sum_{j=1}^{n} q^{j-1}(1-r)^{j-1}$$
$$= (1-q) \frac{1-q^n(1-r)^n}{1-q(1-r)} \frac{1}{1-q^n},$$

i.e.

$$\frac{1}{1-q^n} < \frac{1}{1-(1-r)^n} \frac{(1-q)[1-q^n(1-r)^n]}{[1-q(1-r)](1-q^n)}.$$

It leads to the inequality

$$[1-(1-r)^n][1-q(1-r)] < (1-q)[1-q^n(1-r)^n],$$

or

$$rq[1-(1-r)^n] < (1-q)(1-r)^n - (1-q)q^n(1-r)^n = (1-q)(1-r)^n(1-q^n),$$

i.e.

$$\frac{rq[1-(1-r)^n]}{(1-q)(1-q^n)(1-r)^n} < 1.$$

Earlier, in the case of waiting time (when we started with geometrical distribution), we had the ergodicity condition in the form

$$\frac{rq}{1-q^n}\frac{1-q^n(1-r)^n}{1-q(1-r)} < (1-r)^n.$$

It can be written as

$$rq[1-q^n(1-r)^n] < [1-q+rq][(1-r)^n - q^n(1-r)^n],$$

adding and subtracting 1 in the last brackets

$$rq[1-q^n(1-r)^n] < (1-q)[(1-r)^n - q^n(1-r)^n] + rq[-1+(1-r)^n] + rq[1-q^n(1-r)^n],$$
$$rq[1-(1-r)^n] < (1-q)(1-r)^n(1-q^n)$$

or, finally,

$$\frac{rq[1-(1-r)^n]}{(1-q)(1-q^n)(1-r)^n} < 1,$$

as in the case of queue length.

3 Relation Between $p_0^{(q)}$ and $p_0^{(w)}$

We find the relation between $p_0^{(w)}$ and $p_0^{(q)}$. We consider a moment just before starting the service of a customer and let the system be free or let there be present one customer. The probability of this event is

$$p_0^{(q)} + p_1^{(q)} = p_0^{(q)} + \frac{1-a_0}{a_0}p_0^{(q)} = \frac{p_0^{(q)}}{a_0}.$$

One begins the service of the actual customer and it takes i time units. The waiting time for the next customer will be zero if during the first $i-1$ time slices no customer enters and on the last time slice either no customer enters (the server becomes free) or a new customer appears. So, it is not important that

during this time slice a further customer arrives or not since the server becomes free in both cases, in the case of entry of the next customer the service of the new one can be started on the following time slice, in such sense it will be taken for service without waiting.

We have the probability no customer arrives for the first $i-1$ time slices

$$\sum_{i=1}^{\infty} q_i (1-r)^{i-1} = \frac{Q_1}{1-r},$$

and we obtain

$$p_0^{(w)} = \frac{p_0^{(q)}}{a_0} \cdot \frac{Q_1}{1-r}.$$

We show the fulfilment of this equality. We have

$$p_0^{(q)} = \frac{a_0[1-B'(1)]}{a_0 + A'(1) - B'(1)}, \qquad \frac{p_0^{(q)}}{a_0} = \frac{1-B'(1)}{a_0 + A'(1) - B'(1)}.$$

So, we have to find

$$\frac{1-B'(1)}{a_0 + A'(1) - B'(1)} \cdot \frac{Q_1}{1-r},$$

this second factor appears in the expression for $p_0^{(w)}$

$$p_0^{(w)} = \left[1 - F'_+(1) \frac{(1-r)[1-(1-r)^n]}{Q_1 (1-r)^n}\right] \frac{Q_1}{1-r}.$$

Consequently, it is enough to show that

$$\frac{1-B'(1)}{a_0 + A'(1) - B'(1)} = 1 - F'_+(1) \frac{(1-r)[1-(1-r)^n]}{Q_1 (1-r)^n}.$$

We have

$$F'_+(1) = \sum_{k=1}^{\infty} k \sum_{i=(k-1)n+1}^{kn} q_i - \frac{1}{1-(1-r)^n} \sum_{i=1}^{\infty} q_i \left[(1-r)^{i-1(\mathrm{mod}\ n)} - (1-r)^{i+n-1}\right],$$

$$B'(1) = \sum_{k=0}^{\infty} \sum_{j=1}^{n} q_{kn+j}[1 + (k+1)nr] - \frac{nr}{1-(1-r)^n} \sum_{k=0}^{\infty} \sum_{j=1}^{n} q_{kn+j} (1-r)^{j-1},$$

$$1 - B'(1) = \frac{nr}{1-(1-r)^n} \sum_{k=0}^{\infty} \sum_{j=1}^{n} q_{kn+j} (1-r)^{j-1} - \sum_{k=0}^{\infty} \sum_{j=1}^{n} q_{kn+j} (k+1) nr,$$

$$a_0 + A'(1) - B'(1) = \frac{nr}{1-(1-r)^n} \sum_{i=1}^{\infty} q_i (1-r)^{i-1(\mathrm{mod}\ n)} (1-r)^{\lceil \frac{i}{n} \rceil n}$$

$$= \frac{nr}{1-(1-r)^n} \sum_{i=1}^{\infty} q_i (1-r)^{i+n-1} = \frac{nrQ_1(1-r)^{n-1}}{1-(1-r)^n};$$

$$\frac{\dfrac{1-B'(1)}{a_0+A'(1)-B'(1)}}{=\dfrac{\dfrac{1}{1-(1-r)^n}\displaystyle\sum_{k=0}^{\infty}\sum_{j=1}^{n}q_{kn+j}(1-r)^{j-1}-\sum_{k=0}^{\infty}\sum_{j=1}^{n}q_{kn+j}(k+1)}{\dfrac{Q_1(1-)^{n-1}}{1-(1-r)^n}}}. \tag{11}$$

Furthermore,

$$1-F'_+(1)\frac{1-(1-r)^n}{Q_1(1-r)^{n-1}}=\frac{Q_1(1-r)^{n-1}-[1-(1-r)^n]F'_+(1)}{Q_1(1-r)^{n-1}}. \tag{12}$$

Multiplying in (11) the numerator and denominator by $1-(1-r)^n$, the denominators will be the same. In (12) the numerator will be

$$\begin{aligned}&Q_1(1-r)^{n-1}-\sum_{k=1}^{\infty}k\sum_{i=(k-1)n+1}^{kn}q_i[1-(1-r)^n]\\&+\sum_{i=1}^{\infty}q_i(1-r)^{i-1(\mathrm{mod}\ n)}-\sum_{i=1}^{\infty}q_i(1-r)^{i+n-1}\\=&\sum_{i=1}^{\infty}q_i(1-r)^{i-1(\mathrm{mod}\ n)}-[1-(1-r)^n]\sum_{k=1}^{\infty}k\sum_{i=(k-1)n+1}^{kn}q_i,\end{aligned}$$

i.e. the two numerators coincide.

As example we compute directly these probabilities in the case of geometrical service time distribution. We have from [5,6]

$$\begin{aligned}p_0^{(w)}&=\frac{1-q}{1-q(1-r)}-\frac{rq[1-(1-r)^n]}{[1-q(1-r)](1-q^n)(1-r)^n},\\p_0^{(q)}&=1-r-\frac{rq(1-r)[1-q^n(1-r)^n]}{[1-q(1-r)](1-q^n)(1-r)^n},\\a_0&=\frac{(1-r)(1-q)}{1-q(1-r)}.\end{aligned} \tag{13}$$

We obtain

$$\begin{aligned}p_0^{(w)}&=\frac{p_0^{(q)}}{a_0}\sum_{i=1}^{\infty}(1-q)q^{i-1}(1-r)^{i-1}\\&=\left[1-r-\frac{rq(1-r)[1-q^n(1-r)^n]}{[1-q(1-r)](1-q^n)(1-r)^n}\right]\cdot\frac{1-q(1-r)}{(1-q)(1-r)}\cdot\frac{1-q}{1-q(1-r)}\\&=1-\frac{rq[1-q^n(1-r)^n]}{[1-q(1-r)](1-q^n)(1-r)^n},\end{aligned}$$

it corresponds to the value (13).

4 Mean Values

For the generating function of queue length we obtained the expression

$$P(z) = \frac{p_0}{a_0} \frac{a_0[zA(z) - B(z)] + (1 - a_0)z[A(z) - B(z)]}{[z - B(z)]}.$$

Introducing the notations

$$S(z) = -a_0 B(z) + zA(z) - zB(z) + a_0 zB(z),$$
$$N(z) = z - B(z),$$

for the mean value of queue length we get

$$P'(1) = \frac{p_0}{a_0} \frac{S''(1)N'(1) - S'(1)N''(1)}{2N'^2(1)},$$

or in the terms of generating functions (1) and (2)

$$P'(1) = \frac{p_0}{a_0} \left\{ \frac{A''(1)}{2[1 - B'(1)]} + \frac{a_0 + A'(1) - B'(1)}{1 - B'(1)} - a_0 + \frac{a_0 - 1 + A'(1)}{2[1 - B'(1)]^2} B''(1) \right\}.$$

As example we can consider the case of geometrical service time distribution $[q_i = q^{i-1}(1 - q)]$, then the elements of this expression are

$$1 - B'(1) = \frac{nr}{1 - (1 - r)^n} \frac{(1 - q)(1 - r)^n(1 - q^n) - rq[1 - (1 - r)^n]}{(1 - q^n)[1 - q(1 - r)]};$$

$$\frac{a_0 + A'(1) - B'(1)}{1 - B'(1)} = \frac{(1 - q)(1 - q^n)(1 - r)^n}{\{(1 - q)(1 - r)^n(1 - q^n) - rq[1 - (1 - r)^n]\}};$$

$$a_0 = \frac{(1 - r)(1 - q)}{1 - q(1 - r)};$$

$$\frac{-1 + a_0 + A'(1)}{2[1 - B'(1)]^2} = \frac{q[1 - (1 - r)^n]^2(1 - q^n)[1 - q(1 - r)]}{2n\{(1 - q)(1 - q^n)(1 - r)^n - rq[1 - (1 - r)^n]\}^2};$$

$$A''(1) = \frac{(2nr - nr^2)rq}{[1 - q(1 - r)](1 - q^n)} + \frac{n^2r^2 \cdot rq(1 - q^n)}{[1 - q(1 - r)](1 - q^n)^2};$$

$$B''(1) = 2(1 - r) + \frac{[-2nr + nr^2 - n^2r^2](1 - r)^n}{1 - (1 - r)^n}$$
$$+ \frac{rq}{1 - (1 - r)^n} \left\{ \frac{2n^2r^2q^n[1 - (1 - r)^n]}{(1 - q^n)^2[1 - q(1 - r)]} + \frac{n^2r^2}{1 - q^n} \frac{1 - q^n(1 - r)^n}{1 - q(1 - r)} \right.$$
$$\left. + \frac{nr^2[1 + q(1 - r)]}{1 - q^n} \frac{1 - q^n(1 - r)^n}{[1 - q(1 - r)]^2} \right\}.$$

Numerical results were obtained in [7].

By using the generating function one can compute the mean value of waiting time (measured in cycles). In our case it is equal to

$$\overline{C} = P'(1) = \frac{F''_+(1) + \dfrac{2Q_1(1-r)^n}{(1-r)[1-(1-r)^n]}}{2\left\{\dfrac{Q_1(1-r)^n}{(1-r)[1-(1-r)^n]} - F'_+(1)\right\}} - \frac{1}{1-(1-r)^n},$$

where

$$F''_+(1) = \sum_{k=2}^{\infty} k(k-1) \sum_{i=(k-1)n+1}^{kn} q_i - \sum_{k=1}^{\infty} \frac{2k}{(1-r)^{kn+1}} \sum_{i=kn+1}^{\infty} q_i(1-r)^i.$$

Numerical results concerning the mean value were obtained for the case of exponential service time distribution in [4], there was also possible the optimization of cycle length.

The Second Derivative of $F_+(z)$. Finally, we give an expression for $F''_+(1)$ in a more convenient from the viewpoint of computation form. We have [5]

$$F_+(z) = \sum_{k=1}^{\infty} z^k \left\{ \sum_{i=(k-1)n+2}^{kn} q_i + \frac{1}{(1-r)^{kn+1}} \sum_{i=kn+1}^{\infty} q_i(1-r)^i \right.$$
$$\left. - \frac{1}{(1-r)^{(k-1)n+1}} \sum_{i=(k-1)n+2}^{\infty} q_i(1-r)^i \right\}.$$

From it

$$F''_+(z) = \sum_{k=2}^{\infty} k(k-1) \left\{ \sum_{i=(k-1)n+2}^{kn} q_i + \frac{1}{(1-r)^{kn+1}} \sum_{kn+1}^{\infty} q_i(1-r)^i \right.$$
$$\left. - \frac{1}{(1-r)^{(k-1)n+1}} \sum_{i=(k-1)n+2} q_i(1-r)^i \right\},$$

and find the value $F''_+(1)$.

The sum of first terms is

$$\sum_{k=2}^{\infty} k(k-1) \sum_{i=(k-1)n+2}^{kn} q_i = 2(q_{n+2} + q_{n+3} + \ldots + q_{2n}) + 6(q_{2n+2} + q_{2n+3} + \ldots + q_{3n} + \ldots$$
$$+ k(k-1) \sum_{i=(k-1)n+2}^{kn} q_i + (k+1)k \sum_{i=kn+2}^{(k+1)n} q_i + \ldots$$

The second and third terms may be written in the form of the following table:

$$\frac{2}{(1-r)^{2n+1}}\sum_{i=2n+1}^{\infty} q_i(1-r)^i - \frac{2}{(1-r)^{n+1}}\sum_{i=n+2}^{\infty} q_i(1-r)^i,$$

$$\frac{6}{(1-r)^{3n+1}}\sum_{i=3n+1}^{\infty} q_i(1-r)^i - \frac{6}{(1-r)^{2n+1}}\sum_{i=2n+2}^{\infty} q_i(1-r)^i,$$

$$\frac{12}{(1-r)^{4n+1}}\sum_{i=2n+1}^{\infty} q_i(1-r)^i - \frac{12}{(1-r)^{3n+1}}\sum_{i=3n+2}^{\infty} q_i(1-r)^i,$$

.....

$$\frac{k(k-1)}{(1-r)^{kn+1}}\sum_{i=2n+1}^{\infty} q_i(1-r)^i - \frac{k(k-1)}{(1-r)^{(k-1)n+1}}\sum_{i=(k-1)n+2}^{\infty} q_i(1-r)^i,$$

$$\frac{(k+1)k}{(1-r)^{(k+1)n+1}}\sum_{i=2n+1}^{\infty} q_i(1-r)^i - \frac{(k+1)k}{(1-r)^{kn+1}}\sum_{i=kn+2}^{\infty} q_i(1-r)^i,$$

.....

Adding to the first element of a row the second element from the next one, one gets

$$\frac{k(k-1)}{(1-r)^{kn+1}}\sum_{i=kn+1}^{\infty} q_i(1-r)^i - \frac{(k+1)k}{(1-r)^{kn+1}}\sum_{i=kn+2}^{\infty} q_i(1-r)^i$$

$$= \frac{1}{(1-r)^{kn+1}}\sum_{i=kn+1}^{\infty} q_i(1-r)^i[k^2-k-k^2-k] + \frac{(k+1)k}{(1-r)^{kn+1}}q_{kn+1}(1-r)^{kn+1}$$

$$(k+1)kq_{kn+1} - \frac{2k}{(1-r)^{kn+1}}\sum_{i=kn+1}^{\infty} q_i(1-r)^i.$$

The sum of these values is

$$\sum_{k=2}^{\infty}(k+1)kq_{kn+1} - \sum_{k=2}^{\infty}\frac{2k}{(1-r)^{kn+1}}\sum_{i=kn+1}^{\infty} q_i(1-r)^i - \frac{2}{(1-r)^{n+1}}\sum_{i=n+2}^{\infty} q_i(1-r)^i$$

$$= \sum_{k=2}^{\infty}(k+1)kq_{kn+1} - \sum_{k=2}^{\infty}\frac{2k}{(1-r)^{kn+1}}\sum_{i=kn+1}^{\infty} q_i(1-r)^i$$

$$-\frac{2}{(1-r)^{n+1}}\sum_{i=n+1}^{\infty} q_i(1-r)^i + \frac{2}{(1-r)^{n+1}}q_{n+1}(1-r)^{n+1}$$

$$= \sum_{k=2}^{\infty}(k+1)kq_{kn+1} - \sum_{k=1}^{\infty}\frac{2k}{(1-r)^{kn+1}}\sum_{i=kn+1}^{\infty} q_i(1-r)^i + 2q_{n+1}$$

$$= \sum_{k=1}^{\infty}k(k+1)q_{kn+1} - \sum_{k=1}^{\infty}\frac{2k}{(1-r)^{kn+1}}\sum_{i=kn+1}^{\infty} q_i(1-r)^i.$$

Adding this value to the first term

$$2\sum_{i=n+2}^{2n} q_i + 6\sum_{i=2n+2}^{3n} q_i + \ldots + k(k-1)\sum_{i=(k-1)n+2}^{kn} q_i$$
$$+\sum_{k=1}^{\infty} k(k+1)q_{kn+1} - \sum_{k=1}^{\infty}\frac{2k}{(1-r)^{kn+1}}\sum_{i=kn+1}^{\infty} q_i(1-r)^i$$
$$=\sum_{k=1}^{\infty} k(k+1)\sum_{i=kn+1}^{(k+1)n} q_i - 2\sum_{k=1}^{\infty}\frac{k}{(1-r)^{kn+1}}\sum_{i=kn+1}^{\infty} q_i(1-r)^i = F_+''(1).$$

$F_+''(1)$ can written in the form

$$F_+''(1) = \sum_{k=1}^{\infty} k(k+1)\sum_{i=kn+1}^{(k+1)n} q_i - 2\sum_{k=1}^{\infty} k\sum_{i=kn+1}^{\infty} q_i(1-r)^{i-kn-1}.$$

Its first term is

$$2(q_{n+1}+\ldots+q_{2n}) + 6(q_{2n+1}+\ldots+q_{3n}) + \ldots = \sum_{i=n+1}^{\infty} q_i\left\lceil\frac{i}{n}\right\rceil\left(\left\lceil\frac{i}{n}\right\rceil - 1\right).$$

The second term, $\sum\limits_{k=1}^{\infty} k \sum\limits_{i=kn+1}^{\infty} q_i(1-r)^{i-kn-1}$ is represented by the table

$$\begin{array}{llll} q_{n+1} \,..\, q_{2n}(1-r)^{n-1} & q_{2n+1}(1-r)^n & q_{2n+2}(1-r)^{n+1} & .. \; q_{3n}(1-r)^{2n-1} \;.. \\ & 2q_{2n+1} & 2q_{2n+2}(1-r) & .. \; 2q_{3n}(1-r)^{n-1} \;.. \end{array}$$

From each n columns one can factor out

$$(1-r)^{i-1(\mathrm{mod}\ n)}$$

and there remains

$$\begin{array}{llll} q_{n+1} \,..\, q_{2n} & q_{2n+1}(1-r)^n & .. \; q_{3n}(1-r)^n & q_{3n+1}(1-r)^{2n} \;.. \\ & 2q_{2n+1} & .. \; q_{3n} & 2q_{3n+1}(1-r)^n \;.. \\ & & & 3q_{3n+1} \qquad\quad .. \end{array}$$

Now, let us consider an arbitrary term q_{kn+j}, it gives

$$q_{kn+j}\left[(1-r)^{(k-1)n} + 2(1-r)^{(k-2)n} + \ldots + (k-1)(1-r)^n + k\right].$$

The expression in brackets is

$$(1-r)^{(k-1)n}\left[\,1 + 2(1-r)^{-n} + 3(1-r)^{-2n} + \ldots + (k-1)(1-r)^{-(k-2)n} + k(1-r)^{-(k-1)n}\,\right]$$

or

$$\frac{1}{a^{k-1}}\left[\,1+2a+3a^2+\ldots+(k-1)a^{k-2}+ka^{k-1}\,\right],$$

where

$$a=(1-r)^{-n}.$$

Since

$$a+a^2+\ldots+a^k=a\frac{1-a^k}{1-a}=\frac{a-a^{k+1}}{1-a},$$

$$1+2a+3a^2+\ldots+(k-1)a^{k-2}+ka^{k-1}=\frac{\mathrm{d}}{\mathrm{d}a}\frac{a-a^{k+1}}{1-a}$$

$$=\frac{[1-(k+1)a^k][1-a]+a-a^{k+1}}{(1-a)^2}=\frac{1-(k+1)a^k+ka^{k+1}}{(1-a)^2}.$$

Multiplying by $1/a^{k-1}$ and substituting $a=(1-r)^{-n}$, one obtains

$$(1-r)^{(k-1)n}\frac{1-(k+1)\dfrac{1}{(1-r)^{kn}}+k\dfrac{1}{(1-r)^{(k+1)n}}}{\left(1-\dfrac{1}{(1-r)^n}\right)^2}$$

$$=(1-r)^{(k+1)n}\frac{1-(k+1)\dfrac{1}{(1-r)^{kn}}+k\dfrac{1}{(1-r)^{(k+1)n}}}{[1-(1-r)^n]^2}$$

$$=\frac{1}{[1-(1-r)^n]^2}\left\{k-(k+1)(1-r)^n+(1-r)^{(k+1)n}\right\},$$

and for one element of the second term we get the coefficient

$$\frac{2}{[1-(1-r)^n]^2}(1-r)^{i-1(\mathrm{mod}\ n)}\left\{\left\lceil\frac{i}{n}\right\rceil-1-\left\lceil\frac{i}{n}\right\rceil(1-r)^n+(1-r)^{\left\lceil\frac{i}{n}\right\rceil n}\right\},$$

and finally

$$F''_+(1)=\sum_{i=n+1}^{\infty}q_i\left\{\left\lceil\frac{i}{n}\right\rceil\left(\left\lceil\frac{i}{n}\right\rceil-1\right)\right.$$

$$\left.-\frac{2}{[1-(1-r)^n]^2}(1-r)^{i-1(\mathrm{mod}\ n)}\left[\left\lceil\frac{i}{n}\right\rceil-1-\left\lceil\frac{i}{n}\right\rceil(1-r)^n+(1-r)^{\left\lceil\frac{i}{n}\right\rceil n}\right]\right\}$$

or

$$F''_+(1) = \sum_{i=n+1}^{\infty} q_i \left\{ \left\lceil \frac{i}{n} \right\rceil \left(\left\lceil \frac{i}{n} \right\rceil - 1 \right) \right.$$
$$\left. - \frac{2}{[1-(1-r)^n]^2} (1-r)^{i-1(\mathrm{mod}\ n)} \left[\left\lceil \frac{i}{n} \right\rceil (1-(1-r)^n) - 1 + (1-r)^{\left\lceil \frac{i}{n} \right\rceil n} \right] \right\}.$$

References

1. Koba, E.V.: Queueing system with repetition of requests for service and FCFS service discipline. Dopovidi NAN Ukrainy, pp. 101–103 (2000). (in Russian)
2. Koba, E.V., Pustova, S.V.: Lakatos queuing systems, their generalization and application. Cybern. Syst. Anal. **48**, 387–396 (2012)
3. Lakatos, L., Szeidl, L., Telek, M.: Introduction to Queueing Systems with Telecommunication Applications. Springer, Boston (2013). https://doi.org/10.1007/978-1-4614-5317-8
4. Lakatos, L., Efrosinin, D.: A discrete waiting time model for optical signals. In: Vishnevsky, V., Kozyrev, D., Larionov, A. (eds.) DCCN 2013. CCIS, vol. 279, pp. 114–123. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-05209-0_10
5. Lakatos, L.: On the waiting time in the discrete cyclic–waiting system of $Geo/G/1$ type. In: Vishnevsky, V., Kozyrev, D. (eds.) DCCN 2015. CCIS, vol. 601, pp. 86–93. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30843-2_9
6. Lakatos, L.: On the queue length in the discrete cyclic-waiting system of $Geo/G/1$ type. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds.) DCCN 2016. CCIS, vol. 678, pp. 121–131. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-51917-3_12
7. Lakatos, L., Serebriakova, S.V.: Number of calls in a cyclic waiting system. Reliab.: Theory Appl. **11**, 37–43 (2016)

A Retrial Queueing System with Alternating Inter-retrial Time Distribution

Valentina Klimenok[1], Alexander Dudin[1], and Vladimir Vishnevsky[2(✉)]

[1] Department of Applied Mathematics and Computer Science, Belarusian State University, 220030 Minsk, Belarus
{klimenok,dudin}@bsu.by
[2] Institute of Control Sciences of Russian Academy of Sciences and Closed Corporation "Information and Networking Technologies", Moscow, Russia
vishn@inbox.ru

Abstract. We consider a single-server retrial queuing system with Markovian Arrival Process (MAP) and phase-type (PH) service time distribution. Customers which find the server busy enter the orbit of infinite size and try their luck after some random time. Concerning the retrial process, we suppose that inter-retrial times have PH distribution if the number of customers in the orbit does not exceed some threshold and have exponential distribution otherwise. Such an assumption allows to some extent take into account the realistic nature of retrial process and, at the same time, to avoid a large increase in the dimensionality of the state space of this process. We consider two different policies of repeated attempts and describe the operation of the system by two different multi-dimensional Markov chains: by quasi-Toeplitz Markov chain in the case of a constant retrial rate and by asymptotically quasi-Toeplitz Markov chain in the case of an infinitely increasing retrial rate. Both chains are successfully analyzed in this paper. We derive the ergodicity condition, calculate the stationary distribution and the main performance measures of the system.

Keywords: Single-server retrial queueing system
Phase type and exponential distribution of inter-retrial times
Stationary distribution · Performance measures

1 Introduction

Retrial queueing systems are characterized by the fact that an arrival customer, which find the servers busy, does not enter the queue or leaves the system forever, but tries his/her luck after some random time. Retrial queueing systems describe the operation of many switching telephone systems, modern telecommunication networks, contact centers, etc. Such queues have been extensively studied under a variety of scenarios for single and multiple server cases, for references see,

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 302–315, 2018.
https://doi.org/10.1007/978-3-319-99447-5_26

e.g., surveys [1,2] and books [3,4]. In the most of research the systems with a stationary Poisson input and exponential distribution of inter-retrial times are analyzed.

Analysis of the current situation shows that there is a practical need and theoretical preconditions for the development of theory of systems with non-exponential distributions of intervals between retrials. To the present day, such kind of systems did not find much interest in the literature. A small number of publications deals with $M/G/1$ and $M/M/1$ retrial queue with non-exponential inter-retrial time distribution. But all these publications consider so-called *constant retrial policy*, see, e.g. [5–12]. Under such a policy, all customers from the orbit are allowed to generate repeated attempts so that the total retrial rate is constant. Such a policy arises naturally in problems where the server is required to search for customers (see e.g. [13]) and in communication protocols of type carrier sense multiple access (CSMA) when the base station polls the end stations.

However, in most real-life systems where the effect of retrials is observed, systems operate under the *classical retrial policy*, where each orbital customer generates a flow of repeated attempts independently of the rest of the customers in the orbit. The majority of research in the field of retrial queueing system are devoted precisely to such systems, for references see already mentioned surveys [1,2] and books [3,4]. At the same time, as far as we know, retrial queues with classical retrial policy and non-exponential inter-retrial time distribution were considered only in articles [14–16]. The inherent difficulty of such queues stems from the fact that it is necessary to keep track of the elapsed retrial time for each of a possibly very large number of orbital customers. In [14], the author developed an approximate method for calculating the steady state distribution of $M/G/1$ retrial queue with inter-retrial time that are mixtures of Erlangs. In the paper [15] the authors propose an approximation of stationary distribution of $M/G/1$ retrial queue with non-exponential inter-retrial times by noting that, for most applications, inter-retrial times are significantly shorter than service time. Thus, while the elapsed retrial times for different orbital customers are dependent, the dependence is very weak. Using this feature, the authors of [15] assume that the elapsed retrial time for any orbital customer is a random variable independent of other orbital customer's elapsed retrial times. Such an assumption greatly simplifies the mathematical analysis of the system. The approximation is used to derive the distribution of the number of orbital customers and the mean waiting time and the number of retrials per customer. In the paper [16] an approximation idea of [15] is used for the $M/PH/1$ retrial queue with PH distribution of inter-retrial times. The authors of [16] use this idea to approximate the generator of the queue itself. This allow them to approximate the performance measures of the system.

To the best of our knowledge, all research on retrial queues with non-exponential distribution of inter-retrial times assumes that input flow is a stationary Poisson one. However, the flows in the modern telecommunication networks have lost the nice properties of their predecessors in the old classic net-

works. In opposite to the stationary Poisson flow, the modern real life flows are non-stationary and correlated. The MAP (Markovian Arrival Process), see [17], is one of the most appropriate mathematical models of such flows. Retrial queues with Markovian Arrival process and exponential inter-retrial times were previously investigated in a number of papers, see, e.g., [2,4,18–23] and references therein. At the same time, we can not refer to any research work where queues with MAP and non-exponential inter-retrial times would be investigated.

In the present paper, we consider $MAP/PH/1$ retrial queue with alternating distribution of inter-retrial times. We assume that inter-retrial times have PH distribution if the number of customers in the orbit does not exceed some large threshold K and have exponential distribution otherwise. We suppose that, under a large value of K our model can be considered as a good approximation of the $MAP/PH/1$ retrial queue with PH distribution of inter-retrial times. This supposition is based on our internal convictions which, in turn, are based on the theorems by A. Ya. Khinchin, G.A. Ososkov, B.I. Grigelionis about superpositions of the large number of small flows. Our model allows to some extent take into account the realistic nature of retrial process and, at the same time, to avoid a large increase in the dimensionality of the state space of this process.

2 Model Description

We consider a single-server retrial queueing system. The primary customers arrive to the system according to a MAP (Markovian Arrival Process). The MAP is defined by means of the underlying process $\nu_t, t \geq 0$, which is an irreducible continuous time Markov chain with the state space $\{0, \ldots, W\}$ where W is some finite integer. Arrivals may occur only at the epochs of the process ν_t, $t \geq 0$, transitions. Transition rates, which are accompanied by an arrival, are combined into the matrix D_1 and transition rates, which are not accompanied by an arrival, are combined into the matrix D_0. The matrices D_0, D_1 are of size $(W+1) \times (W+1)$. The matrix $D_0 + D_1$ is an infinitesimal generator of the process $\nu_t, t \geq 0$. The stationary distribution vector $\boldsymbol{\theta}$ of this process satisfies the equations $\boldsymbol{\theta}(D_0 + D_1) = \mathbf{0}, \boldsymbol{\theta}\mathbf{e} = 1$ where $\mathbf{0}$ is a zero row vector and $\mathbf{e}$ is a column vector consisting of $1's$. The average arrival rate (fundamental rate) λ of the MAP is defined as $\lambda = \boldsymbol{\theta} D_1 \mathbf{e}$. More detailed description of the MAP and its properties is given by Lucantoni in [17].

Service time of customers has PH distribution with an irreducible representation $(\boldsymbol{\beta}, S)$. This time can be interpreted as a time until the underlying Markov process $m_t, t \geq 0$, with a finite state space $\{1, \ldots, M, M+1\}$ reaches the single absorbing state $M+1$ conditional the initial state of this process is selected among the states $\{1, \ldots, M\}$ according to probabilistic row vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)$. Transition rates of the process m_t within the set $\{1, \ldots, M\}$ are defined by the sub-generator S and transition rates into the absorbing state (which lead to service completion) are given by the entries of the column vector $\boldsymbol{S}_0 = -S\mathbf{e}$. For more information about PH distributions see, e.g., [24].

If an arriving primary customer meets the server being idle, he/she occupies the server and starts the service. If the server is busy, the customer goes to the

so called orbit. The orbit capacity is assumed to be unlimited. A customer from the orbit repeats his/her attempts to reach the server in random time intervals.

If $i, 0 < i \leq K$, customers stay in the orbit, each orbital customer generates repeated attempts at random intervals having the PH distribution with R dimensional irreducible representation $(\boldsymbol{\tau}, T)$. The underlying Markov process $r_t, t \geq 0$, has the state space $\{1, \ldots, R, R+1\}$ where the state $R+1$ is an absorbing one. Transition rates into the absorbing state are given by the entries of the column vector $\boldsymbol{T}_0 = -T\mathbf{e}$.

If i, $i > K$, customers stay in the orbit, the total rate of retrials is equal to α_i. We will consider two variants of dependence of α_i of i :

— $\alpha_i = \gamma$, $i > 0$;

— $\lim_{i\to\infty} \alpha_i = \infty$.

The latter variant includes the classic retrial strategy ($\alpha_i = i\alpha$) and the linear strategy ($\alpha_i = i\alpha + \gamma$).

3 Process of the System States

Let, at time t,

- i_t be the number of customers in the orbit, $i_t \geq 0$;
- $n_t = 0$, if the server is busy; $n_t = 1$, if the server is idle;
- m_t be the state of the underlying process of the service at the busy server, $m_t = \overline{1, M}$;
- $r_t^{(j)}$ be the state of the underlying process of the inter-retrial time of the jth orbital customer, $r^{(j)} = \overline{1, R}, j = \overline{1, i}$;
- ν_t be the state of the underlying process of the MAP, $\nu_t = \overline{0, W}$.

The process of the system states is described by a regular irreducible continuous time Markov chain ξ_t, $t \geq 0$, with state space

$$\{(i, n, \nu),\ i = 0,\ n = 0;\ \nu = \overline{0, W}\} \bigcup$$

$$\{(i, n, \nu, m),\ i = 0,\ n = 1;\ \nu = \overline{0, W}, m = \overline{1, M}\} \bigcup$$

$$\{(i, n, \nu, r^{(1)}, \ldots, r^{(i)}),\ i = \overline{1, K},\ n = 0,\ \nu = \overline{0, W},\ r^{(j)} = \overline{1, R}, j = \overline{1, i}\} \bigcup$$

$$\{(i, n, \nu, m, r^{(1)}, \ldots, r^{(i)}),\ i = \overline{1, K},\ n = 1,\ \nu = \overline{0, W},\ m = \overline{1, M},\ r^{(j)} = \overline{1, R}, j = \overline{1, i}\} \bigcup$$

$$\{(i, n, \nu),\ i > K,\ n = 0,\ \nu = \overline{0, W}\} \bigcup$$

$$\{(i, n, \nu, m),\ i > K,\ n = 1,\ \nu = \overline{0, W},\ m = \overline{1, M}\}.$$

In what follows we suppose that the states of the process ξ_t, $t \geq 0$, are enumerated in the lexicographic order.

Let $Q_{i,l}$, $i, l \geq 0$, be the matrices formed by the rates of the chain transition from the state corresponding to the value i of the denumerable component i_t to the state corresponding to the value l of this component.

Lemma 1. *Infinitesimal generator of the Markov chain ξ_t, $t \geq 0$, has the following block tridiagonal structure:*

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & O & O & \cdots & O & O & O & O & O \cdots \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & O & \cdots & O & O & O & O & O \cdots \\ O & Q_{2,1} & Q_{2,2} & Q_{2,3} & \cdots & O & O & O & O & O \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \cdots \\ O & O & O & O & \cdots & Q_{K+1,K} & Q_{K+1,K+1} & Q_{K+1,K+2} & O & O \cdots \\ O & O & O & O & \cdots & O & Q_{K+2,K+1} & Q_{K+2,K+2} & Q_{K+2,K+3} & O \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \ddots \end{pmatrix},$$

where the non-zero blocks $Q_{i,l}$ have the following form:

$$Q_{0,0} = \begin{pmatrix} D_0 & D_1 \otimes \boldsymbol{\beta} \\ I_{\bar{W}} \otimes \boldsymbol{S}_0 & D_0 \oplus S \end{pmatrix},$$

$$Q_{i,i-1} = \begin{pmatrix} O_{\bar{W}R^i \times \bar{W}R^{i-1}} & I_{\bar{W}} \otimes \boldsymbol{\beta} \otimes \boldsymbol{T}_0^{\oplus i} \\ O & O_{\bar{W}MR^i \times \bar{W}MR^{i-1}} \end{pmatrix}, \ i = \overline{1, K},$$

$$Q_{i,i} = \begin{pmatrix} D_0 \oplus T^{\oplus i} & D_1 \otimes \boldsymbol{\beta} \otimes I_{R^i} \\ I_{\bar{W}} \otimes \boldsymbol{S}_0 \otimes I_{R^i} & D_0 \oplus S \oplus (T + \boldsymbol{T}_0 \boldsymbol{\tau})^{\oplus i} \end{pmatrix}, \ i = \overline{1, K},$$

$$Q_{i,i+1} = \begin{pmatrix} O_{\bar{W}R^i \times \bar{W}R^{i+1}} & O \\ O & D_1 \otimes I_M \otimes I_{R^i} \otimes \boldsymbol{\tau} \end{pmatrix}, \ i = \overline{1, K-1},$$

$$Q_{K,K+1} = \begin{pmatrix} O_{\bar{W}R^K \times \bar{W}} & O \\ O & D_1 \otimes I_M \otimes \mathbf{e}_{R^K} \end{pmatrix},$$

$$Q_{K+1,K} = \begin{pmatrix} O_{\bar{W} \times \bar{W}R^K} & I_{\bar{W}} \otimes \boldsymbol{\beta} \otimes \boldsymbol{\tau}^{\otimes K} \\ O & O_{\bar{W}M \times \bar{W}MR^K} \end{pmatrix},$$

$$Q_{i,i-1} = \begin{pmatrix} O_{\bar{W} \times \bar{W}} & \alpha_i I_{\bar{W}} \otimes \boldsymbol{\beta} \\ O & O_{\bar{W}M \times \bar{W}M} \end{pmatrix}, \ i > K+1,$$

$$Q_{i,i} = \begin{pmatrix} D_0 - \alpha_i I_{\bar{W}} & D_1 \otimes \boldsymbol{\beta} \\ I_{\bar{W}} \otimes \boldsymbol{S}_0 & D_0 \oplus S \end{pmatrix}, \quad Q_{i,i+1} = \begin{pmatrix} O_{\bar{W} \times \bar{W}} & O \\ O & D_1 \otimes I_M \end{pmatrix}, \ i > K.$$

Here, $\oplus$ and $\otimes$ are symbols of Kronecker's sum and product respectively, see, e.g., [25].

4 Case of Constant Retrial Rate

In this section, we suppose that $\alpha_i = \gamma > 0,\ i > 0,\ \alpha_0 = 0$. This strategy describes the situations where the retrial process is controlled by some decision-maker. For example, this kind of retrial control policy is well known for the stability of the ALOHA protocol in communication systems [26]. When the number of customers in the orbit $i > K$ only one customer is allowed to make the repeated attempts in intervals, which are exponentially distributed with rate γ. Or all customers are allowed to make the retrials while the individual intensity of retrials should be equal to γ/i. In this case for $i > K+1$ the matrices $Q_{i,l}$ depend on the values i, l of the denumerable component i_t only via the difference $l - i$. According to [27,28] this means that

Corollary 1. *The Markov chain $\xi_t, t \geq 0$, belongs to the class of quasi-bath-and-death process (QBD) with $K+2$ boundary macro-states, see [27].*

Corollary 2. *The Markov chain $\xi_t, t \geq 0$, is a partial case of multi-dimensional quasi-Toeplitz Markov chain (QTMC) with $K+2$ boundary macro-states, see [28].*

In this section, we will use the following notation for the non-zero matrices $Q_{i,j}$, $i > K+1$:

$$Q_k = Q_{i,i+k-1},\ i > K+1,\ k = 0, 1, 2.$$

Theorem 1. *The necessary and sufficient condition for ergodicity of the Markov chain ξ_t, $t \geq 0$, is the fulfillment of the inequality*

$$\boldsymbol{\delta}_1 D_1 \mathbf{e} < \gamma(1 - \boldsymbol{\delta}_1 \mathbf{e}) \tag{1}$$

where

$$\boldsymbol{\delta}_1 = \mathbf{x}_1 (I_{\bar{W}} \otimes \mathbf{e}_M), \tag{2}$$

and row vectors $\mathbf{x}_0, \mathbf{x}_1$ form the vector $\mathbf{x} = (\mathbf{x_0}, \mathbf{x_1})$ which is the unique solution to the system of linear algebraic equations

$$\mathbf{x} \begin{pmatrix} D_0 - \gamma I_{\bar{W}} & (D_1 + \gamma I) \otimes \boldsymbol{\beta} \\ I_{\bar{W}} \otimes \boldsymbol{S}_0 & (D_0 + D_1) \oplus S \end{pmatrix} = \mathbf{0},\ \mathbf{x}\mathbf{e} = \mathbf{1}, \tag{3}$$

Proof. As follows from [27], the necessary and sufficient condition for ergodicity of QBD process ξ_t, $t \geq 0$, is formulated in terms of the blocks of the generator Q as follows:

$$\mathbf{x} Q_2 \mathbf{e} < \mathbf{x} Q_0 \mathbf{e} \tag{4}$$

where the vector $\boldsymbol{x}$ is the unique solution to the system of linear algebraic equations

$$\mathbf{x}(Q_0 + Q_1 + Q_2) = \mathbf{0},\ \mathbf{x}\mathbf{e} = \mathbf{1}, \tag{5}$$

Represent the stochastic vector $\mathbf{x}$ in the form $\mathbf{x} = (\mathbf{x_0}, \mathbf{x_1})$ where the vector $\mathbf{x}_n$ is of size $\bar{W} M^n$, $n = 0, 1$. Taking into account the structure of the matrices

Q_0, Q_2 and the relation $(D_0 + D_1)\mathbf{e} = \mathbf{0}^T$, we easily reduce inequality (4) to the following one:

$$\mathbf{x}_1[D_1\mathbf{e} \otimes \mathbf{e}_M] < \mathbf{x}_0\mathbf{e}\gamma. \tag{6}$$

Using notation (2) and mixed product rule, we obtain from (6) inequality (1). System of linear algebraic equations (3) follows from (5) after substitution in (5) expressions for the matrices Q_0, Q_1, Q_2 defined by Lemma 1. □

Remark 1. Ergodicity condition (1) can be interpreted as follows. The νth component of the vector $\boldsymbol{\delta}_1$ is a probability that the server is busy and the underlying process of the MAP is in the state ν conditional the orbit is overloaded. Then the left hand part of inequality (1) is the arrival rate of primary customers to the orbit. The value $1-\boldsymbol{\delta}_1\mathbf{e}$ is a probability that the server is idle under overload condition. Then the right hand part of (1) is the rate of retrials of customers from the orbit which find the server idle. It is intuitively clear that the Markov chain describing queueing model under study is ergodic if and only if the arrival rate to the orbit is less than the rate of "successful" customers from the orbit.

In case of a stationary Poisson input and exponentially distributed service time ergodicity condition (1) takes the simple form given by the following statement.

Corollary 3. *In case of stationary Poisson input and exponentially distributed service time the necessary and sufficient condition for ergodicity of the Markov chain ξ_t, $t \geq 0$, is the fulfillment of the inequality*

$$\frac{\lambda}{\mu} < \frac{\gamma}{\lambda + \gamma}. \tag{7}$$

In what follows we assume that stability condition (1) is fulfilled.

Enumerate the steady state probabilities of the chain $\xi_t, t \geq 0$, in the lexicographic order and form the row vectors $\mathbf{p}_i$ of steady state probabilities corresponding the value i of the first (countable) component, $i \geq 0$. Note that the dimension of the vectors $\mathbf{p}_i$ depends on the value of i as follows: for $i \leq K$ the dimension is equal to $\bar{W}R^i(1 + M)$, and, for $i > K$, the dimension is equal to $\bar{W}(1+M)$. To calculate the vectors $\mathbf{p}_i$, we use a stable algorithm, see [28], developed to calculate the stationary distribution of multi-dimensional quasi-Toeplitz Markov chains. In our case, the chain ξ_t, $t \geq 0$, has $K + 2$ boundary levels for $i_t = 0, 1, \ldots, K + 1$ and the algorithm consists of the following principal steps.

Algorithm 1

(1) Calculate the matrix G as the minimal nonnegative solution of the matrix equation

$$Q_0 + Q_1G + Q_2G^2 = O.$$

(2) Calculate the matrices G_{K+1}, $G_K, \ldots, G_0$ using the backward recursion

$$G_i = -(Q_{i+1,i+1} + Q_{i+1,i+2}G_{i+1})^{-1}Q_{i+1,i}, \; i = K + 1, K, \ldots, 0,$$

with initial condition $G_{K+2} = G$.

(3) Calculate the matrices

$$\bar{Q}_{i,i} = Q_{i,i} + Q_{i,i+1}G_i, \quad \bar{Q}_{i,i+1} = Q_{i,i+1},\ i \geq 0,$$

where $G_i = G$, $i \geq K+2$.

(4) Calculate the matrices Φ_i using the recursive formula

$$\Phi_0 = I,\ \Phi_i = \Phi_{i-1}\bar{Q}_{i-1,i}\left(-\bar{Q}_{i,i}\right)^{-1}, i \geq 1.$$

(5) Calculate the vector $\mathbf{p}_0$ as the unique solution to the system

$$\mathbf{p}_0\bar{Q}_{0,0} = \mathbf{0}, \quad \mathbf{p}_0[\sum_{l=0}^{K+1}(\Phi_l\mathbf{e}) + (\sum_{l=K+2}^{\infty}\Phi_l)\mathbf{e}] = 1.$$

(6) Calculate the vectors $\mathbf{p}_i$ as follows $\mathbf{p}_i = \mathbf{p}_0\Phi_i, i \geq 1$.

5 Case of Infinitely Increasing Retrial Rate

In this section, we assume that $\alpha_i \to \infty$ when $i \to \infty$. This case includes the classic strategy of retrials ($\alpha_i = i\alpha$) and the linear strategy of retrials ($\alpha_i = i\alpha + \gamma$, $i > 0$).

Comparing the definition of the chain ξ_t, $t \geq 0$, given by Lemma 1 with the definition of asymptotically quasi-Toeplitz Markov chain ($AQTMC$) given in [28], we arrive to the following

Corollary 4. *The Markov chain ξ_t, $t \geq 0$, belongs to the class of $AQTMC$, see [28].*

Proof. Let T_i be a diagonal matrix with the diagonal entries defined by modules of the corresponding diagonal entries of the matrix $Q_{i,i}, i \geq 0$. According to [28], the Corollary will be proved if we show that there exist the limits

$$Y_0 = \lim_{i\to\infty} T_i^{-1}Q_{i,i-1},\ Y_1 = \lim_{i\to\infty} T_i^{-1}Q_{i,i} + I,\ Y_2 = \lim_{i\to\infty} T_i^{-1}Q_{i,i+1}, \qquad (8)$$

and the martix $Y_0 + Y_1 + Y_2$ is a stochastic one.

Note that last $\bar{W}M$ diagonal entries of the matrices T_i, $i \geq K+1$, do not depend on i. Denote by T the diagonal matrix with just mentioned diagonal entries.

The simple calculation leads to the following expressions for the matrices Y_k :

$$Y_0 = \begin{pmatrix} O_{\bar{W}} & I_{\bar{W}} \otimes \boldsymbol{\beta} \\ O & O_{\bar{W}M} \end{pmatrix}, \quad Y_1 = \begin{pmatrix} O_{\bar{W}} & O \\ T^{-1}(I_{\bar{W}} \otimes \boldsymbol{S}_0) & T^{-1}(D_0 \oplus S) + I \end{pmatrix},$$

$$Y_2 = \begin{pmatrix} O_{\bar{W}} & O \\ O & T^{-1}(D_1 \otimes I_M) \end{pmatrix}.$$

Thus, limits (8) exist and the sum $Y_0 + Y_1 + Y_2$ is a stochastic matrix. This implies that the Markov chain ξ_t belongs to the class of $AQTMC$. □

The ergodicity condition for $AQTMC\,\xi_t$ is formulated in terms of the matrices Y_0, Y_1, Y_2. Follow to [28], we first obtain the generating function of these matrices.

Corollary 5. *The matrix generating function* $Y(z) = Y_0 + Y_1 z + Y_2 z^2$ *is of the form*

$$Y(z) = \begin{pmatrix} O_{\bar{W}} & I_{\bar{W}} \otimes \boldsymbol{\beta} \\ zT^{-1}(I_{\bar{W}} \otimes \boldsymbol{S}_0) & zT^{-1}(D_0 + D_1 z) \oplus S) + zI \end{pmatrix}.$$

Theorem 2. (*i*) *The Markov chain* ξ_t *is ergodic if the following inequality*

$$\rho = \lambda/\mu < 1 \tag{9}$$

holds.

(*ii*) *The Markov chain* ξ_t *is not ergodic if* $\rho > 1$.

Proof. (*i*) It is evidently seen from Corollary 5 that the matrix $Y(1)$ is irreducible. According to Theorem 1 from [28], this means that sufficient condition for ergodicity of the Markov chain ξ_t is the fulfillment of the inequality

$$[\det(zI - Y(z))]'_{z=1} > 0. \tag{10}$$

Our aim is to prove that inequality (10) reduces to inequality (9).

It is evident that inequality (10) is equivalent to the following inequality:

$$\det(zI - \tilde{Y}(z))'_{z=1} > 0 \tag{11}$$

where

$$\tilde{Y}(z) = \begin{pmatrix} A(z) & B(z) \\ I_{\bar{W}} \otimes \boldsymbol{\beta} & O \end{pmatrix},$$

$$A(z) = zT^{-1}[(D_0 + D_1 z) \oplus S] + zI,\ B(z) = zT^{-1}(I_{\bar{W}} \otimes \boldsymbol{S}_0).$$

The matrix $zI - \tilde{Y}(z)$ has the block structure. Using the known formula for the determinant of block matrix, we get

$$\det(zI - \tilde{Y}(z)) = \det[zI - A(z) - B(z)(I_{\bar{W}} \otimes \boldsymbol{\beta})z^{-1}] \det(zI).$$

Differentiating this relation at the point $z = 1$ we obtain

$$\begin{aligned} \det(zI - \tilde{Y}(z))]'_{z=1} = {} & [\det(zI - A(z) - B(z)(I_{\bar{W}} \otimes \boldsymbol{\beta})z^{-1})]'_{z=1} \\ & + \det[I - A(1) - B(1)(I_{\bar{W}} \otimes \boldsymbol{\beta})]. \end{aligned} \tag{12}$$

It is easy to verify that $[I - A(1) - B(1)(I_{\bar{W}} \otimes \boldsymbol{\beta})]\mathbf{e} = \mathbf{0}^T$. This implies $\det[I - A(1) - B(1)(I_{\bar{W}} \otimes \boldsymbol{\beta})] = 0$. Using this and (12), inequalities (10) and (11) are transformed to the form

$$[\det(zI - A(z) - B(z)(I_{\bar{W}} \otimes \boldsymbol{\beta})z^{-1})]'_{z=1} > 0$$

which is equivalent to

$$[\det(-z[(D_0 + D_1 z) \oplus S] - I_{\bar{W}} \otimes \boldsymbol{S}_0 \boldsymbol{\beta})]'_{z=1} > 0 \tag{13}$$

Because the matrix in (13) is an irreducible generator at the point $z = 1$, following the scheme of the proof of Corollary 1 in [28], it is possible to show that inequality (13) is equivalent to the following inequality:

$$\mathbf{x}\{z[(D_0 + D_1 z) \oplus S]\}'_{z=1}\mathbf{e} < 0 \tag{14}$$

where the vector $\mathbf{x}$ is the unique solution to the system

$$\mathbf{x}\{(D_0 + D_1) \oplus S + I_{\bar{W}} \otimes \boldsymbol{S}_0\boldsymbol{\beta}\} = \mathbf{0}, \quad \mathbf{x}\,\mathbf{e} = 1. \tag{15}$$

Represent the vector $\mathbf{x}$ in the form

$$\mathbf{x} = \boldsymbol{\theta} \otimes \boldsymbol{\sigma}, \tag{16}$$

where $\boldsymbol{\sigma}$ is a stochastic vector of size M. Then, using the relations $\boldsymbol{\theta}D(1) = \mathbf{0}$ $\boldsymbol{\sigma}\mathbf{e} = 1$, we reduce inequality (14) to the form

$$\boldsymbol{\theta}D_1\mathbf{e} < \boldsymbol{\sigma}S\mathbf{e}. \tag{17}$$

In (17), the left hand side is equal to the arrival fundamental rate λ. To simplify the right hand side, we substitute the vector $\mathbf{x}$ of form (16) into the system (15). Then we obtain

$$\boldsymbol{\theta} \otimes \boldsymbol{\sigma}S + \boldsymbol{\theta} \otimes \boldsymbol{\sigma}\boldsymbol{S}_0\boldsymbol{\beta} = \mathbf{0}, \quad \boldsymbol{\sigma}\,\mathbf{e} = 1,$$

or

$$\boldsymbol{\theta} \otimes \boldsymbol{\sigma}[S + \boldsymbol{S}_0\boldsymbol{\beta}] = \mathbf{0}, \quad \boldsymbol{\sigma}\,\mathbf{e} = 1. \tag{18}$$

It follows from (18) that $\boldsymbol{\sigma}$ is the stationary vector of underlying process of service time. Using this fact in inequality (17), we see that the right hand side of this inequality is equal to the service rate μ.

Thus, we proved that inequality (17) is equivalent to inequality (9).

(ii) By [28], the AQTMC $\xi_n, n \geq 1$, is not ergodic if inequality (9) has the opposite sign. From this the statement *(ii)* of the theorem follows immediately. □

In the following, we assume that ergodicity condition (9) is fulfilled. Enumerate the steady state probabilities of the chain $\xi_t, t \geq 0$, in the lexicographic order and form the row vectors $\mathbf{p}_i$ of steady state probabilities corresponding the value i of the first (countable) component, $i \geq 0$. To calculate the vectors $\mathbf{p}_i$, we use a stable algorithm, see [28], developed for calculating the stationary distribution of asymptotically multi-dimensional quasi-Toeplitz Markov chains. The algorithm consists of the following principal steps.

Algorithm 2

1. Compute the matrix G as the minimal non-negative solution to the matrix equation

$$G = Y(G).$$

Remark 3. Taking into account the structure of the matrix $Y(G)$, one can see that the matrix G has the following block form:

$$G = \begin{pmatrix} O_{\bar{W}} & I_{\bar{W}} \otimes \boldsymbol{\beta} \\ G^{(N,N-1)} & G^{(N,N)} \end{pmatrix}$$

where the unknown blocks $G^{(N,N-1)}$, $G^{(N,N)}$ can be calculated by the iterative method.

2. Calculate the matrices G_{i_0-1}, $G_{i_0-2}, \ldots, G_0$ using the equation of the backward recursion

$$G_i = \left(-Q_{i+1,i+1} - Q_{i+1,i+2}G_{i+1}\right)^{-1} Q_{i+1,i},$$

$i = i_0 - 1, i_0 - 2, \ldots, 0$, with boundary condition $G_i = G$, $i \geq i_0$, where i_0 is an integer defined in such a way that, for a preassigned small positive number ϵ (the accuracy of the calculations), the inequality $\|G_{i_0} - G\| < \epsilon$ holds.

3 Calculate the matrices

$$\bar{Q}_{i,i} = Q_{i,i} + Q_{i,i+1}G_i, \quad \bar{Q}_{i,i+1} = Q_{i,i+1},\ i \geq 0,$$

where $G_i = G$, $i \geq i_0$.

4 Calculate the matrices Φ_i using the recursive formula

$$\Phi_0 = I,\ \Phi_i = \Phi_{i-1}\bar{Q}_{i-1,i}\left(-\bar{Q}_{i,i}\right)^{-1}, i \geq 1.$$

5) Calculate the vector $\mathbf{p_0}$ as the unique solution of the system

$$\mathbf{p}_0\bar{Q}_{0,0} = \mathbf{0}, \quad \mathbf{p}_0[\sum_{l=0}^{K+1}(\Phi_l\mathbf{e}) + (\sum_{l=K+2}^{\infty}\Phi_l)\mathbf{e}] = 1.$$

6. Calculate the vectors $\mathbf{p}_l$ as follows: $\mathbf{p}_l = \mathbf{p}_0\Phi_l$, $l \geq 1$.

6 Stationary Performance Measures

In this section, we bring a number of important stationary performance measures of the system. The corresponding formulas are valid both for the case of constant retrial rate and for the case of infinitely increasing retrial rate.

- Probability that $n (n = 0, 1)$ servers are busy and i customers stay in the orbit

$$q(n,i) = \mathbf{p}_i \begin{pmatrix} \mathbf{0}^T \\ \bar{W}R^i \sum\limits_{l=0}^{n-1} M^l \\ \mathbf{e}_{\bar{W}R^iM^n} \\ \mathbf{0}^T \\ \bar{W}R^i \sum\limits_{l=n+1}^{N} M^l \end{pmatrix}, \ i = \overline{0,K},$$

$$q(n,i) = \mathbf{p}_i \begin{pmatrix} \mathbf{0}^T \\ \bar{W} \sum\limits_{l=0}^{n-1} M^n \\ \mathbf{e}_{\bar{W}M^n} \\ \mathbf{0}^T \\ \bar{W} \sum\limits_{l=n+1}^{N} M^l \end{pmatrix}, \ i > K.$$

- Probability that i customers stay in the orbit $q_i = q(0,i) + q(1,i), \ i \geq 0.$
- Probability that $n (n = 0, 1)$ servers are busy $q^{(n)} = \sum\limits_{i=0}^{\infty} q(n,i), \ n = 0, 1.$
- Probability that $n, \ n = 0, 1,$ servers are busy conditional i customers stay in the orbit

$$q(n/i) = \frac{q(n,i)}{q_i}, \ n = 0, 1, \ i \geq 0.$$

- Probability that i customers stay in the orbit conditional $n, \ n = 0, 1,$ servers are busy

$$q(n/i) = \frac{q(n,i)}{q^{(n)}}, \ n = 0, 1, \ i \geq 0.$$

- Mean number of customers in the orbit $L = \sum\limits_{i=1}^{\infty} iq_i.$
- Probability that $n, \ n = 0, 1,$ servers are busy at an arrival epoch

$$P^{(n)} = \frac{1}{\lambda}\left[\sum_{i=0}^{k} \mathbf{p}_i \begin{pmatrix} O_{\bar{W}R^i \sum\limits_{l=0}^{n-1} M^l \times \bar{W}} \\ I_{\bar{W}} \otimes \mathbf{e}_{R^iM^n} \\ O_{\bar{W}R^i \sum\limits_{l=n+1}^{N} M^l \times \bar{W}} \end{pmatrix} + \sum_{i=K+1}^{\infty} \mathbf{p}_i \begin{pmatrix} O_{\bar{W} \sum\limits_{l=0}^{n-1} M^n \times \bar{W}} \\ I_{\bar{W}} \otimes \mathbf{e}_{M^n} \\ O_{\bar{W} \sum\limits_{l=n+1}^{N} M^l \times \bar{W}} \end{pmatrix}\right] D_1\mathbf{e}.$$

7 Conclusion

In this paper, we investigated a retrial single-server queuing system with Markovian arrival process and phase-type service time distribution. We depart from the usual assumptions about exponential distribution of inter-retrial times and suppose that inter-retrial times have PH distribution if the number of customers in the orbit does not exceed some threshold and have exponential distribution otherwise. We consider constant retrial policy and classical retrial policy of repeated attempts and describe the operation of the system by two different multi-dimensional Markov chains. For these chains, we derive the ergodicity condition and present the algorithms for calculation their stationary distributions. We also derive formulas for main performance measures of the system.

Acknowledgments. This work has been financially supported by the Russian Science Foundation and the Department of Science and Technology (India) via grant No 16-49-02021 (INT/RUS/RSF/16) for the joint research project by the V.A. Trapeznikov Institute of Control Problems of the Russian Academy Sciences and the CMS College Kottayam.

References

1. Artalejo, J.: Accessible bibliography on retrial queues. Math. Comput. Model. **30**, 223–233 (1999)
2. Gomez-Corral, A.: A bibliographical guide to the analysis of retrial queues through matrix analytic techniques. Ann. Oper. Res. **141**, 163–191 (2006)
3. Falin, G., Templeton, J.: Retrial Queues. Chapman and Hall, London (1997)
4. Artalejo, J.R., Gomez-Corral, A.: Retrial Queueing Systems: A Computational Approach. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78725-9
5. Choi, B.D., Shin, Y.W., Ahn, W.C.: Retrial queues with collision arising from unslottted CDMA/CD protocol. Queueing Syst. **11**, 335–356 (1992)
6. Gomez-Corral, A.: Stochastic analysis of a single server retrial queue with general retrial time. Naval Res. Logist. **46**, 561–581 (1999)
7. Moreno, P.: An $M/G/1$ retrial queue with recurrent customers and general retrial times. Appl. Math. Comput. **159**, 651–666 (2004)
8. Atencia, I., Moreno, P.: A single server retrial queue with general retrial time and Bernoulli schedule. Appl. Math. Comput. **162**, 855–880 (2005)
9. Choudhury, G.: An $M/G/1$ retrial queue with an additional phase of second service and general retrial time. Int. J. Inf. Manag. Sci. **20**, 1–14 (2009)
10. Kumar, B.K., Vijay Kumar, A., Arivudainambi, D.: An $M/G/1$ retrial queueing system with two phase service and preemptive resume. Ann. Oper. Res. **113**, 61–79 (2002)
11. Wu, X., Brill, P., Hlyanka, M., Wang, J.: An $M/G/1$ retrial queue with balking and retrials during service. Int. J. Oper. Res. **1**, 30–51 (2005)
12. Choudhury, G., Ke, J.-C.: A batch arrival retrial queue with general retrial times under Bernoulli vacation schedule for unreliable server and delaying repair. Appl. Math. Model. **36**, 255–269 (2012)
13. Dudin, A.N., Deepak, T.G., Varghese, C.J., Krishnamoorthy, A., Vishnevsky, V.M.: On a BMAP/G/1 retrial system with two types of search of customers from the orbit. Commun. Comput. Inf. Sci. **800**, 1–12 (2017)
14. Liang, H.M.: Retrial queues (queueing system, stability condition, K-ordering). Ph.D. thesis, University of Noth Carolina, Chapel Hill (1991)
15. Yang, T., Posner, M.J.M., Templeton, J.G.C., Li, H.: An approximation method for the $M/G/1$ retrial queue with general retrial times. Eur. J. Oper. Res. **76**, 110–116 (1994)
16. Diamond, J.E., Alfa, A.S.: An approximation method for the $M/PH/1$ retrial queue with phase type inter-retrial times. Eur. J. Oper. Res. **113**, 620–631 (1999)
17. Lucantoni, D.: New results on the single server queue with a batch Markovian arrival process. Commun. Statist.-Stoch. Models **7**, 1–46 (1991)
18. Breuer, L., Dudin, A., Klimenok, V.: A retrial $BMAP/PH/N$ system. Queueing Syst. **40**, 433–457 (2002)
19. Klimenok, V., Orlovsky, D., Dudin, A.: A $BMAP/PH/N$ system with impatient repeated calls. Asia-Pac. J. Oper. Res. **24**, 293–312 (2007)

20. Breuer, L., Klimenok, V., Birukov, A., Dudin, A., Krieger, U.: Mobile networks modeling the access to a wireless network at hot spots. Eur. Trans. Telecommun. **16**, 309–316 (2005)
21. Kim, C., Klimenok, V., Mushko, V., Dudin, A.: The $BMAP/PH/N$ retrial queueing system operating in Markovian random environment. Comput. Oper. Res. **37**, 1228–1237 (2010)
22. Klimenok, V.I., Orlovsky, D.S., Kim, C.S.: The $BMAP/PH/N$ retrial queue with Markovian flow of breakdowns. Eur. J. Oper. Res. **189**, 1057–1072 (2008)
23. Kim, C.S., Park, S.H., Dudin, A., Klimenok, V., Tsarenkov, G.: Investigaton of the BMAP/G/1'./PH/1/M tandem queue with retrials and losses. Appl. Math. Model. **34**, 2926–2940 (2010)
24. Neuts, M.: Structured Stochastic Matrices of $M/G/1$ Type and Their Applications. Marcel Dekker, New York (1989)
25. Graham, A.: Kronecker Products and Matrix Calculus with Applications. Ellis Horwood, Cichester (1981)
26. Bertsekas, D., Gallager, R.: Data Network. Prentice Hall, Englewood Cliffs, New York (1987)
27. Neuts, M.: Matrix-Geometric Solutions in Stochastic Models. The Johns Hopkins University Press, Baltimore (1981)
28. Klimenok, V.I., Dudin, A.N.: Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory. Queueing Syst. **54**, 245–259 (2006)

Implementation of Unlimited Anticollision for RFID System by Multilateration Method

Sergey Suchkov(✉), Viktor Nikolaevtsev, Dmitry Suchkov, Sergey Komkov, Aleksey Pilovets, and Sergey Nikitov

Saratov State University named after N.G.Chernyshevsky, Astrakhanskaya str. 83, 410012 Saratov, Russia
suchkov.s.g@mail.ru, nikolaevcev@yandex.ru, suchkovds@yandex.ru

Abstract. Novel method of the anticollision problem solution in radio frequency identification systems was proposed. It allows to accomplish an unlimited anticollision. The identification of objects is based on using of multiband radio frequency identification tags with time discrete coding, on the introduction of an extended code position, on the application of the multilateration method and the multi-antenna receiving system. The implementation of this method allows the simultaneous identification of an unlimited number of objects, which are marked by radio frequency identification tags, in real time. The method can be implemented in systems using radio frequency identification tags both on surface acoustic waves and on integrated circuits.

Keywords: Radio frequency identification · RFID tag
Anticollision problem · Time discrete coding
Tags on surface acoustic waves · Tags on integrated circuits

1 Introduction

The problem of simultaneous processing of large arrays of coded data is becoming more important every year due to the increasing temps of scientific and technological development, the expansion of the production activities robotization scale, the rapid change of not only production but also social technologies, the dynamism of modern society and its globalization. To solve the emerging problems, radio frequency identification (RFID) technology is widely applied. However, the expanded application of RFID systems is limited by the so-called problem of anticollision. Currently, many RFID systems use correlation methods that allows the identification of not more than 40 codes simultaneously [1–3], which is a fundamental limitation of such systems. Therefore, the system operates only in a semi-automatic mode with using of a portable reader.

There are RFID systems that allow simultaneously identification of up to 200 tags. They use tags on integrated circuits (ICs) and code identification using

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 316–326, 2018.
https://doi.org/10.1007/978-3-319-99447-5_27

probabilistic algorithms. In this case, the tags in the reader radiation zone reradiate the code signals at random moments. There is a large number of solutions in which the reader controls tags in different ways. A significant number of such solutions is based on the Aloha protocol, which was designed for the multiple network access [4–9]. The main disadvantage of such systems is a considerable time of identification, which is from several minutes to several hours.

Since the known methods (correlation and probabilistic methods) do not allow an unlimited anticollision to be performed for a limited time (up to 1 s), this paper is devoted to another approach which is based on using of the multilateration method, in which, in addition to individual codes identifying, the coordinates of the corresponding objects are also determined. Isolation of individual codes from a mixture of code signals is based on digital processing of interference patterns of code pulses, which have the difference in the delays of the code pulses, which are received by different antennas of the multi-antenna receiving system.

The basis of this approach accomplishing is using of a new type of RFID tags. These are the multiband RFID tags on surface acoustic waves (SAW) with time discrete coding [10,11]. We describe this approach in this article.

2 Calculation of the Characteristics of Multiband RFID Tags on SAW

The design of the multiband SAW tag is specific for the microwave band and allows the transmission of electromagnetic input radio frequency pulse from the rider antenna to the system of the IDTs which are tuned to adjacent non-overlapping frequency bands (subbands) within the allowed frequency band. And it also allows to return the RFID SAW tag code pulse back to the rider antenna. Shown in the Fig. 1 is the triband RFID tag on SAW. Three parallelly arranged acoustic channels are formed on its piezoelectric substrate [10,11]. Each acoustic channel contains the IDT and the reflector. Each acoustic channel is conditionally divided to some number of code positions and the reflector of each acoustic channel is located in one of the code positions. All the IDTs are connected in series and they are the elements of the microstrip transmission line, which is connected to the antenna [10,11].

Multiband RFID SAW tag operates as follows. The antenna receives an interrogation signal in the form of a radio frequency pulse with a frequency f_n (where n is the number of the subband). The current which was generated in the microwave antenna comes into a IDTs microstrip transmission line. SAW is exited if f_n is equal to the resonance frequency of one of the IDTs. In this case impedance of IDT have an active component of some tens of Ohm.

The SAW signal which propagates in the acoustic channel is reflected by the reflector. Then it transformes into a IDT and generates one reflected pulse in the antenna. The RFID SAW tag antenna emits signals from each acoustic channel to the space, and then they are received by the reader. Thus, based on the results of the first interrogation in a first subband, the first part of the code

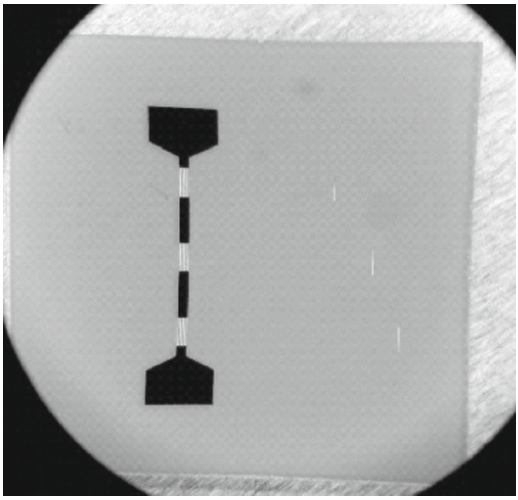


Fig. 1. A micrograph of a triband RFID SAW tag with time discrete coding with code 1, 3, 5.

of the RFID SAW tag is determines. Then, a second pulse is sent to the antenna with a frequency in the next operating frequency subband, and the next part of the RFID SAW tag code is determined in a similar way, and so on. SAW that has passed all the reflectors in the acoustic channel is absorbed by the edge absorber.

In the described construction, due to the presence of at least two acoustic channels in one RFID SAW tag, it is possible to obtain the total number of codes which is determined by the multiplicative law

$$Q_M = \prod_{i=1}^{N} M_i, \tag{1}$$

where N is the number of subbands, M_i is the number of different codes in the i-th acoustic channel.

For a triband RFID SAW tag with 100 code positions in each acoustic channel, there are 1 million codes according to Eq. 1. Four subbands give us 100 millions codes.

The characteristics of multiband RFID tags on SAW were calculated by a modified quasi-field method, which was described in [12]. Shown in Fig. 2 is the RFID SAW tag quasi-field equivalent circuit which was used in the calculation.

The velocity of the SAW under the electrode structure and the coefficient of SAW scattering on the electrodes into the volume were determined by FEM-BEM method [13].

Any IDT in such triband RFID SAW tags are multisectional IDT. Shown in Fig. 3 are the time responses of acoustic channels on lithium niobate with

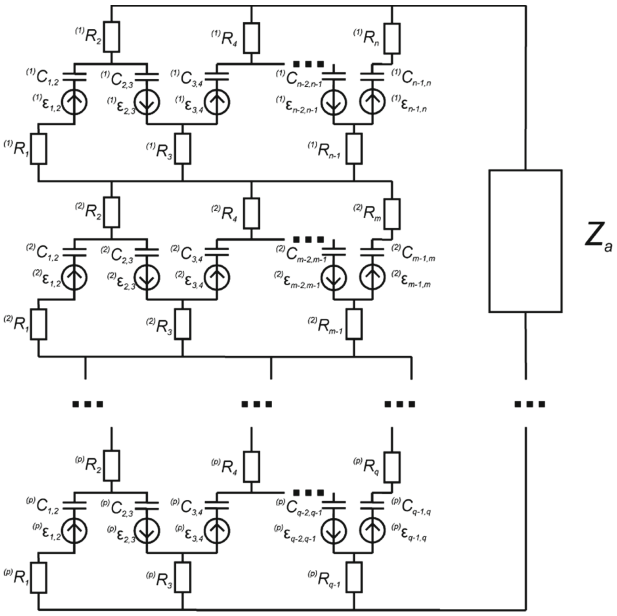


Fig. 2. Quasi-field equivalent circuit of multiband RFID tag on SAW.

the conventional (solid line) and the multisectional (dashed line) IDT of similar bandwidths. We can see that multisectional IDT supress the false signal much better (−30 dB) than conventional (−20 dB).

For the experimental study, experimental samples of triband radio frequency identification tags on lithium niobate substrate which operate in the frequency band 860–960 MHz were made. The electrode structures of triband RFID SAW tags were fabricated using the e-beam lithography unit CABL-9000C and the vacuum magnetron deposition unit ULVAC C-400-2C. To test the fabricated electrode structures of triband RFID SAW tags and to verify the results of calculations, frequency and time characteristics of the RFID SAW tags were measured. Time and frequency characteristics of the RFID SAW tags were measured by the measurement facility including the MPI TS 150 probe station and the Agilent Technologies N5242A PNA-X network analyzer.

Shown in Fig. 4 are the experimental frequency characteristics of the parameter $S_{11}(f)$ of RFID tags. The three minima of $S_{11}(f)$ on the characteristics correspond to the maximum excitation of the SAW by each of the IDTs that occurs in each of the three frequency subbands.

Shown in Fig. 5 is the calculated time response of a triband RFID tag with code 1, 3, 5 on SAW operating in allowed radio frequency band of 860–960 MHz on the 33 ns gauss pulse with a carrier frequency of 875 MHz (the first subband). The solid line shows the calculation results, the triangles show the results of the

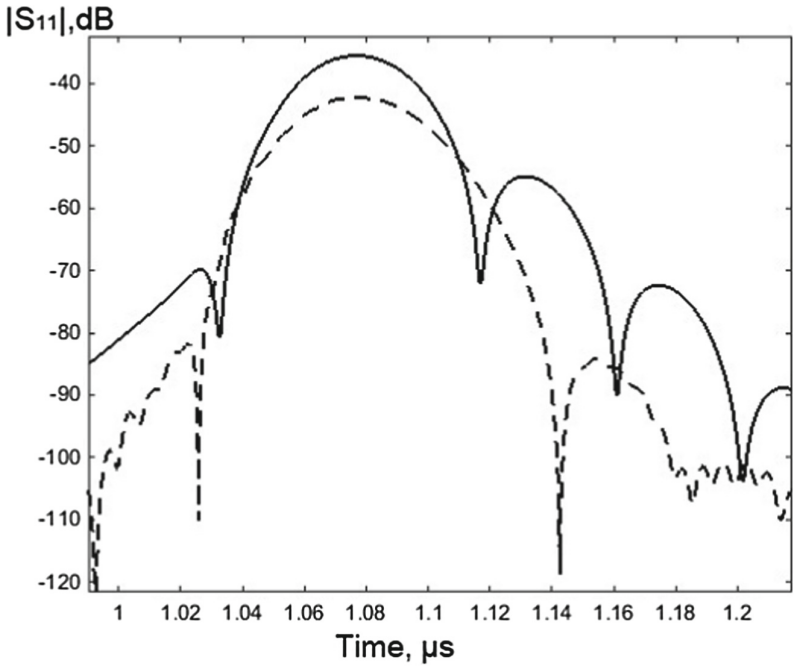


Fig. 3. Time responses of the acoustical channel with the conventional (solid line) and the multisectional (dashed line) IDT

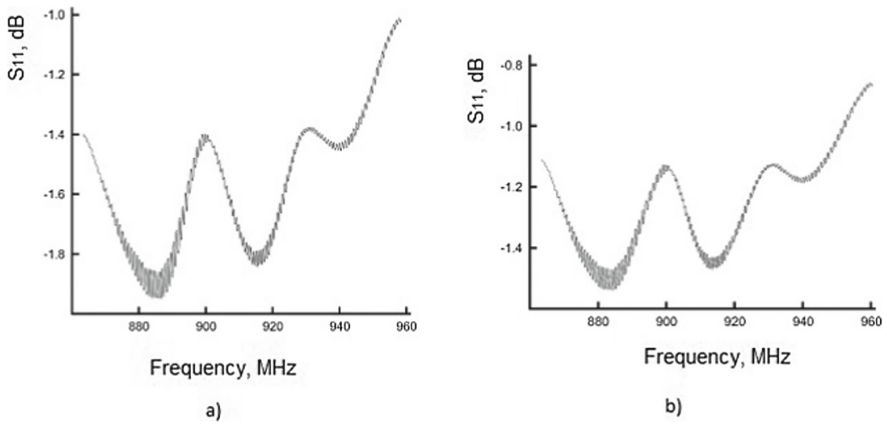


Fig. 4. Frequency responses of the triband RFID SAW tag with code (a) 1, 3, 5; (b)100, 100, 100.

experiment. The comparison shows good agreement with the calculation of both time (less than 0.01%) and signal level (less than 3 dB).

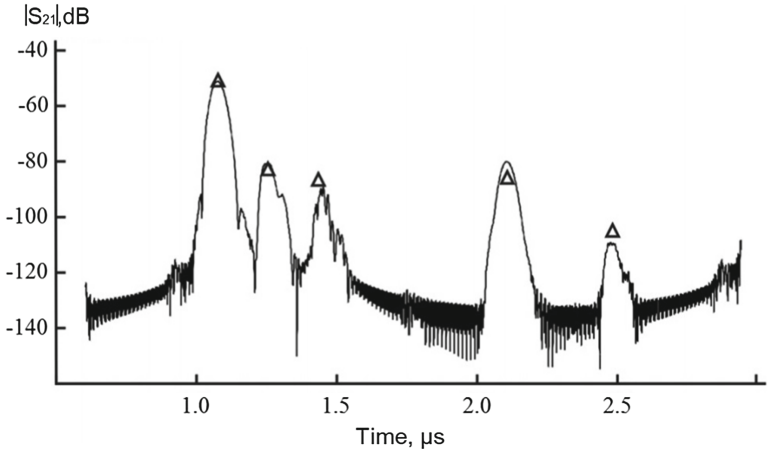


Fig. 5. Time response on the input pulse in the first subband of the triband RFID SAW tag with code 1, 3, 5.

The first peak is the code response of the first subband, the second and third peaks correspond to the suppressed responses in the second and the third subbands (cross signals), the subsequent peaks correspond to the double-pass signals of SAW in the first and the second acoustic channels. Losses of true signal in the RFID SAW tag are −50 dB, and suppression of false signals exceeds 25 dB. If such RFID SAW tags have an external antenna with an efficiency of 90%, a sensitivity level of the reader of −140 dB/mW and a maximum allowable electromagnetic wave energy flow at the reader antenna of 1 mW/cm^2 (to meet demands of the Sanitary rules and norms) then the identification distance can achieve the value of 20 m for a tag with 100 codes in each frequency subband.

For the RFID SAW tag with the code number (100, 100, 100) (see Fig. 6), no false signal is observed, since the cross signals have the same time positions, and they have a low level (less than −80 dB), and double-pass signals are not observed, because they are much more (more than 10 μs) delayed in time relative to the true signal.

3 Method of Codes Identification for RFID SAW Tags with Known Coordinates

It was shown in [14,15] that the spatial position of the RFID SAW tag relative to the reader antenna influences to the code pulses delay time. This delay time consists of the signal processing times in the RFID SAW tag and the time of

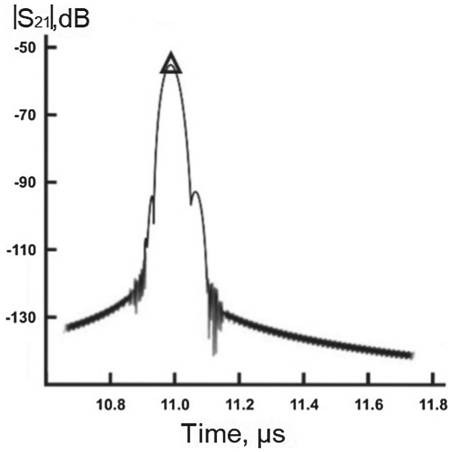


Fig. 6. Time response on the input pulse in the first subband of the triband RFID SAW tag with code 100, 100, 100.

double-pass of electromagnetic signal from the reader antenna to the tag and on the vice versa. These delay times are comparable in magnitude for real RFID systems applications, therefore, to exclude any code identification errors, an extension of the time code position is necessary (an extension of the length of time within which the pulse determines the specific code). Indeed, we suppose that there are all the identifiable RFID tags on SAW in some area D. In Fig. 7 the values of t_n and t_{n+1} denote the times of the beginning and the end of the n-th time code position, T_0 is the initial delay, R_0 is the minimum distance from the reader antenna, ΔR is the maximum allowable RFID SAW tag removal interval from the reader antenna within the area D, c is the speed of light. The extreme positions of the pulse within the time code position are shown without color, the shaded pulse is some realization of the code pulse. For the indicated time code position, the length of the expanded spatial code position at the center of which the reflecting RFID SAW tag electrode is disposed has the form

$$l_n = \frac{1}{2}\Delta t V_s = \frac{V_s}{2}\left(\frac{2\Delta R_n}{c} + \tau\right), \tag{2}$$

where V_s is the SAW velocity on the free surface of the crystal, and τ is the duration of the code pulse of the RFID tag response. Formula Eq. 2 shows that the structure of the RFID tag on SAW is determined by the size of the area D.

The set of radiating tags from the area D create code responses in the receiver of the reader within the time code positions. According to their position with the known RFID SAW tags coordinates and, therefore, with the known time delays, the codes are compared and then they are identified. Shown, for example, in Fig. 8 is a simplified two-dimensional topological scheme of the multilateration method for the analysis of the simultaneous identification of three RFID SAW tags in a three-antenna system of the reader.

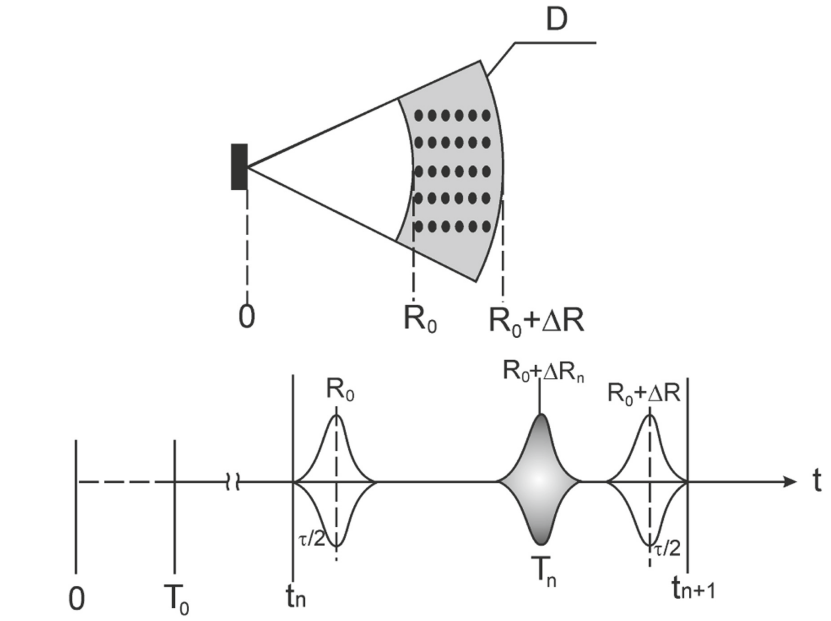


Fig. 7. Time position of the code pulse in the time code position.

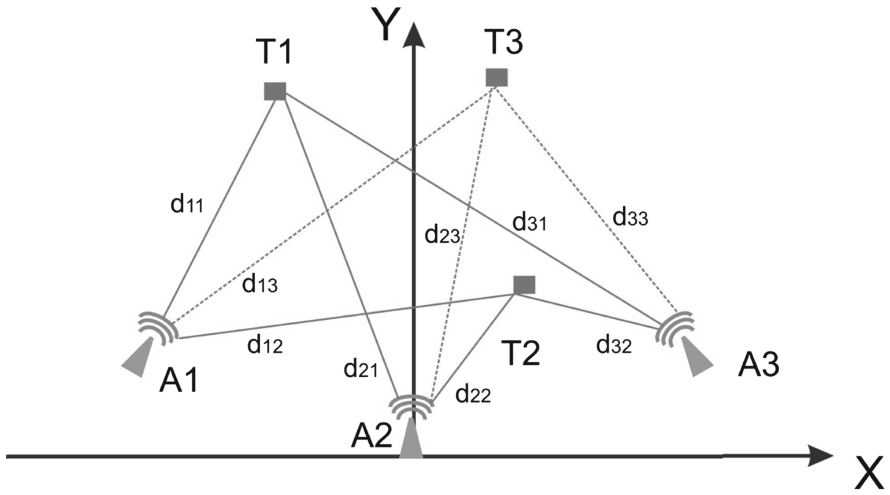


Fig. 8. Topological scheme for determining of the coordinates of the RFID SAW tags by the method of multilateration.

For the known tag coordinates, the pulse delay time in the i-th tag (for the position in the i-th time code position) is calculated by the formula

$$T_{ij} = t_i + \frac{2\,(d_{ij} - R_0)}{c}, \tag{3}$$

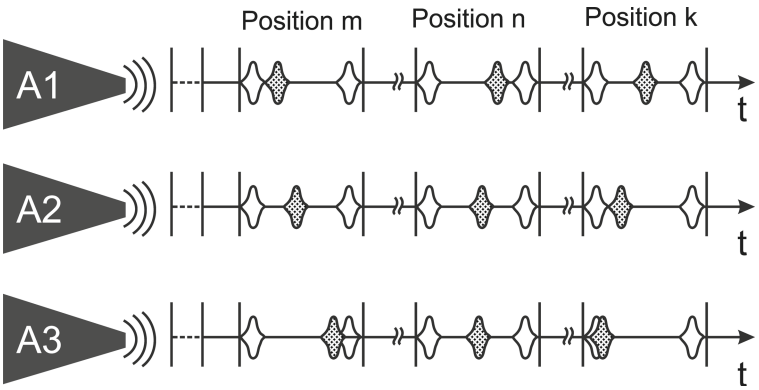


Fig. 9. Time positions of the code pulses in the corresponding code positions.

where i is the number of the tag, j is the number of the antenna, T_{ij} is the delay time of the i-th tag, which is measured by the j-th antenna, d_{ij} is the distance between the i-th tag and the j-th antenna.

Shown in Fig. 9 are the time positions of the response of the RFID SAW tag to the request pulse in one of the frequency subbands on the different antennas. The correspondence of the RFID SAW tags coordinates to the emitted code is determined by searching of the known coordinates of the marked objects using Eq. 3. Then each object is coupled to a code position number for each subband. This couples allow to determine the full codes of multiband tags.

4 Method of Codes Identification for RFID SAW Tags with Unknown Coordinates

We suppose that in the special case of three antennas the coordinates of the RFID SAW tags are unknown. Then, to determine the coordinates of the RFID SAW tags, a set of nonlinear equations of the multilateration method which was modified by introducing of the own delay of the signal in the tag must be solved [16,17]

$$\begin{cases} d_{ij}^2 = (X_i - x_j)^2 + (Y_i - y_j)^2 \\ T_{ij} = t_i + \frac{2(d_{ij} - R_0)}{c} \end{cases} \tag{4}$$

where X_i and Y_i are the unknown coordinates of the i-th tag, x_j and y_j are the known coordinates of the j-th antenna. The set *set of equations* 4 reduces to the form

$$(X_i - x_j)^2 + (Y_i - y_j)^2 = \frac{c^2}{4} (T_{ij} - t_i)^2 . \tag{5}$$

Set of equations 5 is a set of three equations with three unknowns (X_i, Y_i and t_i). When we solving the set of equations for each of the three subbands, we found coordinates of the objects generally do not coincide. The central point of

the error triangle is assumed as the coordinates of the RFID SAW tag. For more RFID SAW tags quantity, more antennas are required. This redundancy of the antenna system allows to solve a set of equations for each triple of the antennas, which will allow us to separate several pulses in one time position, while all the error triangles are used to determine the real coordinates.

If the RFID SAW tags are located in 3D space then the *set of equations* 5 has an additional variable Z_i and an additional equation appears. In this case, at least four antennas are required to determine the coordinates of the tags.

After RFID SAW tags coordinates determination the identification is carried out according to the algorithm, which was described in the previous paragraph.

5 Conclusion

A new demand in the anticollision problem of simultaneous identification of multiple RFID SAW tags is the necessity of taking into account of their spatial locations associated with the size of the identification objects and the remoteness of the RFID SAW tags from each other. This problem was investigated theoretically and experimentally. To solve this problem, it was proposed to use multiband SAW tags with realization of the time discrete coding method. In addition, an extended code position, which is determined by the size of the localization area of the marked objects, was introduced.

The number of simultaneously identified tags is practically unlimited, and the maximum number of codes of the considered triband system is one million. With an increase in the number of subbands, it is possible to increase the maximum number of codes by the multiplicative law. For example, for a quadband system, it will be one hundred million codes.

Thus, the presented results show that using of the multiband RFID SAW tags, the extended code position and the multilateration method allows to solve the problem of unlimited anticollision in RFID systems using both tags which are on SAW and on IC if we use the discrete time coding.

References

1. Hines, J.: Review of recent passive wireless saw sensor and sensor-tag activity. In: 2011 4th Annual Caneus Fly by Wireless Workshop (FBW), pp. 1–2 (2011)
2. Dudzik, E., Abedi, A., Hummels, D., da Cunha, M.P.: Wireless sensor system based on saw coded passive devices for multiple access. In: 2008 20th International Symposium on Power Semiconductor Devices and IC's, pp. 1116–1119 (2008)
3. Brown, P., et al.: Asset tracking on the international space station using global SAW tag RFID technology. In: 2007 IEEE Ultrasonics Symposium Proceedings, pp. 72–75 (2007)
4. Bing, B.: Broadband Wireless Access. Kluwer Academic Publishers, Boston (2000)
5. Furuta, S.: Patent application 08/540.092 US, IPC G06K17/00, G07B15/02, G06F13/00, G07C9/00, G06K7/00. Method and system for identifying and communicating with a plurality of contactless 1C cards. applicant and patent holder of Mitsubishi Denki Kabushiki Kaisha. – applications, 06 October 1995, (published 28 October 1994)

6. EM Microelectronic-Marin SA, P4022: Multi frequency contactless identification device. http://pdf1.alldatasheet.com/datasheetpdf/view/220419/EMMICRO/P4022.html
7. Microchip Technology Inc.: MicroID 125 kHz system design guide. (http://ww1.microchip.com/downloads/en/devicedoc/51115f.pdf
8. Black, D., Yornes, D.: Patent application 09/395.999 US, IPC G06K7/00. Method for resolving signals collisions between multiple RFID transponders in a field. applicant and patent holder of Micron Technology, Inc. – applications, 13 September 1999, (published 18 July 2000)
9. Mahdavi, P.: A system and method for communicating with multiple transponders. applicant and patent holder of Integrated Sensor Solutions. – applications, 13 November 1999, (published 05 July 2000)
10. Suchkov, S., et al.: Multiband radio frequency identification mark on surface acoustic waves. Patent for invention of the Russian Federation No. 2609012. BI, 30 January 2017 (in Russian)
11. Suchkov, S., et al.: Anticollision multiband RFID tag on surface acoustic waves. In: 2016 International Conference on Actual Problems of Electron Devices Engineering (APEDE), pp. 358–364 (2016)
12. Suchkov, S., et al.: Quasi-field method for calculation of characteristics of radio-frequency identification tags on the basis of surface acoustic waves. J. Commun. Technol. Electron. **60**, 1333–1337 (2015)
13. Suchkov, S., Yankin, S., Nikitov, S., Shatrova, Y.: Scattering of surface acoustic waves by a system of topographical irregularities comparable to a wavelength. J. Commun. Technol. Electron. **59**, 373–378 (2014)
14. Suchkov, S., et al.: Anticollision protection of distant radio-frequency identification tags based on surface acoustic waves. J. Commun. Technol. Electron. **61**, 932–936 (2016)
15. Suchkov, S., et al.: Anticollision radiofrequency identification tag on surface acoustic waves. Patent for the utility model of the Russian Federation No. 168220. BI, 30 January 2017 (in Russian)
16. Kaplan, E., Hegarty, C.: Understanding GPS: Principles And Applications. Artech House Mobile Communications Series. Artech House, Boston (2006)
17. Bechteler, T., Yenigun, H.: 2-D localization and identification based on SAW. IEEE Trans. Microw. Theory Tech. **51**, 1584–1590 (2003)

Characteristics of Lost and Served Packets for Retrial Queueing System with General Renovation and Recurrent Input Flow

E. V. Bogdanova[1], I. S. Zaryadov[1,2(✉)], T. A. Milovanova[1], A. V. Korolkova[1], and D. S. Kulyabov[1,3]

[1] Department of Applied Probability and Informatics, RUDN University, 6 Miklukho-Maklaya Str., Moscow 117198, Russia
official_kb@mail.ru,
{zaryadov_is,milovanova_ta,korolkova_av,kulyabov_ds}@rudn.university
[2] Institute of Informatics Problems of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilova Str., Moscow 119333, Russia
[3] Laboratory of Information Technologies, Joint Institute for Nuclear Research, Joliot-Curie 6, Dubna, Moscow Region 141980, Russia

Abstract. The retrial queuing system with general renovation is under investigation. The mechanism of general renovation with retrials means that the packet at the end of its service in accordance with a given probability distribution discards a certain number of other packets from the buffer and itself stays in the system for another round of service, or simply leaves the system without any effect on it. In order to obtain some probability and time related performance characteristics the embedded Markov chain technique is applied. Under the assumption of the existence of a stationary regime, the steady-state probability distribution (for the embedded Markov chain) of the number of packets in the system is obtained, as well as some other characteristics, such as the probability of the accepted task to be served or the probability of the accepted task to be dropped from the buffer, the probability distribution of number of repeated services. Also time characteristics are given.

Keywords: Retrial queueing system · General renovation
Recurrent input flow · Repeated service
Probability–time characteristics · Lost packet · Served packet

1 Introduction

Even though mathematical modelling of telecommunication systems with possible losses of information has been the subject of numerous research papers, this topic still attracts attention from the research community. The main research directions, to name a few, are:

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 327–340, 2018.
https://doi.org/10.1007/978-3-319-99447-5_28

- dropping mechanisms which regulate queue (buffer) lengths by discarding the incoming packets (see [1–8]);
- disaster arrivals, when the incoming signals cause the buffer to drop some or all the packets (see [9–20]);
- unreliable servers, which cause the packet dropping (see [21–25]);
- repairable and reliability systems (see [26–30])
- renovation, when the queue (partially or fully) empties out upon service completions (see [31–37]).

In [31] the authors have introduced the so-called renovation mechanism, when the packet at the end of its service empties the buffer with the probability q and leaves the system or with the complementary probability $p = 1 - q$ leaves the system, having no effect on it. The application of the renovation mechanism in finance and some other application fields was shown in [32]. In [33] Bocharov proposed the mathematical model of renovation mechanism with retrials (or repeated service): if the served packet empties the buffer it enters the server for another round of service. Later on the renovation mechanism was further generalized by A. V. Pechinkin, who proposed the mathematical model of general renovation: at the end of service the packet discards from the buffer of capacity $0 < r < \infty$ exactly i, $i \geq 1$, other packets with probability $q(i)$ and leaves the system, or just leaves the system without any effect on it with the complementary probability $p = 1 - \sum_{i=1}^{r} q(i)$.

In [34–36] the $GI|M|n|r$ queueing system with various types of service disciplines and renovation was studied. The $M|G|1|r$ queue was analysed in [37,38]. The first paper to analyse the queue with the general renovation and retrials is apparently [33], where the authors obtained the main steady-state characteristics. In [39] the $GI|M|1|\infty$ queueing system with renovation (when the buffer is fully emptied, in case of renovation) was thoroughly investigated: the expressions for the steady-state probabilities, the probabilities of incoming packet to be served (or dropped from the buffer) as well as main stationary waiting and sojourn time characteristics were derived in analytic form.

In [8] such performance characteristics as stationary loss rate, moments of the number in the system for $M/D/1/N$ queue were obtained in order to compare the renovation mechanism with well known active queue mechanisms like RED.

The first attempt to apply the general renovation to systems with repeated service (retrials) was done in [40] for the $M|M|1|\infty$ system. In [41] some possible approaches to the investigation of the $GI|M|1|\infty$ system were formulated. The main goal of this article is to present some new analytic results concerning the steady-state analysis of $GI|M|1|\infty$ queue with general renovation and retrials.

The main goal of this article is to present some new analytic results concerning the steady-state analysis of $GI|M|1|\infty$ queue with general renovation and retrials.

The structure of the article is follows. In Sect. 2 the description of retrial queueing system with general renovation is presented Sect. 2.1, some auxiliary probabilities are formulated Sect. 2.2, in Sect. 2.3 the embedded Markov chain and transition probabilities matrix are defined and the formulas of the steady-

state probability distribution of embedded Markov chain are obtained in Subsect. 2.4. The Sect. 3 is devoted to characteristics of served (lost) packets: in Sect. 3.1 the probability p^{serv} that the incoming into the system packet will be served is defined; the probability p^{loss} that the arriving into the system packet will be dropped from the buffer is obtained in Sect. 3.2; some other probabilities (such as the probability $\tilde{p}_1$ that none of the incoming packets will be lost, the probability $\tilde{p}_2$ that all the incoming packets will be served only once) are presented in Sect. 3.3; in Sect. 3.4 the probability distribution of number of repeated services is obtained; and in Sect. 3.5 the Laplace-Stieltjes transformation of waiting time steady-state distribution $\omega^{\text{serv}}(s)$ of accepted and served packet as well as the Laplace-Stieltjes transformation of waiting time steady-state distribution $\omega^{\text{loss}}(s)$ of accepted and lost packet are defined. In Sect. 4 the future goals are formulated.

2 The Description Of the Retrial Queueing System, the Steady-State Probability Distribution

2.1 The General Renovation with Retrials Mechanism

Consideration is given to the queueing system $GI|M|1|\infty$ with recurrent input flow, exponentially distributed service times, unlimited buffer capacity and general renovation mechanism with retrials.

The general renovation with retrial is defined as follows. The packet at the end of its service with probability $q(i)$, $i \geq 0$, drops exactly i packets from the buffer (if there are more the i packets present in it) or empties out the buffer (if there are i or less packets in it) and stays in the system for another round of service (retrial). There are two possible types of retrials: either the served packet occupies the first free place in the buffer, or remains in the server for repeated service. With probability $p = 1 - \sum_{i=0}^{\infty} q(i)$ the served packet leaves the system having no effect on it.

In order to analyse this system we use the embedded Markov chain technique. Before we can proceed some auxiliary probabilities are needed.

2.2 The Auxiliary Probabilities

In order to obtain time-probability characteristics of the system some auxiliary probabilities are needed. The first type of auxiliary probabilities—$\pi(\cdot)$—will be defined for the case, when the exact number of customers leaves the system (from the server or (and) from the buffer) and system is not empty. The second type auxiliary probabilities—$\pi^*(\cdot)$—when the buffer is emptied by one of the served packets (the server remains busy). The results of [40] are used for auxiliary probabilities deriving.

The probability of the first type: $\pi(k, n, m)$—is the probability that between successive arrival moments exactly k ($k \geq 0$) packets will be served, exactly m ($m \geq 0$) packets will be dropped from the buffer, and n ($0 \leq n \leq k$) served

packets will leave the system if at the previous arrival moment there were $n+m$ packets. For $\pi(k,n,m)$ the following formulas are valid:

$$\pi(0,0,0)=1, \qquad \pi(0,n,m)=0, \quad n\geq 1, m\geq 1; \tag{1}$$

$$\pi(k,n,0)=C_k^n p^n q^{k-n}(0), \quad p+q(0)\neq 1, \quad k\geq 1, \quad n=\overline{0,k}; \tag{2}$$

$$\pi(1,0,m)=q(m), \quad m\geq 0; \tag{3}$$

$$\pi(k,0,m)=\sum_{i=0}^{m}\pi(1,0,i)\pi(k-1,0,m-i), \quad k>1, \quad m>1; \tag{4}$$

$$\pi(k,n,m)=C_k^n p^n \pi(k-n,0,m), \quad k>1, \quad n=\overline{0,k}, \quad m\geq 1. \tag{5}$$

For the second type probability $\pi^*(k,n,m)$—the probability that k $(k\geq 1)$ served packets will empty the buffer of the system and exactly n $(0\leq n\leq k-1)$ served packets will leave the system, the following relations are valid:

$$\pi^*(0,0,0)=1; \quad \pi^*(1,0,m)=\sum_{j=m}^{\infty}q(j)=Q(m), \quad m\geq 0; \tag{6}$$

$$\pi^*(k,0,0)=(1-p)^k, \quad k\geq 1; \tag{7}$$

$$\pi^*(k,0,m) = \sum_{i=0}^{m-1}\pi(1,0,i)\pi^*(k-1,0,m-i)+\pi^*(1,0,m)\pi^*(k-1,0,0), \quad k\geq 2, m\geq 1; \tag{8}$$

$$\pi^*(k,n,0)=C_k^n p^n (1-p)^{k-n}, \quad k\geq 1, \quad n=\overline{0,k}; \tag{9}$$

$$\pi^*(k,n,m)=C_k^n p^n \pi^*(k-n,0,m), \quad k>1, n=\overline{0,k}, m\geq 0. \tag{10}$$

Also we need to define the following transformations:

$$\pi(g)=\sum_{i=0}^{\infty}g^i\pi(1,0,i), \quad \pi^*(g)=\sum_{i=0}^{\infty}g^i\pi^*(1,0,i), \tag{11}$$

where g is some variable, $0<g<1$, which will be defined in the Subsect. 2.4.

Also we may define:

$$\pi_k(g)=\sum_{i=0}^{\infty}g^i\pi(k,0,i)=\pi^k(g), \quad \pi_k^*(g)=\sum_{i=0}^{\infty}g^i\pi^*(k,0,i)=(\pi^*(g))^k, k>1. \tag{12}$$

For $\pi(g)$ and $\pi^*(g)$ the following relation is true:

$$\pi^*(g)=\frac{1-p-g\pi(g)}{1-g}. \tag{13}$$

Now we may define the embedded Markov chain and transition ptobabilities matrix.

2.3 The Embedded Markov Chain, Transition Probabilities Matrix

To investigate our system we will construct the embedded upon arrival times Markov chain $\nu_n = \nu(\tau_n - 0)$ (τ_n—the moment of the n-th task arrival) with enumerable number of states $\mathcal{X} = \{0, 1, 2, \ldots\}$ and the matrix $P = (p_{ij})_{i,j\geq 0}$ of transition probabilities.

Now the transition probability matrix $P = (p_{i,j})_{i,j=\overline{1,n+r}}$ of embedded Markov chain may be defined in the following form:

$$P = \begin{pmatrix} P_1^* & \tilde{P}_0 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots \\ P_2^* & \tilde{P}_1 & P_0 & 0 & \ldots & 0 & 0 & 0 & \ldots \\ P_3^* & \tilde{P}_2 & P_1 & P_0 & \ldots & 0 & 0 & 0 & \ldots \\ P_4^* & \tilde{P}_3 & P_2 & P_1 & \ldots & 0 & 0 & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots \\ P_{k-1}^* & \tilde{P}_{k-2} & P_{k-3} & P_{k-4} & \ldots & P_1 & P_0 & 0 & \vdots \\ P_k^* & \tilde{P}_{k-1} & P_{k-2} & P_{k-3} & \ldots & P_2 & P_1 & P_0 & \vdots \\ P_{k+1}^* & \tilde{P}_k & P_{k-1} & P_{k-2} & \ldots & P_3 & P_2 & P_1 & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{14}$$

The probability $P_{i+1}^* = p_{i,0}$ $(i \geq 0)$ corresponds to the transition probability from the state i $(i \geq 0)$ to the state 0 of the system being empty.

$$P_1^* = 1 - \alpha(p\mu), \tag{15}$$

$$P_{i+1}^* = \int_0^\infty \left(\int_0^x \left(\sum_{n=0}^{i} \sum_{k=n+1}^{\infty} \pi^*(k, n, i-n) \frac{(\mu y)^k}{k!} e^{-\mu y} \right) p\mu e^{-\mu(x-y)} dy \right) dA(x), \quad i > 1. \tag{16}$$

The probability $\tilde{P}_i = p_{i,0}$ $(i \geq 0)$ is the transition probability from the state i $(i \geq 0)$ to the state 1 when the buffer of the system is empty and the server is occupied.

$$\tilde{P}_0 = \alpha(p\mu), \tag{17}$$

$$\tilde{P}_i = \int_0^\infty \left(\sum_{n=0}^{i} \sum_{k=n+1}^{\infty} \pi^*(k, n, i-n) \frac{(\mu x)^k}{k!} e^{-\mu x} \right) dA(x), \quad i \geq 1. \tag{18}$$

The probability $P_k = p_{i+k,i+1}$ $(k \geq 0,\ i \geq 0)$—the transition probability from the state $i + k$ to the state $i + 1$ (the exactly k packets leave the system and the buffer is nor empty):

$$P_0 = \alpha\left(\mu(1 - q(0))\right), \tag{19}$$

$$P_k = \int_0^\infty \left(\sum_{j=k}^{\infty} \pi(j,k,0) \frac{(\mu x)^j}{j!} e^{-\mu x} \right) dA(x) + \int_0^\infty \left(\sum_{n=0}^{k-1} \sum_{j=n+1}^{\infty} \pi(j,n,k-n) \frac{(\mu x)^j}{j!} e^{-\mu x} \right) dA(x), \quad k \geq 1. \tag{20}$$

Here, $\alpha(s)$ is the Laplace-Stieltjes transformation of an interarrival time probability distribution function $A(x)$.

Now we may derive the formulas for steady-state probability distribution of embedded Markov chain.

2.4 The Steady-State Probability Distribution

Let's define the steady-state probability distribution of the embedded Markov chain (in assumption that the steady-state regime exists) as p_k^- $(k \geq 0)$. Here, the probability p_k^- means that there were k $(k \geq 0)$ packets in the system at the moment of arrival.

The steady-state probabilities p_k^-, $k \geq 0$, satisfy the following system of equations

$$p_0^- = \sum_{i=0}^{\infty} P_{i+1}^* p_i^-, \quad p_1^- = \sum_{i=0}^{\infty} \tilde{P}_i p_i^-, \tag{21}$$

$$p_k^- = \sum_{i=0}^{\infty} P_i p_{k-1+i}^-, \quad k \geq 2, \tag{22}$$

with the normalization requirement:

$$\sum_{k=0}^{\infty} p_k^- = 1. \tag{23}$$

The steady-state probabilities p_i^- for $i \geq 2$ may be written down in the geometric form as in [34,35,39]:

$$p_i^- = p_2^- \cdot g^{i-2}, \quad i \geq 2, \tag{24}$$

where the constant g is the unique solution $(0 < g < 1)$ of the following equation (obtained by substituting (24) in (22):

$$g = \sum_{i=0}^{\infty} P_i g^i = \alpha\left(\mu(1 - pg - \pi(g))\right). \tag{25}$$

From the (21) and (22) (by using (24)) we may derive the probabilities p_1^- and p_2^- via p_0^-:

$$p_1^- = p_0^- \tilde{P}_0 \frac{\tilde{P}(g) + P^*(g)}{P_2^* \tilde{P}(g) + \left(1 - \tilde{P}_1\right) P^*(g)}, \tag{26}$$

$$p_2^- = p_0^- \frac{\tilde{P}_0}{\tilde{P}(g)} \left(\frac{\left(1-\tilde{P}_1\right)\left(\tilde{P}(g)+P^*(g)\right)}{P_2^*\tilde{P}(g)+\left(1-\tilde{P}_1\right)P^*(g)} - 1 \right), \tag{27}$$

where $P^*(g) = \sum_{i=2}^{\infty} P_{i+1}^* g^{i-2}$, $\tilde{P}(g) = \sum_{i=1}^{\infty} \tilde{P}_i g^{i-2}$.

Finally, by using the equations (23), (25), (26) and (27), the probability p_0^- of the system being empty is derived:

$$p_0^- = \left(1 + \tilde{P}_0 \frac{\tilde{P}(g)+P^*(g)}{P_2^*\tilde{P}(g)+\left(1-\tilde{P}_1\right)P^*(g)} + \frac{\tilde{P}_0}{(1-g)\tilde{P}(g)} \left(\frac{\left(1-\tilde{P}_1\right)\left(\tilde{P}(g)+P^*(g)\right)}{P_2^*\tilde{P}(g)+\left(1-\tilde{P}_1\right)P^*(g)} - 1 \right) \right)^{-1}. \tag{28}$$

In the next section the main probabilistic and time characteristics of lost packets and served packets will be presented.

3 The Characteristics of Lost and Served Packets

In this section some additional probability and time characteristics will be presented. But first let's make some assumptions about the service discipline and the discipline of the reset of packets from the buffer, as well as the behavior of the remaining packet for possible repeated service.

- packets are served from the queue in the FCFS (First-Come-First-Served) order;
- packets to be dropped from the buffer are chosen successively starting from the head of the queue;
- the served packet occupies the first free space in the buffer.

These assumptions determine the characteristics presented below.

3.1 The Probability that the Incoming Packet Will Be Served

The probability p^{serv} that the incoming into the system packet will be served is defined as follows:

$$p^{\text{serv}} = p_0^- + \sum_{i=1}^{\infty} p_i^- \sum_{n=0}^{i} \pi(i,n,0) + \sum_{i=2}^{\infty} p_i^- \sum_{k=1}^{i-1} \sum_{n=0}^{k-1} \pi(k,n,i-k),$$

and with the help of (1)–(10), (11) and (12), (25) takes the form:

$$p^{\text{serv}} = p_0^- + p_1^-(p+q(0)) + p_2^- \frac{(p+q(0))^2}{1-g(p+q(0)} + p_2^- \frac{1}{g}\left(\frac{p+\pi(g)}{1-g(p+\pi(g))} - \frac{p+q(0)}{1-g(p+q(0))}\right) = p_0^- + p_1^-(p+q(0)) + p_2^- \frac{\pi(g)-q(0)+g(p+\pi(g))(p+q(0))}{g(1-g(p+\pi(g)))} \quad (29)$$

3.2 The Probability that the Incoming Packet Will Be Dropped from the Buffer

The probability p^{loss} that the arriving into the system packet will be dropped from the buffer by one of the served packets is defined as:

$$p^{\text{loss}} = \sum_{i=1}^{\infty} p_i^- \pi^*(1,0,i) + \sum_{i=2}^{\infty} p_i^- \sum_{k=1}^{i-1}\sum_{n=0}^{k} \pi(k,n,0)\pi^*(1,0,i-k) + \sum_{i=3}^{\infty} p_i^- \sum_{k=1}^{i-2}\sum_{n=0}^{k-1}\sum_{j=1}^{i-k-1} \pi(k,n,j)\pi^*(1,0,i-k-j),$$

and with the help of (1)–(10), (11) and (12), (25) takes the form:

$$p^{\text{loss}} = p_1^- Q(1) + p_2^- \frac{\pi^*(g)-Q(0)-gQ(1)}{g^2} + p_2^- \frac{\pi^*(g)-Q(0)}{g}\frac{p+\pi(g)}{1-g(p+\pi(g))} = p_1^- Q(1) + p_2^- \frac{\pi^*(g)-Q(0)-gQ(1)+g^2Q(1)(p+\pi(g))}{g^2(1-g(p+\pi(g)))}. \quad (30)$$

3.3 Some Other Probability Characteristics

The probability $\tilde{p}_1$ that none of the incoming packets will be lost is

$$\tilde{p}_1 = p_0^- + p_1^-(p+q(0)) + p_2^- \frac{(p+q(0))^2}{1-g(p+q(0)}.$$

The probability $\tilde{p}_2$ that all the incoming packets will be served (and only once) is

$$\tilde{p}_2 = \sum_{i=0}^{\infty} p_i^- \pi(i+1,i+1,0) = pp_0^- + p^2 p_1^- + p_2^- \frac{p^3}{1-pg}.$$

The probability $\tilde{p}_3$ that the incoming packet will be dropped by the first served packet is

$$\tilde{p}_3 = \sum_{i=1}^{\infty} p_i^- \pi^*(1,0,i) = p_1^- Q(1) + p_2^- \frac{\pi^*(g)-Q(0)-Q(1)}{g^2}.$$

3.4 The Probability Distribution of Number of Repeated Services

Let's define the probability $\tilde{q}_k$, $k \geq 0$, that the incoming packet will be served exactly k times. Then the following expressions are valid:

$$\tilde{q}_0 = p^{\text{loss}}, \tag{31}$$

$$\tilde{q}_k = \left(p^{\text{serv}}\right)^k (1-p)^{k-1} p + \left(p^{\text{serv}}\right)^k (1-p)^k p^{\text{loss}}, \quad k \geq 1. \tag{32}$$

It's easy to see that $\sum\limits_{k=0}^{\infty} \tilde{q}_k = 1$.

The mean number of services of accepted packets $\tilde{N}$ is

$$\tilde{N} = \sum_{k=0}^{\infty} k\tilde{q}_k = \frac{p^{\text{serv}}}{p^{\text{loss}} + pp^{\text{serv}}}.$$

3.5 The Time Characteristics

The Laplace-Stieltjes transformation of waiting time steady-state distribution $\omega^{\text{serv}}(s)$ of accepted and served packet is defined by the formula:

$$\omega^{\text{serv}}(s) = \frac{1}{p^{\text{serv}}} \left(p_0^- + \sum_{i=1}^{\infty} p_i^- \omega^i(s) \sum_{n=0}^{i} \pi(i,n,0) + \sum_{i=2}^{\infty} p_i^- \sum_{k=1}^{i-1} \omega^k(s) \sum_{n=0}^{k-1} \pi(k,n,i-k) \right),$$

where $\omega(s)$—Laplace-Stieltjes transformation of service time distribution function.

By using relations (1)–(10), (11) and (12), $\omega^{\text{serv}}(s)$ takes form:

$$\begin{aligned}
\omega^{\text{serv}}(s) = \frac{1}{p^{\text{serv}}} \Bigg(& p_0^- + p_1^- \omega(s)\,(p + q(0)) + p_2^- \frac{\omega^2(s)\,(p+q(0))^2}{1 - g\omega(s)\,(p+q(0))} \\
& + p_2^- \frac{\omega(s)}{g} \left(\frac{p + \pi(g)}{1 - g\omega(s)\,(p+\pi(g))} - \frac{p+q(0)}{1-g\omega(s)\,(p+q(0))} \right) \Bigg) \\
= \frac{1}{p^{\text{serv}}} \Big(& p_0^- + p_1^- \omega(s)\,(p+q(0)) \\
& + p_2^- \omega(s) \frac{\pi(g) - q(0) + g\omega(s)\,(p+q(0))\,(p+\pi(g))}{g\,(1 - g\omega(s)\,(p+\pi(g)))} \Bigg).
\end{aligned} \tag{33}$$

The Laplace-Stieltjes transformation of waiting time steady-state distribution $\omega^{\text{loss}}(s)$ of accepted and lost packet is defined as

$$\begin{aligned}
\omega^{\text{loss}}(s) = \frac{1}{p^{\text{loss}}} \Bigg(& \omega(s) \sum_{i=1}^{\infty} p_i^- \pi^*(1,0,i) \\
& + \sum_{i=2}^{\infty} p_i^- \sum_{k=1}^{i-1} \omega^{k+1}(s) \sum_{n=0}^{k} \pi(k,n,0) \pi^*(1,0,i-k) \\
& + \sum_{i=3}^{\infty} p_i^- \sum_{k=1}^{i-2} \omega^{k+1}(s) \sum_{n=0}^{k-1} \sum_{j=1}^{i-k-1} \pi(k,n,j) \pi^*(1,0,i-k-j) \Bigg),
\end{aligned}$$

and due to relations (1)–(10), (11) and (12), $\omega^{\text{loss}}(s)$ takes form:

$$\omega^{\text{loss}}(s) = \frac{\omega(s)}{p^{\text{loss}}}\Big(p_1^- Q(1) + p_2^- \frac{\pi^*(g) - Q(0) - gQ(1) + g^2 Q(1)\omega(s)\,(p + \pi(g))}{g^2\,(1 - g\omega(s)\,(p + \pi(g)))}\Big), \quad (34)$$

If w^{serv} and w^{loss} are mean waiting times for a served packet and a lost packet (they may be easily found from (33) and (34)), then the mean dwell time (time in system) w for an arbitrary packet is

$$w = w^{\text{loss}} p^{\text{loss}} + \sum_{k=1}^{\infty} (p^{\text{serv}})^k\,(1-p)^{k-1}\,pk\left(w^{\text{serv}} + \mu^{-1}\right) + \sum_{k=1}^{\infty} (p^{\text{serv}})^k\,(1-p)^k\,p^{\text{loss}}\left(k\left(w^{\text{serv}} + \mu^{-1}\right) + w^{\text{loss}}\right) = w^{\text{loss}} p^{\text{loss}} + \frac{p^{\text{serv}}\left(w^{\text{serv}} + \mu^{-1}\right) + w^{\text{loss}} p^{\text{loss}}}{1 - p^{\text{serv}}(1-p)}. \quad (35)$$

4 Conclusion

The conception of the retrial queueing system with general renovation was introduced in this article.

The main probability-time characteristics of retrial queueing system with general renovation such as the probability distribution (24), (26), (27) and (28), as well as the probability of arrival packet to be served p^{serv} (29) or to be dropped from the queue p^{loss} (30), as well as Laplace-Stieltjes transformation of waiting time steady-state distribution $\omega^{\text{serv}}(s)$ of accepted and served packet (33) and Laplace-Stieltjes transformation of waiting time steady-state distribution $\omega^{\text{loss}}(s)$ of accepted and lost packet (34) are presented in analytical form. Also the mean dwell time (time in system) w (35) for an arbitrary packet via mean waiting times for a served packet w^{serv} and a lost packet w^{loss} is presented in analytical form.

The future goals are to obtain the same probability-time characteristics for different combination of initial assumptions:

- packets are served from the queue in theFCFS or LCFS (Last-Come-First-Served) order;
- packets to be dropped from the buffer are chosen successively starting from the head or from the end of the queue;
- the served packet (if it remains in the system) occupies the first free space in the buffer or immediately goes to the server for repeated service.

It is also of interest to combine renovation mechanism and hysteretic overload control policies [42–51] in order to construct more adequate mathematical models of real telecommunication systems (for example, SIP server [42,45,46,49,52,53] or RED-like AQM algorithms).

Acknowledgments. The publication has been prepared with the support of the "RUDN University Program 5-100" and has been funded by Russian Foundation for Basic Research (RFBR) according to the research project No. 18-07-00692 and No. 16-07-00556.

References

1. Floyd, S., Jacobson, V.: Random early detection gateways for congestion avoidance. IEEE/ACM Trans. Netw. **4**(1), 397–413 (1993)
2. Ramakrishnan, K., Floyd, S., Black, D.: The Addition of Explicit Congestion Notification (ECN) to IP. RFC 3168. Internet Engineering Task Force (2001). https://tools.ietf.org/html/rfc3168
3. Korolkova, A.V., Zaryadov, I.S.: The mathematical model of the traffic transfer process with a rate adjustable by RED. In: International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), Moscow, Russia, pp. 1046–1050. IEEE (2010)
4. Sharma, V., Purkayastha, P.: Performance analysis of TCP connections with RED control and exogenous traffic. Queueing Syst. **48**(3), 193–235 (2004)
5. Velieva, T.R., Korolkova, A.V., Kulyabov, D.S.: Designing installations for verification of the model of active queue management discipline RED in the GNS3. In: 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), pp. 570–577. IEEE Computer Society (2015)
6. Korolkova, A.V., Kulyabov, D.S., Sevastianov, L.A.: Combinatorial and operator approaches to RED modeling. Math. Model. Geom. **3**, 1–18 (2015)
7. Zaryadov, I., Korolkova, A., Kulyabov, D., Milovanova, T., Tsurlukov, V.: The survey on Markov-modulated arrival processes and their application to the analysis of active queue management algorithms. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds.) DCCN 2017. CCIS, vol. 700, pp. 417–430. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66836-9_35
8. Konovalov, M., Razumchik, R.: Queueing Systems with Renovation vs. Queues with RED. Supplementary Material (2017). https://arxiv.org/abs/1709.01477
9. Gelenbe, E.: Product-form queueing networks with negative and positive customers. J. Appl. Probab. **28**(3), 656–663 (1991)
10. Pechinkin, A., Razumchik, R.: Waiting characteristics of queueing system $Geo/Geo/1/\infty$ with negative claims and a bunker for superseded claims in discrete time. In: International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, ICUMT 2010, pp. 1051–1055 (2010). https://doi.org/10.1109/ICUMT.2010.5676508
11. Pechinkin, A.V., Razumchik, R.V.: A method for calculating stationary queue distribution in a queuing system with flows of ordinary and negative claims and a bunker for superseded claims. J. Commun. Technol. Electr. **57**(8), 882–891 (2012)
12. Pechinkin, A.V., Razumchik, R.V.: The stationary distribution of the waiting time in a queueing system with negative customers and a bunker for superseded customers in the case of the LAST-LIFO-LIFO discipline. J. Commun. Technol. Electr. **57**(12), 1331–1339 (2012)
13. Razumchik, R.V.: Analysis of finite capacity queue with negative customers and bunker for ousted customers using chebyshev and gegenbauer polynomials. Asia-Pacific J. Oper. Res. **31**(04), 1450029 (2014). https://doi.org/10.1142/S0217595914500298

14. Semenova, O.V.: Multithreshold control of the $BMAP/G/1$ queuing system with MAP flow of Markovian disasters. Autom. Remote Control **68**(1), 95–108 (2007)
15. Li, J., Zhang, L.: $M^X|M|c$ queue with catastrophes and state-dependent control at idle time. Front. Math. China **12**(6), 1427–1439 (2017)
16. Gudkova, I., et al.: Modeling and analyzing licensed shared access operation for 5G network as an inhomogeneous queue with catastrophes. In: International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, December 2016, 7765372, pp. 282–287 (2016)
17. Sudhesh, R., Savitha, P., Dharmaraja, S.: Transient analysis of a two heterogeneous servers queue with system disaster, server repair and customers impatience. TOP **25**(1), 179–205 (2017)
18. Li, J., Zhang, L.: $M^X/M/c$ queue with catastrophes and state-dependent control at idle time. Front. Math. China **12**(6), 1427–1439 (2017)
19. Suranga Sampath, M.I.G., Liu, J.: Transient analysis of an $M/M/1$ queue with reneging, catastrophes, server failures and repairs. Bull. Iran. Math. Soc. (2018). https://doi.org/10.1007/s41980-018-0037-6
20. Azadeh, A., Naghavi Lhoseiny, M.S., Salehi, V.: Optimum alternatives of tandem $G/G/K$ queues with disaster customers and retrial phenomenon: interactive voice response systems. Telecommun. Syst. **68**(3), 535–562 (2018)
21. Dudin, A., Klimenok, V., Vishnevsky, V.: Analysis of unreliable single server queueing system with hot back-up server. Commun. Comput. Inf. Sci. **499**, 149–161 (2015)
22. Krishnamoorthy, A., Pramod, P.K., Chakravarthy, S.R.: Queues with interruptions: a survey. TOP **22**(1), 290–320 (2014)
23. Vishnevsky, V.M., Kozyrev, D.V., Semenova, O.V.: Redundant queuing system with unreliable servers. In: International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, pp. 283–286. IEEE Xplore (2014)
24. Xu, B., Xu, X.: Equilibrium strategic behavior of customers in the $M/M/1$ queue with partial failures and repairs. Oper. Res. **18**(2), 273–292 (2018)
25. Nazarov, A., Sztrik, J., Kvach, A., Berczes, T.: Asymptotic analysis of finite-source $M/M/$ retrial queueing system with collisions and server subject to breakdowns and repairs. Ann. Oper. Res. 1–17 (2018). https://doi.org/10.1007/s10479-018-2894-z
26. Ometov, A., Kozyrev, D., Rykov, V., Andreev, S., Gaidamaka, Y., Koucheryavy, Y.: Reliability-centric analysis of offloaded computation in cooperative wearable applications. Wirel. Commun. Mob. Comput. **2017**, 9625687 (2017)
27. Rykov, V., Kozyrev, D., Zaripova, E.: Modeling and simulation of reliability function of a homogeneous hot double redundant repairable system. In: Proceedings - 31st European Conference on Modelling and Simulation, ECMS 2017, pp. 701–705 (2017)
28. Houankpo, H.G.K., Kozyrev, D.V.: Sensitivity analysis of steady state reliability characteristics of a repairable cold standby data transmission system to the shapes of lifetime and repair time distributions of its elements. In: CEUR Workshop Proceedings, vol. 1995, pp. 107–113 (2017)
29. Rykov, V.V., Kozyrev, D.V.: Analysis of renewable reliability systems by Markovization method. In: Rykov, V.V., Singpurwalla, N.D., Zubkov, A.M. (eds.) ACMPT 2017. LNCS, vol. 10684, pp. 210–220. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-71504-9_19
30. Rykov, V., Kozyrev, D.: On sensitivity of steady-state probabilities of a cold redundant system to the shapes of life and repair time distributions of its elements. In:

Pilz, J., Rasch, D., Melas, V., Moder, K. (eds.) IWS 2015. Springer Proceedings in Mathematics & Statistics, vol. 231, pp. 391–402. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76035-3_28
31. Kreinin, A.: Queueing systems with renovation. J. Appl. Math. Stochast. Anal. **10**(4), 431–443 (1997)
32. Kreinin, A.: Inhomogeneous random walks: applications in queueing and finance. In: CanQueue 2003. Fields Institute, Toronto (2003)
33. Bocharov, P.P., Zaryadov, I.S.: Probability distribution in queueing systems with renovation. Bull. Peoples' Friendsh. Univ. Russia. Ser. Math. Inf. Sci. Phys. **1–2**, 15–25 (2007)
34. Zaryadov, I.S., Pechinkin, A.V.: Stationary time characteristics of the $GI/M/n/\infty$ system with some variants of the generalized renovation discipline. Autom. Remote Control **70**(12), 2085–2097 (2009)
35. Zaryadov, I.S.: Queueing systems with general renovation. In: International Conference on Ultra Modern Telecommunications, ICUMT 2009, St.-Petersburg. IEEE (2009). https://doi.org/10.1109/ICUMT.2009.5345382
36. Zaryadov, I., Razumchik, R., Milovanova, T.: Stationary waiting time distribution in $G|M|n|r$ with random renovation policy. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds.) DCCN 2016. CCIS, vol. 678, pp. 349–360. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-51917-3_31
37. Bogdanova, E.V., Milovanova, T.A., Zaryadov, I.S.: The analysis of queueing system with general service distribution and renovation. Bull. Peoples' Friendsh. Univ. Russia. Ser. Math. Inf. Sci. Phys. **25**(1), 3–8 (2017)
38. Zaryadov, I.S., Bogdanova, E.V., Milovanova, T.A.: Probability-time characteristics of $M|G|1|1$ queueing system with renovation. In: CEUR Workshop Proceedings, vol. 1995, pp. 125–131 (2017)
39. Zaryadov, I.S., Scherbanskaya, A.A.: Time characteristics of queuing system with renovation and reservice. Bull. Peoples' Friendsh. Univ. Russia. Ser. Math. Inf. Sci. Phys. **2**, 61–66 (2014)
40. Matskevich, I.A.: Time-probabilistic characteristics of queueing system with general renovation and repeated service. In: Proceedings of the VI Conference on Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems, Moscow, Russia, pp. 37–39 (2016)
41. Zaryadov, I.S., Matskevich, I.A., Scherbanskaya, A.A.: The queueing system with general renovation and repeated service – time-probability characteristics. In: Proceedings of the Nineteenth International Scientific Conference on Distributed Computer and Communication Networks: Control, Computation, Communications (DCCN 2016), vol. 3, pp. 458–462 (2016)
42. Abaev, P., Pechinkin, A., Razumchik, R.: On analytical model for optimal SIP server hop-by-hop overload control. In: International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, 6459680, pp. 286–291 (2012)
43. Abaev, P., Gaidamaka, Y., Samouylov, K.E.: Modeling of hysteretic signaling load control in next generation networks. In: Andreev, S., Balandin, S., Koucheryavy, Y. (eds.) NEW2AN/ruSMART -2012. LNCS, vol. 7469, pp. 440–452. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-32686-8_41
44. Abaev, P., Pechinkin, A., Razumchik, R.: On mean return time in queueing system with constant service time and bi-level hysteric policy. In: Dudin, A., Klimenok, V., Tsarenkov, G., Dudin, S. (eds.) BWWQT 2013. CCIS, vol. 356, pp. 11–19. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-35980-4_2

45. Pechinkin, A.V., Razumchik, R.V.: Approach for analysis of finite $M_2|M_2|1|R$ with hysteric policy for SIP server hop-by-hop overload control. In: Proceedings - 27th European Conference on Modelling and Simulation, ECMS 2013, pp. 573–579 (2013)
46. Abaev, P., Razumchik, R.V.: Queuing model for SIP server hysteretic overload control with bursty traffic. In: Balandin, S., Andreev, S., Koucheryavy, Y. (eds.) NEW2AN/ruSMART -2013. LNCS, vol. 8121, pp. 383–396. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40316-3_34
47. Pechinkin, A.V., Razumchik, R.V.: Stationary characteristics of $M_2|G|1|r$ system with hysteretic policy for arrival rate control. J. Commun. Technol. Electron. **58**(12), 1282–1291 (2013)
48. Gaidamaka, Y., Pechinkin, A., Razumchik, R., Samouylov, K., Sopin, E.: Analysis of an $M|G|1|R$ queue with batch arrivals and two hysteretic overload control policies. Int. J. Appl. Math. Comput. Sci. **24**(3), 519–534 (2014)
49. Samouylov, K.E., Abaev, P.O., Gaidamaka, Y.V., Pechinkin, A.V., Razumchik, R.V.: Analytical modelling and simulation for performance evaluation of SIP server with hysteretic overload control. In: Proceedings - 28th European Conference on Modelling and Simulation, ECMS 2014, pp. 603–609 (2014)
50. Gaidamaka, Y., Pechinkin, A., Razumchik, R.: Time-related stationary characteristics in queueing system with constant service time under hysteretic policy. In: International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, January 2015, 7002158, pp. 534–540 (2015)
51. Abaev, P., Khachko, A., Beschastny, V.: Queuing model for SIP server hysteretic overload control with K-state MMPP bursty traffic. In: International Congress on Ultra Modern Telecommunications and Control Systems and Workshops 2015-January, 7002151, pp. 495–500 (2015)
52. Hilt, V., Widjaja, I., Labs, B.: Controlling overload in networks of SIP servers. In: Proceedings - International Conference on Network Protocols, ICNP, art. no. 4697027, pp. 83–93 (2008)
53. Abdelal, A., Matragi, W.: Signal-based overload control for SIP servers. In: 7th IEEE Consumer Communications and Networking Conference, CCNC 2010, art. no. 5421642 (2010). https://doi.org/10.1109/CCNC.2010.5421642

Using Predictive Monitoring Models in Cloud Computing Systems

Kristina Kucherova, Serg Mescheryakov(✉), and Dmitry Shchemelinin

Peter the Great St. Petersburg Polytechnic University,
Polytekhnicheskaya 29, 195251 St. Petersburg, Russia
kristina.mylife@gmail.com, serg-phd@mail.ru, dshchmel@gmail.com,
http://english.spbstu.ru/

Abstract. Predictive modeling is an important part of the monitoring process in cloud computing systems that helps to improve the service availability for the customers. This paper describes two industrial examples of predictive monitoring models for database disk space utilization and Java memory leaks. Practical recommendations are given to improve the forecast accuracy, which can also be used in the other similar cases. The results of this work are validated in the open source monitoring system and are implemented in three big International telecommunications companies.

Keywords: Predictive monitoring · Forecast accuracy
Cloud computing · Globally distributed infrastructure
High loaded system · Big data · Database · Disk space utilization
File autogrowth · Java memory leak · Internet applications
Service availability

1 Introduction

Modern world has reached enormous speed of IT systems growth. They develop exponentially and the number of servers and applications running on servers is dramatically high. In early 2000th IT companies usually had about 10 servers and one system administrator who was looking after them. Nowadays, nobody is surprised when you say that your company runs on 1000 servers. No wonder, these servers have to be maintained automatically since it is obviously impossible to do it manually. As a result, cloud service technologies are now offering Infrastructure as a Service (IaaS) and Platform as a Service (PaaS) allowing companies to focus on the business processes rather than on their hardware problems.

Businesses and markets know very well that competition between companies is higher than ever, and that clients and users are the key factor for commercial success. We cannot afford system being out of service because the clients are

K. Kucherova—The publication has been prepared with the partial support of Genesys [4] and RingCentral [7] Telecommunications Companies, USA.

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 341–352, 2018.
https://doi.org/10.1007/978-3-319-99447-5_29

constantly looking for new better applications to meet their growing needs. This is the reason why big IT companies target the 99.999% level of service availability (SA) in 24/7 mode. Monitoring systems are used as an automated tool to meet this high demand.

In this article, we describe several approaches to predictive monitoring and fault detection of the highly loaded production system. For data evaluation and validation, we used Zabbix monitoring system because it is open source and is now in top 5 most popular monitoring solutions in the world [1] (Fig. 1). The proposed approaches can be applied to any other monitoring tool having prediction functionality.

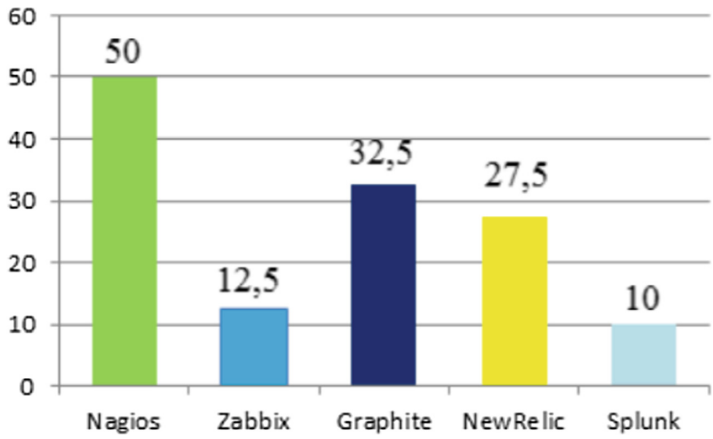


Fig. 1. Top 5 most popular monitoring solutions in the world

2 Disk Space Prediction for Databases

One of the important issues in the highly loaded system with high SA is to have enough free disk space for new data in databases, logs, etc. Lack of free disk space may lead to termination and stop processing of all incoming custom requests. The case of disk space prediction for databases (DBs) is described below. The example is shown for MS SQL Server database but the given approach can be applied to other SQL systems like Oracle, MySQL, PostgreSQL. There are usually several challenges with DB space prediction:

1. DB files are stored on several hardware disks. This means that even if you are running out of space on a particular disk, it does not mean that you have no space for data as the data will automatically be stored on another different disk available for the same file group. It is an important thing to remember for struggling with false alerting.

2. System DBs should also be taken into consideration because we need to monitor space for them too. If you have not enough space for system DBs, the whole system won't work.
3. Transaction log is the other important thing that should have enough space for growing. No space for logging means no new action in DBs.
4. DB files do not grow like usual files but stepwise. This means that each time a file reserves some space on a disk ahead its actual usage, and after that DB is gradually filled up with data. DB files are usually configured for a particular extent in growing. Such file growth algorithm makes monitoring a bit harder because DB file changes in an abrupt way and the operator will see a spike on the graph. Although the trend displays fast growing of the DB file in future, it is not going to happen in reality. Just in opposite, when the file has an additional space it is very unlikely to see autogrowth in the nearest future. This should be taken into account when setting up the forecasting parameters.

An example of stepwise change of DB file size at Distillery Company, USA [2] is shown in Fig. 2. Linear function is used for prediction. The next graph (see Fig. 3) shows that the forecast of DB size has dropped because the system predicts the same amount of space needed on a daily basis. This is a great way to bring the whole company to panic. It is a classical example of false alert that may lead to extra financial cost due to paid overtime for expensive specialists.

To improve the forecast accuracy, the following recommendations are proposed [3]. Zabbix look back period should cover at least 5–10 extent steps like it is shown in sample Fig. 4 at Genesys Company, USA [4]. Look back history for forecasting should be extended to have several (at least 5) autogrowth events and also to correlate with forecast horizon at the ratio of about 10:1 [5].

The result of this tuning is shown in Fig. 5. Now the trend looks more gradual and the forecast is more accurate. This approach allows preventing false alerts during DB file autogrowth and gives true information about disk space resources. Operations teams can plan maintenance beforehand and increase system fault tolerance. Consequently, planned maintenance helps to reduce operational expenses for the company, prevent service outages with extra payments for experts' overtime, and finally improve SA to the customers that is an extremely important metric of global IT services.

3 Prediction of Available Memory

Memory leak in Java based web applications running on Internet servers is worldwide known problem [6]. Memory leak is mostly observed on a high loaded system under heavy user load when new Java objects are generated dynamically faster than old unused objects are removed from memory by Garbage Collector (GC). Long term solution is to rewrite the application and fix the code that leads to memory leak. If the application is large and has legacy code of a 3rd party company, this could be challenging and expensive task for a company. Also, the

Fig. 2. Actual monitoring trend of DB file size at Distillery

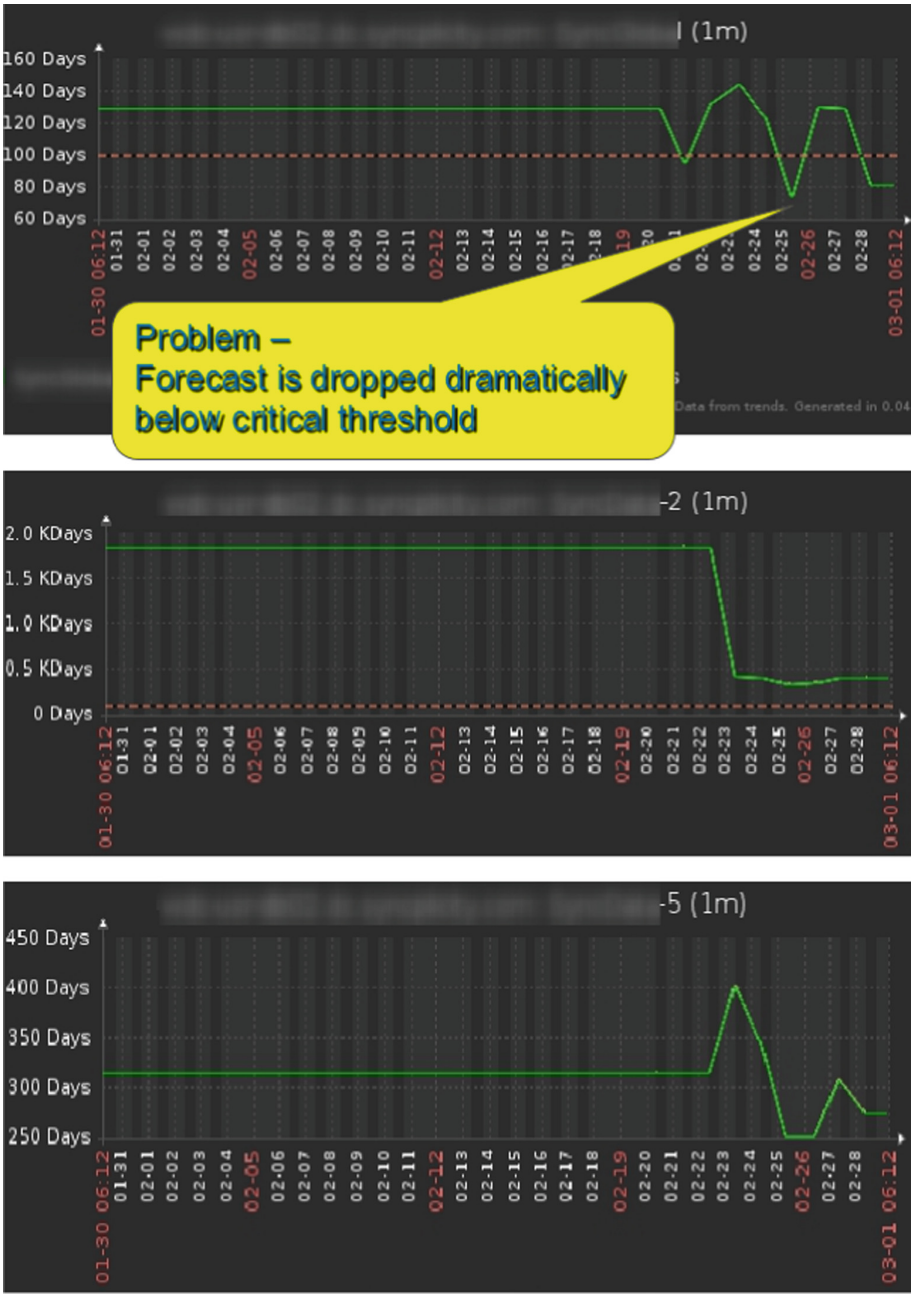


Fig. 3. The forecast of DB file at Distillery

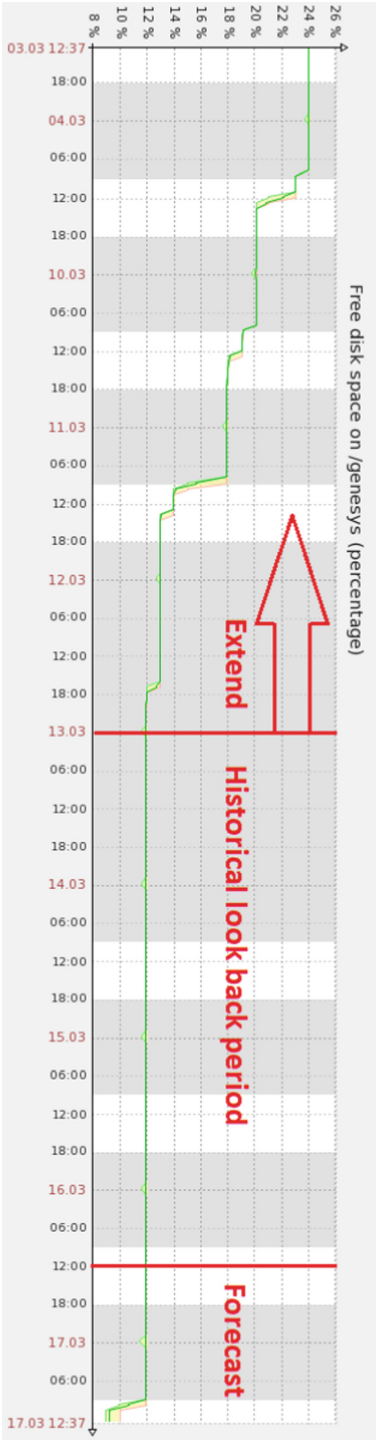


Fig. 4. Actual monitoring trend of free disk space at Genesys

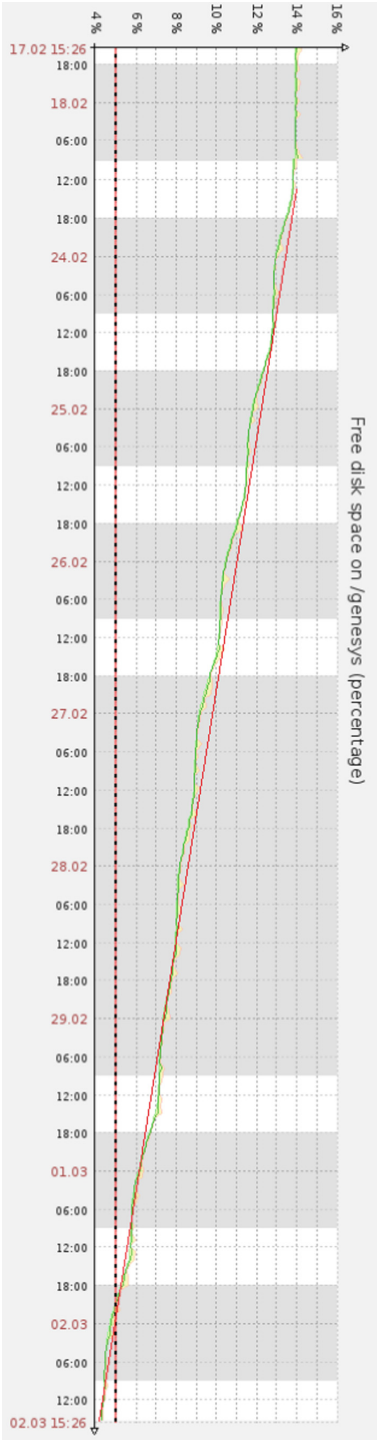


Fig. 5. The forecast of free disk space at Genesys

more business value for IT company is developing new advanced functionality for the clients rather than refactoring the old outdated code. In terms of customer support and operations, much more efficient approach is to monitor available memory, predict the time of critical threshold and restart the application automatically, safely, during low user workload or regular maintenance.

Here are the challenges with available memory prediction:

1. Monitoring system observes amount of available memory on the server with assumption that most of the memory is used by the application. But memory can be temporarily consumed by some internal server processes (log rotation, security scan, etc.) that may happen unexpectedly and cause "Out of memory" (OOM) error on the server. Such data produces much noise and makes prediction less reliable.
2. Web applications consume memory depending on the number of current users online. During peak time the total amount of available memory is reduced faster than the same metrics during non-working hours. This means that we need to monitor and predict the growing trend of peak values to deploy more memory beforehand and avoid OOM outages in peak time.
3. GC releases memory for unused objects with a preliminary specified frequency. Each time GC is initiated, available memory is rapidly increased for some time but then is slowly consumed back and it is hard to predict the next possible OOM outage. GC non-linear function makes the whole prediction process more difficult.
4. Monitoring of Java based web applications requires almost immediate action in case of insufficient memory or entire OOM detected. This means that either 24/7 operator should be on duty to manually restart the application in case of OOM alarm or better the service may be restarted automatically and safely to release memory.
5. Once application has been restarted, either manually or automatically, the graph with actual available memory metrics shows a spike, which breakes the whole previous historical statistics and the forcasting trend is no longer reliable, using neither linear nor any other prediction model (see Figs. 6 and 7 as the examples of actual available Java memory statistics, predictive monitoring and auto-restarting Java applications at RingCentral Company, USA [7]).

The best solution in such cases is to ignore the previous historical look back period and start forecasting from scratch using the most recent time point of restart. Once the safety auto-restart procedure is implemented, we may also automate the logic to reset the forecasting historical statistics to a corresponding shorter time interval and forecasting horizon in the monitoring system, thus preventing a monitoring operator from manual human errors.

The analysis [8] showed that the linear model of prediction is still more reliable independently of Java memory reset jumps if use reasonable configuration of prediction function and at least 5:1 rate between historical look back period and forecasting horizon.

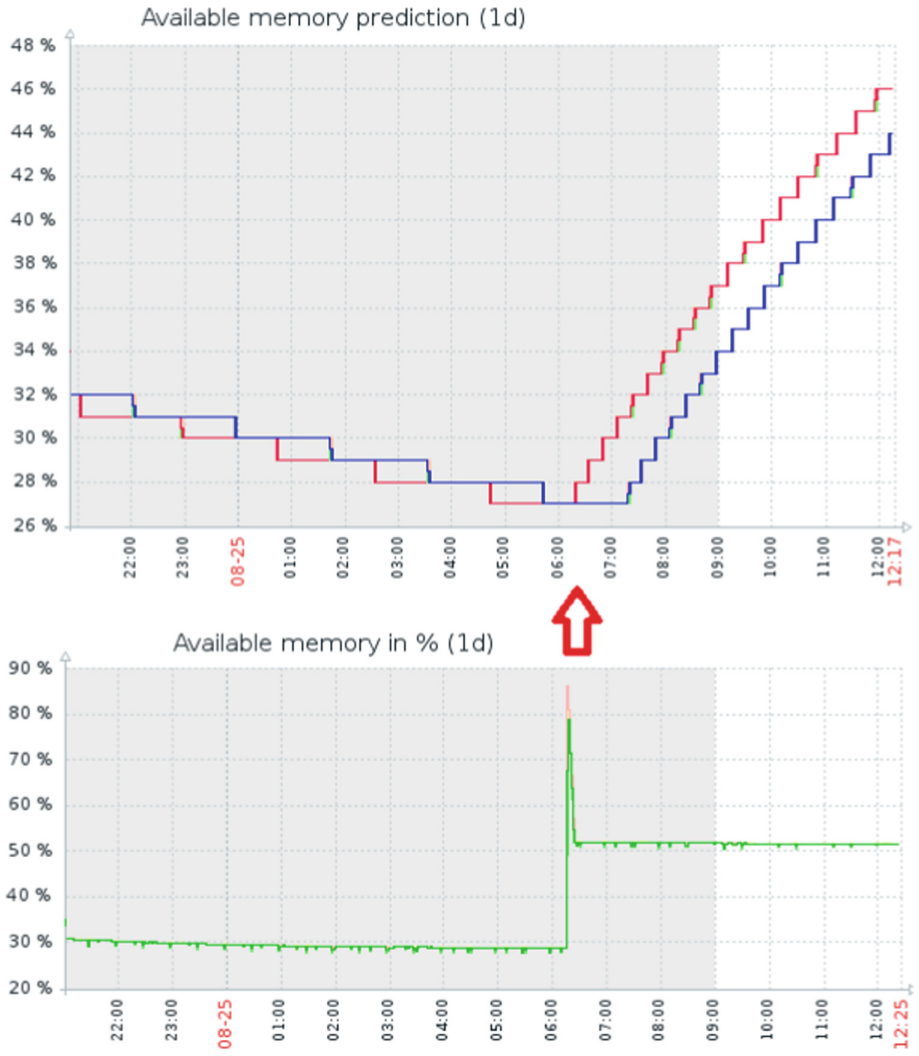


Fig. 6. Memory prediction (upper graph) and actual available memory for Java applications with auto-restart jump (lower graph) at RingCentral

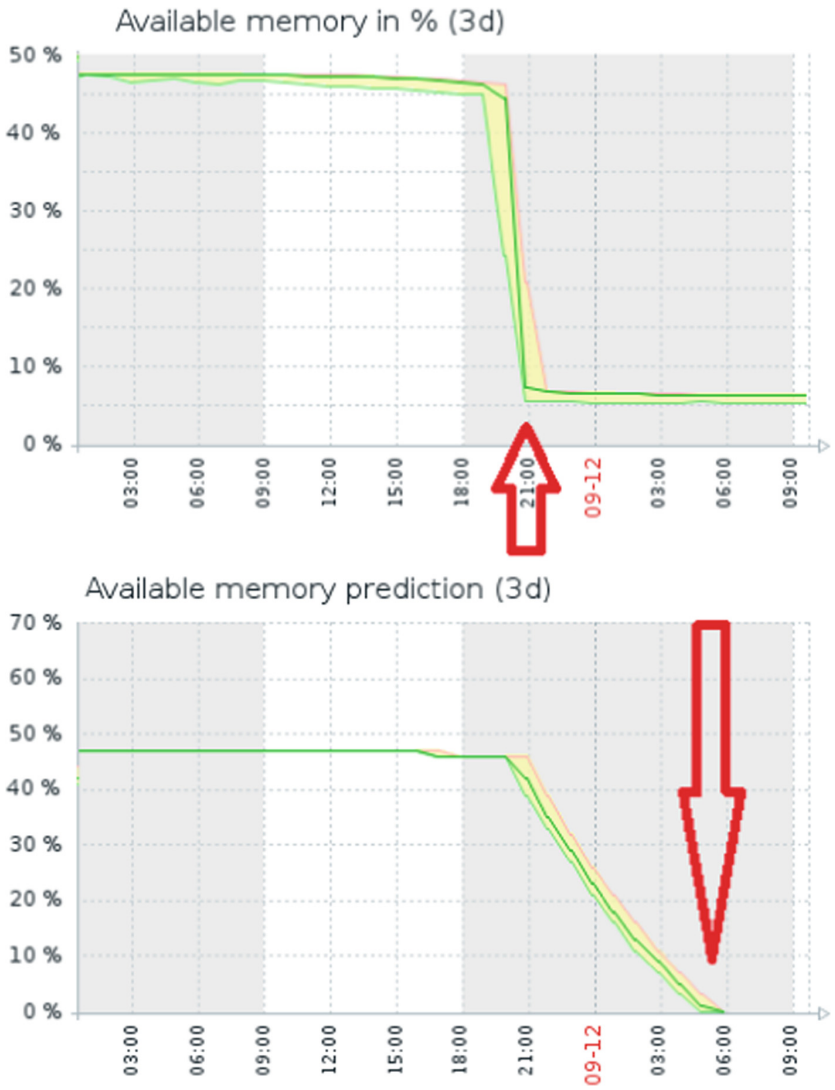


Fig. 7. Actual Java memory degradation drop (upper graph) and memory prediction (lower graph) at RingCentral

Another option is to apply triggers correlation and dependencies. For example, we can link two trigger events – "Available memory prediction is below critical threshold" and "Application has just been restarted". Dependency between these events means that the first trigger will not fire in case of the second one is activated. The idea is to trigger only one event, which is the root cause of the problem, and in this way reduce alarms noise.

One more solution of auto-adjusting prediction model after application restart is to monitor its process identifier (PID). Each time the application restarted its PID number is being changed. Therefore it is possible to auto-reset historical look back period when application PID has been changed for some reason. After that we may start prediction from the moment of restart, prevent false alarms, and thus make forecast more accurate.

The proposed solutions are implemented in the cloud infrastructure of big International IT companies – Genesys Telecommunications Laboratories [4] and RingCentral Inc. [7], USA, providing Internet telecommunications services in globally distributed regions – USA and Canada, West Europe and South East Asia. Monitoring and prediction of available memory with auto-restart feature allowed to avoid service outage due to memory leak on company servers and increase SA quality to worldwide level of 99.999%.

4 Conclusion

In this paper, we have described two approaches to predictive monitoring in cloud computing systems. All use cases are based on real monitoring statistics of the production cloud infrastructure at big International IT companies – Distillery [2], Genesys [4], and RingCentral [7]. The results are validated in Zabbix monitoring system that is now in top 5 of the most popular monitoring solutions in the world based on the analytic report [1].

Practical recommendations to predictive monitoring are described through various examples of monitoring metrics – DB space utilization, free disk space, available memory. It was found that linear model is good enough in many cases, even when the monitoring trend shows periodic jumps or one time spike. Different solutions of how to adjust the prediction model and its parameters, which allow to prevent false alarms and improve the forecast accuracy, are provided.

Predictive monitoring with application auto-restart implementation is a good way to increase customers SA up to the highest level of 99.999% for global IT services working in 24/7 mode. Prediction can also be successfully used for capacity planning and analytics [9].

One more challenging topic of prediction modeling is monitoring metrics with cyclic workload [5], such as calls count, the number of network connections, of cloud system processes, of customer web requests, etc. This is the subject of future research since the standard prediction models like linear or sinusoidal or polynomial do not give accurate results as needed.

Acknowledgments. The results of this research are based on real production statistical data collected by Zabbix monitoring system [3] at big International IT companies

– Distillery LLC [2], Genesys Telecommunications Laboratories [4], and RingCentral Inc. [7].

References

1. What We Learnt Talking to 60 Companies about Monitoring. Dataloop.IO. https://dataloopio.wordpress.com/2014/01/30/what-we-learnt-talking-to-60-companies-about-monitoring/
2. Distillery LLC. https://www.distillery.com/
3. Predictive trigger functions. In: Documentation Zabbix 3.0. https://www.zabbix.com/documentation/3.0/manual/config/triggers/prediction
4. Genesys Telecommunications Laboratories. http://www.genesys.com/
5. Kucherova, K., Mescheryakov, S., Shchemelinin, D.: Prediction experience and new model. In: The 7th Annual International Zabbix Conference, Riga, Latvia (2017). http://www.zabbix.com/conf2017_agenda.php
6. Ardulov, Y., Shchemelinin, D., Mescheryakov, S.: Monitoring and remediation of cloud services based on 4R approach. In: Proceedings of the 41st International IT Capacity and Performance Conference by Computer Measurement Group (CMG 2015), San Antonio, TX, USA (2015). http://www.cmg.org/publications/conference-proceedings/conference-proceedings2015/
7. RingCentral Inc. https://www.ringcentral.com/
8. Kucherova, K.N., Mescheryakov, S.V., Shchemelinin, D.A.: Prediction Modeling and Visualization in Cloud Monitoring System. In: Distributed Computer and Communication Networks: Control, Computation, Communications (DCCN-2016): Proceedings of the 19th International Scientific Conference, Moscow, Russian University of People's Friendship, vol. 1, pp. 222–230 (2016). https://www.dccn.ru/
9. 86 Percent of Predictive Analytics Users Report Tangible Gains to Their Bottom Line. https://www.forbes.com/sites/forbespr/2015/10/27/86-percent-of-predictive-analytics-users-report-tangible-gains-to-their-bottom-line

A Functional Approach to Estimation of the Parameters of Generalized Negative Binomial and Gamma Distributions

Andrey Gorshenin[1,2(✉)] and Victor Korolev[1,2,3]

[1] Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia
agorshenin@frccsc.ru
[2] Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Moscow, Russia
vkorolev@cs.msu.su
[3] Hangzhou Dianzi University, Hangzhou, China

Abstract. The generalized negative binomial distribution (GNB) is a new flexible family of discrete distributions that are mixed Poisson laws with the mixing generalized gamma (GG) distributions. This family of discrete distributions is very wide and embraces Poisson distributions, negative binomial distributions, Sichel distributions, Weibull–Poisson distributions and many other types of distributions supplying descriptive statistics with many flexible models. These distributions seem to be very promising for the statistical description of many real phenomena. GG distributions are widely applied in signal and image processing and other practical problems. The statistical estimation of the parameters of GNB and GG distributions is quite complicated. To find estimates, the methods of moments or maximum likelihood can be used as well as two-stage grid EM-algorithms. The paper presents a methodology based on the search for the best distribution using the minimization of ℓ^p-distances and L^p-metrics for GNB and GG distributions, respectively. This approach, first, allows to obtain parameter estimates without using grid methods and solving systems of nonlinear equations and, second, yields not point estimates as the methods of moments or maximum likelihood do, but the estimate for the density function. In other words, within this approach the set of decisions is not a Euclidean space, but a functional space.

Keywords: Generalized negative binomial distributions
Generalized gamma distributions · Estimation of parameters
Optimization problems · Mixed probability models

1 Introduction

The generalized negative binomial distribution (GNB) is a new flexible family of discrete distributions that are mixed Poisson laws with the mixing

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 353–364, 2018.
https://doi.org/10.1007/978-3-319-99447-5_30

generalized gamma (GG) distributions. The GNB distributions were introduced and studied in [1] under the name of GG mixed Poisson distributions. This family of discrete distributions is very wide and embraces Poisson distributions (as limit points corresponding to a degenerate mixing distribution), negative binomial (Polya) distributions including geometric distributions (corresponding to the gamma mixing distribution, see [2]), Sichel distributions (corresponding to the inverse gamma mixing distributions, see [3]), Weibull–Poisson distributions (corresponding to the Weibull mixing distributions, see [4]) and many other types supplying descriptive statistics with many flexible models. These distributions seem to be very promising for the statistical description of many real phenomena being very convenient and almost universal models. It is quite natural to expect that, having introduced one more free parameter into the pure negative binomial model, namely, the power parameter in the exponent of the original gamma mixing distribution, instead of the negative binomial model one might obtain a more flexible GNB model that provides even better fit with the statistical data. For example, GNB distributions can be successfully applied to modeling statistical regularities in duration of specific periods in data.

The GG distributions are proposed in order to have a flexible Bayesian model with a mixing (prior) distribution which is "responsible" for the description of statistical regularities of the manifestation of external stochastic factors. The class of GG distributions was first described as a unitary family in 1962 by Stacy [5]. The family of GG distributions contains practically all the most popular absolutely continuous distributions concentrated on the non-negative half-line including Weibull and gamma distributions.

GG distributions are widely applied in many practical problems. There are dozens of papers dealing with the application of GG distributions as models of regularities observed in practice. As an example, the following research areas involving models based on GG distributions can be mentioned:

- climatic and hydrological problems: drop size distributions [6], drought data [7], phenomena in warm clouds [8];
- synthetic-aperture radar (SAR) image processing and various applications: distribution for the real and imaginary parts of the complex SAR backscattered signal [9], flexible model for the SAR images with different land-cover typologies [10], statistical modeling of SAR images [11,12];
- astrophysical problems, for example, new galaxy luminosity functions [13];
- speech signal processing: parametric characterization of speech spectra [14], modelling speech samples [15], real-time implementations of algorithms [16].

Apparently, the popularity of GG distributions is due to that most of them can serve as adequate asymptotic approximations, since all the representatives of the class of GG distributions listed above appear as limit laws in various limit theorems of probability theory in rather simple limit schemes.

The problem of statistical estimation of the parameters of GNB and GG distributions (for example, the search for maximum likelihood (ML) estimates) is quite complicated. To find the estimates of the parameters, the method of moments or ML method [17] for the GG distribution as well as the two-stage

grid EM-algorithm for the GNB, can be used. It should be noted that the implementations of the methods of moments and ML method for GG distribution are difficult computational tasks, moreover, the efficiency depends on the sample size (ML method is better for large volumes).

The paper presents a methodology based on finding the best distribution using minimization of ℓ^1-, ℓ^2- and ℓ^∞-distances and L^1-, L^2- and L^∞-metrics for GNB and GG distributions, respectively. This approach, first, allows to obtain parameter estimates without using grid methods and solving systems of nonlinear equations and, second, yields not point estimates as the methods of moments or maximum likelihood do, but the estimate for the density function. In other words, within this approach the set of decisions is not a Euclidean space, but a functional space.

2 The GNB and GG Distributions

It will be assumed that all the random variables are defined on the same probability space $(\Omega, \mathfrak{F}, \mathbb{P})$.

A random variable having the gamma distribution with shape parameter $r > 0$ and scale parameter $\mu > 0$ will be denoted $G_{r,\mu}$,

$$\mathbb{P}(G_{r,\mu} < x) = \int_0^x g(z; r, \mu)dz, \quad \text{with} \quad g(x; r, \mu) = \frac{\mu^r}{\Gamma(r)} x^{r-1} e^{-\mu x}, \ x \geq 0, \quad (1)$$

where $\Gamma(r)$ is Euler's gamma-function, $\Gamma(r) = \int_0^\infty x^{r-1}e^{-x}dx$, $r > 0$.

A GG distribution is the absolutely continuous distribution defined by the density

$$g^*(x; r, \gamma, \mu) = \frac{|\gamma|\mu^r}{\Gamma(r)} x^{\gamma r-1} e^{-\mu x^\gamma}, \qquad x \geq 0, \quad (2)$$

with $\gamma \in \mathbb{R}$, $\mu > 0$, $r > 0$. The distribution function corresponding to the density $g^*(x; r, \gamma, \mu)$ can be denoted $F^*(x; r, \gamma, \mu)$.

The properties of GG distributions were described in [5,18]. A random variable with the density $g^*(x; r, \gamma, \mu)$ will be denoted $\overline{G}_{r,\gamma,\mu}$. It can be easily made sure that

$$\overline{G}_{r,\gamma,\mu} \stackrel{d}{=} G_{r,\mu}^{1/\gamma}, \quad (3)$$

and hence,

$$(\overline{G}_{r,\gamma,\mu})^\gamma \stackrel{d}{=} G_{r,\mu}. \quad (4)$$

The symbol $\stackrel{d}{=}$ in (3) and (4) denotes the coincidence of distributions.

A random variable $N_{r,p}$ is said to have the negative binomial (NB) distribution with parameters $r > 0$ ("shape") and $p \in (0, 1)$ ("success probability"), if

$$\mathbb{P}(N_{r,p} = k) = \frac{\Gamma(r + k)}{k!\Gamma(r)} \cdot p^r(1 - p)^k, \qquad k = 0, 1, 2, ... \quad (5)$$

Let $r > 0$, $\gamma \in \mathbb{R}$ and $\mu > 0$. We say that the random variable $N_{r,\gamma,\mu}$ has the GNB distribution, if

$$\mathbb{P}(N_{r,\gamma,\mu} = k) = \frac{1}{k!} \int_0^\infty e^{-z} z^k g^*(z; r, \gamma, \mu) dz, \quad k = 0, 1, 2..., \tag{6}$$

and $g^*(z; r, \gamma, \mu)$ is determined by formula (2).

3 A Functional Approach to Estimation of the Parameters of GNB Distributions

The problem of statistical estimation of the parameters of GNB distribution (for example, the search for maximum likelihood estimates) is extremely complicated. To find estimators, the two-stage grid EM-algorithm for the GNB distribution $F(x; r, \gamma, \mu)$ can be used. At the first stage, the main part of the support of the mixing distribution is determined. That is, a bounded interval is determined such that the probability of a GG distributed mixing random variable to fall into this interval is insignificantly less than one. This interval is covered by a finite grid containing (possibly, a very large number) $K \in \mathbb{N}$ of known nodes $\lambda_1, ..., \lambda_K$. The GNB distribution under study is approximated by the finite mixture of Poisson distributions:

$$F(x; r, \gamma, \mu)(x + 0) \approx \sum_{j=0}^{[x]} \frac{1}{j!} \sum_{i=1}^{K} p_i e^{-\lambda_i} \lambda_i^j, \quad x \in \mathbb{R}. \tag{7}$$

In the mixture on the right-hand side of (7), only the parameters $p_1, ..., p_K$ are unknown. At the second stage, it remains to use some standard method for fitting the GG distribution to the histogram-type data $(\mu_1, p_1), ..., (\mu_K, p_K)$, obtained at the first stage. For example, the parameters r, γ and μ can be determined as the point minimizing the corresponding chi-square statistic or some special least squares problem.

However, with a fixed grid, the two-stage method yields only approximate estimates of the parameters of GG distributions. Moreover, the accuracy of the approximation depends on the choice of the grid. The estimates can be consistent in the traditional sense only if the grid mesh becomes infinitely small as the sample size infinitely increases in an appropriate way. Moreover, the conditions unifying the rate of decrease of the grid mesh with the rate of increase of the sample size that provide the statistical consistency of the estimators are very cumbersome and practically unverifiable.

In this section we present an alternative methodology based on finding the best GNB distribution using minimization of ℓ^1-, ℓ^2- and ℓ^∞-distances (they correspond to the spaces of sequences whose series are absolutely convergent, the space of square-summable sequences and the space of bounded sequences, respectively). Namely, the histogram of the initial data should be obtained. The integer rule is used as bining algorithm (due to that the observations in the

sample are integer), so bins are created for each value. Let N_b be the number of histogram bins (with a uniform width that equals 1), $\mathbf{h}$ be the vector of bar heights ($h_i \in [0,1]$ for all $i = 1, \ldots, N_b$). The value of each component h_i is equal to the ratio of a number of observations in the bin to a total number of observations, the sum of the bar areas is 1. So, the bars of empirical distribution can be approximated by ones of GNB. For finding estimations of unknown parameters of generalized negative binomial distributions the following optimization problems should be solved (the density $g^*(x; r, \gamma, \mu)$ is determined by (2) and the probability $\mathbb{P}(N_{r,\gamma,\mu} = k)$ is determined by (6)).

– If the target function is based on ℓ^1-distance:

$$(r^*, \gamma^*, \mu^*) = \arg\min_{r,\gamma,\mu} \sum_{k=1}^{N_b} \left| \frac{1}{k!} \int_0^\infty e^{-z} z^k g^*(z; r, \gamma, \mu) dz - h_k \right|. \tag{8}$$

– If the target function is based on ℓ^2-distance:

$$(r^*, \gamma^*, \mu^*) = \arg\min_{r,\gamma,\mu} \sqrt{\sum_{k=1}^{N_b} \left(\frac{1}{k!} \int_0^\infty e^{-z} z^k g^*(z; r, \gamma, \mu) dz - h_k \right)^2}. \tag{9}$$

– If the target function is based on ℓ^∞-distance:

$$(r^*, \gamma^*, \mu^*) = \arg\min_{r,\gamma,\mu} \max_{k=\overline{1,N_b}} \left| \frac{1}{k!} \int_0^\infty e^{-z} z^k g^*(z; r, \gamma, \mu) dz - h_k \right|. \tag{10}$$

Formulas (8)–(10) allow to obtain parameter estimates without using grid methods. It should be noted that this methodology can also be used for the classical negative binomial distribution (5) (the ratio (5) should be used in formulas (8)–(10) instead of (6)).

A special `MATLAB` program is implemented for finding GNB approximations and plotting figures. The numerical optimization is based on the simplex search method [19]. The functions for estimating the values of all three unknown parameters of the GNB distribution or two parameters provided the shape parameter r estimate based on NB distribution is given are created.

The histogram and approximating graphs of the NB and GNB distributions as well as errors in the corresponding metrics are plotted. The examples of results are shown on Figs. 1, 2 and 3. They demonstrate a high quality of the approximation of the histogram of the initial data by each type of distributions. Table 1 represents approximation errors, the parameters are estimated by each of the metrics. Obviously, the results for GNB distributions are better (see the bold marked items).

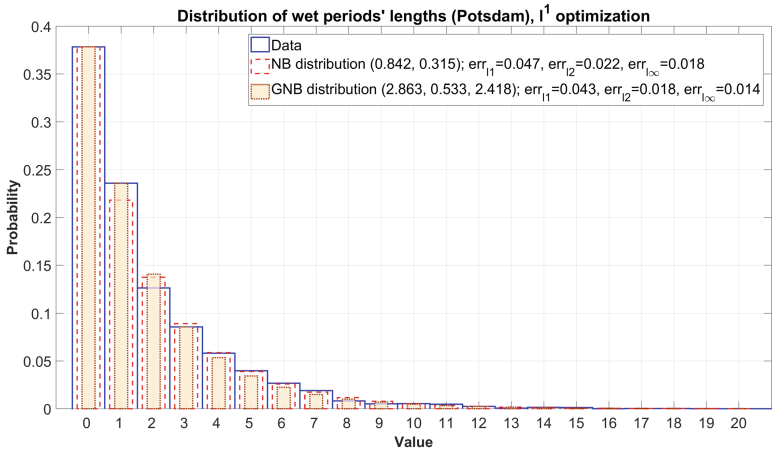


Fig. 1. Approximation of the initial data distribution by optimization of ℓ^1-distance.

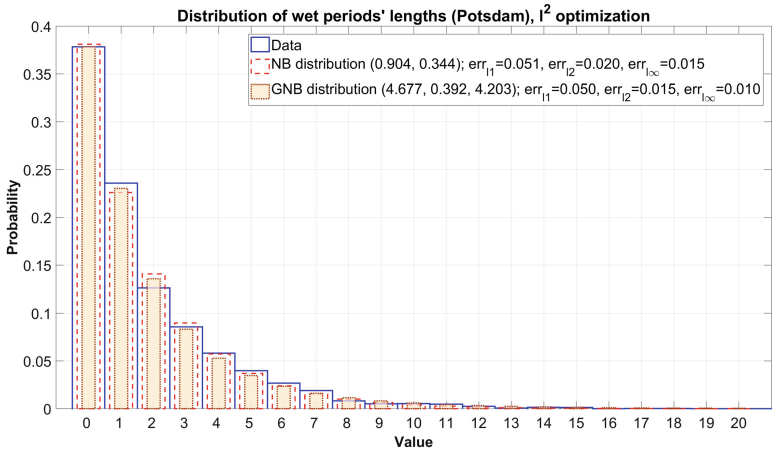


Fig. 2. Approximation of the initial data distribution by optimization of ℓ^2-distance.

Table 1. Approximation errors for negative binomial and generalized negative binomial distributions, test sample.

Distribution	Error (ℓ^1)	(ℓ^2)	Error (ℓ^∞)
NB (ℓ^1-optimization)	0,047	0,022	0,018
GNB (ℓ^1-optimization)	**0,043**	0,018	0,014
NB (ℓ^2-optimization)	0,051	0,0195	0,015
GNB (ℓ^2-optimization)	0,05	**0,015**	0,0097
NB (ℓ^∞-optimization)	0,061	0,022	0,012
GNB (ℓ^∞-optimization)	0,061	0,017	**0,007**

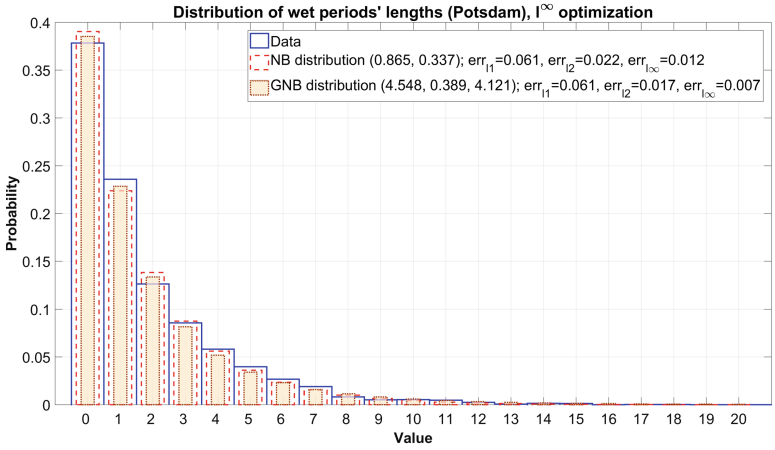


Fig. 3. Approximation of the initial data distribution by optimization of ℓ^∞-distance.

4 Recurrence Formulas for GNB Distributions

Using formulas (2) and (6) we can obtain the following results:

$$\mathbb{P}(N_{r,\gamma,\mu}=k)=\frac{|\gamma|\mu^r}{\Gamma(r)k!}\int\limits_0^\infty e^{-z-\mu z^\gamma}\,z^{\gamma r+k-1}\,dz=\frac{|\gamma|\mu^r}{\Gamma(r)k!}\int\limits_0^\infty e^{-z-\mu z^\gamma}\,d\frac{z^{\gamma r+k}}{\gamma r+k}$$

$$=\frac{|\gamma|\mu^r}{\Gamma(r)k!}\times\left[\frac{z^{\gamma r+k}}{\gamma r+k}\,e^{-z-\mu z^\gamma}\Big|_0^\infty-\int\limits_0^\infty e^{-z-\mu z^\gamma}\,\frac{z^{\gamma r+k}}{\gamma r+k}(-1-\mu\gamma z^{\gamma-1})\,dz\right]$$

$$=\frac{|\gamma|\mu^r}{\Gamma(r)k!}\times\left[\frac{1}{\gamma r+k}\int\limits_0^\infty e^{-z-\mu z^\gamma}\,z^{\gamma r+k}\,dz+\frac{\mu\gamma}{\gamma r+k}\int\limits_0^\infty e^{-z-\mu z^\gamma}\,z^{\gamma r+\gamma+k-1}\,dz\right]$$

$$=\frac{k+1}{\gamma r+k}\,\mathbb{P}(N_{r,\gamma,\mu}=k+1)+\frac{\gamma^2\mu^{r+1}}{(\gamma r+k)\Gamma(r)k!}\int\limits_0^\infty e^{-z-\mu z^\gamma}\,z^{\gamma r+k-1+\gamma}\,dz$$

$$=\frac{k+1}{\gamma r+k}\,\mathbb{P}(N_{r,\gamma,\mu}=k+1)+\frac{|\gamma|\mu}{\gamma r+k}\mathbb{P}(N_{r+1,\,\gamma,\mu}=k).$$

So, the recurrence formulas for GNB distributions can be represented as follows:

$$(\gamma r+k)\,\mathbb{P}(N_{r,\gamma,\mu}=k)=(k+1)\,\mathbb{P}(N_{r,\gamma,\mu}=k+1)+|\gamma|\mu\,\mathbb{P}(N_{r+1,\,\gamma,\mu}=k),$$

or

$$\mathbb{P}(N_{r,\gamma,\mu}=k+1)=\frac{\gamma r+k}{k+1}\,\mathbb{P}(N_{r,\gamma,\mu}=k)-\frac{|\gamma|\mu}{k+1}\,\mathbb{P}(N_{r+1,\,\gamma,\mu}=k).\qquad(11)$$

Unfortunately, the representation (11) does not significantly simplify the computational process, since in addition to the value $\mathbb{P}(N_{r,\gamma,\mu} = k)$ a value $\mathbb{P}(N_{r+1,\gamma,\mu} = k)$ should be known.

5 A Functional Approach to the Estimation of the Parameters of GG Distributions

In this section we present a methodology based on the search for the best GG distribution using minimization of L^1-, L^2- and L^∞-metrics (they correspond to the spaces of functions for which the p^{th} power of the absolute value is Lebesgue integrable, where functions that agree almost everywhere are identified).

The histogram of the initial data should be obtained. The Freedman–Diaconis rule [20] is used as bining algorithm due to its suitableness for data with heavy-tailed distributions. It uses a bin width of

$$2\,\frac{x_{0.75} - x_{0.25}}{\sqrt[3]{n}}, \tag{12}$$

where $x_{0.25}$, $x_{0.75}$ are 0.25- and 0.75-quantiles, numerator of fraction (12) represents an interquartile range and n is a sample size.

Let N_b be the number of histogram bins, $\mathbf{h}$ be the vector of bar heights ($h_i \in [0,1]$ for all $i = 1,\ldots,N_b$). The value of each component h_i is equal to the ratio of the number of observations in the bin and the total number of observations, the sum of the bar areas is 1. Let $\mathbf{b}$ be the vector of bin edges. The bars of empirical distribution should be approximated by GG distribution.

To find the estimates of unknown parameters of GG distributions the following optimization problems should be solved (the density $g^*(x; r, \gamma, \mu)$ is determined by (2)).

– If the target function is based on L^1-metric:

$$(r^*, \gamma^*, \mu^*) = \arg\min_{r,\gamma,\mu} \sum_{k=1}^{N_b-1} \int_{b_k}^{b_{k+1}} |g^*(z; r, \gamma, \mu) - h_k|\, dz. \tag{13}$$

– If the target function is based on L^2-metric:

$$(r^*, \gamma^*, \mu^*) = \arg\min_{r,\gamma,\mu} \sqrt{\sum_{k=1}^{N_b-1} \int_{b_k}^{b_{k+1}} \left(g^*(z; r, \gamma, \mu) - h_k\right)^2 dz}. \tag{14}$$

– If the target function is based on L^∞-metric:

$$(r^*, \gamma^*, \mu^*) = \arg\min_{r,\gamma,\mu} \max_{k\in[1,N_b-1]} \int_{b_k}^{b_{k+1}} |g^*(z; r, \gamma, \mu) - h_k|\, dz. \tag{15}$$

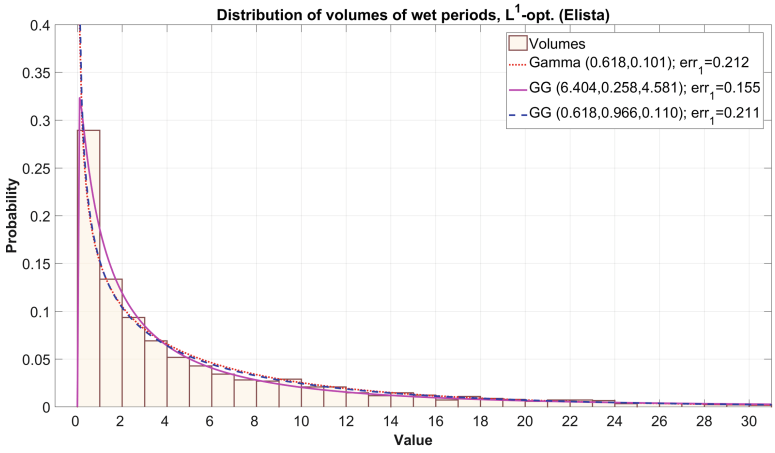


Fig. 4. Approximation of the initial data distribution by optimization of L^1-metric.

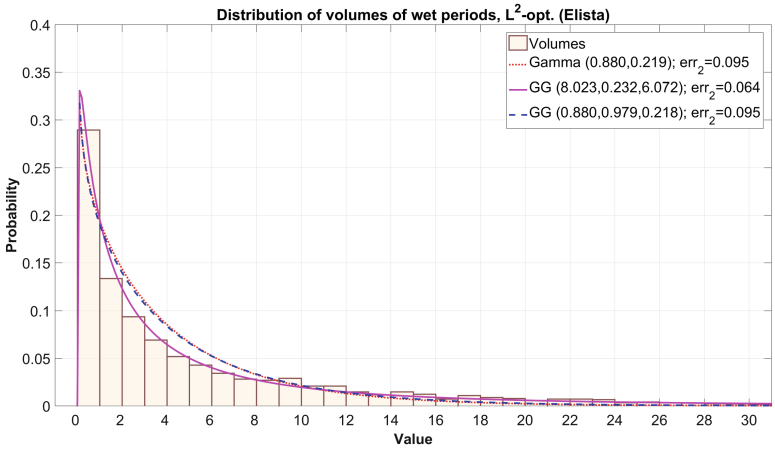


Fig. 5. Approximation of the initial data distribution by optimization of L^2-metric.

Formulas (13)–(15) allow to obtain parameter estimates without using grid methods. It should be noted that this methodology can also be used for the classical gamma distribution (1).

A special `MATLAB` program is implemented for finding GG approximations and plotting figures. The numerical optimization is based on the simplex search method [19]. The functions for estimating values of all three unknown parameters of GG distribution or two parameters provided the shape parameter r is given based on the gamma distribution model are created. The histogram and approximating probability density functions of gamma and GG distributions as well as errors in the corresponding metrics are plotted. The examples of results are shown on Figs. 4, 5 and 6.

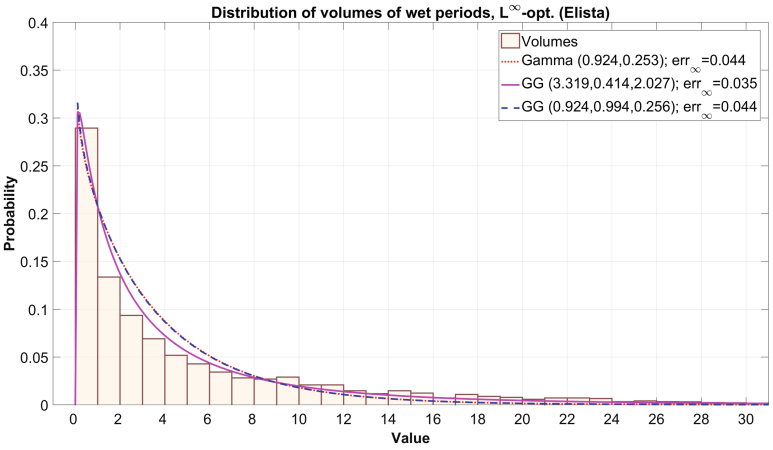


Fig. 6. Approximation of the initial data distribution by optimization of L^∞-metric.

They demonstrate a high quality of the approximation of the histogram of the initial data by each type of distributions. Table 2 represents approximation errors, the parameters are estimated by each of the metrics. Obviously, the results for GG distributions are better (see the bold marked items).

Table 2. Approximation errors for gamma and generalized gamma distributions, test sample.

Distribution	Error (L^1)	Error (L^2)	Error (L^∞)
Gamma	0,212	0,095	0,044
GG	**0,155**	**0,064**	**0,035**
GG, fixed r	0,211	0,095	0,044

6 Conclusion

The classical negative binomial distribution was successfully used as a model for the number of subsequent wet days in precipitation problems for the data registered in climatically different points (see, for example, [21–23]). It was demonstrated that the fluctuations of the data with very high confidence fit the negative binomial distribution. Obviously, a more flexible GNB model could provide even better fit with the statistical data. Herewith the GG distribution can be effectively used to model aggregated data (for example, volumes accumulated over a period) and can be useful for statistical testing of hypotheses about their extremality.

Moreover, such types of mixed probability models are quite adequate for information systems (for example, in insurance [24,25], financial mathematics [26,27], physics [28–30], data flows [31] and many other fields). The developed functional methods for the estimation of the unknown distribution parameters can be implemented as numerical procedures in the research support system for stochastic data processing [32,33] to analyze events in various information flows.

Acknowledgments. The research is partially supported by the Russian Foundation for Basic Research (project 17-07-00851) and the RF Presidential scholarship program (No. 538.2018.5).

References

1. Korolev, V.Yu., Zeifman, A.I.: GG mixed Poisson distributions as mixed geometric laws and related limit theorems. arXiv:1703.07276v2 [math.PR], 11 December 2017
2. Greenwood, M., Yule, G.U.: An inquiry into the nature of frequency-distributions of multiple happenings, etc. J. R. Stat. Soc. **83**, 255–279 (1920)
3. Sichel, H.S.: On a family of discrete distributions particularly suited to represent long tailed frequency data. In: Proceedings of the 3rd Symposium on Mathematical Statistics, pp. 51–97. CSIR, Pretoria (1971)
4. Korolev, V.Yu., Korchagin, A.Yu., Zeifman, A.I.: Poisson theorem for the scheme of Bernoulli trials with random probability of success and a discrete analog of the Weibull distribution. Informatika i Ee Primeneniya **10**(4), 11–20 (2016)
5. Stacy, E.W.: A generalization of the gamma distribution. Ann. Math. Stat. **33**, 1187–1192 (1962)
6. der Maur, A.N.F.: Statistical tools for drop size distributions: moments and generalized gamma. J. Atmos. Sci. **58**(4), 407–418 (2001)
7. Nadarajah, S., Gupta, A.K.: Statistical tools for drop size distributions: moments and generalized gamma. Math. Comput. Simul. **74**(1), 1–7 (2007)
8. Xie, X., Liu, X.: Analytical three-moment autoconversion parameterization based on generalized gamma distribution. J. Geophys. Res.-Atmos. **114**, D17201 (2009)
9. Li, H.-C., Hong, W., Wu, Y.-R., Fan, P.-Z.: An efficient and flexible statistical model based on generalized gamma distribution for amplitude SAR images. IEEE Trans. Geosci. Remote Sens. **48**(6), 2711–2722 (2010)
10. Li, H.-C., Hong, W., Wu, Y.-R., Fan, P.-Z.: On the empirical-statistical modeling of SAR images with generalized gamma distribution. IEEE J. Sel. Top. Sig. Process. **5**(3), 386–397 (2011)
11. Qin, X., Zou, H., Zhou, S., Ji, K.: Region-based classification of SAR images using Kullback-Leibler distance between generalized gamma distributions. IEEE Geosci. Remote Sens. Lett. **12**(8), 1655–1659 (2015)
12. Sportouche, H., Nicolas, J.-M., Tupin, F.: Mimic capacity of fisher and generalized gamma distributions for high-resolution SAR image statistical modeling. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **10**(12), 5695–5711 (2017)
13. Zaninetti, L.: The luminosity function of galaxies as modeled by the generalized gamma distribution. Acta Phys. Polonica B **41**(4), 729–751 (2010)
14. Shin, J., Chang, J., Kim, N.: Statistical modeling of speech signals based on generalized gamma distribution. IEEE Sig. Process. Lett. **12**(3), 258–261 (2005)

15. Almpanidis, G., Kotropoulos, C.: Phonemic segmentation using the generalised Gamma distribution and small sample Bayesian information criterion. Speech Commun. **50**(1), 38–55 (2008)
16. Song, K.-S.: Globally convergent algorithms for estimating generalized gamma distributions in fast signal and image processing. IEEE Trans. Image Process. **17**(8), 1233–1250 (2008)
17. Huang, P.-H., Hwang, T.Y.: New moment estimation of parameters of the generalized gamma distribution using it's characterization. Taiwanese J. Math. **10**(4), 1083–1093 (2006)
18. Zaks, L.M., Korolev, V.Yu.: Generalized variance gamma distributions as limit laws for random sums. Informatika i Ee Primeneniya **7**(1), 105–115 (2013)
19. Lagarias, J.C., Reeds, J.A., Wright, M.H., Wright, P.E.: Convergence properties of the nelder-mead simplex method in low dimensions. SIAM J. Optim. **9**(1), 112–147 (1998)
20. Freedman, D., Diaconis, P.Z.: On the histogram as a density estimator: L_2 theory. Zeitschrift Fur Wahrscheinlichkeitstheorie und Verwandte Gebiete. **57**(4), 453–476 (1981)
21. Korolev, V.Yu., Gorshenin, A.K., Gulev, S.K., Belyaev, K.P., Grusho, A.A.: Statistical analysis of precipitation events. In: AIP Conference Proceedings, vol. 1863, p. 090011 (2017)
22. Korolev, V.Yu., Gorshenin, A.K.: The probability distribution of extreme precipitation. Dokl. Earth Sci. **477**(2), 1461–1466 (2017)
23. Gorshenin, A.K.: Pattern-based analysis of probabilistic and statistical characteristics of precipitations. Informatika i Ee Primeneniya **11**(4), 38–46 (2017)
24. Grandell, J.: Mixed Poisson Processes. Chapman and Hall, London (1997)
25. Bening, V.E., Korolev, V.Yu.: Generalized Poisson Models and Their Applications in Insurance and Finance. VSP, Utrecht (2002)
26. Gorshenin, A.K., et al.: Modelling stock order flows with non-homogeneous intensities from high-frequency data. In: AIP Conference Proceedings, vol. 1558, pp. 2394–2397 (2013)
27. Korolev, V.Yu., Chertok, A.V., Korchagin, A.Y., Zeifman, A.I.: Modeling high-frequency order flow imbalance by functional limit theorems for two-sided risk processes. Appl. Math. Comput. **253**, 224–241 (2015)
28. Korolev, V.Yu., Skvortsova, N.N. (eds.): Stochastic Models of Structural Plasma Turbulence. VSP, Utrecht (2006)
29. Batanov, G.M., Gorshenin, A.K., Korolev, V.Yu., Malakhov, D.V., Skvortsova, N.N.: The evolution of probability characteristics of low-frequency plasma turbulence. Math. Models Comput. Simul. **4**(1), 10–25 (2012)
30. Gorshenin, A.K., Korolev, V.Yu., Skvortsova, N.N., Malakhov, D.V.: On nonparametric methodology of the plasma turbulence research. In: AIP Conference Proceedings, vol. 1558, pp. 2377–2380 (2013)
31. Gorshenin, A., Korolev, V.: Modelling of statistical fluctuations of information flows by mixtures of gamma distributions. In: Proceedings of 27th European Conference on Modelling and Simulation, pp. 569–572. Digitaldruck Pirrot GmbHP, Dudweiler (2013)
32. Gorshenin, A., Kuzmin, V.: On an interface of the online system for a stochastic analysis of the varied information flows. In: AIP Conference Proceedings, vol. 1738, p. 220009 (2016)
33. Gorshenin, A.K., Kuzmin, V.Yu.: Research support system for stochastic data processing. Pattern Recogn. Image Anal. **27**(3), 518–524 (2017)

Reliability of a Discrete-Time System with Investment

Ekaterina Bulinskaya(✉) and Andrey Kolesnik

Lomonosov Moscow State University, Leninskie Gory 1, 119234 Moscow, Russia
ebulinsk@yandex.ru, kolesnik.and1997@gmail.com

Abstract. We consider a discrete-time model describing the capital of an input-output system of mixed type. Such models can arise in various applications of probability theory, e.g. queuing and reliability theory, telecommunication, inventories and dams, insurance and many others. The inflow consists of constants, whereas outflow is a sequence of independent identically distributed random variables with a known distribution function. It is also assumed that at the beginning of each period the company under consideration invests a certain quota of the available capital in a non-risky asset for a fixed number of periods. The objective function is the system reliability. Thus, we establish the formula for company ruin probability, in other words, the probability that sooner or later its capital becomes negative. Some numerical results are also provided.

Keywords: Reliability · Investment · Discrete-time model

1 Introduction

It is well known that many systems studied by methods of probability theory have input-output character (see, e.g., [1]). In order to perform a system analysis one needs an appropriate mathematical model. There exist a lot of models describing the system more or less precisely. Moreover, the same model frequently arises in different application areas, such as insurance and finance, queuing and reliability, telecommunication and information, inventory and dams, population dynamics and many others. So, the methods used in one domain can turn out fruitful in others.

It is interesting that in some areas investigating risks (e.g., inventory, dams) the researchers proposed the discrete-time models from the beginning (see, e.g., [2]). On the other hand, in insurance, finance, queuing, telecommunication and others dominated continuous-time models (see, e.g., [3,4]). However it turned out that sometimes the discrete-time models are more appropriate in these domains as well, describing the situation more precisely. Thus, dividends payments or reinsurance treaties are usually discussed at the end of financial year. A review of discrete-time risk models (until 2009) is provided by [5]. Some further results are given in [6–20].

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 365–376, 2018.
https://doi.org/10.1007/978-3-319-99447-5_31

2 Model Description

We consider a discrete-time model of a company capital dynamics. For certainty, we proceed in terms of insurance company. During the ith period (year, month or week) the company gets a fixed premium amount h and pays a random indemnity X_i, $i = 1, 2, \ldots$. It is supposed that $\{X_i\}$ form a sequence of independent identically distributed random variables with a known distribution function $F(x)$. Let the initial capital S_0 be fixed. It is possible to place a quota δ of this amount in a bank for s periods, the interest rate being β per period. The same procedure is repeated each period. Thus,

$$S_1 = (1-\delta)S_0 + h - X_1.$$

For simplicity sake, we put $s = 1$ and $u = 1 + \beta$ getting the following recurrent formula

$$S_n = (1-\delta)S_{n-1} + h - X_n + u\delta S_{n-2}. \tag{1}$$

Now we are able to formulate

Theorem 1. *The explicit form of S_n is as follows*

$$S_n = S_0 \cdot \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} a_{i,n}(\delta u)^i (1-\delta)^{n-2i} + h \cdot \sum_{k=0}^{\lfloor \frac{n-1}{2} \rfloor} \sum_{j=0}^{n-2k-1} a_{k,j+2k}(\delta u)^k (1-\delta)^j \tag{2}$$
$$- \sum_{k=0}^{\lfloor \frac{n-1}{2} \rfloor} \sum_{j=0}^{n-2k-1} a_{k,j+2k}(\delta u)^k (1-\delta)^j \cdot X_{n-2k-j},$$

where the coefficients $a_{i,j}$ provided by the Table 1.

Proof. Clearly, for $n = 2, \ldots, 6$ formula (1) leads to explicit expressions of the form

$$S_2 = (1-\delta)^2 S_0 + (1-\delta)h - (1-\delta)X_1 + h - X_2 + \delta u S_0,$$
$$S_3 = (1-\delta)^3 S_0 + 2(1-\delta)\delta u S_0 + h(1-\delta)^2 + h(1-\delta) + h + \delta u h$$
$$- (1-\delta)^2 X_1 - (1-\delta)X_2 - X_3 - \delta u X_1,$$
$$S_4 = (1-\delta)^4 S_0 + 3(1-\delta)^2(\delta u)S_0 + (\delta u)^2 S_0 + h(1-\delta)^3 + h(1-\delta)^2 + h(1-\delta) + h$$
$$+ 2h(\delta u)(1-\delta) + (\delta u)h - (1-\delta)^3 X_1 - (1-\delta)^2 X_2 - (1-\delta)X_3 - X_4$$
$$- 2X_1(\delta u)(1-\delta) - (\delta u)X_2,$$
$$S_5 = (1-\delta)^5 S_0 + 4(1-\delta)^3(\delta u)S_0 + 3(1-\delta)(\delta u)^2 S_0 + h(1-\delta)^4 + h(1-\delta)^3 + h(1-\delta)^2$$
$$+ h(1-\delta) + h + 3h(\delta u)(1-\delta)^2 + 2h(\delta u)(1-\delta) + (\delta u)h + (\delta u)^2 h$$
$$- (1-\delta)^4 X_1 - (1-\delta)^3 X_2 - (1-\delta)^2 X_3 - (1-\delta)X_4 - X_5$$
$$- 3(\delta u)(1-\delta)^2 X_1 - 2(\delta u)(1-\delta)X_2 - (\delta u)X_3 - (\delta u)^2 X_1,$$
$$S_6 = (1-\delta)^6 S_0 + 5(1-\delta)^4(\delta u)S_0 + 6(1-\delta)^2(\delta u)^2 S_0 + (\delta u)^3 S_0$$

$$+h(1-\delta)^5+h(1-\delta)^4+h(1-\delta)^3+h(1-\delta)^2+h(1-\delta)+h$$
$$+4(\delta u)(1-\delta)^3h+3(\delta u)(1-\delta)^2h+2(\delta u)(1-\delta)h+3(\delta u)^2(1-\delta)h+(\delta u)^2(1-\delta)h$$
$$+(\delta u)h-(1-\delta)^5X_1-(1-\delta)^4X_2-(1-\delta)^3X_3-(1-\delta)^2X_4-(1-\delta)X_5-X_6$$
$$-4(\delta u)(1-\delta)^3X_1-3(\delta u)(1-\delta)^2X_2-2(\delta u)(1-\delta)X_3-(\delta u)X_4$$
$$-3(\delta u)^2(1-\delta)X_1-(\delta u)^2X_2.$$

We see that the summands in expression of S_n are of three types, namely, having factors S_0, h and X_i. Thus, S_0 is multiplied by the sum of $a_{k,l}(1-\delta)^k(\delta u)^l$ where k begins by taking value n and decreasing at each step by 2, whereas l starts from 0 and increases at each step by 1. Moreover, it is not difficult to construct the Table 1 for the coefficients calculation with elements $a_{i,j}$, $i,j = 0,1,2,\ldots$, where $a_{0,j}=1\ \forall j=0,1,\ldots$ and $a_{i,j}=a_{i,j-1}+a_{i-1,j-2}$. Furthermore, the degree of δu cannot be greater than $\lfloor\frac{n}{2}\rfloor$.

The same Table 1 is useful for calculation of coefficients by h. Namely, factor $(1-\delta)^k(\delta u)^l$ is multiplied by $a_{l,k+2l}$. The highest degree of $1-\delta$ is $n-1$ and it decreases to 0 by jumps equal to 1. If a factor δu is included the same process begins from $n-3$ and so on.

The same coefficients appear in terms with X_i. However instead of h one has to multiply $a_{l,k+2l}(1-\delta)^k(\delta u)^l$ by X_{n-2k-l}.

Table 1. Table for coefficients construction

Factor/n	0	1	2	3	4	5	6	
$(1-\delta)^n$	1	1	1	1	1	1	1	
$(\delta u)(1-\delta)^{n-2}$	0	0	1	2	3	4	5	
$(\delta u)^2(1-\delta)^{n-4}$	0	0	0	0	1	3	6	
$(\delta u)^3(1-\delta)^{n-6}$	0	0	0	0	0	0	1	
$(\delta u)^4(1-\delta)^{n-8}$	0	0	0	0	0	0	0	
...								

Hence, it is possible to write the explicit expression (2) for S_n which is rather complicated.

3 Algorithm for Ruin Probability Calculation

Our next aim is to prove the following result

Theorem 2. *The ultimate ruin probability is represented by*

$$\sum_{n=1}^{\infty}\int_0^{f^{(1)}}\cdots\int_0^{f^{(n-1)}}\int_{f^{(n)}}^{+\infty}\prod_{k=1}^{n}p_{X_k}(v_k(y_1,\ldots,y_n))dy_1\ldots dy_n,$$

where $f^{(n)} = (1-\delta)f^{(n-1)} + u\delta f^{(n-2)} + h$ *for* $n > 1$ *and* $f^{(0)} = S_0$, $f^{(1)} = (1-\delta)S_0 + h$. *The functions* $v_k(y_1, \dots, y_n)$, $k \geqslant 1$, *are specified in the proof.*

Proof. The ruin time is defined by $T = \min\{n : S_1 > 0, \dots, S_{n-1} > 0, S_n \leqslant 0\}$. So, we calculate the probability of ruin at the nth step obtaining the distribution of the ruin time. To this end, we rewrite (2) as a nonhomogeneous linear combination of X_i, $i = 1, \dots, n$, in the form

$$S_n = f^{(n)}(S_0, \delta, u, h) - \sum_{i=1}^{n} g_i^{(n)}(\delta, u) \cdot X_i. \tag{3}$$

Combining (1) and (3) we easily establish the recurrent relations for calculation of $f^{(n)}$ (given in the Theorem 2 statement) and $g_i^{(n)}$. Thus,

$$g_i^{(n)} = (1-\delta)g_i^{(n-1)} + u\delta g_i^{(n-2)}, \quad 1 \leqslant i \leqslant n, \quad g_n^{(n)} = 1, \quad g_{n-1}^{(n)} = 1-\delta, \quad n \geqslant 1.$$

Hence, the probability under consideration $P(T = n)$ is given by $P(U_n)$ with

$$U_n = \left\{ g_1^{(1)} X_1 < f^{(1)}, \dots, \sum_{i=1}^{n-1} g_i^{(n-1)} X_i < f^{(n-1)}, \sum_{i=1}^{n} g_i^{(n)} X_i \geqslant f^{(n)} \right\}.$$

Performing the change of variables $\xi_k = \sum_{i=1}^{k} g_i^{(k)} X_i$, $1 \leqslant k \leqslant n$, we obtain a one-to-one correspondence between $(\xi_i)_{i=1}^n$ and $(X_i)_{i=1}^n$. It can be written in a matrix form as follows

$$\begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} = \begin{pmatrix} g_1^{(1)} & 0 & 0 & \dots & 0 \\ g_1^{(2)} & g_2^{(2)} & 0 & \dots & 0 \\ g_1^{(3)} & g_2^{(3)} & g_3^{(3)} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ g_1^{(n)} & g_2^{(n)} & g_3^{(n)} & \dots & g_n^{(n)} \end{pmatrix} \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$

or $\overrightarrow{\xi} = J \cdot \overrightarrow{X}$, where $J = \left(\frac{\partial \xi_i}{\partial X_j} \right)_{n \times n}$ is a Jacobi matrix of partial derivatives. Thus, Jacobian is equal to $|\prod_{i=1}^{n} g_i^{(i)}|$. The inverse transformation has the form

$$\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} \frac{1}{g_1^{(1)}} & 0 & 0 & \dots & 0 \\ \frac{-g_1^{(2)}}{g_1^{(1)} g_2^{(2)}} & \frac{1}{g_2^{(2)}} & 0 & \dots & 0 \\ \frac{g_1^{(2)} g_2^{(3)}}{g_1^{(1)} g_2^{(2)} g_3^{(3)}} & \frac{-g_2^{(3)}}{g_2^{(2)} g_3^{(3)}} & \frac{1}{g_3^{(3)}} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{(-1)^{n+1} \prod_{i=1}^{n-1} g_i^{(i+1)}}{\prod_{j=1}^{n} g_j^{j}} & \frac{(-1)^{n} \prod_{i=2}^{n-1} g_i^{(i+1)}}{\prod_{j=2}^{n} g_j^{j}} & \frac{(-1)^{n+1} \prod_{i=3}^{n-1} g_i^{(i+1)}}{\prod_{j=3}^{n} g_j^{j}} & \dots & \frac{1}{g_n^{(n)}} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix}.$$

Clearly, $P(T = n)$ is a multiple integral over the domain

$$U_n' = \left\{ g_1^{(1)} x_1 < f^{(1)}, \dots, \sum_{i=1}^{n-1} g_i^{(n-1)} x_i < f^{(n-1)}, \sum_{i=1}^{n} g_i^{(n)} x_i \geqslant f^{(n)} \right\}$$

of $p_{X_1,\ldots,X_n}(x_1,\ldots,x_n)$. Performing the change of variables $\overrightarrow{y} = J \cdot \overrightarrow{x}$, we see that the integral transforms into

$$\int_0^{f^{(1)}} \cdots \int_0^{f^{(n-1)}} \int_{f^{(n)}}^{+\infty} \frac{p_{X_1,\ldots,X_n}(v_1(y_1,\ldots,y_n),\ldots\ldots v_n(y_1,\ldots,y_n))}{|\det J(v_1(y_1,\ldots,y_n),\ldots\ldots v_n(y_1,\ldots,y_n))|} dy_1 \ldots dy_n.$$

Obviously, $\overrightarrow{x} = J^{-1} \cdot \overrightarrow{y}$, so $v_i(y_1,\ldots,y_n) = (J^{-1} \cdot \overrightarrow{y})_i$.

Using the fact that the determinant is a product of diagonal elements $g_i^{(i)} = 1$ and the sequence X_i consists of i.i.d. r.v.'s we establish the desired form of probability

$$\int_0^{f^{(1)}} \cdots \int_0^{f^{(n-1)}} \int_{f^{(n)}}^{+\infty} \prod_{k=1}^{n} p_{X_k}(v_k(y_1,\ldots,y_n)) dy_1 \ldots dy_n.$$

Summing these expressions over all $n \geqslant 1$ we obtain the ultimate ruin probability.

4 Numerical Analysis

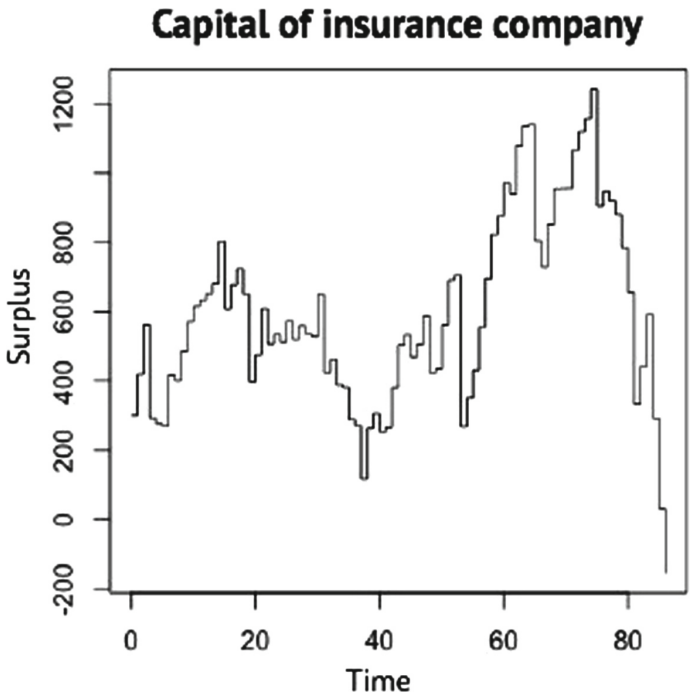
Now we supply some numerical results (obtained by programming in language R). We would like to establish the combinations of parameters leading to ruin or to rapid increase of capital. Take $\beta = 0.0725$ and $S_0 = 300$ (according to the law of Russian Federation on 01.01.2018 the initial capital of insurance company cannot be less than $300 \cdot 10^6$ roubles). We suppose also that the distribution of X_i is exponential with parameter α.

Simulation was carried out with the help of the following program.

```
s = 300.                # initial value
h = [parameter]         # annual premium
beta = 0.0725           # bank extra return
alpha = [parameter]         # parameter of exponential distribution
delta = [fixed param.]  # fraction, invested at each step
s1 = 0.                 # 1 year ago
s2 = 0.                 # 2 years ago
a <- c(s)
i = 0
s1 = s
x = rexp(1, alpha)
s = (1 - delta) * s1 + h - x
a <- c(a, s)
i = i + 1
while (s > 0) {
```

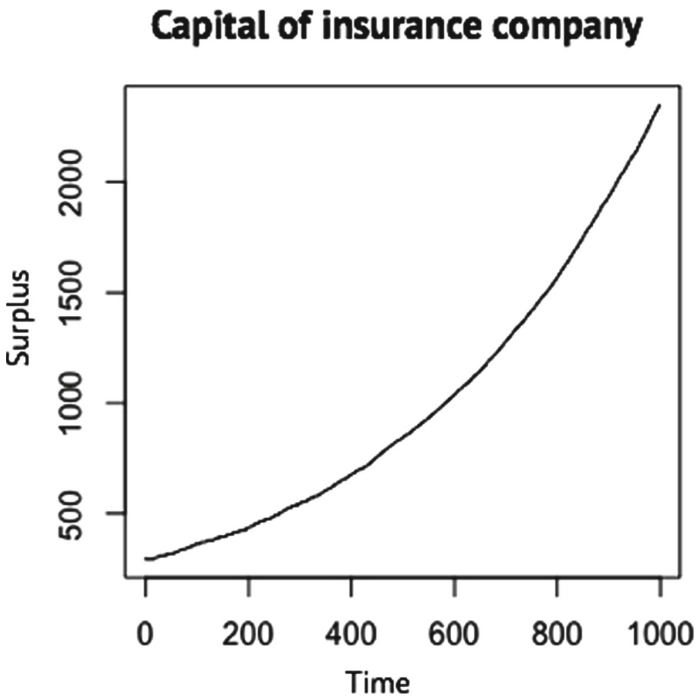
```
    s2 = s1
    s1 = s
    x = rexp(1, alpha)
    s = (1 - delta) * s1 + h - x +
        (1 + beta) * delta * s2
    a <- c(a, s)
    i = i + 1
    if ((i == 1000)&(s >= 300)) break
}
plot(0:i, a, type = "s",
     xlab = "time", ylab = "surplus",
     main = "Capital of company")
```

Here are some of the results in graphical form.



In this case the company was ruined at the 86-th step, $\delta = 0$

In this case growth of surplus is almost linear, $\delta = 0.01$



In this case growth of surplus is exponential, $\delta = 0.03$

Six tables below illustrate the dependence of the simulated ruin probability on parameters (h, α) and δ. Cells of the first row list the values of α, whereas first column stands for values of h. We assume that the company is not ruined if its capital is not less than the initial one after the first 1000 periods. For each set of parameters we performed 1000 simulations according to the program.

If we don't invest anything, we have the following results (Table 2):

Table 2. Table for $\delta = 0$

	0.05	0.10	0.25	0.50	1.00
1	1	1	1	1	0.048
5	1	1	0	0	0
10	1	0.374	0	0	0
25	0.002	0	0	0	0
50	0	0	0	0	0
100	0	0	0	0	0

If we invest just one percent of our surplus, then (Table 3)

Table 3. Table for $\delta = 0.01$

	0.05	0.10	0.25	0.50	1.00
1	1	1	1	1	0
5	1	1	0	0	0
10	1	0.219	0	0	0
25	0.001	0	0	0	0
50	0	0	0	0	0
100	0	0	0	0	0

Let us now increase δ five times. In this case, we have (Table 4)

Table 4. Table for $\delta = 0.05$

	0.05	0.10	0.25	0.50	1.00
1	1	1	1	0.348	0
5	1	1	0	0	0
10	1	0.012	0	0	0
25	0.002	0	0	0	0
50	0	0	0	0	0
100	0	0	0	0	0

After doubling δ, the results are (Table 5)

Table 5. Table for $\delta = 0.1$

	0.05	0.10	0.25	0.50	1.00
1	1	1	1	0	0
5	1	1	0	0	0
10	1	0.003	0	0	0
25	0.001	0	0	0	0
50	0	0	0	0	0
100	0	0	0	0	0

If we invest a quarter of the surplus, then (Table 6)

Table 6. Table for $\delta = 0.25$

	0.05	0.10	0.25	0.50	1.00
1	1	1	0.001	0	0
5	1	0.784	0	0	0
10	0.998	0	0	0	0
25	0	0	0	0	0
50	0	0	0	0	0
100	0	0	0	0	0

Finally, for $\delta = 0.5$ we have (Table 7)

Table 7. Table for $\delta = 0.5$

	0.05	0.10	0.25	0.50	1.00
1	1	0.967	0.001	0	0
5	1	0.028	0	0	0
10	0.921	0	0	0	0
25	0	0	0	0	0
50	0	0	0	0	0
100	0	0	0	0	0

It can be seen that the probability decreases while δ is increasing. The most difficult situation to forecast is the case $h\alpha$ being in the neighborhood of 1.

Another research direction is the study of the model sensitivity to small parameters fluctuations and perturbations of the underlying probability distributions (see, e.g., [10]).

The Sobol' Sensitivity Indices. Let the model output $Y = h(X_1, \ldots, X_k)$ be the function of k parameters. In case of uncertainty in parameters values it is important to establish the most influential parameter (or groups of parameters) and those that have a negligible effect on the output and may be fixed. For this purpose the Sobol' method of decomposition is useful.

Given a square integrable function h over the k-dimensional unit hypercube, Sobol' considers an expansion of h into terms of increasing dimensions:

$$h = h_0 + \sum_i h_i + \sum_i \sum_{j>i} h_{ij} + \ldots + h_{12\ldots k} \tag{4}$$

in which each individual term is also square integrable over the domain of existence and is a function only of the factors in its index, i.e. $h_i = h_i(X_i), h_{ij} = h_{ij}(X_i, X_j)$ and so on.

This expansion, called high-dimensional model representation (HDMR), is not unique, meaning that, for a given model h, there could be infinite number of choices for its terms. However, Sobol' proved that, if each term in the expansion above has zero mean, then all the terms of the decomposition are orthogonal in pairs. As a consequence, these terms can be uniquely calculated using the conditional expectations of the model output Y. In particular, $h_0 = E(Y)$, $h_i = E(Y|X_i) - E(Y)$, $h_{ij} = E(Y|X_i, X_j) - h_i - h_j - E(Y)$.

The variances of the terms in the decomposition (4) are the measures of importance being sought. In particular, $V(h_i(X_i))$ is $V[E(Y|X_i)]$, so dividing this by the unconditional variance $V(Y)$, we obtain the first-order sensitivity index. In short:

$$S_i = \frac{V[E(Y|X_i)]}{V(Y)}.$$

The first-order index represents the main effect contribution of each input factor to the variance of the output.

We can write the so-called ANOVA-HDMR decomposition of variance

$$V(Y) = \sum_i V_i + \sum_i \sum_{j>i} V_{ij} + \ldots + V_{12\ldots k}$$

giving immediately

$$\sum_i S_i + \sum_i \sum_{j>i} S_{ij} + \sum_i \sum_{j>i} \sum_{l>j} S_{ijl} + \ldots + S_{123\ldots k} = 1.$$

The total effect index accounts for the total contribution to the output variation due to factor X_i, i.e. its first-order effect plus all higher-order effects due

to interactions. It can be written (by conditioning this time with respect to all the factors but one, i.e. $X_{\sim i}$) in the form

$$S_{T_i} = \frac{E[V(Y|X_{\sim i})]}{V(Y)} = 1 - \frac{V[E(Y|X_{\sim i})]}{V(Y)}.$$

5 Conclusion and Further Research Directions

First of all, we have obtained the explicit expression for calculation of company capital for each n under assumption that a certain quota of the capital is placed in a bank. Although the expression is complicated, it enabled us to establish the form of probabilities $P(S_1 > 0, \ldots, S_{n-1} > 0, S_n \leqslant 0)$ for all n. Summing these probabilities we got the ultimate ruin probability. The numerical results provided in the paper show the dependence of this probability on the system parameters. Due to the space restrictions we present here only the first step of our research.

For simplicity, we considered only the case of one-period investment ($s = 1$), the further investigation includes the case $s > 1$. Moreover, for practical applications it is interesting to add the possibility of investment in the risky assets and establish the optimal investment policy maximizing the non-ruin probability. It is possible to consider the other objective functions in the framework of cost approach (see, e.g., [21]) and take into account not only investment but dividends and reinsurance as well (see, e.g., [8,9,15,20]).

Acknowledgement. This research is partially supported by RFBR grant No. 17-01-00468.

References

1. Bulinskaya, E.: New research directions in modern actuarial sciences. In: Panov, V. (ed.) MPSAS 2016. PROMS, vol. 208, pp. 349–408. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-65313-6_15
2. Arrow, K.J., Karlin, S., Scarf, H.: Studies in the Mathematical Theory of Inventory and Production. Stanford University Press, Stanford (1958)
3. Afanaseva, L., Bulinskaya, E.: Asymptotic analysis of traffic lights performance under heavy traffic assumption. Methodol. Comput. Appl. Probab. **15**, 935–950 (2013)
4. Vishnevskii, V.M., Andronov, A.M.: Estimating the throughput of wireless hybrid systems operating in a semi-Markov stochastic environment. Autom. Remote Control **78**, 2154–2164 (2017)
5. Li, S., Lu, Y., Garrido, J.: A review of discrete-time risk models. Revista de la Real Academia de Ciencias Naturales. Serie A, Matemáticas **103**, 321–337 (2009)
6. Li, S., Sendova, K.P.: The finite-time ruin probability under the compound binomial model. Eur. Actuar. J. **3**, 249–271 (2013)
7. Picard, P., Lefèvre, C.: Probabilité de ruine éventuelle dans un modèle de risque à temps discret. J. Appl. Probab. **40**, 543–556 (2003)

8. Bulinskaya, E., Gusak, J., Muromskaya, A.: Discrete-time insurance model with capital injections and reinsurance. Methodol. Comput. Appl. Probab. **17**, 899–914 (2015)
9. Bulinskaya, E., Gromov, A.: Asymptotic behavior of the processes describing some insurance models. Commun. Stat. - Theory Methods **45**, 1778–1793 (2016)
10. Bulinskaya, E., Gusak, J.: Optimal control and sensitivity analysis for two risk models. Commun. Stat. - Simul. Comput. **45**, 1451–1466 (2016)
11. Alfa, A.S., Drekic, S.: Algorithmic analysis of the Sparre Andersen model in discrete time. ASTIN Bull. **37**, 293–317 (2007)
12. Bao, Zh., Liu, H.: On the discounted factorial moments of the deficit in discrete time renewal risk model. Int. J. Pure Appl. Math. **79**(2), 329–341 (2012)
13. Blaževičius, K., Bieliauskienė, E., Šiaulys, J.: Finite-time ruin probability in the nonhomogeneous claim case. Lith. Math. J. **50**(3), 260–270 (2010)
14. Bulinskaya, E.: Asymptotic analysis of insurance models with bank loans. In: Bozeman, J.R., Girardin, V., Skiadas, C.H. (eds.) New Perspectives on Stochastic Modeling and Data Analysis, pp. 255–270. ISAST, Athens (2014)
15. Bulinskaya, E., Yartseva, D.: Discrete time models with dividends and reinsurance. In: Proceedings of SMTDA 2010, Chania, Greece, 8–11 June 2010, pp. 155–162 (2010)
16. Cai, J.: Discrete time risk models under rates of interest. Probab. Eng. Inf. Sci. **16**, 309–324 (2002)
17. Li, S.: On a class of discrete time renewal risk models. Scand. Actuar. J. **4**, 241–260 (2005)
18. Li, S.: Distributions of the surplus before ruin, the deficit at ruin and the claim causing ruin in a class of discrete time risk models. Scand. Actuar. J. **4**, 271–284 (2005)
19. Picard, P., Lefèvre, C., Coulibaly, I.: Problèmes de ruine en théorie du risque à temps discret avec horizon fini. J. Appl. Probab. **40**, 527–542 (2003)
20. Qi, X.: Analysis of the generalized Gerber-Shiu function in discrete-time dependent Sparre Andersen model. Thesis, The University of Hong Kong Libraries, University of Hong Kong (2016)
21. Bulinskaya, E.: Cost approach versus reliability. In: Proceedings of International Conference DCCN 2017, 25–29 September 2017, pp. 382–389. Technosphera, Moscow (2017)

Model of Next-Generation Optical Switching System

K. A. Vytovtov[1], E. A. Barabanova[1(✉)], and V. S. Podlazov[2]

[1] Astrakhan State Technical University, Tatischeva 16, Astrakhan, Russia
elizavetaalexb@yandex.ru
[2] V. A. Trapeznikov Institute of Control Sciences of RAS, Profsoyuznaya 65, Moscow, Russia

Abstract. The new types of photonic switching systems based on the newly developed fundamentally new 4x4-switching cell are presented. Using this cell instead well-known 2x2-switching cell allows us to significantly improve the characteristics of switching systems. The double-stage 16x16-switching system and 256x256-switching system are described in detail for the first time. The expressions for the numbers of stages and basic elements on the number of inputs for the developed schemes are obtained for the first time also. Numerical calculations and comparison with well-known schemes is carried out.

1 Introduction

Fiber and optical communication lines are able to transmit information flows at rate of some tens and even hundreds terabits per a second. In its return carrying capacity of the most efficient switches does not exceed some hundreds *Gbits* per a second up to date. Hence switch nodes restrict optical network carrying capacity. For today there are two switching technologies namely information signal transfer from optical form into electrical one and vice versa, and completely optical switches. The second technology allows increasing switch speed exponentially and it is considered to be advanced for next-generation high-speed communication systems designs [1–3]. In the modern literature we distinguish: mechanical optical switches, electro-optic switches, thermo-optical switches, optoelectronic switches based on SOA (semiconductor optical amplifiers), integrated active-waveguide switches, switches on photonic crystals, switches on multilayer light-wave liquid-crystal matrices, switches on integrated circuits IC) with a set of matrices of optoelectronic gates connected by interaction with an optical beam [1]. Currently, for constructing optical switching systems, as a rule, the matrix scheme, the Benes scheme, the Klose scheme [2], the Batcher scheme, the Benes-Shpanke scheme and the Shpanke scheme, the banyan-like schemes have been used [1–3]. In all these schemes a 2x2 switching cell has been used. A switching

The reported study was funded by RFBR according to the research project 18-37-00059/18.

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 377–386, 2018.
https://doi.org/10.1007/978-3-319-99447-5_32

cell with four inputs and four outputs has been proposed in [4] only. The operation principle of this cell is based on the spatial deviation of light beam due to the change in the parameters of a ferrite under influence of an external control signal. The increase in the spatial diversity of the outputs has been achieved due to the use of a re-reflected signal in the structure. New structures for completely optical switch development are offered by a number of scientists too. These are nanosurfaces, metasurfaces and bulk metamaterials [5,6]. It has been note in those papers that the switching cells based on these materials give an acceptable speed (Tbit/s). But delays of a control system due to change speed of structure parameter (for example, heat capacity, etc.) and also delay due to imperfections in circuit solutions have not been taken into account by the authors. Moreover these cells are in fact only optical switches with two outputs. Such a switch provides either transmission or reflection of an information beam in the presence of a control signal. Schemes constructed on such elements are multistage and thus its increase processing time of a signal. Therefore the development of simpler schemes of next-generation optical switching systems and new elements for their construction are actual and significant.

2 The Next-Generation Switching System

The operating principal of next-generation optical switching system is based on packet switching. According to this technology sequence of bits from the transmitter fits in the container which is called the packet. The packet consists of the packet header and the information. The packet header contains the control information. It is a source address and a destination address, the method of check of integrity of contents of a packet, etc. In this work the control information is of interest. Basic purpose of the optical switches is packet switching from one network node to another. This process is under the processer control or can be self-turning. In this paper, self-turning control process is offered. But to use such type of control, the optical switching system must be built on special principles. In this paper the new types of photon switching systems based on the newly developed fundamentally new 4x4-switching cell [4] are presented. It is the very important fact that this cell is self-adjusting and does not require external control. Using this cell instead a well-known 2x2-switching one allows us to significantly improve the characteristics of switching systems. The double-stage 16x16- and 256x256-switching systems are developed and described in detail for the first time. The number of stages and base elements in the presented switches is significantly reduced in comparison with the previously known typical schemes due to the new type of switching cell. It is important that the proposed systems are fully accessible and non-blocking. It should be noted that the systems are fundamentally new and there is no standard methods for calculating the parameters of such systems to date. However in our treatment the expressions for the numbers of stages and basic elements as the functions of the inputs (outputs) number are obtained for the first time also. Additionally the numerical calculations and comparison the obtained with well-known schemes [2–4] is carried out

too. As a result it is shown that the presented scheme has great advantages in comparison with existing ones.

3 The 16x16 Switching System

At first note that the presented 16x16 switching system (Fig. 1) is actually a full bipartite directed graph $G(32, 32)$. Let us consider the structure and functioning principle of the developed 16x16 switching system (Fig. 1). This circuit is fully accessible and also the one is unidirectional non-blocking switching system.

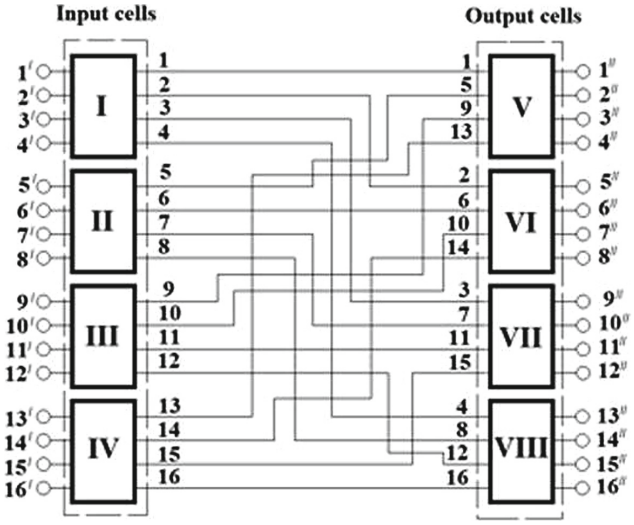


Fig. 1. The 16x16-switching system

This double-stage system (Fig. 1) is built on the basis of the 4x4 switching cells described in [4]. In the considered case each input cell is connected with each output cell. For this, the first output of the input stage cell I is connected to the first input of the output stage cell V; the second output of the input stage cell I is connected to the first input of the output stage cell VI; the third output of the input stage cell I is connected to the first input of the output stage cell VII; the fourth output of the input stage cell I is connected to the first input of the output stage cell VIII. Similarly, the outputs of the other cells of the input stage (II - IV) are connected with the inputs of the output cells (V-VIII). Therefore an information signal can transmit from any input to any output of the system. For example, let us show the path of an information signal from the input 2 to the output 10 (Fig. 1). The signal at input 2 is switched to output 3 of the input cell I. Then this signal transmits to the input 1 of the switching cell VII (see the path 3 in Fig. 1). Finally the cell VII commutes this signal to the output 10 of this switching system. Additionally note that it has a dimension of 8x8 if we really need to reserve this system as 1+1.

4 The 256x256 Switching System

First of all the presented system is a bipartite full directed graph $G(512, 512)$. Obviously, the eccentricity of this graph is equal to 1. The model of the double-stage switching system based on the new 16x16 optical units is presented in Fig. 1. The system is supposed to have 256 inputs and 256 outputs. It is very important that this double-stage system is fully accessible and non-blocking. The system contains the 16 input blocks and the 16 output blocks I, II, III, IV, ... (Fig. 2). Each block contains 16 inputs and 16 outputs. Let's consider the path of an information signal in this system. We assume that the signal has arrived at the input 2 of the system, and it must exit the output 125. The second input of the system is simultaneously the second input of the first level switching unit I in the second level switching unite A (Figs. 1, 2). The switching unit I is actually the 16x16 unidirectional switch. And it can commute any itself inputs with any outputs. In the considered case the block commutes the input 2 with the output 8. The signal exits from the output 8 of the switching unit I and it transmits to the first input of the first level block IV of the second level block B (the input 8 in Fig. 2).

5 The Structural Characteristics

The main structural characteristics by which the proposed optical switching system and existing switching schemes can be compared are following:

- The number of basic cells in the circuit.
- The number of system stages that determines the maximum length of a connecting path and also affects the switching speed and the probability of blocking [1–5].

Note that the problem of constructing a mathematical model of systems based on 4x4 switching cells requires further in-depth study. However, we can present the first results of the analysis of the developed systems. Here the expressions for finding the number of basic elements that are required to construct the proposed optical system as well as the number of the stages in the commutation scheme depending on the number of inputs are obtained by using the mathematical induction method. In our investigations a switching cell is chosen as a basic element. For the 16x16 switching system (Fig. 1) the number of the basic elements can be calculated as

$$R = \begin{cases} \frac{N}{2} \text{ if } N \leq 16, \\ N \text{ if } N > 16. \end{cases} \tag{1}$$

Here N is the numbers of the inputs (outputs). Indeed the number of cells is 4 for 8x8-system, the number of cells is 6 for 12x12-system, the number of cells is 8 for 16x16-system, the number of cells is 256 for 256x256-system, the number of cells is 65536 for 65536x65536-system. Now let us consider the problem of

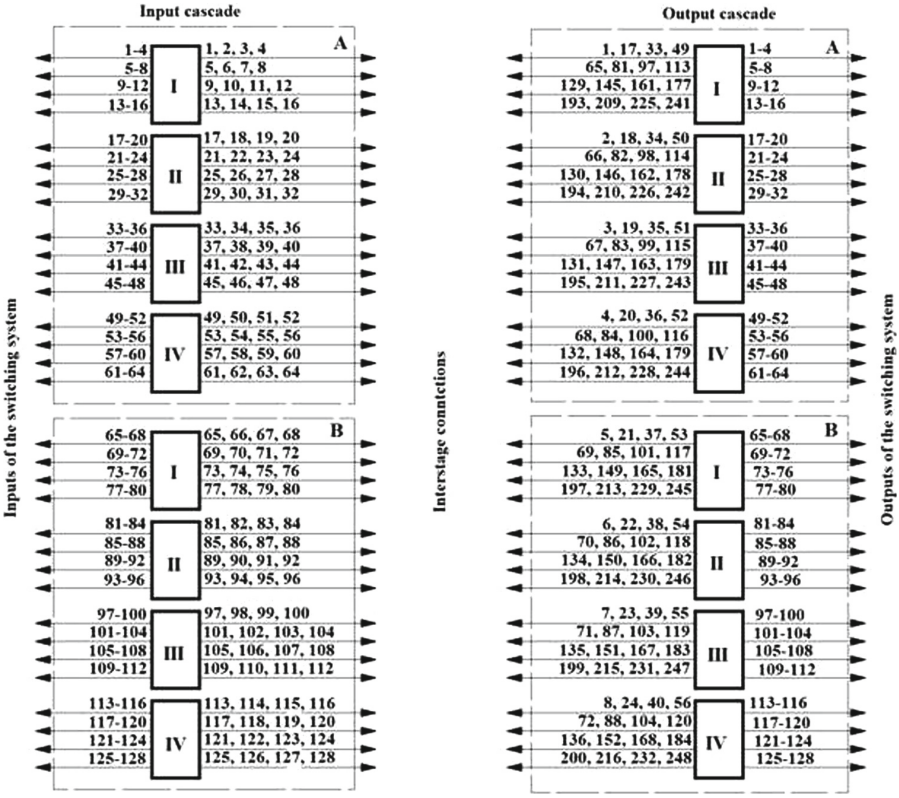


Fig. 2. The 256x256-switching system (the first part)

the cascade number of the switching systems based on 4x4 cells. For the considered case we find the relation between the number of inputs and the number of cascades:

$$K = \frac{log_2 N}{2} \tag{2}$$

The dependences of the number R of basic elements on the number of inputs (outputs) for Benes, Benes-Shpanke, and our double-stage schemes are presented in Fig. 4. To calculate these dependences we use the well-known formulas [1–5] for the above mentioned schemes and the expression (1). The curve 1 corresponds to Shpanke scheme, the curve 2 corresponds to Benes scheme, the curve 3 corresponds to Benes-Shpanke scheme, and the curve 4 corresponds to our double-stage schemes. The comparison of the calculation results shows that the proposed switching system gives us multiple winnings in the basic elements. For example, the number of the elements are 130560, 32640, 2047, and 256 for Shpanke scheme, Benes scheme, BenesShpanke scheme, and our double-stage scheme correspondingly. Thus the proposed scheme gives us a win of eight times

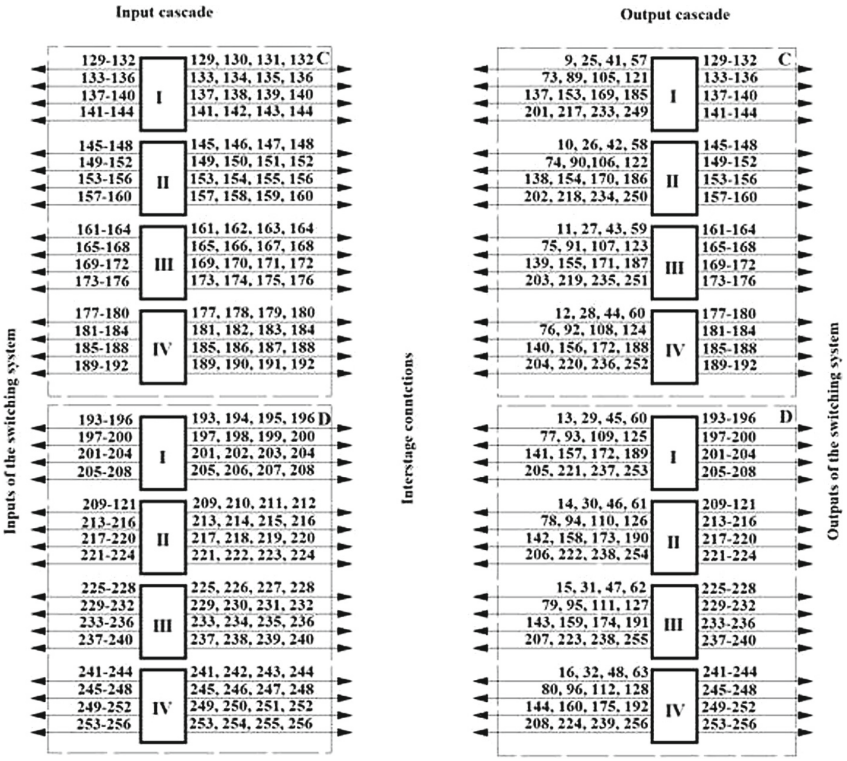


Fig. 3. The 256x256-switching system (the second part)

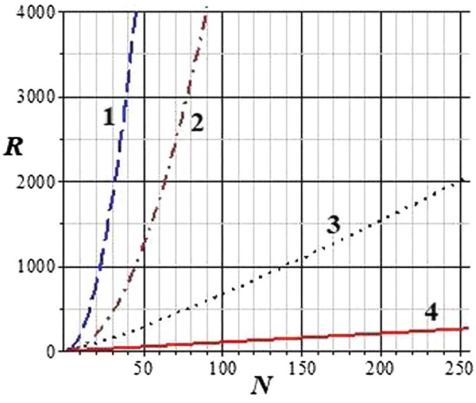


Fig. 4. The dependence of the number of elements on the number of inputs

in comparison with Benes-Shpanke scheme and it gives a gain of five hundred and ten times in comparison with Shpanke scheme.

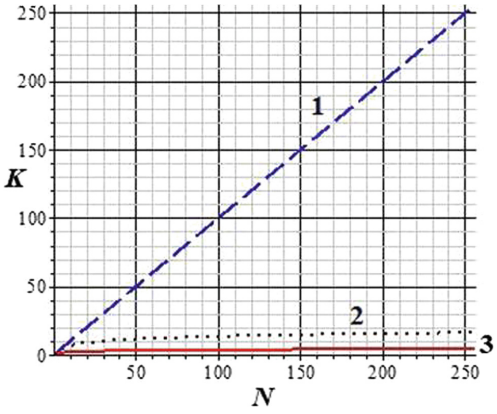


Fig. 5. The dependence of the number of cascades on the number of inputs

The dependences of the number of stages on the number of inputs are shown in Fig. 5. The curve 1 corresponds to Benes scheme, the curve 2 corresponds to Benes-Shpanke scheme, and the curve 3 corresponds to our double-stage scheme. Here we obtain that Benes scheme includes 16 stages, Shpanke scheme includes 16 stages, Benes-Shpanke scheme includes 256 stages, and our scheme has 4 stages. Therefore the offered switching system gives us a win in the number of stages too. These advantages will reduce the probability of blocking in the circuit and increase its speed.

6 The Structure of Optical Switching Cell

In this section we consider the base element of the proposed above system: switching cell. The optical switching cell consists the buffer device and the switching unit (Fig. 6) [4]. The functional principle of the cell is based on frequency separation of control and information signals [5]. An input optical signal consists a control signal with the wavelength λ_s and an information signal with the wavelength λ_i. These signals are separated in the Bragg filter (BF) of switching unit. The used Bragg filter is actually multilayered periodic structure. Thus a control signal with λ_s is reflected from surface of the one and transmitted to the frequency detector (FD), and an information signal transmits through the structure and goes to the deflection system (DS). The deflection system is a controlled photonic crystal. Actually it is multilayered structure including ferromagnetic, optoelectric, thermoelectric or ferroelectric films. The properties of these films can be charged by an external control signal (voltage, current, thermal or optical radiation) and therefore a transmitted angle can be controlled by an external signal. In our system we choose an optically controlled material.

The frequency detector converts the frequency of the control signal into its amplitude. It is a multilayered isotropic structure with linear dependence of the transmitting coefficient on a frequency in the operating domain [9,10].

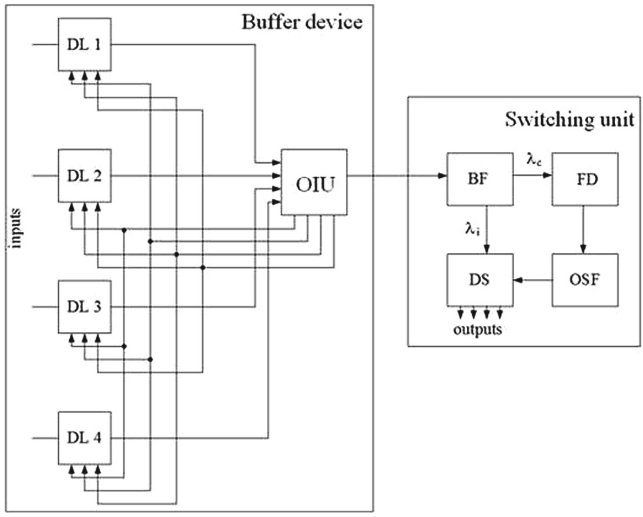


Fig. 6. The structure of an optical switching cell

The amplitude-modulated optical signal from the frequency detector is transmitted to the optical signal former (OSF) that actually is a controlled light emission diode.

The operating principle of the deflecting system is described in [4] in detail. This system has one input and four outputs. To effective control of function of the cell, it is necessary to use two signals with different frequency. The combination of these two control signals determinate the necessary output for given input signal.

It obviously must be used the buffer-multiplexing device in this scheme as the deflecting system has only one input (Fig. 6). The buffer device includes the optical integrated device (OIU) and four controlled delay lines (DL). Let us consider the functioning principle of the buffer device. It is assumed that an input signal is arrived at one of the delay lines. If there are no signals at other inputs at this time, it transmits to the integrated optical device of the buffer. The intermittent device generates a prohibition signal for the other inputs. Thus, the maximum delay within the buffer is determined by the time of the information packet propagation through the three inputs. Note additionally that the optical integrated device performs spatial multiplexing of signals from four inputs as the switching cell has only one input.

The most important problem in the development of such devices is redundancy. In our case, when integrated optical circuits are used, it is advisable to use anisotropic materials and its phenomenon of birefraction.

Obviously, the speed of the represented cell is determined by a delay time in the buffer device and a delay time in the switching unit. Here we must take into account a delay time in each elements of the scheme and duration of the information packet.

In order to calculate the total time for setting up the switching cell, it is necessary to find the time for the information and control signals to transmit through the cell elements

$$t_t = \begin{cases} t_{FD} + t_{OSF} + t_{DS} & \text{if } t_{BF} < t_{FD} + t_{OSF}, \\ t_{BF} + t_{DS} & \text{if } t_{BF} > t_{FD} + t_{OSF}. \end{cases} \quad (3)$$

where t_{BF} is a time of information signal transmission through the Bragg filter, t_{FD} is a time of control signal transmission through the frequency detector, t_{OSF} is a time of control signal transmission through optical signal former, t_{DS} is a time of information signal passing through the deflection system.

The each delay time in (3) is determined by light velocity $v = c/\sqrt{\mu\varepsilon}$ and thickness of the structure. In the presented devices we obtain that the velocity is approximately $v_{anis} = (1 - 1.1) \cdot 10^9$ m/s for the anisotropic materials and $v_{is} = (3.5 - 4.5) \cdot 10^9$ m/s for the isotropic materials.

Here we consider a typical Bragg filter as the isotropic multilayer slab with the thickness $100\mu m$. Thus the time of information signal passing through the one is about $4 \cdot 10^{-14}$ s. Analogously the delay of control signal in the frequency detector is $0.9 \cdot 10^{-14}$ as the thickness of this device to be not more than $40\mu m$. The optical signal is delayed in the optical signal former $3 \cdot 10^{-14}$ s because we intend to use a light-emitting diode (LED) as a control element in this device. The delay time of the deflecting system is determined by the time of the change in the structure parameters and it depends on the material characteristic.

7 Conclusions

In this paper the new types of photon switching systems based on the newly developed fundamentally 4x4-switching cell are presented. Using this cell instead well-known 2x2-switching cells allows us to significantly improve the characteristics of switching systems. The double-stage 16x16-switching system and 256x256-switching system are described in detail for the first time. And the detail description of the switching sell is presented too. It is important that this cell is self-tuning. The expressions for the dependencies of the numbers of stages and basic elements on the number of inputs for the developed schemes are obtained for the first time also. Additionally the expression for delay time in the switching cell is presented here. Numerical calculations and comparison with well-known schemes are carried out.

References

1. Bawab, E.: Optical Switching. Springer, Heidelberg (2006). https://doi.org/10.1007/0-387-29159-8
2. Podlazov, V.S.: A new form of an unblockable network. Avtom. Telemekhanika **75**(10), 139–152 (2014)
3. Chai, Z., Hu, X., Wang, F., Niu, X., Xie, J., Gong, Q.: Ultrafast all-optical switching, revised paper. In: Advanced Optical Materials, p. 21 (2017)

4. Barabanova, E.A., Maltseva, N.S.: Switching systems with parallel processing. Astrakhan, ASTU (2012)
5. Barabanova, E.A., Vytovtov, K.A., Barabanov, I.O., Maltseva, N.S.: Photon switching cell cellutility model. Patent 171015 (2017)
6. Barabanov, I.O., Maltseva, N.S., Barabanova, E.A.: Switching cell for information transmission optical systems. In: Conference Proceedings - 2016 International Conference on Actual Problems of Electron Devices Engineering. APEDE 2016, pp. 343–347 (2016)
7. Gholipour, B., Zhang, J., Maddock, J., MacDonald, K., Hewak, D., Zheludev, N.: All-optical, non-volatile, chalcogenide phase-change meta-switch. In: Proceedings of the European Conference on Lasers and Electro-Optics, Germany (2015)
8. Dabidian, N., et al.: Switching of mid-infrared light using plasmonic fano-resonant meta-surfaces integrated with graphene. In: Proceedings of CLEO: QELS Fundamental Science, US (2014)
9. Vytovtov, K.A., Gnatushenko, V.V., Wojcik, W.: Frequency detector of the terahertz domain based on stratified structure. Elektronika **54**(8), 58–60 (2013)
10. Vytovtov, K.A., Mospan, L.P.: Penetration effect in gyrotropic slab: theory and applications. J. Opt. Soc. Am. A **29**(6), 877–882 (2012)

On Some Properties of Smoothly Irregular Waveguide Structures Critical for Information Optical Systems

A. A. Egorov[1], G. Andler[2], A. L. Sevastianov[3], and L. A. Sevastianov[3(✉)]

[1] Prokhorov General Physics Institute of the Russian Academy of Sciences, Moscow, Russia
yegorov@kapella.gpi.ru
[2] Stockholm University, Stockholm, Sweden
guillermo.andler@fysik.su.se
[3] Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow 117198, Russian Federation
{sevastianov_al,sevastianov_la}@rudn.university

Abstract. Two types of optical smoothly irregular waveguide structures promising for application in optical information systems are studied by numerical simulation and experimentally: the thin film generalised waveguide Luneburg lens and the liquid thin waveguide lens. The importance of the statistical analysis of functioning of optical components in information optical systems is emphasised.

Keywords: Information optical system · Planar lens · Waveguide
Optical radiation · Computer-aided design · Numerical simulation

1 Introduction

Data, software, and hardware are essential components of an information system [1]. Using its facilities, the information system provides processing and transmission of information.

During the last decades, extensive development of integrated optics, fibre optics, and waveguide optoelectronics facilitated the progress of promising integrated optical and fibre optical information systems, in which the most important constituent parts are optical detectors (sensors) and various integrated optical processors, executing different transformations of the input information [2–7]. The progress of technology stimulates further interest to the development and improvement of integrated optical and fibre optical sensors and integrated processors, intended for the use in different fields of science, engineering, and industry, particularly, in promising infocommunication technologies, such as intelligent monitoring of environment and objects using the Internet (remote access

L. A. Sevastyanov—The publication has been prepared with the support of the RUDN University Program 5-100.

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 387–398, 2018.
https://doi.org/10.1007/978-3-319-99447-5_33

via an optical fibre/twisted pair, wireless IR connection), the wireless (cellular, satellite) radio connection, etc. Using the remote sensors one can receive the information about the chemical composition, shape and structure, position and dynamics. The most important advantage of such information optical systems as compared to the electronic ones is that their construction can use only a few optical elements for executing the functions that could require hundreds of electronic components. Undoubtedly, the solution of these problems has both the fundamental and the applied significance for the development of nanotechnologies in the above fields. The modern integrated optical processors are the key part of such distributed systems of data acquisition and processing [2], in particular, the systems that collect, transmit and process the information about the condition of different systems, e.g., by means of rapidly developing Internet technologies, mobile and satellite-aided communications [3]. Various optical waveguides are the base of integrated optical processors. Such processors can be combined with different integrated optical devices and components, e.g., sensors, prisms, lenses, muldems, etc. (see, e.g., [2–7]). One of the major stages of developing an optical integrated system is the analysis and synthesis of the optical components, necessary for its normal functioning, by means of computer modelling and computer-aided design [2–7] with the usage of modern numerical methods [7,10,11,15,16], including that considering probability models.

The description of coherent polarised monochromatic radiation propagating in (smoothly) irregular integrated optical waveguides, considering the possible effects of depolarisation and hybridisation of waveguide modes, as well as the electromagnetic fields matching in the waveguide interfaces of different integrated optical devices and components [2–14], is important for the aims of modelling and computer-aided design of a number of basic optical elements of information systems, e.g., the thin film generalised waveguide Luneburg lens (Luneburg TFGWL) and the horn-type waveguide [2–7,15]. There are at least two important cases, in which the consideration of vector character of electromagnetic fields is necessary. First, in the synthesis of different 3D interface elements (prisms, lenses, etc.) it allows the implementation of efficient transfer of energy via the interface. The second case is related to the operation of an integrated optical spectrum analyser operation in real time, e.g., on board an aircraft. The purpose of such spectrum analyser is to implement instantaneous spectral analysis of the input signal, e.g., the radar one, in order to determine whether the carrier aircraft is traced by another aircraft, a missile, or a ground-based radar. The main characteristic of the integrated optical spectrum analyser is its resolution power, largely dependent on the resolution of waveguide lenses [7,14]. The method of adiabatic waveguide modes proposed by us is suitable for computer simulation of propagation of coherent polarised monochromatic radiation through a thin-layer waveguide lens and for the mathematical synthesis (computer-aided design) of integrated optical devices (processors), implementing a specified amplitude-phase transformation of optical signals [7–15]. In the present paper we study the Luneburg TFGWL and the thin-layer waveguide lens (TLWL), implemented using different liquids, namely, nematic liquid crys-

tal (NLC), aniline and glycerol. The urgency and transformative potential of these studies is also caused by the necessity to develop new methods of studying the waveguide structures formed by liquid media, including liquid crystals (LCs) (see, e.g., [17–19]).

2 Objects of Study

The objects of study were the liquid-based TLWL (Fig. 1) and the Luneburg TFGWL (Fig. 2). In detail we consider only the case of using NLC as a waveguide layer. The liquid crystal TLWL is based on the NLC 4-Cyano-4'-pentylbiphenyl (5CB), well known by multiple scientific and engineering publications (see, e.g., [17–20]). Figure 2 presents the studied smoothly irregular multilayer integrated optical structure and the three-dimensional synthesised thickness profile of the Luneburg TFGWL (see, e.g., [7,10]). The left-hand part of Fig. 2 shows the three-layer regular integrated optical waveguide formed by the media 1-3. The propagation direction of the specified mode is indicated by a thick arrow in the left-hand part of Fig. 2. The four-layer Luneburg TFGWL is presented in the right-hand part of Fig. 2. In Fig. 2 1 is the surrounding medium (air) with the refractive index n_c; 2 is the first (basic) waveguide layer with the refractive index n_f; 3 is the substrate with the refractive index n_s; 4 is the second (additional) waveguide layer with the refractive index n_l.

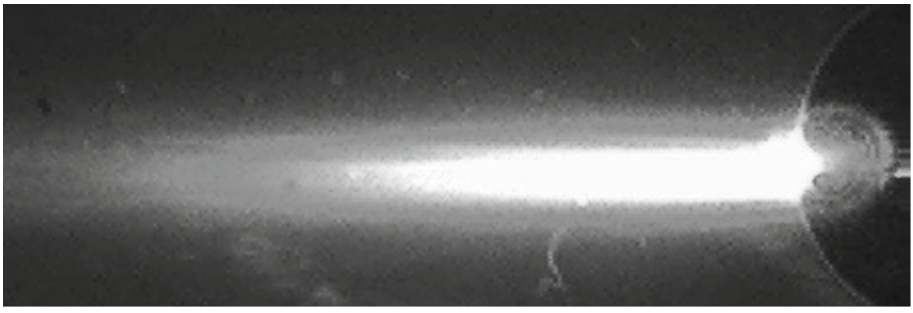


Fig. 1. Photograph of liquid-based TLWL (right); waveguide mode tracks (left) in the integrated optical liquid-based waveguide

The Luneburg TFGWL was fabricated on a silicon substrate, coated with the first (regular) waveguide layer (the Corning 7059 glass), over which the second waveguide layer (Ta_2O_5) having the variable thickness $h\,(y, z)$ was applied. The covering layer was air.

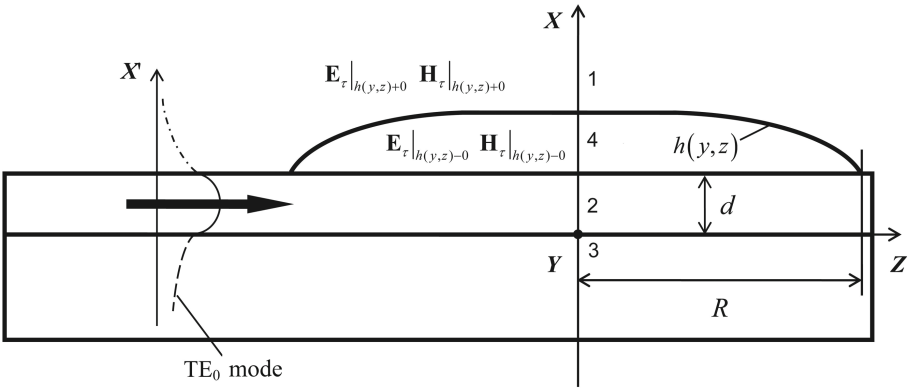


Fig. 2. Cross section (in the xz plane) of the studied smoothly irregular integrated optical waveguide structure

3 Results of the Study of Smoothly Irregular Waveguide Structures

In the experiments we studied multimode NLC-based TLWLs, formed by two glass plates and a layer of liquid, e.g., LC between them. The NLC possessed homogeneous planar orientation with the optical axis along the director (coincident with the z-axis). The NLC layer had the ordinary refractive index 1.53 and the extraordinary refractive index 1.70 (for the laser radiation wavelength $\lambda = 0.64$ μm and the temperature $\approx 25°C$). The glass plates had the refractive index $n_1 = n_3 = 1.52$. Thus, in the experiments and in the numerical calculations the symmetric three-layer waveguide structures were studied having the following refractive indices: $n_1 = n_3 = 1.52$, $n_2 = 1.53$. We emphasise that in the experiments the thickness profile $h\,(y, z)$ of the liquid crystal lenses was not known; only their maximal width was known, actually determined by the thickness h of the waveguide, which in the studied samples varied from 25 to 125 μm. Figure 4 presents one of the measured profiles of the radiation intensity (curve 1) not far from the back focal plane of the planar LC lens. In Fig. 4 the curve 2 is obtained by smoothing the distribution 1; 3 is the fitted Gaussian curve; the double arrow 4 shows the error (less than 7%). The half-width of the distributions like 1 in the vicinity of the focal plane, measured at the half-maximum level, approached (10–45)λ. This fact is due to the small aperture of the lenses; besides that, we studied the multimode liquid crystal waveguide structures with the number of modes 14 and more. Note, that in the case of using aniline and glycerol, the studied waveguide structures mainly had a small number of modes (less than 10). The focal length estimation of the studied liquid crystal lenses has shown that it lies within the range from $2R$ to $8R$, where R is the radius of the planar LC lens, which in the present case was equal nearly to 2 mm, the appropriate focal length being about 5 mm. We studied the liquid crystal TLWLs with R from about 1.5 to 4 mm. It was found that the half-width of the inten-

sity profiles at the half-maximum level exceeds the theoretical diffraction limit $\delta = 0.61\lambda/NA$ ($NA < 1$ being the numerical aperture) at least by a few times. Note, that in the case of using aniline and glycerol the waveguide structures with small number of modes were mainly studied (less than 10 modes), and the liquid-based TLWLs had the values of R from nearly 3 to 8 mm, which allowed us to get narrower lines than in Fig. 4.

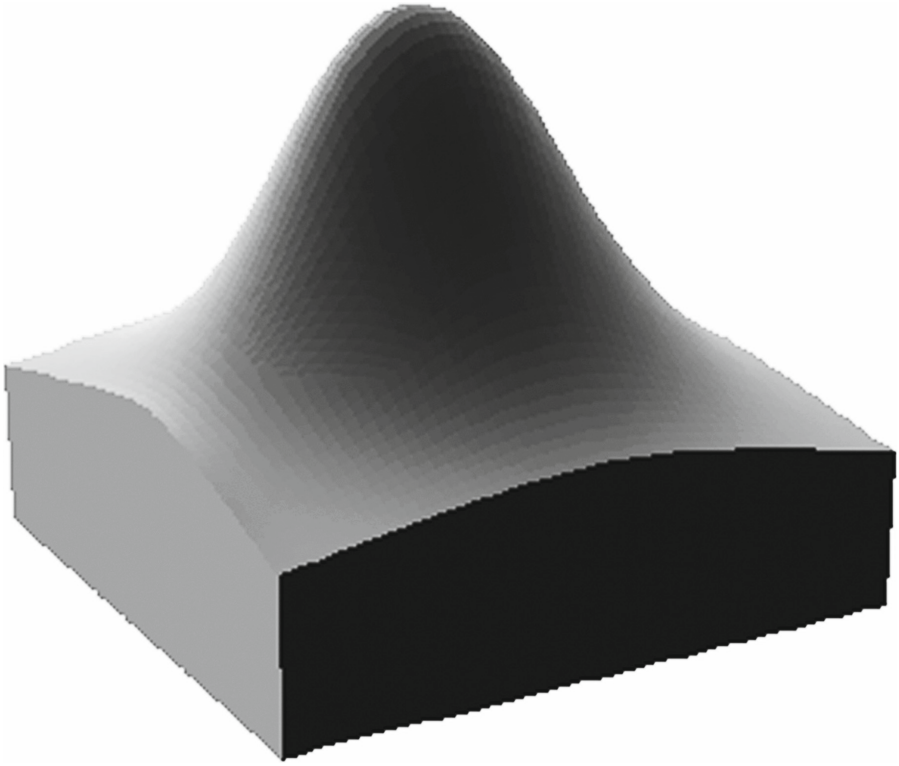


Fig. 3. Three-dimensional synthesised thickness profile $h(y, z)$ of the Luneburg TFGWL

To study the possible modes of the integrated optical waveguide, it is necessary to solve the appropriate equations of wave optics [7] together with the equation of director motion [18–20]. Then it is necessary to solve numerically the dispersion equations for TE and TM modes of the three-layer integrated optical waveguide. The numerical solution of the dispersion equations yields a set of permitted discrete values of the phase velocity slowing factor γ, corresponding to the guided modes of the waveguide. The calculations of dispersion dependences have shown that at the thickness of the waveguide layer $h \approx 25\ \mu$ in a three-layer waveguide up to 14 TE_m and TM_m modes can coexist (see Fig. 5), the curves for TE and TM modes being practically coincident. For comparison

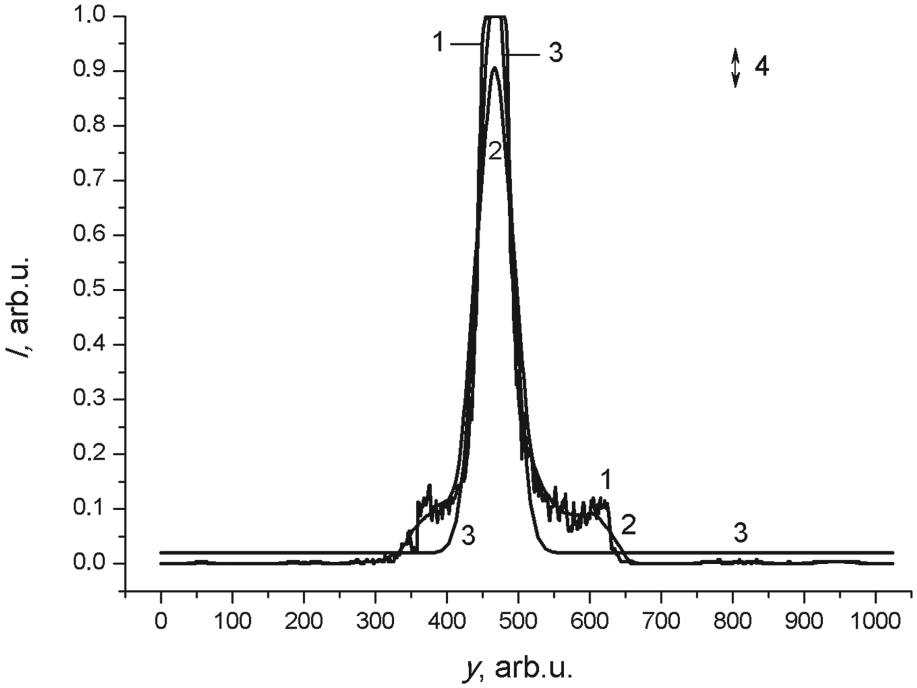


Fig. 4. Intensity profile of laser radiation behind the liquid crystal TLWL

let us present the results of numerical modelling for the Luneburg TFGWL with the radius $R = 5\,\text{mm}$ and the focal length $F = 7.5$ mm. The Luneburg TFGWL had the following parameters: the refractive index of the substrate (SiO_2): $n_s = 1.470$, the refractive index of the first (regular) waveguide layer (Corning 7059 glass) $n_f = 1.565$; the thickness of the regular waveguide layer $d \approx 0.96\,\mu\text{m}$; the refractive index of the second waveguide layer (Ta_2O_5) having the variable thickness $h\,(y, z)$ $n_l = 2.100$; the refractive index of the covering layer (air) $n_c = 1.000$. The parameters of the media forming the waveguide structure are given for the wavelength of laser radiation $\lambda = 0.9$ μm. We used the formulae presented in Ref. [14] and performed the calculations as described in Ref. [9]. The numerical results are presented in Figs. 6 and 7. After the passage through the Luneburg TFGWL, the (adiabatic) waveguide mode experiences an amplitude-phase transformation, so that the profile of the electromagnetic field distribution along the y-axis having the step-like shape at the lens input in the focal plane (along the y'-axis) acquires the form shown in Fig. 7.

It is seen that the Luneburg TFGWL with the given profile and focal length executes the necessary amplitude-phase transformation with superresolution, exceeding the classical diffraction limit. As seen from this characteristic, the considered solid-state Luneburg TFGWL excels the studied liquid crystal TLWLs.

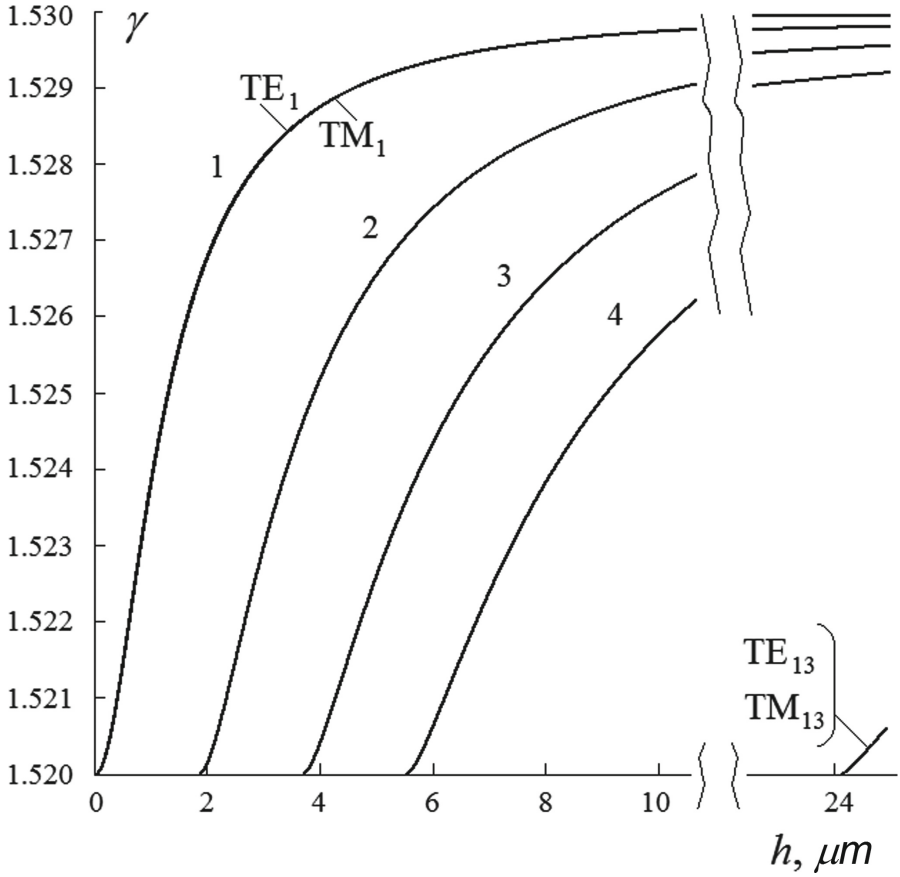


Fig. 5. Dispersion dependences for TE and TM modes of the NLC waveguide

Let us also present the calculated data on the dispersion dependences for the studied Luneburg TFGWL. The description of algorithms for computing the dispersion dependences is given in Refs. [7–14]. We recall that the substitution of the known field solutions in each layer of a multilayer waveguide into the boundary conditions yields a system of linear algebraic equations for the amplitude coefficients that determine the fields in the waveguide layers. A condition for the solution existence in a system of linear algebraic equations is that its determinant equals zero. For example, Fig. 8 presents one of the obtained typical dispersion dependences for the TM_0 mode of the studied smoothly irregular four-layer integrated optical structure. Thus, the left-hand part of Fig. 7 (approximately from 0.5 to 3.0) is the dispersion dependence of a three-layer regular waveguide, and the right-hand part (from 3.0 to 3.8, i.e., nearly to 3.4 μm) is the dispersion dependence for the four-layer smoothly irregular waveguide, including the Luneburg TFGWL.

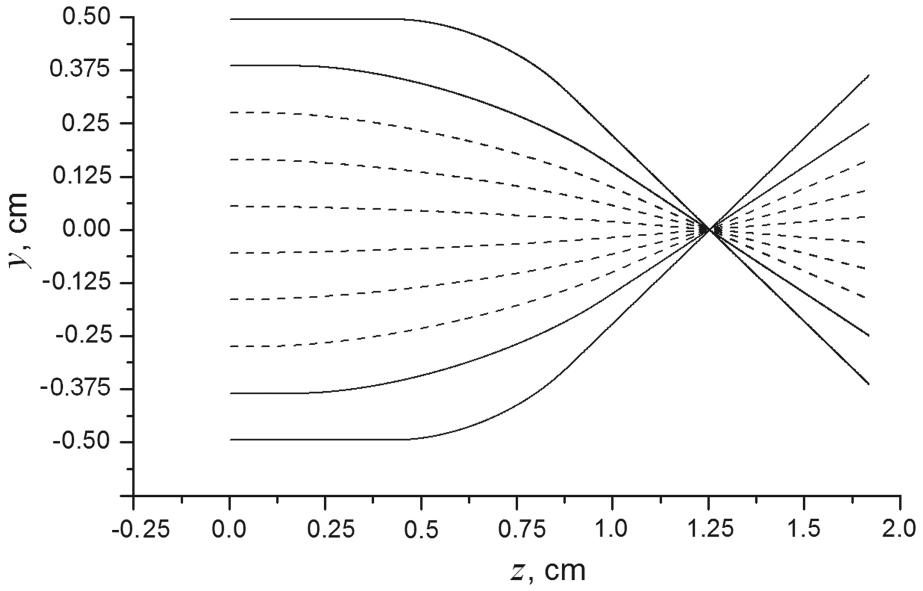


Fig. 6. Ray trajectories in the Luneburg TFGWL. The solid line corresponds to using 99% and the dash-dotted line to 60% of the lens aperture

The data on the electromagnetic field of the propagating waveguide mode obtained by numerical modelling allow the calculation of the full electromagnetic field distribution and the intensity in the vicinity of the focal points, located at the circle with the radius R, drawn around the centre of the Luneburg TFGWL. We should emphasise that the requirements to the accuracy of calculations strongly increase with the transition to the nanometre range due to the appearance of restrictions related to diffraction effects [7–14]. The latter strongly determine the accuracy of the Fourier transform performed by the lens and, as a consequence, the resolution, e.g., of a waveguide spectrum analyser entering an optical information system. In conclusion, note that our estimation of power losses in the considered Luneburg TFGWL has shown good agreement with the experimental data [10]: for the considered lens with the radius 1 cm the losses amount to about 0.9 dB, which corresponds to the power damping coefficient about $0.22\,\mathrm{cm}^{-1}$. With such damping coefficient the power losses of waveguide modes at such Luneburg TFGWL with full aperture will not exceed 20%. The sequential restriction of the lens aperture (at the edges, where the contribution of leaky modes is high) by 10 and 20% allows the reduction of power losses to 15 and 13%, respectively. In the case of the liquid-crystal TLWL, the losses can be a few times or even an order of magnitude higher (2–15 dB/cm), which should be taken into account in the design of optical information systems using such materials. In the cases of using aniline and glycerol as waveguide layers, the losses were comparable or smaller, than for the Luneburg TFGWL. It is important to emphasise that the damping of optical power, as well as the

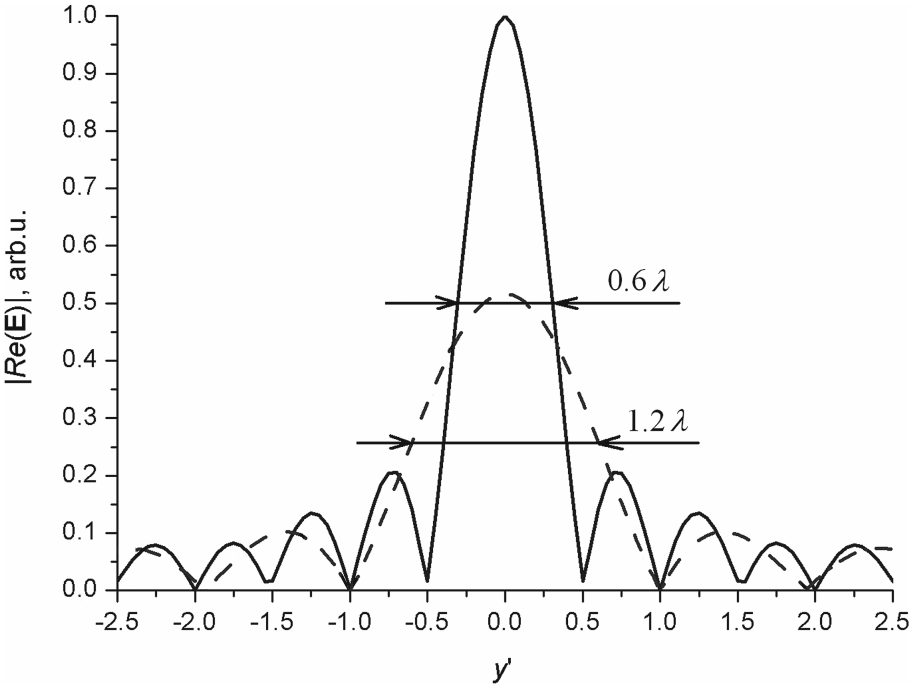


Fig. 7. Field distribution in the focal plane of the Luneburg TFGWL. The parameters are the same as in Fig. 6

half-width of the electromagnetic field distribution in the focal plane of the planar lens, is a key critical parameter restricting the dynamic range on the optical spectrum analyser. For example, for the damping 2.5 dB/cm the experimentally measured dynamic range amounts to about 17 dB [21,22]. Further expansion of the dynamic range of an integrated optical spectrum analyser is possible by reducing these two parameters. Thus, to reduce the optical power damping both in individual optical elements and in the entire information optical system (IOS), one has to use the materials weakly absorbing the radiation in the specified frequency range, having minimal losses due to the scattering by the material inhomogeneities and interface roughness, and minimal losses at the bendings of optical elements. It is also necessary to note that the minimisation of focal spot size is stimulated by the necessity to reduce the size of detector pixels forming a photodetector array of an integrated optical spectrum analyser, as well as the separation between them, in connection with the nanotechnology development in these fields. Undoubtedly, all this essentially increases the requirements to both the stage of computer simulations and computer-aided design (analysis and synthesis) of optical components necessary for normal IOS operation, and the stage of the further fabrication technology.

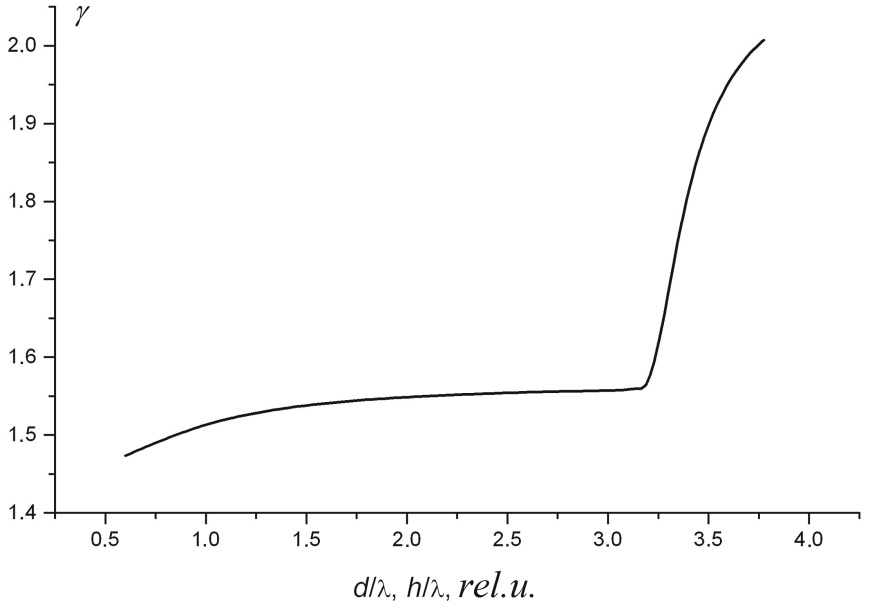


Fig. 8. Dispersion dependences for the TM_0 mode in the integrated optical four-layer structure presented in Fig. 2

One of the applications of this kind of waveguide optical lenses could be connecting the output of optical circuits written in glasses for quantum information purposes (see, e.g., [23,25]) to an array of fibres or to a sole fibre. The goal is to increase the efficiency of modes transmission in the coupling between the facet of the glass and the facet of the fibre array or a single fibre, which connect one circuit to another. Since the refractive index of the written waveguides in the bulk of the glass is different from that of the substrate just by 10^{-3}, some field of the transmitted mode is propagating in the surrounding air above the sample, where a waveguide lens can be added. The samples generally used for quantum optics have a thickness of about 1 mm. This kind of lenses can concentrate the laser radiation into the core of the fibres. The same could be done when the output of the quantum information circuits written in the bulk of a glass sample have to connect a detector, especially in the case of single photons carrying information that have to be decrypted and a high visibility is needed.

The final stage of IOS construction is the dynamic analysis of functioning of the optical components, e.g., under the conditions of static polarisation fluctuation, when the signals propagate through distributed optical systems [24], including those in the presence of noise [7], by means of computer modelling.

References

1. ISO/IEC 2382:2015 Information technology - Vocabulary. https://www.iso.org/obp/ui/#iso:std:iso-iec:2382:ed-1:v1:en
2. Roth, B.: Integrated planar-optical networks in thin polymer foils - a new approach to large-area distributed sensing. In: Lasers and Electro-Optics Europe & European Quantum Electronics Conference. CLEO/Europe-EQEC (2017). https://doi.org/10.1109/CLEOE-EQEC.2017.8087744
3. Braun, S., Meng, X.M.: Advanced optical network. Fiber Integr. Opt. **25**(4), 257–278 (2006). https://doi.org/10.1080/01468030600692818
4. Egorov, A.A., Egorov, M.A., Tsareva, Y.I., Chekhlova, T.K.: Study of the integrated-optical concentration sensor for gaseous substances. Laser Phys. **17**(1), 50–53 (2007)
5. Egorov, A.A., Egorov, M.A., Chekhlova, T.K., Timakin, A.G.: Application of integrated optical sensors for the control of dangerous gaseous substances. Datchiki i sistemy, no. 1, p. 25–28 (2008) (in Russian)
6. Egorov, A.A., Egorov, M.A., Stavtsev, A.V., Timakin, A.G., Chekhlova, T.K.: A fast integrated optical sensor of gaseous substances. J. Russ. Laser Res. **31**(1), 12–21 (2010)
7. Egorov, A.A., Lovetskiy, K.P., Sevastyanov, A.L., Sevastyanov, L.A.: Intgral Optics: Theory and Computer Modelling. A monograph. Izdatel'stvo RUDN, Moscow (2015). (in Russian)
8. Egorov, A.A., Sevastyanov, L.A.: Mode structure of smoothly irregular integrated optical four-layer three-dimensional waveguide. Quant. Electron. **39**(6), 566–574 (2009)
9. Egorov, A.A., Sevastyanov, A.L., Ayryan, E.A., Lovetskiy, K.P., Sevastyanov, L.A.: Adiabatic modelling of smoothly irregular optical waveguide: zero-order approximation of vector theory. Mat. Model. **22**(8), 42–54 (2010). (in Russian)
10. Egorov, A.A., Lovetskiy, K.P.: Modelling of guided modes (eigenmodes) and synthesis of thin film generalised waveguide Luneburg lens in the zero-order vector approximation. Quant. Electron. **40**(9), 830–836 (2010)
11. Ayryan, E.A., Egorov, A.A., Sevastyanov, L.A., Lovetskiy, K.P., Sevastyanov, A.L.: Mathematical modeling of irregular integrated optical waveguides. In: Adam, G., Buša, J., Hnatič, M. (eds.) MMCP 2011. LNCS, vol. 7125, pp. 136–147. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28212-6_12
12. Sevastianov, L.A., Sevastyanov, A.L., Egorov, A.A.: Method of adiabatic modes in studying problems of smoothly irregular open waveguide structures. Phys. At. Nucl. **76**(2), 224–239 (2013)
13. Egorov, A.A., Sevast'yanov, L.A., Sevast'yanov, A.L.: Method of adiabatic modes in research of smoothly irregular integrated optical waveguides: zero approximation. Quant. Electron. **44**(2), 167–173 (2014)
14. Egorov, A.A., Sevastyanov, A.L., Ayryan, E.A., Sevastyanov, L.A.: Stable computer modelling of thin film generalised waveguide Luneburg lens. Mat. Model. **26**(11), 37–44 (2014). (in Russian)
15. Sevastyanov, L.A., Sevastyanov, A.L., Tyutyunnik, A.A.: Analytical calculations in maple to implement the method of adiabatic modes for modelling smoothly irregular integrated optical waveguide structures. In: Gerdt, V.P., Koepf, W., Seiler, W.M., Vorozhtsov, E.V. (eds.) CASC 2014. LNCS, vol. 8660, pp. 419–431. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10515-4_30

16. Gusev, A.A., Gerdt, V.P., Hai, L.L., Derbov, V.L., Vinitsky, S.I., Chuluunbaatar, O.: Symbolic-numeric algorithms for solving BVPs for a system of ODEs of the second order: multichannel scattering and eigenvalue problems. In: Gerdt, V.P., Koepf, W., Seiler, W.M., Vorozhtsov, E.V. (eds.) CASC 2016. LNCS, vol. 9890, pp. 212–227. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-45641-6_14
17. Wang, T.-J., Yang, S.-C., Chen, T.-J., Chen, B.-Y.: Wide tuning of SiN microring resonators by auto-realigning nematic liquid crystal. Opt. Express **20**(14), 15853–15858 (2012)
18. Egorov, A.A., Maslyanitsyn, I.A., Shigorin, V.D., Ayryan, A.S., Ayryan, E.A.: Study of the effect of pulsed-periodic electric field and linear polarisation of laser radiation on the properties of NLS waveguide. In: Proceedings of V International Conference on Problems of Mathematical and Theoretical Physics and Mathematical Modelling, 5–7 April 2016, pp. 51–53. MEPhI, Moscow (2016). (in Russian)
19. Egorov, A.A.: Influence of fluctuations of local orientation of nematic liquid crystal molecules on the coefficient of damping of waveguide modes. In: Proceedings of the VI International conference on Problems of Mathematical Physics and Mathematical Modelling, 25–27 May 2017, pp. 128–129. MEPhI, Moscow (2017)
20. Blinov, L.M.: Electro- and Magnetooptics of Liquid Crystals. Nauka, Moscow (1978). (in Russian)
21. Boyd, J.T., Anderson, D.B.: Effect of waveguide optical scattering on the integrated optical spectrum analyser dynamic range. IEEE J. QE **14**(6), 437–443 (1978)
22. Egorov, A.A.: Effect of quality of substrate surface processing and imperfection of waveguide layers on some characteristics of optical integrated circuits. Science and Techology Conference on Optical Commutation and Optical Communication Networks. Book of Abstracts, pp. 33–34. TsNIIS, Moscow (1990)
23. Liu, J.-M.: Photonic Devices. Cambridge University Press, Cambridge (2005)
24. Czegledi, C.B., Karlsson, M., Agrell, E., Johannisson, P.: Polarization Drift Channel Model for Coherent Fibre-Optic Systems // Scientific Reports. **6**, 21217 (2016). https://doi.org/10.1038/srep21217
25. Nanophotonics / Ed. Rigneault H., Lourtioz J.-M., Delalande C., Levenson A. London: ISTE Ltd., 2006

Stability of a Two-Pool N-Model with Preemptive-Resume Priority

Evsey Morozov[1,2(✉)]

[1] Institute of Applied Mathematical Research of the Karelian Research Centre of the Russian Academy of Sciences, Pushkinskaya st. 11, Petrozavodsk, Russia
[2] Petrozavodsk State University, Lenin str. 33, Petrozavodsk, Russia
emorozov@karelia.ru

Abstract. The regenerative methodology is applied to find stability conditions of the so-called N-model which consists of two pools of the interacting servers with two classes of external customers following a renewal input. Service times are assumed to be pool-dependent and, in each pool, are i.i.d. with a general distribution. If the queue size in pool 1 exceeds a given threshold, then a class-1 customer jumps to pool 2 and becomes class-(1,2) customer with the preemptive-resume priority. The stability analysis of this model has been developed in [5] by a modified fluid approach. In addition to the results obtained in work [5], we find the conditions when the 1st pool is stable solely, and when the 1st pool is stable, while the 2nd pool is unstable.

Keywords: Stability analysis · Two-pool N-model · Regeneration
Preemptive - resume priority

1 Introduction

We consider a two-pool queueing system with N_1 and N_2 parallel servers, respectively and infinite-capacity buffers. The system is fed by a general renewal input with the rate λ. With a probability p_i, a new customer is class-i and directed to pool i, regardless of the state of system, $p_1 + p_2 = 1$. We denote $Q_i(t)$ the *queue size* (the number of customers waiting service) in pool i at instant t^-, $i = 1, 2$. Each server in pool 2, when free at an instant t, accepts class-1 customer, provided $Q_1(t) > K_1$, where $K_1 \geq 0$ is a given threshold (constant). That is, if $Q_1(t) > K_1$, then a class-1 customer jumps to pool 2, becomes a *class-(1,2) customer* and starts service immediately, provided the number of class-(1,2) customers in pool 2 is less than N_2. Thus, we apply *preemptive class-(1,2) priority* in pool 2, provided $Q_1(t) > K_1$. However, the opposite jumps (from pool 2 to pool 1) are forbidden. We assume an arbitrary *work-conserving service discipline* in each pool, in particular, an *arbitrary waiting class-1 customer* may jump to pool 2 (provided $Q_1(t) > K_1$), because it is unimportant for the stability analysis. This model is a variation of the single-server $N-$model from [2] in which server 2 accepts class-1 customers, when free, but gives preemptive priority to

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 399–409, 2018.
https://doi.org/10.1007/978-3-319-99447-5_34

class-2 customers. By this reason, stability conditions of both models are quite different. Following [5], we call this model *N-model with static priority*. In this N-model, which is very well motivated, see again [5] and also [6,7], pool 1 can be treated as *beneficiary*, while pool 2 is the *donor*. This research is motivated by the work [5], where a modified fluid approach has been applied to develop stability analysis of a similar model with non-preemptive priority. The main contribution of the present work is that we develop regenerative stability analysis of N-model with no extra "Assumption 1" which is used in [5]. Our analysis allows to establish stability condition of the 1st pool solely and also to distinguish the transience and null recurrent conditions for each pool. Note that the latter analysis is impossible in the framework of the fluid stability analysis. We apply regenerative method which is very effective and powerful tool of stability analysis of a wide class of the queueing systems [3]. A novelty of this work is also that it shows that the regenerative stability analysis of a complex queueing system with interacting components can be based on simple balance relations (like (1), (11) below). This approach considerably extends previously developed method [3].

The paper is organized as follows. In Sect. 2, using regenerative approach, we prove Theorem 1 which contains necessary stability condition of the N-model under consideration. Then, in Sect. 3, we first find sufficient stability condition of the 1st pool solely, called partial stability (Sect. 3.1), and then find sufficient stability condition of the entire system (Sect. 3.2). Indeed we prove that necessary condition coincides with sufficient condition, implying stability criterion. This criterion is consistent with what has been proved by a modified fluid analysis in [5] for N-model with non-preemptive priority.

2 Necessary Stability Conditions

We assume that the service times of class-i customers $\{S_k^{(i)}, k \geq 1\}$ are i.i.d. with rate $\mu_i = 1/\mathsf{E}S^{(i)} \in (0, \infty)$, $i = 1, 2, (1, 2)$. (In what follows, we omit serial index to denote a generic element of an i.i.d sequence.) All sequences are assumed to be independent. By construction, the external input in pool i is renewal with rate $\lambda_i = \lambda\, p_i$, $i = 1, 2$.

2.1 Stability of the 1st Pool

Because class-(1,2) customers have the preemptive (absolute) priority in pool 2, a class-(1,2) customer starts service immediately after jump to pool 2. It is worth mentioning that an isolated pool 1 is a *regenerative* N_1-server *infinite-buffer* $GI/G/N_1$-type system with a generic regeneration period length T_1. The regenerations of pool 1 are generated by class-1 customers arriving to an empty pool. Denote $V_1(t)$, $B_1(t)$ the arrived and departed work, respectively, in interval $(0, t]$. Note that $B_1(t) = \sum_{i=1}^{N_1} B_1^{(i)}(t)$, where $B_1^{(i)}(t)$ is the busy time of server i in interval $(0, t]$, $i = 1, \ldots, N_1$. Also denote $L_1(t)$ the work *lost by pool 1* in interval in $(0, t]$. In other words, it is the summary (not realized) work in pool 1

of class-(1,2) customers. Denote $A_1(t)$ the number of class-1 arrivals and $A_{12}(t)$ the number of class-(1,2) customers in interval $(0,\, t]$. Let $\hat{A}_{12}(t)$ be the number of class-(1,2) customers *among all arrivals* $A_1(t)$, so $\hat{A}_{12}(t) \geq A_{12}(t)$. Denote $W_1(t)$ the remaining work (workload) at instant t^+, and let $\rho_i = \lambda_i/\mu_i$, $i = 1, 2$.

To find the necessary stability condition of the 1st pool, we assume that it is stable (*positive recurrent*), that is $\mathsf{E}T_1 < \infty$ [3]. Now we obtain an important relation between the stationary number of the busy servers in pool 1 and the stationary probability that a class-1 customer jumps to pool 2. Note that $B_1^{(i)}(t) = t - I_1^{(i)}(t)$, where $I_1^{(i)}(t)$ is an empty time of server i (in pool 1) in interval $[0,\, t)$, $i = 1, \ldots, N_1$. We have the following balance equation for any $t \geq 0$:

$$V_1(t) = W_1(t) + L_1(t) + B_1(t) = W_1(t) + L_1(t) + N_1 t - I_1(t), \tag{1}$$

where $I_1(t) := \sum_{i=1}^{N_1} I_1^{(i)}(t)$ is the summary idle time of all servers in interval $[0,\, t]$. By the Strong Law of Large Numbers (SLLN), with probability (w.p.) 1,

$$\frac{V_1(t)}{t} = \frac{\sum_{k=1}^{A_1(t)} S_k^{(1)}}{A_1(t)} \, \frac{A_1(t)}{t} \to \rho_1, \quad t \to \infty. \tag{2}$$

Denote $\mathcal{L}(t)$ the set of numbers of class-1 customers which jump to pool 2 in interval $[0,\, t)$, so $\mathcal{L}(t) \subseteq \{1, \ldots, A_1(t)\}$. We assume that the service times $S_k^{(1)}$ of the customers jumping to pool 2 are assigned *after the jump*. Then, by assumption, the jump of a class-1 customer does not dependent on its service times in both pools. (The corresponding service times $S_k^{(1)}$ of class-1 customers which jump in pool 2 are lost for pool 1.) Then the summary (potential) work related to these customers,

$$L_1(t) = \sum_{k \in \mathcal{L}(t)} S_k^{(1)}, \tag{3}$$

is *independent of the summands* $S_k^{(1)}$ in (3). As a result, the random set $\mathcal{L}(t)$ and its capacity $|\mathcal{L}(t)| =: A_{12}(t)$ are also independent of the summands. Now, by the SLLN, w.p.1 as $t \to \infty$,

$$\frac{1}{t} L_1(t) = \frac{1}{t} \sum_{k \in \mathcal{L}(t)} S_k^{(1)} = \frac{\sum_{k \in \mathcal{L}(t)} S_k^{(1)}}{A_{12}(t)} \, \frac{A_{12}(t)}{A_1(t)} \, \frac{A_1(t)}{t} \to \rho_1 \mathsf{P}_\ell, \tag{4}$$

where the limit

$$\mathsf{P}_\ell := \lim_{t \to \infty} \frac{A_{12}(t)}{A_1(t)} \tag{5}$$

exists and is the stationary probability (more exactly, the limiting fraction) a class-1 customer jumps from pool 1 to pool 2. Note that, under preemptive discipline, the probability P_ℓ does not depend on the dynamics of the 2nd pool.

Denote indicator $1_i(t) = 1$, if server i is free at instant t, and $1_i(t) = 0$, otherwise, and let $\mathsf{P}_b^{(i)}$ be the stationary busy probability of server i. By the positive recurrence of the *cumulative process* $\{B_1(t),\, t \geq 0\}$, the limit w.p.1

$$\lim_{t\to\infty} \frac{B_1(t)}{t} = \lim_{t\to\infty} \frac{1}{t} \int_0^t \sum_{i=1}^{N_1} 1_i(u)du = \sum_{i=1}^{N_1} \mathsf{P}_b^{(i)} := \mathcal{B}_1, \tag{6}$$

exists and is the *mean stationary number of the busy servers in the 1st pool*. If servers are (stochastically) equivalent, then $\mathsf{P}_b^{(i)} \equiv \mathsf{P}_b$, implying $\mathcal{B}_1 = N_1\, \mathsf{P}_b$. It is well known that in the positive recurrent system [8],

$$W(t) = o(t),\, t \to \infty \quad \text{w.p.1},$$

and (1)–(6) imply the following important relation,

$$\mathsf{P}_\ell = 1 - \frac{\mathcal{B}_1}{\rho_1}. \tag{7}$$

Assume the generic interarrival time τ and service times of the external customers satisfy conditions

$$\mathsf{P}(\tau > S^{(i)}) > 0,\, i = 1, 2. \tag{8}$$

Then the mean summary idle time $\mathsf{E}I_0^{(1)}$ of all servers in pool 1, within regeneration period, is positive:

$$\mathsf{E}I_0^{(1)} \geq \mathsf{E}\Big(I_0^{(1)} | \tau \geq S^{(1)} + \delta\Big)\mathsf{P}(\tau \geq S^{(1)} + \delta) \geq \delta \mathsf{P}(\tau \geq S^{(1)} + \delta) > 0,$$

for some $\delta > 0$. Because the idle time $\{I_1(t),\, t \geq 0\}$ is positive recurrent *cumulative process*, then the positive limit

$$\lim_{t\to\infty} \frac{I_1(t)}{t} = \frac{\mathsf{E}I_0^{(1)}}{\mathsf{E}T} > 0, \tag{9}$$

exists [8]. Obviously, the following balance equation holds:

$$N_1 t = \mathcal{B}_1(t) + I_1(t),\; t \geq 0.$$

Then it follows from (6) and (9) that $\mathcal{B}_1 < N_1$. Now (1)–(7) give

$$\rho_1 = \rho_1 \mathsf{P}_\ell + N_1 - \frac{\mathsf{E}I_0^{(1)}}{\mathsf{E}T} < \rho_1 \mathsf{P}_\ell + N_1.$$

Thus the *necessary stability condition of the 1st pool* is

$$\rho_1(1 - \mathsf{P}_\ell) < N_1. \tag{10}$$

Note that the probability P_ℓ is analytically unavailable in the most of cases, and simulation remains the only way to estimate it. However this estimation faces with a difficulty because the stationary regime is assumed in advance. In part this problem has been addressed in [5] and in [4] for the non-preemptive priority.

2.2 Stability of Two-Pool System

Now we present the necessary stability conditions of the original two-pool system. Let now T be the generic regeneration period length of this system generated by an arrival to an empty system. More precisely, we distinguish two type of the regenerations: generated by class-1 customers arriving in an empty system with generic regeneration period length T_1; and by class-2 customers arriving in an empty system, with generic regeneration period length T_2. Assume the indicator $1_1 = 1$ if a class-1 customer starts regeneration period, and $1_1 = 0$, otherwise. Then

$$T =_{st} T_1 1_1 + T_2(1 - 1_1),$$

where $=_{st}$ means stochastic equality, so $\mathsf{E}T = p_1\mathsf{E}T_1 + p_2\mathsf{E}T_2$. Assume the positive recurrence (stability), that is $\mathsf{E}T < \infty$, and write down the following balance equation for the work arriving in pool 2 in interval $(0, t]$:

$$V_2(t) + V_{12}(t) = W_2(t) + B_2(t) = W_2(t) + N_2 t - I_2(t), \tag{11}$$

where $W_2(t)$ is the current workload in pool 2 at instant t, $B_2(t) = N_2 t - I_2(t)$ is the accumulated busy time in pool 2, $I_2(t)$ is the summary idle time of all server in pool 2, and $V_{12}(t)$ is the work arrived in pool 2 from pool 1, in interval $(0, t]$. Note that

$$V_{12}(t) = \sum_{k \in \mathcal{L}(t)} S_k^{(12)}.$$

We note that if the generic service times of class-1 and class-(1,2) customers equal, $S^{(1)} =_{st} S^{(12)}$, then $V_{12}(t) =_{st} L_1(t)$, but in general $V_{12}(t) \neq_{st} L_1(t)$.

By the positive recurrence, $W_2(t) = o(t)$, $t \to \infty$ [8], and as above, we obtain that the following limits w.p.1 exist

$$\lim_{t\to\infty} \frac{V_{12}(t)}{t} = \frac{\lambda_1}{\mu_{12}} \mathsf{P}_\ell, \tag{12}$$

$$\lim_{t\to\infty} \frac{V_2(t)}{t} = \rho_2, \tag{13}$$

$$\lim_{t\to\infty} \frac{I_2(t)}{t} = \frac{\mathsf{E}I_0^{(2)}}{\mathsf{E}T} > 0, \tag{14}$$

where $I_0^{(2)}$ is the summary idle time of all servers of pool 2 within regeneration period, and we apply (8) to prove strict inequality in the last relation (14), cf. (9). Using (7) we obtain from (11), (12)–(14) the *necessary stability condition of the entire system*:

$$\rho_2 + \frac{\lambda_1}{\mu_{12}} \mathsf{P}_l = \rho_2 + \frac{\lambda_1 - \mu_1 \mathcal{B}_1}{\mu_{12}} < N_2. \tag{15}$$

Finally we note that previous analysis holds true for arbitrary initial states $W_1(0) = x_1$, $W_2(0) = x_2$, the amount of work present in the system at instant 0.

We summarize previous analysis as the following statement.

Theorem 1. Consider the two-pool N-model with the preemptive-resume static priority discipline under assumption (8) and with arbitrary initial state $W_1(0) = x_1$, $W_2(0) = x_2$. Then, if the 1st pool is positive recurrent, then condition (10) holds. If the 2nd pool is positive recurrent, then the 1st pool is positive recurrent as well and both conditions (15), (10) hold.

3 Sufficient Stability Conditions

To find *sufficient* stability conditions, we first assume that the 1st pool is *not positive recurrent*, that is $\mathsf{E}T_1 = \infty$. Note that the case $\mathsf{E}T_1 = \infty$ includes (i) null-recurrence, $\mathsf{P}(T_1 < \infty) = 1$, and (ii) transience, $\mathsf{P}(T_1 < \infty) < 1$. In this case, using assumption (8), we easily obtain by a contradiction that $\mathsf{P}(Q_1(t) \geq k) \to 1$, $k \to \infty$, for more detail see [3]. Thus, we assume that

$$\mathsf{P}(Q_1(t) > K_1) \to 1,\ t \to \infty. \tag{16}$$

Denote

$$Z(t) = \int_0^t \mathsf{1}(Q_1(u) > K_1)du,$$

the time the 1st pool *can generate* the input to the 2nd pool in interval $(0, t]$, which is it only possible when the number of class-(1,2) customers in pool 2 is less than N_2. By (16),

$$\lim_{t\to\infty} \frac{1}{t}\mathsf{E}Z(t) = \lim_{t\to\infty} \frac{1}{t} \int_0^t \mathsf{P}(Q_1(u) > K_1)du = 1. \tag{17}$$

We expect that class-(1,2) customers completely capture the 2nd pool, and the stream of class-(1,2) customers to pool 2 approaches the superposed renewal process with the summary rate $N_2\mu_{12}$.

To prove it, we consider N_2 independent (stochastically) equivalent iid sequences $\{S_i^{(12)}(k), k \geq 1\}$, $i = 1, \ldots, N_2$, distributed, for all i, k, as service time $S^{(12)}$ of class-(1,2) customers. Define (zero-delayed) renewal processes

$$\hat{A}_{12}^{(i)}(t) = \min(k : S_i^{(12)}(1) + \cdots + S_i^{(12)}(k) \geq t),\ t \geq 0, \tag{18}$$

so in particular, $\hat{A}_{12}^{(i)}(0) = 1$, $i = 1, \ldots, N_2$. For each t, $\hat{A}_{12}^{(i)}(t)$ is the number of customers which could be served by server i in interval $(0, t]$, provided all time is devoted to these customers. Denote $\hat{A}_{12}(t) = \sum_{i=1}^{N_2} \hat{A}_{12}^{(i)}(t)$, then by the elementary renewal theorem,

$$\lim_{t\to\infty} \frac{\mathsf{E}\hat{A}_{12}(t)}{t} = N_2\mu_{12}. \tag{19}$$

Denote $A_{12}(t)$ the real number of class-(1,2) customers in interval $(0, t]$. Now, using a coupling, we assign service time $S_i^{(12)}(k)$, realized in the process $\hat{A}_{12}^{(i)}(t)$, $t \geq 0$, for the kth class-(1,2) customer served by server i in the 2nd

pool. Obviously, $\hat{A}_{12}(t) \geq A_{12}(t)$, $t \geq 0$. To obtain a lower bound for $A_{12}(t)$, we introduce $U_i(t)$, the time when server i of the 2nd pool does not serve class-(1,2) customers in interval $(0, t]$. It is easy to see that

$$U_i(t) \leq Y(t) := \int_0^t 1(Q_1(u) \leq K_1)du = t - Z(t), \;\; i = 1, \ldots, N_2. \tag{20}$$

Now, for given t and each server i, we past together all time periods, in interval $(0, t]$, when server i is occupied by class-(1,2) customer, and then shift this merged (busy) interval to the origin. It is easy to verify that the magnitude of this shift is exactly $U_i(t)$. By construction of coupling, the service times of class-(1,2) customers $S_i^{(12)}(j)$ realized to build the process (18) in interval $(t-U_i(t), t]$ are independent of $U_i(t)$, $i = 1, \ldots, N_2$. Now we obtain the following two-sided bounds

$$\begin{aligned} \hat{A}_{12}(t) \quad &\geq A_{12}(t) = \sum_{i=1}^{N_2} \hat{A}_{12}^{(i)}(t - U_i(t)) \\ &\geq_{st} \hat{A}_{12}(t) - \sum_{i=1}^{N_2} \tilde{A}_{12}^{(i)}(U_i(t)) - N_2, \end{aligned} \tag{21}$$

where, for each i, the term $\tilde{A}_{12}^{(i)}(U_i(t))$ denotes the number of the service times of class-$(1, 2)$ customers realized in the process (18) in interval $(t - U_i(t), t]$ and *independent* of $\hat{A}_{12}^{(i)}(t)$. On the other hand, by (17), (20),

$$\lim_{t\to\infty} \frac{1}{t} \mathsf{E}U_i(t) \to 0, \;\; i = 1, \ldots, N_2, \tag{22}$$

and, by the mentioned independence, for all $t \geq 1$,

$$\begin{aligned} &\int_1^t \frac{\mathsf{E}\tilde{A}_{12}^{(i)}(u)}{u} u\mathsf{P}(U_i(t) \in du) \leq \mathsf{E}\tilde{A}_{12}^{(i)}(U_i(t)) \\ &\leq \mathsf{E}\tilde{A}_{12}^{(i)}(1) + \int_1^t \frac{\mathsf{E}\tilde{A}_{12}^{(i)}(u)}{u} u\mathsf{P}(U_i(t) \in du). \end{aligned} \tag{23}$$

Because

$$\mathsf{E}\tilde{A}_{12}^{(i)}(1) = o(t), \;\; \lim_{u\to\infty} \frac{\mathsf{E}\tilde{A}_{12}^{(i)}(u)}{u} = \mu_{12},$$

then we obtain from (22), (23) that

$$\frac{1}{t}\mathsf{E}\tilde{A}_{12}^{(i)}(U_i(t)) \sim \mu_{12}\frac{\mathsf{E}U_i(t)}{t} \to 0, \;\; t \to \infty, \;\; i = 1, \ldots, N_2.$$

($a \sim b$ means $a/b \to 1$). It gives, by (19), (21),

$$\lim_{t\to\infty} \frac{\mathsf{E}A_{12}(t)}{t} = N_2\mu_{12}. \tag{24}$$

By (16),

$$\mathsf{E}I_1(t) \le N_1 \int \mathsf{P}(Q_1(u) = 0)du = o(t),\ t \to \infty,$$

and thus, by (1), (2), the stationary average number of busy servers in the 1st pool is

$$\lim_{t\to\infty} \frac{\mathsf{E}B_1(t)}{t} = N_1 = \mathcal{B}_1. \tag{25}$$

We note that (25) means that, in the limit, all servers of pool 1 are occupied by class-1 customers. Because the number of summands $|\mathcal{L}(t)| = A_{12}(t)$ in (3) is independent of the summands $\{S_k^{(1)}\}$, then by the Wald's identity and (24),

$$\frac{\mathsf{E}L_1(t)}{t} = \frac{1}{t}\mathsf{E}\Big\{\sum_{k=1}^{A_{12}(t)} S_k^{(1)}\Big\} = \frac{1}{\mu_1}\frac{\mathsf{E}A_{12}(t)}{t} \to \frac{N_2\mu_{12}}{\mu_1},\ t \to \infty. \tag{26}$$

Rewriting (1) as

$$W_1(t) = V_1(t) - L_1(t) - N_1 t + I_1(t) \ge 0,$$

using (2), (25), (26) and relation $N_1 t - B_1(t) = I_1(t)$, we obtain

$$\lim_{t\to\infty} \frac{1}{t}\mathsf{E}\Big[V_1(t) - L_1(t) - N_1 t + I_1(t)\Big] = \rho_1 - \frac{N_2\mu_{12}}{\mu_1} - N_1 \ge 0.$$

Thus the *necessary instability condition* of the 1st pool is

$$\lambda_1 \ge N_1\mu_1 + N_2\mu_{12}. \tag{27}$$

As a result, under condition

$$\lambda_1 < N_1\mu_1 + N_2\mu_{12}, \tag{28}$$

the following inequality

$$\inf_j \mathsf{P}(Q_1(z_j) \le C) \ge \delta$$

holds for a deterministic sequence $z_j \to \infty$ and some constants $C < \infty$ and $\delta > 0$. Then the positive recurrence of the 1st pool follows by the standard method, see [3]. To outline the proof of this part, we define $T_1(t)$, the remaining regeneration time of the process Q_1 at instant t. In the proof we show, using assumption (8), that the queue size $Q_1(t)$ hits a regeneration point with a positive probability q within a finite interval of time $[z_j,\, z_j + C_0]$. At that, neither q nor the interval length C_0 depend on instant z_j and j. Equivalently, $T_1(t) \not\Rightarrow \infty$ (in probability). The latter result implies $\mathsf{E}T_1 < \infty$, and thus (28) is *sufficient stability condition* of the regenerative process $Q_1(t)$, $t \ge 0$ [3].

3.1 Partial Stability

Now we find conditions when the 1st pool remains *stable* (under condition (28)), while the 2nd pool is not positive recurrent, meaning that for each k, $\mathsf{P}(Q_2(t) > k) \to 1,\ t \to \infty$. Then, in particular,

$$\mathsf{E}I_2(t) \le N_2 \int_0^t \mathsf{P}(Q_2(u) = 0)du = o(t),\ t \to \infty. \tag{29}$$

It follows from (11), that

$$W_2(t) = V_2(t) + V_{12}(t) - N_2 t + I_2(t) \ge 0. \tag{30}$$

By the positive recurrence of the 1st pool, the limit w.p.1,

$$\frac{V_{12}(t)}{t} = \frac{\sum_{k\in\mathcal{L}(t)} S_k^{(12)}}{A_{12}(t)} \frac{A_{12}(t)}{A_1(t)} \frac{A_1(t)}{t} \to \rho_1 \mathsf{P}_\ell,\ \ t \to \infty, \tag{31}$$

exists, where, recall,

$$\mathsf{P}_\ell = \lim_{t\to\infty} \frac{A_{12}(t)}{A_1(t)} = \lim_{t\to\infty} \frac{\mathsf{E}A_{12}(t)}{\mathsf{E}A_1(t)},$$

and the 2nd limit exists by the elementary renewal theorem, $\lim \mathsf{E}A_1(t)/t = \lambda_1$ (or, alternatively, by the limit theorem for the positive recurrent cumulative processes, [8]). Although the limit P_ℓ is now obtained under the assumption that $\mathsf{P}(Q_2(t) > k) \to 1$ for any k, it is the same as in (5), because the state of the 1st pool is independent of the state of the 2nd pool. Now, by (29)–(31), we obtain the following *necessary instability condition* of the 2nd pool:

$$\lim_{t\to\infty} \frac{\mathsf{E}W_2(t)}{t} = \rho_2 + \frac{\lambda_1}{\mu_{12}} \mathsf{P}_\ell = \rho_2 + \frac{\lambda_1 - \mu_1 \mathcal{B}_1}{\mu_{12}} - N_2 \ge 0. \tag{32}$$

Now we show that the inequalities (32) and (28) indeed coexist. Really, as $\lambda_1 \uparrow N_1\mu_1 + N_2\mu_{12}$, the 1st pool remains stable, while the inequality (32) becomes, in the limit (with $\lambda_1 = N_1\mu_1 + N_2\mu_{12}$) the following strict inequality:

$$\rho_2 + N_2 + \frac{N_1\mu_1}{\mu_{12}} > \frac{\mu_1 \mathcal{B}_1}{\mu_{12}} + N_2. \tag{33}$$

Thus there is the following *non-empty region* of parameter λ_1,

$$\mathcal{B}_1\mu_1 + N_2\mu_{12} - \rho_2\,\mu_{12} \le \lambda_1 < N_1\mu_1 + N_2\mu_{12}, \tag{34}$$

where the 2nd pool is *not positive recurrent*, while the 1st pool is positive recurrent.

Assume the 2nd pool is *not positive recurrent*. It is worth mentioning that we can distinguish *null-recurrence* of the 2nd pool, when equality in (32) holds, and the *transience (strong instability)* of pool 2, when

$$\lim_{t\to\infty} \frac{\mathsf{E}W_2(t)}{t} = \rho_2 + \frac{\lambda_1}{\mu_{12}} \mathsf{P}_\ell - N_2 > 0. \tag{35}$$

Analogously, for the 1st pool, we can distinguish the null-recurrence, when $\lambda_1 = N_1\mu_1 + N_2\mu_{12}$, and the transience, when inequality (27) is strict.

3.2 Stability of the Whole System

Now we consider the reverse inequality to inequality (32), namely,

$$\rho_2 + \frac{\lambda_1}{\mu_{12}}\mathsf{P}_\ell = \rho_2 + \frac{\lambda_1 - \mu_1\mathcal{B}_1}{\mu_{12}} < N_2. \tag{36}$$

By (32), under inequality (36), $Q_2(t)$ does not go to infinity, $Q_2(t) \not\Rightarrow \infty$. Rewriting inequality (36) as follows,

$$\lambda_1 < \mu_{12}N_2 + \mu_1\mathcal{B}_1 - \rho_2\mu_{12} < \mu_{12}N_2 + \mu_1 N_1, \tag{37}$$

we see that (36) implies (28) as well, and hence, the positive recurrence of the 1st pool takes place. Thus, under condition (36), the process $\{Q_1(t)\}$ is positive recurrent, and the limits P_ℓ, $\mathcal{B}_1$ exist. In particular, the positive recurrent regenerative process $\{Q_1(t)\}$ is tight. Then, by the standard way [3], we can deduce the positive recurrence of the process $\{Q_2(t)\}$, and hence the positive recurrence of the whole system. To be more precise, we define the process $\mathcal{W}(t) = W_1(t) + W_2(t)$, the summary remaining work (workload) in the system at instant $t \geq 0$.

Then previous analysis can be summarized as the following statement (where the workload process $W_i(t)$ can be replaced by any particular process describing pool $i = 1, 2$).

Theorem 2. Consider the initially empty preemptive-resume two-pool N-model under assumption (8). Then the process $\mathcal{W}(t)$ is positive recurrent if condition (36) holds. If condition (28) holds, then the process $\{W_1(t)\}$ is positive recurrent. If condition (34) is satisfied, then the process $\{W_1(t)\}$ is positive recurrent, and the process $\{W_2(t)\}$ is either null recurrent or transient.

Assume that

$$\mu_2 = \mu_{12},\ K_1 = 0,\ N_1 = N_2 = 1.$$

Then $\mathcal{B}_1 = \mathsf{P}_l = 1 - \mathsf{P}_0$ is the stationary busy probability and P_0 is the stationary empty probability of the 1st server, and stability condition (36) becomes, by (7),

$$\lambda_1 + \lambda_2 + \mu_1\mathsf{P}_0 < \mu_1 + \mu_2.$$

This expression coincides with condition (4.1) in [5], where P_0 is given in an explicit form for a 2-server Markov model with the non-preemptive priority. In general, to apply condition (36) in practice, a simulation is required to estimate the unknown parameter P_l (or $\mathcal{B}_1$).

4 Conclusion

Using regenerative methodology, we develop the stability analysis of the so-called N-model containing two interacting pools of servers with preemptive-resume priority of class-1 customers in pool 2. The fluid stability analysis of the similar

N-model with non-preemptive priority has been earlier developed in the paper [5]. We find stability conditions of the 1st pool solely, stability conditions of the whole system, and also the conditions when the 1st pool is stable, while the 2nd pool remains unstable.

Acknowledgement. The study was carried out under state order to the Karelian Research Centre of the Russian Academy of Sciences (Institute of Applied Mathematical Research KRC RAS). This research is partially supported by Russian Foundation for Basic Research, projects 18-07-00147, 18-07-00156.

References

1. Morozov, E.: The tightness in the ergodic analysis of regenerative queueing processes. Queueing Syst. **27**, 179–203 (1997)
2. Delgado, R., Morozov, E.: Stability analysis of cascade networks via fluid models. Perform. Eval. **82**, 39–54 (2014)
3. Morozov, E., Delgado, R.: Stability analysis of regenerative queues. Autom. Remote Control **70**(12), 1977–1991 (2009)
4. Morozov, E., Maltseva, M., Steyaert, B.: Verification of the stability of a two-server queueing system with static priority. In: Proceedings of the 22th FRUCT Conference, 15–18 May, Finland (2018, in Print)
5. Tezcan, T.: Stability analysis of N-model systems under a static priority rule. Queueing Syst. **73**, 235–259 (2013)
6. Whitt, W.: Blocking when service is required from several facilities simultaneously. AT&T Tech. J. **64**(8), 1807–1856 (1985)
7. Wong, D., Paciorek, N., Walsh, T., DiCelie, J., Young, M., Peet, B.: *Concordia*: an infrastructure for collaborating mobile agents. In: Rothermel, K., Popescu-Zeletin, R. (eds.) MA 1997. LNCS, vol. 1219, pp. 86–97. Springer, Heidelberg (1997). https://doi.org/10.1007/3-540-62803-7_26
8. Smith, W.L.: Regenerative stochastic processes. Proc. R. Soc. Ser. A **232**, 6–31 (1955)

Myopic Channel Switching Strategies for Stationary Mode: Threshold Calculation Algorithms

A. Mandel[1(✉)] and V. Laptin[2]

[1] V.A. Trapeznikov Institute of Control Sciences RAS,
Profsoyuznaya 65, Moscow, Russia
almandel@yandex.ru
[2] M.V. Lomonosov Moscow State University,
Lenin's Mountings 1, Moscow, Russia
straqker@bk.ru

Abstract. The controllable multiple queuing system (QS) with a switching number of service channels is considered. Such switches are possible at special periodical moments of time (control points) equidistant in time. The intensity of simple input flow at control points randomly changes and these changes ate governed by Markov chain process. It is assumed that during one step (an interval between neighbor control points) QS is in time for a stationary mode. The switching strategy is constructed to maximize a mean one-step profit. Threshold feature of optimal strategy is proved and algorithms for threshold values calculating are derived.

Keywords: Controllable queuing systems · Markovian input flow
Switching strategies

1 Introduction

We consider one of the modifications of the problem of the controllable queuing systems (QS) theory, to which the review article by Rykov [1] is devoted. According to [1], which investigates different types of goal quality functions, this report studies the use of a specific nonlinear goal quality function. Lower described formulation of the problem develops the stochastic programming models and Markov decision processes proposed in [2]. In fact, we investigate the case of the so-called myopic (one-step) quality criterion function. More general multistep criterion function has been considered in the reports [3,4] that precede this paper in [3,4] general algorithms for forming QS control strategies consisting of changing the number of active (working) service channels were constructed. Namely, we use an assumption that at periodical (equidistant) moments of time,

This research is supported by two grants of RFFI (projects No. 16-29-12895 and 17-07-00492 a).

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 410–420, 2018.
https://doi.org/10.1007/978-3-319-99447-5_35

the number of switched service channels may be changed. This time step τ will be used as the main time unit. So we take $\tau = 1$. We take that at control moments the intensity of simple input flow has random jumps governed by finite (or countable) homogeneous Markov chain with given transition probabilities. It is supposed that step duration is enough to settle the QS in stationary probabilistic mode. The problem is to construct the switching strategy (switching off the excessive channels or switching on spare channels) to maximize a mean one-step QS profit. Such strategies are called myopic in operations research [5].

The method of problem-solving is suggested to construct threshold control strategies. This result confirms a fundamental conclusion of Rykov firstly derived at [6], see also [1].

2 Problem Setting for Myopic Case

So, the multiple QS is considered with the number of active service channels to be the control variable. It can be changed periodically at so-called control moments of time (with step 1). It is considered that QS input flow is simple but the intensity λ of simple input flow endures random jumps at control moments to finite or countable number of values λ_j from discrete set Λ. The transition probabilities matrix of corresponding homogeneous Markov chain is $P = \|p_{ij}\|$, where p_{ij} is a probability of transition from the intensity value λ_i (at previous step) to the intensity value λ_j.

It is assumed that step duration is enough to settle the QS in stationary probabilistic mode. If for the current control step the input flow intensity is equal to λ_j and the one service channel intensity is equal μ[1], then, it is evident [7,8], that number of active service channels should satisfy the next inequality:

$$m \geq m_{crit}(\lambda_j) = \left[\frac{\lambda_j}{\mu}\right] + 1. \tag{1}$$

Let the cost components of QS functioning:

c_1 - operating cost of one active service channel per one step;
c_2 - cost of one active service channel switching off ($c_1 > c_2$);
A_1 - fixed price of switching on new service channels ("switching on");
A_2 - fixed price of switching off some service channels ("switching off");
d - cost of the time unit of one demand awaiting service;
h - income obtained when one demand is served;
m_1 - current number of active service channels (before any control decision is made);
m - chosen number of active service channels (control).

[1] All service devices are assumed to have exponential mutually independent service times with a service rate μ.

The goal is to maximize the on-step average income of the system while it operates under given flow input intensity λ_j, current number of active service channels m_1 and chosen number of active service channels m. Note this function as $G^{(1)}(\lambda_j, m_1, m)$:

$$G^{(1)}(\lambda_j, m_1, m) = G^{(1)} - C^{(1)}(\lambda_j, m_1, m), \tag{2}$$

where, under inequality (1), $G^{(1)} = h\lambda_j$, so $C^{(1)}$ is an one-step income, and $C^{(1)}(\lambda_j, m_1, m)$ is the sum expenditures at one step:

$$C^{(1)}(\lambda_j, m_1, m) = C_{exploit} + C_{queue} + C_{sw\_on} + C_{sw\_off}. \tag{3}$$

Here one-step exploitation costs are equal to $C_{exploit} = c_1 m$, stationary mode one-step queue costs are equal to $C_{queue} = d\bar{l}_{queue}$, where $\bar{l}_{queue}$ is equal to an average queue length, and switching costs $C_{switching} = C_{sw\_on} + C_{sw\_off}$ may be represented as:

$$C_{\text{switching}} = \begin{cases} A_1, & \text{if } m > m_1; \\ 0, & \text{if } m = m_1; \\ A_2 + c_2(m_1 - m), & \text{if } m < m_1. \end{cases} \tag{4}$$

Let us use classical results [7], [8] to get:

$$\bar{l}_{queue} = \left[\sum_{i=1}^{m-1} \frac{(m\rho_j)^i}{i!} + \frac{(m\rho_j)^m}{m!(1-\rho_j)} \right]^{-1} \frac{(m\rho_j)^m \rho_j}{m!(1-\rho_j)^2}, \tag{5}$$

where $\rho_j = \frac{\lambda_j}{m\mu}$. Simple to show that an average queue length $\bar{l}_{queue}$ is convex function of the control variable m.

It is evident from formula (2) that it possible to use as the goal function $C^{(1)}(\lambda_j, m_1, m)$ that should be minimized unlike the function $G^{(1)}(\lambda_j, m_1, m)$ to be maximized.

So to construct the optimal myopic switching strategy it is necessary to find:

$$\min_{m \geq m_{crit}} C^{(1)}(\lambda_j, m_1, m)$$

$$= \min_{m \geq m_{crit}} \{c_1 m + d\bar{l}_{queue} + \min \begin{cases} A_1 \mathbb{1}(m - m_1), \\ 0, \\ A_2 \mathbb{1}(m_1 - m) + c_2(m_1 - m), \end{cases} \}, \tag{6}$$

where $\mathbb{1}(u)$ is the Heaviside unit step function which is equal to 1, if $u > 0$, and is equal to 0 otherwise.

3 The Scheme for the Formation of Optimal Switching Strategies

Studying the Eq. (6) solutions with the active service channels number to be the control variable several cases should be considered.

Case $m_1 < m_{crit}$. In this case the number m of switched channels should be increased as minima to m_{crit}. Thus this case is a case of switching on additional service channels. So the switching-on process is realized, and formula (6) leads immediately to the necessity of minimizing the next functional:

$$\min_{m \geq m_{crit}} C^{(1)}(\lambda_j, m_1, m) = \min_{m \geq m_{crit}} \{A_1 + c_1 m + d\bar{l}_{queue}\}. \tag{7}$$

Let us use the next notation $B_{sw\_on}(m) = c_1 m + d\bar{l}_{queue}$. There are possible two function $B_{sw\_on}(m)$ behavior variants dependently of parameters values and control variable m value. Both variants are represented in Fig. 1 (solid red line).

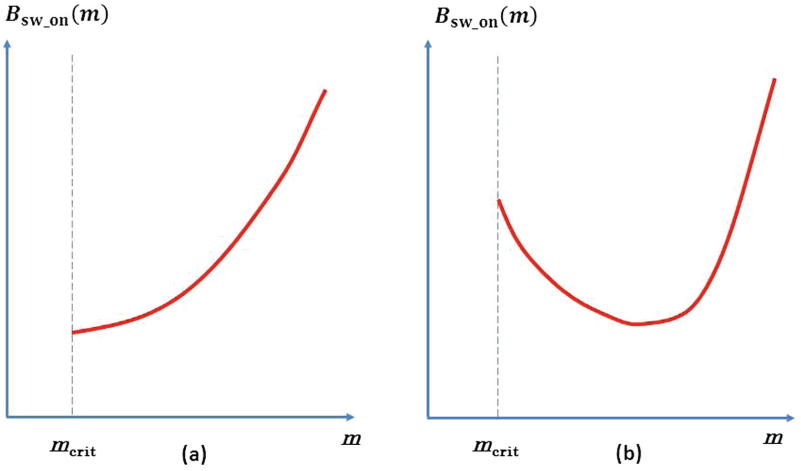


Fig. 1. Two variants of function $B_{sw\_on}(m)$ graphs. (Color figure online)

In case (a), to the left at Fig. 1, the solution is evident: $m = m_{crit}$. In turn, case (b), to the right at Fig. 1, is subdivided for two subcases. See Fig. 2.

In subcase (a), to the left at Fig. 2, the solution is evident: $m = R_1$ where $R_1$is the point of function $B_{sw\_on}(m)$ absolute minimum. In subcase (b), to the right at Fig. 2, the solution is more complicated. Not only the case $m_1 < m_{crit}$ could be analyzed but also the case $m_1 \geq m_{crit}$ (may be, partly). In fact, if r_1 is the intersection point to the left of the absolute minimum of the function $B_{sw\_on}(m)$ graph and the horizontal $A_1 + B_{sw\_on}(R_1)$ then if $m_{crit} \leq m < r_1$ it is necessary to set $m = R_1$, If $r_1 \leq m < R_1$. At last, if $R1 \leq m$ t it should be thought, is it not necessary to turn off unnecessary active service channels.

To do this, we investigate the behavior of the functional (6) in the case of disconnection of active service channels. Indeed, in this case, formula (6) takes the form:

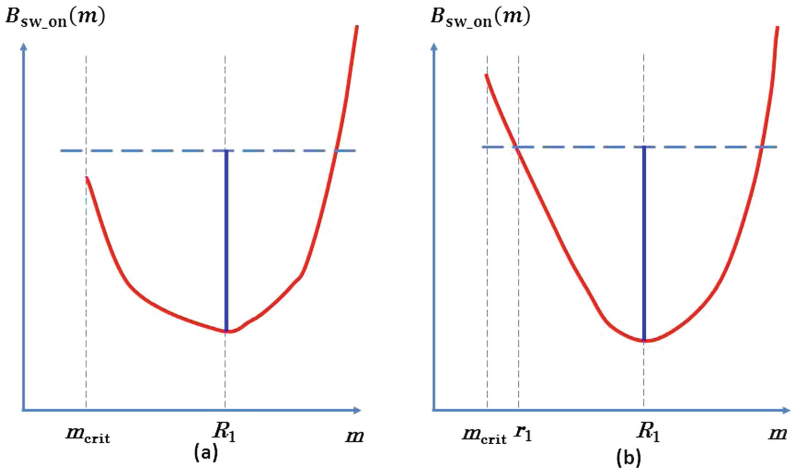


Fig. 2. Two subcases of function $B_{sw\_on}(m)$ graphs.

$$\min_{m \geq m_{crit}} C^{(1)}(\lambda_j, m_1, m) = \min_{m \geq m_{crit}} \{A_2 + (c_1 - c_2)m + c_2 m + d\bar{l}_{queue}\}. \quad (8)$$

By analogy with the function $B_{sw\_on}(m)$ we introduce the function $B_{sw\_off}(m) = (c_1 - c_2)m + c_2 m_1 + d\bar{l}_{queue}$. Again, there are two possible versions of the form of the function $B_{sw\_off}(m)$, shown in Fig. 3.

It is not hard to see from the formulas for the functions $B_{sw\_on}(m)$ and $B_{sw\_off}(m)$ that the absolute minimum point of the function $B_{sw\_off}(m)$, which we denote by R_2, is to the right of the point R_1 of the absolute minimum of the function $B_{sw\_on}(m)$. The case when this minimum is reached at the point mcrit is shown in the left part of Fig. 2. Now enter the point that is found as the abscissa of the intersection of the function $B_{sw\_off}(m)$ graph and the horizontal $A_2 + B_{sw\_off}(R_2)$ to the right of the minimum point R_2.

However, both of them shown in Fig. 3 variants differ only in that in the variant (a) in Fig. 3, it is necessary to switch off to the level of m_{crit} when the number of m_1 channels included is greater than r_2. If the current number m_1 of active channels is less than r_2 (or equal to it), then you do not need to disconnect anything.

For the case shown in the right half of Fig. 3 (option (b)), for $m_1 > r_2$, the number of active channels included is brought about by shutting down to R_2. Otherwise, there is no trip.

This completes the proof of the threshold nature of channel switching strategies in the case of a myopic quality criterion function.

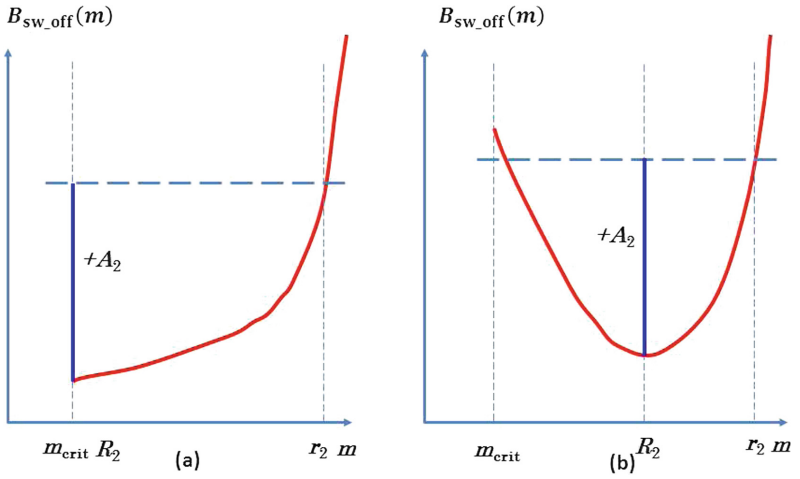


Fig. 3. Two variants of the function $B_{sw\_on}(m)$ graphs.

4 Simulation

We will simulate the behavior of the QS in one step. We choose the initial number of working channels m_1 such that the optimal solution is to turn off part of the active service channels.

In order to simulate the behavior of the QS in one step, first we split the function of average costs into component parts and consider them separately. According to the formula (3), the function of average costs consists of the operating costs $C_{exploit}$ of the active (working) channel, the cost C_{queue} of maintaining the queue, and the cost $C_{sw} = C_{sw\_on} + C_{sw\_off}$ of enabling/disabling service channels.

Since the average operating costs of the working devices at one step are in the form $C_{exploit} = c_1 m$, this function is linear and its dependence on the number of active channels is shown in Fig. 4.

The average cost of a queue in stationary mode is $C_{queue} = d\bar{l}_{queue}$, where $\bar{l}_{queue}$ is the average queue length, which is calculated by the formula (5). This function is convex and monotonically decreasing (see Fig. 5).

The cost of switching $C_{sw} = C_{sw\_on} + C_{sw\_off}$ is as follows (see Fig. 6).

Here, to the left of the minimum point equal to 0, the chart branch is responsible for the cost of disconnecting the working channels, and to the right – for switching the channels from the reserve to the working ones.

Now let's simulate the behavior of the QS in one step. Select the initial number of working channels m_1 so that the optimal solution is to include an additional number of service channels. Below is a graph of the dependence of the average one-step costs on the number of channels built in the simulation environment MATLAB R2017a.

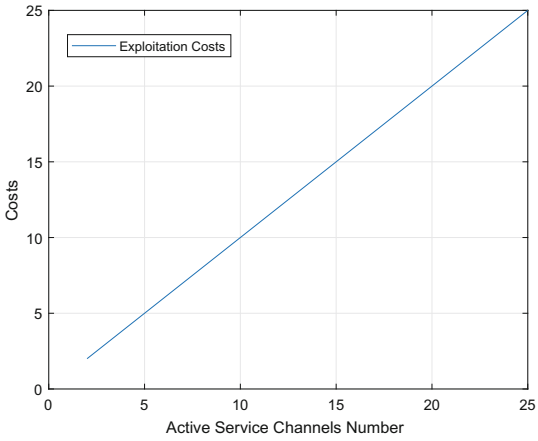


Fig. 4. Function of average operating costs

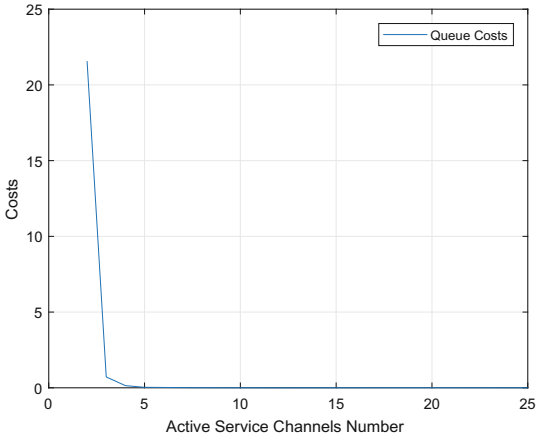


Fig. 5. Average cost of queue maintenance

For Fig. 7 line of low cost crossing with the schedule $B_{sw\_on}(m)$ gives a value of the number of operating channels, at which it is necessary to include an additional number of standby service channels (similar to r_2 in Fig. 3). In this instance case $m_1 < m_{crit}$.

And now consider the case when you need to turn off the extra channels. In order to comply with the concept formulated above, within which it is necessary to carry out disconnection only when the number of channels m_1 is greater than r_2. If the number of m_1 channels is less than (or equal to) r_2, then nothing should be disabled. Now, if you plot the previously introduced function of the average cost of disabling $B_{sw\_off}(m) = (c_1 - c_2)m + c_2 m_1 + d\bar{l}_{queue}$, then its schedule will not be difficult to determine the threshold value r_2, going through

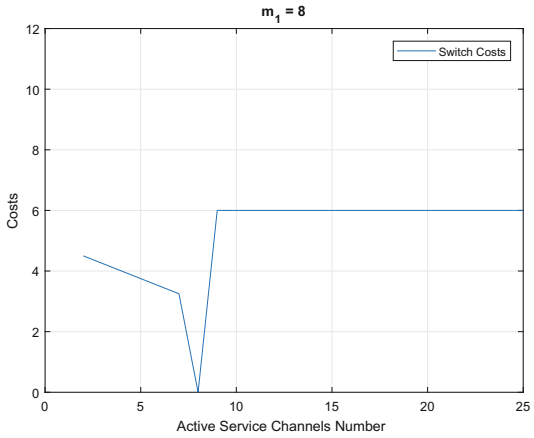


Fig. 6. Average cost of switching service channels on/off

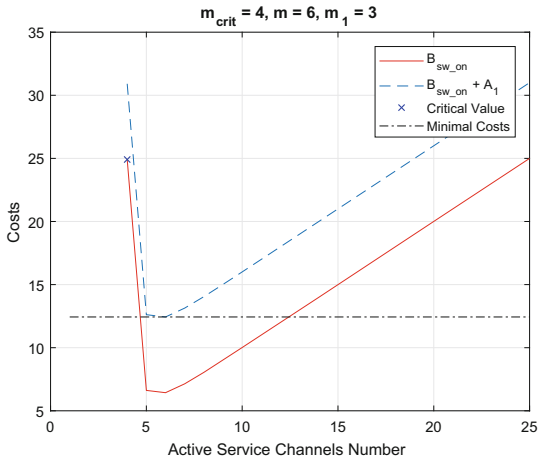


Fig. 7. Average cost of switching service channels on/off

which the number of active service channels must be reduced. However, because the number of service channels can only be integer, the minimum number of channels from which you want to disconnect is

$$m_{sw\_off} = [r_2] + 1. \tag{9}$$

The results of computer simulation are shown in Fig. 8.

By drawing a line of the minimum cost level (dotted line with a dot in Fig. 8), you can easily determine the desired value r_2, and find it and m_{sw_off}.

The following Fig. 9 the results of numerical comparison of the minimum of two functions of average costs are presented $B_{sw\_on}(m)$ and $B_{sw\_off}(m)$.

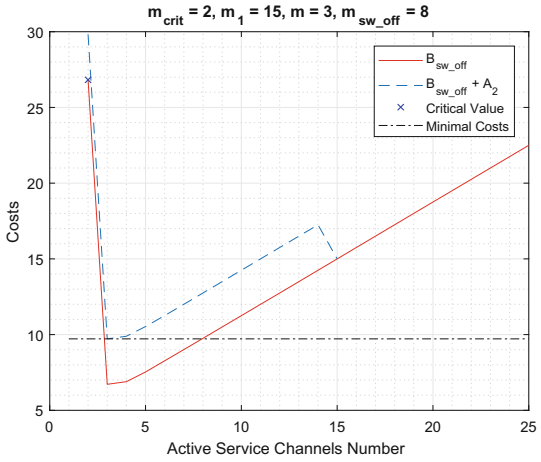


Fig. 8. Determine m_{sw_off}

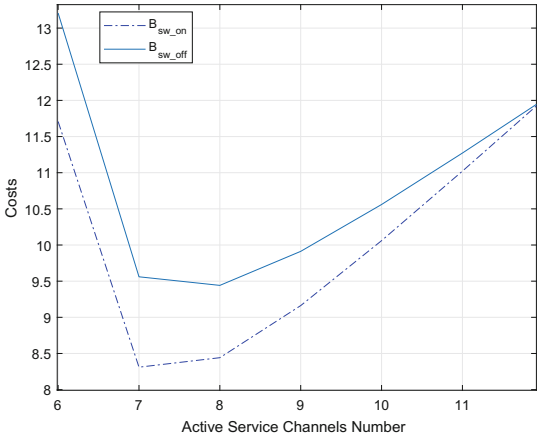


Fig. 9. A comparison of the minima of two functions $B_{sw\_on}(m)$ and $B_{sw\_off}(m)$

We now investigate how the optimal number of working channels depends on the intensity of the incoming flow. We define the service intensity $\mu = 6$. And let's first consider how the number of channels m_{sw_off} for the function $B_{sw\_off}(m)$ will behave. In Fig. 10 the graph of dependence of the minimum number of channels at which achievement it is expedient to carry out shutdown is presented.

And now using the formulas and algorithms presented in Sect. 3, we can calculate how the optimal number of service channels will depend on the intensity of the incoming flow λ. The initial number of channels $m_1 = 11$. In Fig. 11 the graph shows that when the intensity of the incoming flow within certain limits changes, the number of working channels does not change. In other words, in the

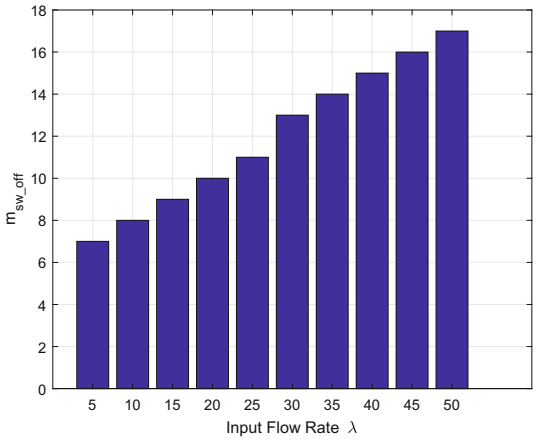


Fig. 10. Number of m_{sw_off} depending on intensity λ

corresponding interval of values of the intensity of the incoming flow is beneficial to leave the number of channels without change.

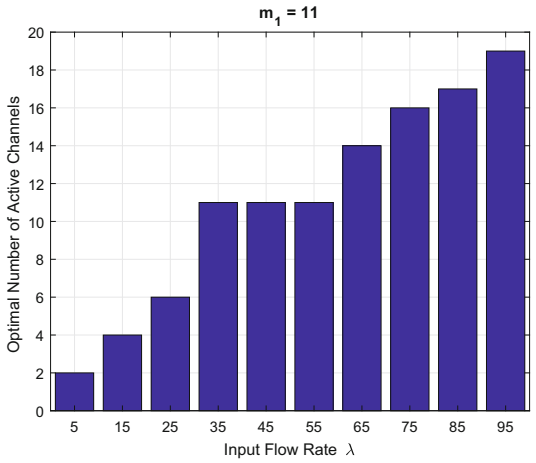


Fig. 11. Dependence of the optimal number of working channels on the intensity λ

5 Summary

The strategies of channel switching in a multiple queuing system under the Markovian intensity of the input flow and a myopic goal function are investigated. At the same time, it was assumed that there are fixed costs, which are

associated with making decisions both on the inclusion of additional ones and on the deactivation of unnecessary channels. It is this circumstance that has a decisive influence on the switching strategies, which, as a result, turn out to be threshold ones. The generalization of this result for non-myopic (multi-step) goal function seems to the authors rather believable.

References

1. Rykov, V.: Controllable queueing systems: from the very beginning up to nowadays. RT&A **12**(2(45)), 39–61 (2017)
2. Sennott, L.I.: Stochastic Dynamic Programming and the Control of Queueing Systems, p. 358. Wiley, Hoboken (1999)
3. Mandel, A.: Econometric models of controllable multiple queuing systems. In: Vishnevsky, V., Kozyrev, D. (eds.) DCCN 2015. CCIS, vol. 601, pp. 296–304. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30843-2_31
4. Mandel, A., Bakulin, K.: Models of controllable multiple queuing systems for channel switching myopic strategies. In: Proceedings of the 20th International Conference, Distributed Computer and Communication Networks. DCCN 2017, Moscow, Russia, pp. 534–542. Technosphera, Moscow (2017). (in Russian)
5. Hadley, G., Whitin, T.M.: Analysis of Inventory Systems, p. 512. Prentice Hall, Inc., Englewood Cliffs (1969)
6. Rykov, V.: Controllable queuing systems. In: Probability Theory, Mathematical Statistics and Theoretical Cybernetics, vol. 12, pp. 43–153 (1975). (in Russian)
7. Gnedenko, B., Kovalenko, I.: Introduction to the Queue Systems Theory. Nauka, Moscow (1966). (in Russian)
8. Vishnevsky, V.: Theoretical Foundations of Computer Networks Design. Technosphera, Moscow (2003). (in Russian)

A Novel Slice-Oriented Network Model

Samuel Muhizi[1(✉)], Abdelhamied A. Ateya[1], Ammar Muthanna[1,2], Ruslan Kirichek[1,2], and Andrey Koucheryavy[1]

[1] The Bonch-Bruevich Saint-Petersburg State University of Telecommunications, 22/1 Prospekt Bolshevikov, Saint-Petersburg 193232, Russian Federation
samno1@yandex.ru, a.ashraf@zu.edu.eg, ammarexpress@gmail.com, kirichek@sut.ru, akouch@mail.ru

[2] Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Street, Moscow 117198, Russian Federation
http://www.sut.ru

Abstract. Network Function Virtualization (NFV), Software-Defined Networks (SDN), and Mobile Edge Computing (MEC) are recent technologies that enable new features and functionalities for 5G networks. These technologies are used to provide flexible, scalable and on-demand services for the vast growing array of applications with diverging requirements such low latency, data transmission security, energy efficiency, mobility, massive connectivity, reliability, guaranteed QoS, throughput etc. The introduction of network slicing offers new solutions to manage challenges of application-tailored services in 5G and optimize business model for network operators. In this paper, we present a novel slice-oriented network model and develop a feasible information network demo intended to describe the managerial characteristics and behavior of network slices. The model is mainly depends on the MEC paradigm.

Keywords: Network slicing · 5G · SDN · NFV · MEC
Network orchestration

1 Introduction

The future will be defined by advances in artificial intelligence, autonomous IoT, big data analytics, machine learning and augmented/virtual reality, supported by high-speed, low-latency, secure connectivity that is ubiquitous and reliable.

According to 5G/IMT-2020 prediction, the 5G network will introduce new features and functionalities that will nearly change the network fundamentals like it have been ever before by transforming network infrastructures to respond to new challenges and application service requirements [1]. 5G networks are designed for mobile broadband networks as well as for supporting the Internet of Things, providing a platform for connecting a large number of sensors and other visualization devices, and allowing the emergence of unprecedented new business models in the future telecommunication industries. One of the discussed challenges for telecommunication networks is the provision of user/application-tailored services, as current network infrastructures do not simplify the task of

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 421–431, 2018.
https://doi.org/10.1007/978-3-319-99447-5_36

dynamic network resources allocation due to lack of flexibility to respond to series of application requirements. These requirements include:

- Near real-time latency (end-to-end delay) for services with requirements on very low and stable latencies,
- Stable and reliable high upload and download speeds,
- Guaranteed SLA: the capability of a network slice to provide certain level of E2E assurance to the requested system functional and performance requirements with appropriate Service Level Reporting (SLR) method,
- Coverage, to ensure seamless service experience across networks and country boundaries,
- Connected device management from only a few devices, up to extreme high density of devices/connections; also including very specific Device to Device (D2D) connectivity and/or hardware requirements [12,13],
- Seamless mobility for uninterrupted service delivery and stable quality in scenarios with medium to high velocity (e.g. high-speed train, aviation), across heterogeneous networks that may also belong to multiple different service providers,
- Energy efficiency could be provided in the case where ultra-low energy utilization is required (e.g. NB-IoT scenarios) on the network side, as well as on the device terminal side (e.g. very long battery life), and
- Data security to satisfy security and privacy requirements beyond today's capabilities and for extremely sensitive data transmission (e.g. National security, fraud/cybercrime sensitive).

Introduction of such network services as high security, very low latencies and global network coverage is key success for 5G networks and opens new opportunities for vertical industries and sectors, such as healthcare, security, energy, transports, automotive, etc. [2]. At the same time, expected QoS/QoE should be preserved with proper Quality of Security. Through technologies like Software-Defined Networking (SDN) [12,16] and Network Function Virtualization (NFV) [11,17], network acquires the programmability, flexibility, and modularity required for service automation and isolation allowing optimizing network resources management by grouping services/applications in isolated network slices. SDN and NFV enable the creation of logical networks over a common physical network infrastructure. A network slice consists of a set of NFs and the corresponding resources [3].

In this paper, we present a novel slice-oriented network model and develop a feasible information network demo intended to describe the manageable characteristics and behavior of network slices.

2 The Concept of Network Slicing from the Structure Point of View

Future 5G communication networks will continue to develop, reaching segments of the network industry such as automotive, manufacturing, logistics, energy, as

well as sectors such as financial, health-care and others that are not currently fully exploiting the potential of network services.

5G networks are expected to create new service capabilities relying on recent advancements in the Internet of things (IoT) area [20]. In particular, analysts forecast that by 2025 the number of IoT devices could grow to a stunning figure of about 100 billion devices connected [1], supporting a wide range of services spanning from low cost sensor-based metering services and delay tolerant vehicle services to critical communications including e-health, e-business and automotive. For mobile operators, IoT does not mean only support for many more devices and massive connectivity, but also defines a promising opportunity for offering novel services and business solutions within the IoT value chain beyond simple connectivity. To this end, 5G enables open interfaces to support vertical segments, i.e. third parties not owning network infrastructure and requiring networking services with specific needs, as well as new business solutions. The automotive industry defines one of the most significant 5G vertical segments. It requires efficient networking capabilities combined with IoT and edge cloud to facilitate a number of services including autonomous driving and real-time assessment of road conditions for example.

The suboptimal use of the network resources is due to the diversity, and even conflicting, network service requirements of such network applications (Fig. 1). One application, for example, may require ultra-reliable services, whereas other applications may need ultra-high-bandwidth communication or extremely low latency. The 5G network is designed to be able to offer a different mix of capabilities to meet all these diverse requirements at the same time [1].

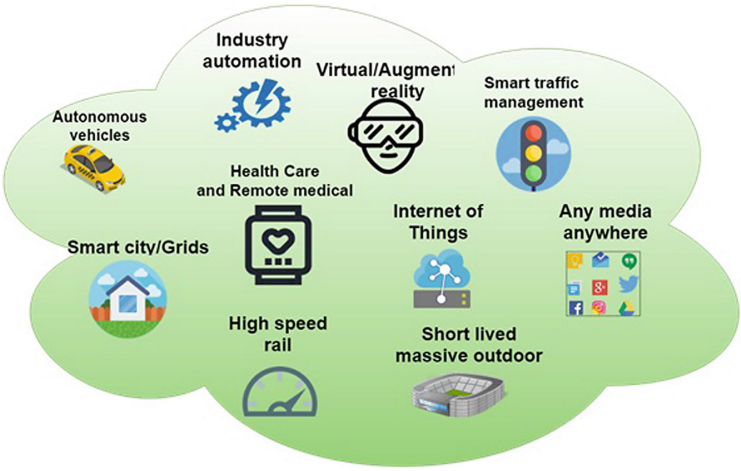


Fig. 1. 5G communication networks application scope

From a functional point of view, the most logical approach is to build a set of dedicated networks each adapted to serve one type of application to allow the

implementation of tailor-made functionality and network operation specific to the needs of each application and business customer, rather than a one-size-fits-all approach as the case in current and previous generations of communication networks.

This approach of operating multiple dedicated networks on a common platform is effectively what network slicing allows.

Network slicing consist of creation and management of multiple logical self-contained networks on top of a common physical infrastructure platform enabling a flexible stakeholder ecosystem that allows technical and business innovation integrating physical and/or logical network and cloud resources into a programmable, open software-oriented multi-tenant network environment. 3GPP defines network slicing as a technology that "enables the operator to create networks, customized to provide optimized solutions for different market scenarios which demand diverse requirements, e.g. in terms of functionality, performance and isolation [3].

Network slicing enables value creation for vertical segments, application providers and third parties that lack physical network infrastructure, by offering radio, networking and cloud resources, allowing a customized network operation and true service differentiation. The VNFs, which constitute a network slice, may vary drastically depending on the service requirements of that particular slice. The type of service associated with a network slice would determine the resources and service treatment the network slice would receive, e.g. a real-time communication network slice would receive the appropriate resources and service treatment to meet ultra-low latency demands [1]. Network slicing builds on top of the following seven main principles that shape the concept and related operations: automation, isolation, customization, elasticity, programmability, end-to-end, hierarchical abstraction.

Automation: enables an on-demand configuration of network slicing without the need of fixed contractual agreements and manual intervention. Such convenient operation relies on signaling-based mechanisms, which allow third parties to place a slice creation request indicating besides the conventional SLA which would reflect the desired capacity, latency, jitter, etc., timing information considering the starting and ending time, and duration of a network slice. Isolation: is a fundamental property of network slicing that assures performance and guarantees security for each tenant even when different tenants use network slices for services with conflicting performance requirements. However, isolation may come at the cost of reducing multiplexing gain, depending on the means of resource separation for explicit use, which may result in inefficient network resource utilization. The notion of isolation involves not only the data plane but also the control plane, while its implementation defines the degree of resource separation. Isolation can be deployed by using a different physical resource, when separating via virtualization means a shared resource and through sharing a resource with the guidance of a respective policy that defines the access rights for each tenant [15]. Customization: assures that the resources allocated to a particular tenant are efficiently utilized in order to meet best the respective service require-

ments. Slice customization can be realized in a network wide level considering the abstracted topology and the separation of data and control plane, on the data plane with service-tailored network functions and data forwarding mechanism, on the control plane introducing programmable policies, operations and protocols and through value-added services such as big data and context awareness. Elasticity: is an essential operation related with the resource allocated to a particular network slice, in order to assure the desired SLA under varying radio and network conditions, amount of serving users, or geographical serving area because of user mobility. Such resource elasticity can be realized by reshaping the use of the allocated resources by scaling up/down or relocating VNFs and value-added services, or by adjusting the applied policy and re-programing the functionality of certain data and control plane elements. Elasticity can also take the form of altering the amount of initially allocated resources by modifying physical and virtual network functions, e.g. by adding a different RAN technology or a new VNF, or by enhancing the radio and network capacity. However, this process requires an inter-slice negotiation since it may influence the performance of other slices that share the same resources. Programmability: allows third parties to control the allocated slice resources, i.e. networking and cloud resources, via open APIs that expose network capabilities facilitating on-demand service-oriented customization and resource elasticity. End-to-end: is an inherent property of network slicing for facilitating a service delivery all the way from the service providers to the end-user/customer(s). Such a property has two extensions, it stretches across different administrative domains, i.e. a slice that combines resources that belong to distinct infrastructure providers, and it unifies various network layers and heterogeneous technologies, e.g. considering RAN, core network, transport and cloud. In particular, an end-to-end network slicing consolidates diverse resources enabling an overlaid service layer, which provides new opportunities for efficient networking and service convergence. Hierarchical abstraction: is a property of network slicing that has its roots on recursive Virtualization, wherein the resource abstraction procedure is repeated on a hierarchical pattern with each successively higher level, offering a greater abstraction with a broader scope. In other words, the resources of a network slice, allocated to a particular tenant, can be further traded either partially or fully to yet another third player, which relates to the network slice tenant facilitating in this way another network slice service on top of the prior one. For example, a virtual mobile operator who acquired a network slice from an infrastructure provider, offers a partial amount of such resources to enable a utility provider that uses its virtual network to form an IoT slice.

Network slicing concept consists of organizing and running multiple logical networks as virtually independent business operations on a common physical infrastructure in an efficient and economical way.

Network Slicing is an End-to-End solution to network management across different administrative domains and technologies with a hierarchical recursive business nature. With network slicing, 5G network acquires the automation and programmability needed to respond to business requirements [6].

3 Slice Management Architectural Model

In multiservice network systems slices are designed for service optimization by organizing multiple per-service tailored networks, each of them represents an isolated independent end-to-end logical network to ensure required QoS on a shared network infrastructure. The considered network slicing model is depicted in Fig. 2. The slicing model consists of three layers: System infrastructure resource layer, System resource management layer, and slice management layer (slice controller) in Fig. 3.

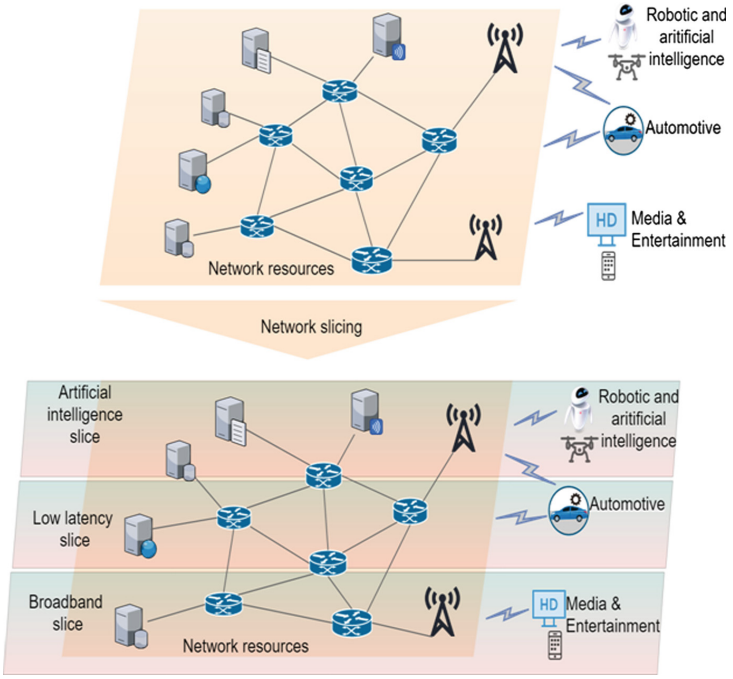


Fig. 2. Architectural conception of network slicing

The slice controller handles network slice requests and collects slices related information: SLA, group of users, location, timing, duration, service type, CPU, network slice workload, priority index, etc.

Resource management layer consists of network resources orchestrator and virtual network manager. The orchestrator creates slices and associates network functions to them. It sets up connectivity between network functions and coordinate seemingly disparate network processes for creating, managing and delivering services.

Respectively, once a slice request arrives, the orchestrator performs:

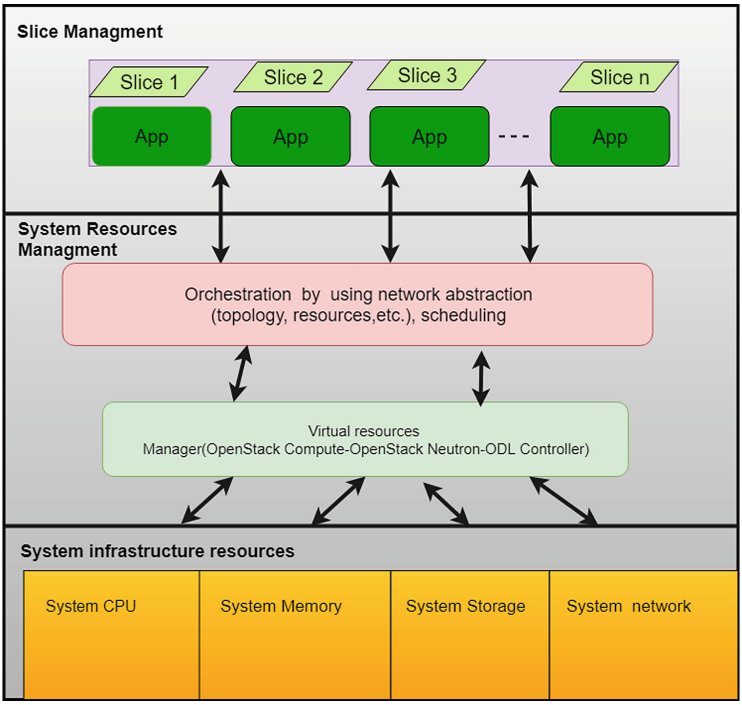


Fig. 3. Slice management architectural model

- admission control - virtual resource negotiation to create logical isolated virtual networks,
- service mapping - assigns slice template, and
- creation of related virtual network graph.

For a more detailed description of orchestrator functions, see [4].

In this regard, OpenStack Compute can be responsible for the creation of the overlay virtual network and the configuration of the network infrastructure to achieve the desired connectivity. The control plane can be based on the SDN controller [19] (OpenDaylight controller in this context) [5,18]. The ODL controller includes several modules and features that cooperate to operate the virtual switches, establish connectivity between switches and create overlay virtual networks to serve applications services.

In the proposed model, the MEC server is fed with a slice controller, which is able to distinguish between different slices. The slice controller extracts the application type that the received data belongs to and as a response, it estimates which virtual machines dedicated with kind of data.

4 Experimental Demonstration of the Network Model. Simulation and Results

In this part the performance and evaluation of the proposed work is conducted over a reliable environment. Many simulation environments are prober for implementing edge-computing systems and check their performance. These environments vary in the provided facilities and associated capabilities [7,8]. We use our developed tool kit introduced in [3]. This is a java-based environment, built over the CloudSim framework [10]. The simulation environment enables the creation of edge clouds with different number of virtual machines (VM). In addition, remote execution of web-based services is enabled. The simulation is run on a machine with an Intel core i5 processor, with 3.07 GHz of speed, and memory of 8 GB. Simulation results are illustrated in Table 1.

Table 1. Simulation parameters

Parameter	Description	Value
S	Number of source nodes	20
N_s	Total number of slices	6 (applications)
N_{VM}	Total number of virtual machines	8
n	Size of header identifier	3 bits
W^*_{mmax}	Maximum workload of MEC server	100 events/s
λ	Arrival rate	15 Mbps
M	Service rate of the MEC server	8 Mbps
RAM, HDD	RAM, Storage	2048 Mb, 10 Gb

Table 2. Resources allocation for different applications

APP1	APP2	APP3	APP4	APP5	APP6
VM7	VM0 and VM1	VM6	VM2 and VM3	VM4	VM5

An edge server with eight virtual machines is constructed. The edge server works based on the model for Micro-cloud edge servers presented in [9]. Twenty heterogeneous source nodes produces data for six different applications, each application is considered as a slice. Each application reserves a part of resources in the MEC server. Resources for applications are assumed heterogeneous, thus some applications are allocated more resources than others, based on a pre-allocated scheme. This scheme is set based on the probability of requests and data flow dedicated with each application. For our implementation, the number of virtual machines allocated to each application is included in Table 2. Application 2 and 3 are assumed to have a high flow and thus allocated both virtual machines.

For the performance evaluation, two main performance metrics are considered; the latency and the blocking probability. We measure each metric for each application in two cases. The first case represents the proposed framework, in which the resources are allocated for each application. The second case represents the alternative case, in which MEC server handle serves without any classifications. All MEC server resources serve for all application tasks.

Figure 4 presents the average latency for each application in both considered cases. For all applications, it is clear that slicing achieves higher performance intermesh of latency.

Figure 5 illustrates the average blocking probability for each application in both indicated cases. The blocking probability is much better in case of slicing for all considered applications.

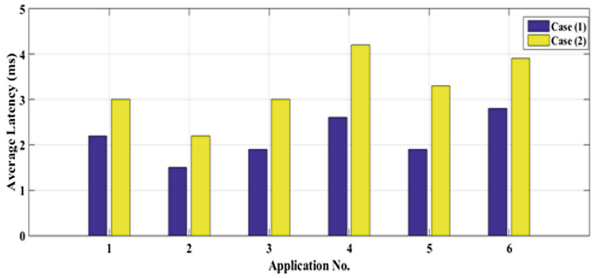


Fig. 4. Average latency for each application in the considered cases

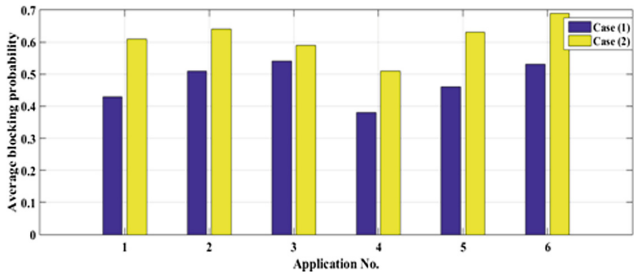


Fig. 5. Average blocking probability for each application in the considered cases

5 Conclusion

NFV along with SDN technologies facilitates a flexible network, where slices can be dynamically provisioned and migrated at need. Newer applications can be easily deployed into the network as per customer requirements. Network slicing is a

radical change of paradigm in the 5G networks compared to current implementations by allowing resource customization to accommodate SLA requirements: traffic forwarding considering service requirements, traffic steering considering physical condition of links, combining cloud and network capacity resources. With network slicing, 5G cellular system is able to adapt to the external environment rather than the other way around, allowing new business models based on user/application requirements. In this paper, we reviewed the requirements to 5G/IMT-2020 implementation and proposed an architectural model for network slicing. Furthermore, we will extend the model to analyze the possible challenges and solutions for full implementation of network slicing in 5G networks.

Acknowledgments. The publication has been prepared with the support of the "RUDN University Program 5-100" and funded by RFBR according to the research projects No. 17-07-00845 and No.18-07-00576.

References

1. Y.3101: Requirements of the IMT-2020 network
2. 5G-PPP, ERTICO, EFFRA, EUTC, NEM, CONTINUA, Networld2020 ETP: 5G empowering vertical industries. White Paper, February 2016
3. de Foy, X., Rahman, A.: Network Slicing - 3GPP Use Case. Internet Requests for Comments, RFC Editor, RFC, October 2017. https://tools.ietf.org/id/draft-defoy-netslices-3gpp-network-slicing-02.html
4. ONF TR-521: SDN Architecture, February 2016
5. ONF TR-526: Applying SDN Architecture to 5G Slicing, April 2016
6. ITU-T Y.3111: IMT-2020 network management and orchestration framework
7. Wang, S., et al.: Mobile micro-cloud: application classification, mapping, and deployment. In: Proceedings of Annual Fall Meeting of ITA 2013, October 2013
8. Bahwaireth, K., Tawalbeh, L., Benkhelifa, E., Jararweh, Y.: Experimental comparison of simulation tools for efficient cloud and mobile cloud computing applications. EURASIP J. Inf. Secur. **2016**, 15 (2016)
9. Ateya, A.A., Vybornova, A., Samouylov, K., Koucheryavy, A.: System model for multi-level cloud based tactile internet system. In: Koucheryavy, Y., Mamatas, L., Matta, I., Ometov, A., Papadimitriou, P. (eds.) WWIC 2017. LNCS, vol. 10372, pp. 77–86. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61382-6_7
10. Kumar, R., Sahoo, G.: Cloud computing simulation using CloudSim. arXiv preprint arXiv:1403.3253 (2014)
11. Ateya, A., Muthanna, A., Gudkova, I., Abuarqoub, A., Vybornova, A., Koucheryavy, A.: Development of intelligent core network for tactile internet and future smart systems. J. Sens. Actuator Netw. **7**, 1 (2018)
12. Ateya, A.A., Muthanna, A., Koucheryavy, A.: 5G framework based on multi-level edge computing with D2D enabled communication. In: 20th International Conference on Advanced Communication Technology (ICACT), pp. 507–512 (2018)
13. Muthanna, A., et al.: Analytical evaluation of D2D connectivity potential in 5G wireless systems. In: Galinina, O., Balandin, S., Koucheryavy, Y. (eds.) NEW2AN/ruSMART -2016. LNCS, vol. 9870, pp. 395–403. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46301-8_33

14. Vladyko, A., Muthanna, A., Kirichek, R.: Comprehensive SDN testing based on model network. In: Galinina, O., Balandin, S., Koucheryavy, Y. (eds.) NEW2AN/ruSMART -2016. LNCS, vol. 9870, pp. 539–549. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46301-8_45
15. Kotulski, Z.: Towards constructive approach to end-to-end slice isolation in 5G networks. EURASIP J. Inf. Secur. **2018**, 2 (2018)
16. Muthanna, A., Khakimov, A., Gudkova, I., Paramonov, A., Vladyko, A., Kirichek, R.: Openflow switch buffer configuration method. In: Proceedings of the International Conference on Future Networks and Distributed Systems, ICFNDS 2017 (2017)
17. Sahoo, K.S., Mohanty, S., Tiwary, M., Mishra, B.K., Sahoo, B.: A comprehensive tutorial on software defined network: the driving force for the future internet technology. In: Proceedings of the International Conference on Advances in Information Communication Technology and Computing, p. 114. ACM (2016)
18. Muhizi, S., Shamshin, G., Muthanna, A., Kirichek, R., Vladyko, A., Koucheryavy, A.: Analysis and performance evaluation of SDN queue model. In: Koucheryavy, Y., Mamatas, L., Matta, I., Ometov, A., Papadimitriou, P. (eds.) WWIC 2017. LNCS, vol. 10372, pp. 26–37. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61382-6_3
19. Hock, D., Hartmann, M., Gebert, S., Jarschel, M., Zinner, T., Tran-Gia, P.: Pareto-optimal resilient controller placement in SDN-based core networks. In: 2013 25th International Teletraffic Congress (ITC), pp. 1–9. IEEE (2013)
20. Masek, P., Fujdiak, R., Zeman, K., Hosek, J., Muthanna, A.: Remote networking technology for IoT: cloud-based access for Alljoyn-enabled devices. In: Proceedings of the 18th Conference of Open Innovations Association FRUCT and Seminar on Information Security and Protection of Information Technology 2016, pp. 200–205 (2016)

Reliability of the Information System with Intermediate Storage Devices

Yuriy E. Obzherin(✉), Stanislav M. Sidorov, and Mikhail M. Nikitin

Sevastopol State University, Universitetskaya str., 33, 299053 Sevastopol, Russia
objsev@mail.ru, xaevec@mail.ru, m.nikitin.1979@gmail.com

Abstract. Using of intermediate information storage devices is an important mean to increase the reliability and effectiveness of information system operation. In this paper the semi-Markov model of a single-stream information system with intermediate information storage devices is built. With the aid of phase merging algorithm the stationary characteristics of system operation are approximately found. The analysis of storage devices capacity influence into system characteristics is carried out.

Keywords: Information system · Intermediate storage device
Semi-Markov model · Phase merging algorithm
Reliability characteristics

1 Introduction

Time redundancy (see [1–8] for details) is one of the ways to increase the reliability and operation effectiveness of information systems. Time reserving is spoken about in the cases where the opportunity to spend some additional time (time reserve) is provided to the system to restore its characteristics. The sources of time reserve in information system may be represented by intermediate storage devices, productivity reserve, etc.

In this paper the semi-Markov model with discrete-continuous phase space of states (see [9–12]) of a single-streamed information system with intermediate storage devices is built. To find approximately the reliability characteristics of the system phase merging algorithms (Koroluk and Turbin [9], Koroluk [10], and in [13–15]) are used allowing to solve effectively the problem of high dimension of the system.

The analysis of intermediate storage devices (see [1,16–19]) capacity influence into system reliability and efficiency is carried out, the results of analytic and simulation modeling of the system are compared.

The research was carried out within the state assignment of the Minobrnauki of Russia No. 1.10513.2018/11.12, with financial support by RFBR (project No. 18-01-00392a).

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 432–444, 2018.
https://doi.org/10.1007/978-3-319-99447-5_37

2 Semi-Markov Model Building and Obtaining of Reliability Characteristics

We consider a multiphase, single-stream system consisting of serving devices and intermediate storage devices, the connections between which are depicted in Fig. 1.

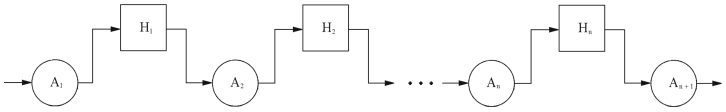


Fig. 1. Structural diagram of a multiphase single-flow structure with storage devices.

The following notation is used in the scheme: $A_i, i = \overline{1, n+1}$ - service devices; $H_i, i = \overline{1, n}$ - intermediate storage devices. The model is being constructed under the following assumptions.

1. The possible states of each of the serving devices A_i are: operable, recovery and shutdown.
2. Time to failure (renewal time) for the device A_i is a random variable (RV) $\alpha_i^{(0)}\left(\alpha_i^{(1)}\right)$ with distribution function $F_i^{(0)}(x)\left(F_i^{(1)}(x)\right)$. RV $\alpha_i^{(0)}\left(\alpha_i^{(1)}\right)$ are independent and have finite expectations; there are distribution densities $f_i^{(0)}(x)\left(f_i^{(1)}(x)\right)$ for $F_i^{(1)}(x)\left(F_i^{(0)}(x)\right)$.
3. The storage devices H_i are absolutely reliable devices that have limited capacity $h_i \geq 0$ (The storage capacity of the device H_i is expressed in units of time that device A_{i+1} will need to fully free this storage device).
4. An operable device A_i disconnects, being in operable state, if the storage device H_{i-1} is empty or the storage device H_{i+1} is full.
5. The productivity of device A_i is constant and equal to c_i, herewith $c_i \geq c_{i+1}$.
6. The system is in failed state if the output device does not produce products; the renewal of the device A_i is considered as infinite.

We suppose the system under consideration to be real.

To describe the operation of real system let us introduce the next space of states:

$$E = \left\{ i\bar{d}\bar{x}\bar{z} : i = \overline{1, n+1}, \bar{d} = (d_1, ..., d_{n+1}), \bar{x} = (x_1, ..., x_{n+1}), \bar{z} = (z_1, ..., z_{n+1}) \right\},$$

where i is the number of device A_i that failed or was recovered by the last. The element d_k of the vector $\bar{d}$ fixes the state of the device: operable ($d_k = 0$), recovery ($d_k = 1$), shut-down ($d_k = 1$). The value of the element x_k of the vector $\bar{x}$ is the time lasted from the instance of device A_k change of state, it should be noted that $x_i = 0$. The component z_k of the vector $\bar{z}$ defines the time

within which the storage device H_k can supply the device A_{k+1} with production, $0 \le z_k \le h_k, k = \overline{1,n}$.

Let us introduce additional notation. Vectors $\bar{0}_{(i)}, \bar{h}, \bar{0}$ are defined by the following manner

$$\bar{0}_{(i)} = (\underbrace{2,...,2}_{\text{i}-1},1,0,...,0),\ \ i = \overline{1,n+1},\ \ \bar{h} = (h_1,h_2,...,h_n)\,,\ \ \bar{0} = (0,0,...,0).$$

To find approximately the reliability characteristics of the system considered let us use the phase merging algorithm (see [9,10,13–15]).

Let us suppose that stochastic kernel of embedded Markov chain (EMC) $\{\xi_n; n \ge 0\}$ for the initial system semi-Markov process $\xi(t)$ is close to the stochastic kernel of EMC $\left\{\xi_n^{(0)}; n \ge 0\right\}$ for some supporting system S_0 with the single stationary distribution $\rho(dx)$. Then the next approximate formulas represented in Korlat et al. [14] can be used for approximate finding of the average stationary time to failure T_+, the average stationary renewal time T_- and the stationary availability factorK_a:

$$T_+ \approx \frac{(\rho,\bar{m}_1)}{\left(\rho,P^{(r)}\bar{1}_0\right)},\qquad T_- \approx \frac{\left(\rho,P^{(r)}\bar{m}_0\right)}{\left(\rho,P^{(r)}\bar{1}_0\right)},\qquad K_a = \frac{T_+}{T_+ + T_-}, \tag{1}$$

where

$$\bar{m}_1(x) = \begin{cases} m(x), x \in E_+, \\ 0, x \in E_- \end{cases},\qquad \bar{m}_0(x) = \begin{cases} 0, x \in E_+, \\ m(x), x \in E_- \end{cases},$$

$$\bar{1}_0(x) = \begin{cases} 0, x \in E_+, \\ 1, x \in E_- \end{cases},\qquad (\rho,f) = \int\limits_E f(x)\rho(dx)\,,$$

$\rho(dx)$ is supporting EMC $\left\{\xi_n^{(0)}; n \ge 0\right\}$ stationary distribution; $m(x)$ are the average sojourn times for the states x of the initial system; $P^{(r)}(x,B)$ are the initial system EMC $\{\xi_n; n \ge 0\}$ transition probabilities; r is the minimum number of steps through which the system can transit to the failure states subset E_- from the operable states subset E_+ where both disjoint subsets are the members of the ergodic class E_0.

An important moment of using of the method applied is the choice of the supporting system S_0. Suppose that devices A_i,$i = \overline{1,n}$ have fast renewal, i.e. their renewal times $\alpha_i^{(1)}$ depends on small positive parameter ε in such a manner that

$$\lim_{\varepsilon \to 0} E\alpha_i^{(1,\varepsilon)} = 0,\ \ i = \overline{1,n}, \tag{2}$$

and output device A_{n+1} has fixed time to failure and recovery time. This leads to the fact that recovery system S_0 will be the system with instant recovery of the devices $A_i, i = \overline{1,n}$ and completely filled storage devices H_i.

Let us determine embedded Markov chain (EMC) $\left\{\xi_0^{(n)}; n \ge 0\right\}$ transitions probabilities for a semi-Markov process (SMP) describing the supporting system operation.

1. The system transits from the state $i\bar{0}_{(i)}\bar{x}\bar{h}$ to the state $i\bar{0}\bar{x}\bar{h}$ with unit probability.
2. In the case of states $i\bar{0}\bar{x}\bar{h}$, transitions to the states $j\bar{0}_{(j)}\bar{y}\bar{h}$ are possible for $j = \overline{1, n+1}$, transition distribution density is calculated by formula

$$p_{i\bar{0}\bar{x}\bar{h}}^{j\bar{0}_{(j)}\bar{y}\bar{h}} = \begin{cases} \dfrac{f_j^{(0)}(x_j+t)\prod\limits_{\substack{l=1\\ l\neq j}}^{n+1}\bar{F}_l^{(0)}(x_l+t)}{\prod\limits_{k=1}^{n+1}\bar{F}_k^{(0)}(x_k)}, & i \neq j, \quad y_k = x_k + t, \quad k \neq j, \\ \dfrac{f_i^{(0)}(t)\prod\limits_{\substack{l=1\\ l\neq i}}^{n+1}\bar{F}_l^{(0)}(x_l+t)}{\prod\limits_{k=1}^{n+1}\bar{F}_k^{(0)}(x_k)}, & i = j. \end{cases} \tag{3}$$

Sojourn times in the states of the supporting system are

1. $\theta_{i\bar{0}_{(i)}\bar{x}\bar{h}} = 0, i = \overline{1, n}$, $\theta_{(n+1)\bar{0}_{(n+1)}\bar{x}\bar{h}} = \alpha_{n+1}^{(1)}$,
2. $\theta_{i\bar{0}\bar{x}\bar{h}} = \min\left\{\left[\alpha_j^{(0)} - x_j\right]^+, j = \overline{1, n+1}\right\}, \quad i = \overline{1, n+1},$

hence

$$E\theta_{i\bar{0}\bar{x}\bar{h}} = \int_0^\infty \frac{\bar{F}_i^{(0)}\prod\limits_{\substack{j=1\\ j\neq i}}^{n+1}\bar{F}_j^{(0)}(x_j+t)}{\prod\limits_{k=1}^{n+1}\bar{F}_k^{(0)}(x_k)}dt. \tag{4}$$

In Koroluk and Turbin [9] it is shown that EMC stationary distribution density $\left\{\xi_n^{(0)}; n \geq 0\right\}$ is defined by the formulas

$$\rho\left(i\bar{0}\bar{x}\bar{h}\right) = \rho\left(i\bar{0}_i\bar{x}\bar{h}\right) = \rho_0\prod_{k=1}^{n+1}\bar{F}_k^{(0)}(x_k), \tag{5}$$

where the constant ρ_0 is determined from the normalization requirement.

Thus, the ergodic class E^0 of the supporting system comprises the next states

$$\mathrm{E}^0 = \left\{i\bar{0}\bar{x}\bar{h}, i\bar{0}_{(i)}\bar{x}\bar{h}\right\}.$$

Let us proceed to the determination of real system stationary characteristics: time to failure $\mathrm{T}_+^{(h_1,\ldots,h_n)}$, average stationary renewal time $\mathrm{T}_-^{(h_1,\ldots,h_n)}$, stationary availability factor $\mathrm{K}_{\mathrm{a}}^{(h_1,\ldots,h_n)}$ and productivity $\mathrm{Pr}^{(h_1,\ldots,h_n)}$. Let us use the formulas (1).

Let us calculate $\left(\rho, \mathrm{P}^{(r)}\bar{1}_0\right)$, to do this let us determine beforehand real system transition probabilities for transits from ergodic states comprised by operable states subset E_+ to failure states. In this case $r = 1$ as the real system can

transit to failure subset of states E_- from operable subset E_+ comprised by the ergodic class E^0 during one step. E_- comprises the states with $d_{n+1} = 1$ or $d_{n+1} = 2$.

1. Let us consider the states $i\bar{0}\bar{x}\bar{h}$,$i = \overline{1, n+1}$. It is possible the one-step transition to failure states caused by the device A_n failure:
$i\bar{0}\bar{x}\bar{h} \;\to\; (n+1)\,\bar{0}_{(n+1)}\bar{y}\bar{h}, \bar{y} \;=\; (x_1 + t, ..., x_{i-1} + t, t, x_{i+1} + t, ..., x_n + t, 0)$,
the transition distribution density is defined by the formula (3) for $j = 1$.
2. In the case of the states $i\bar{0}_{(i)}\bar{x}\bar{h}, i \neq n+1$ let us consider the next transitions to the subset E_- of the real system:

(a) one-step transition to failure subset caused by the device A_n failure what is determined by the conditions:
$i\bar{0}_{(i)}\bar{x}\bar{h} \to (n+1)\,\bar{d}\bar{y}\bar{z}$,
$\bar{d} = \left(\underbrace{2, ..., 2}_{i-1}, 1, 2, ..., 2, 1\right)$, $\bar{y} = (x_1, ..., x_{i-1} + t, ..., x_n + t, 0)$,
$\bar{z} = (h_1, h_2, ..., h_{i-1}, h_i - t, h_{i+1}, ..., h_n)$, $0 \le t \le h_i$;
transition distribution density is calculated by the formula

$$p^{(n+1)\bar{d}\bar{y}\bar{z}}_{i\bar{0}_{(i)}\bar{x}\bar{h}} = \frac{f^{(0)}_{n+1}(x_{n+1} + t)\,\bar{F}^{(1,\varepsilon)}_i \prod\limits_{l=i+1}^{n} \bar{F}^{(0)}_l (x_l + t)}{\prod\limits_{s=i+1}^{n+1} \bar{F}^{(0)}_s (x_s)}, \tag{6}$$

(b) transition to the subset E_- along the chain of states caused by consequent depletion of the storage devices; this causes the failure of only one device; the probability of this transition is determined by the equity

$$\mathrm{P}^{\mathrm{E}_-}_{i\bar{0}_{(i)}\bar{x}\bar{h}} = \frac{\bar{F}^{(1,\varepsilon)}_i \left(\sum\limits_{k=i}^{n} h_k\right) \prod\limits_{l=i+1}^{n+1} \bar{F}^{(0)}_l \left(x_l + \sum\limits_{k=i}^{l-1} h_k\right)}{\prod\limits_{s=i+1}^{n+1} \bar{F}^{(0)}_s (x_s)}. \tag{7}$$

Using (6), (7) and the form of the stationary distribution (5) we can obtain that

$$\left(\rho, \mathrm{P}^{(r)}\bar{1}_0\right) = \int\limits_{\mathrm{E}_+} \mathrm{P}^{(r)}(z, \mathrm{E}_-)\,\rho(dz)$$

$$= \rho_0 \left[\sum_{i=1}^{n} \int\limits_0^\infty ... \int\limits_0^\infty \prod_{l=1}^{n} \bar{F}^{(0)}_l (x_l + t)\, f^{(0)}_{n+1}(x_{n+1} + t)\, d\bar{x}^{(i)} dt \right.$$

$$\left. + \underbrace{\int\limits_0^\infty ... \int\limits_0^\infty}_{n+1} \prod_{l=1}^{n} \bar{F}^{(0)}_l (x_l + t) f^{(0)}_{n+1}(t)\, d\bar{x}^{(n+1)} dt + \sum_{i=1}^{n} \int\limits_0^\infty \bar{F}^{(0)}_1 (x_1) dx_1 ... \int\limits_0^\infty \bar{F}^{(0)}_{i-1}(x_{i-1})\, dx_{i-1}\right.$$

$$\int_0^\infty dx_{i+1} \ldots \int_0^\infty dx_{n+1} \int_0^{h_i} \bar{F}_i^{(1,\varepsilon)}(t) \bar{F}_{i+1}^{(0)}(x_{i+1}+t) \ldots \bar{F}_n^{(0)}(x_n+t) f_{n+1}^{(0)}(x_{n+1}+t) dt$$

$$+\sum_{i=1}^{n} \bar{F}_i^{(1,\varepsilon)} \left(\sum_{k=i}^{n} h_k \right) \prod_{k=1}^{i-1} \mathrm{E}\alpha_k^{(0)} \int_{h_i}^{\infty} \bar{F}_{i+1}^{(0)}(x_{i+1})\, dx_{i+1} \int_{h_i+h_{i+1}}^{\infty} \bar{F}_{i+2}^{(0)}(x_{i+2})\, dx_{i+2}$$

$$\ldots \int_{h_i+\ldots+h_n}^{\infty} \bar{F}_{n+1}^{(0)}(x_{n+1})\, dx_{n+1} \Bigg] \approx \rho_0 \left[\prod_{k=1}^{n} \mathrm{E}\ \alpha_k^{(0)} + \sum_{i=1}^{n} \prod_{\substack{k=1 \\ k}}^{n} \mathrm{E}\alpha_k^{(0)} \int_0^{h_i} \bar{F}_i^{(1,\varepsilon)}(t) F_{n+1}^{(0)}(t) dt \right.$$

$$\left. +\sum_{i=1}^{n} \prod_{r=1}^{i-1} \mathrm{E}\alpha_r^{(0)} \bar{F}_i^{(1,\varepsilon)} \left(\sum_{k=i}^{n} h_k \right) \prod_{m=i+1}^{n+1} \int_{\sum_{l=i}^{m-1} h_l}^{\infty} \bar{F}_m^{(0)}(x_m) dx_m \right] \approx \rho_0 \left[\prod_{k=1}^{n} \mathrm{E}\ \alpha_k^{(0)} \right.$$

$$\left. +\sum_{i=1}^{n} \prod_{\substack{k=1 \\ k\neq i}}^{n} \mathrm{E}\alpha_k^{(0)} \int_0^{h_i} \bar{F}_i^{(1,\varepsilon)}(t) dt + \sum_{i=1}^{n} \prod_{r=1}^{i-1} \mathrm{E}\alpha_r^{(0)} \bar{F}_i^{(1,\varepsilon)} \left(\sum_{k=i}^{n} h_k \right) \int_{\sum_{l=i}^{m-1} h_l}^{\infty} \bar{F}_m^{(0)}(x_m) dx_m \right].$$

Let us calculate $(\rho, \bar{m}_1)$, to do this let us find the average sojourn times in the states $m(z)$ of the real system.

1. Sojourn time in the state $i\bar{0}\bar{x}\bar{h}$ is calculated by the formula:

$$\theta_{i\bar{0}\bar{x}\bar{h}} = \min \left\{ \left[\alpha_k^{(0)} - x_k \right]^+, \quad k = \overline{1, n+1} \right\},$$

therefore

$$\mathrm{E}\theta_{i\bar{0}\bar{x}\bar{h}} = \int_0^\infty \frac{\bar{F}_i^{(0)}(t) \prod\limits_{\substack{k=1 \\ k\neq i}}^{n+1} \bar{F}_k^{(0)}(x_k+t)}{\prod\limits_{k=1}^{n+1} \bar{F}_k^{(0)}(x_k)} dt. \tag{8}$$

2. For the states $i\bar{0}_{(i)}\bar{x}\bar{h}, \quad i = \overline{1, n}$

$$\theta_{i\bar{0}_{(i)}\bar{x}\bar{h}} = \min \left\{ h_i, \alpha_i^{(1,\varepsilon)}, \quad \left[\alpha_k^{(0)} - x_k \right]^+, \quad k = \overline{i+1, n+1} \right\}$$

and

$$\mathrm{E}\theta_{i\bar{0}_{(i)}\bar{x}\bar{h}} = \int_0^\infty \frac{\bar{F}_i^{(1,\varepsilon)}(t) \prod\limits_{k=i+1}^{n+1} \bar{F}_k^{(0)}(x_k+t)}{\prod\limits_{k=i+1}^{n+1} \bar{F}_k^{(0)}(x_k)} dt. \tag{9}$$

Hence, using the Eqs. (5), (8), (9) we can obtain that

$$(\rho,\bar{m}_1)=\int\limits_{\mathrm{E}_+} m\,(z)\,\rho\,(dz)=\rho_0\left[\sum_{i=1}^{n+1}\int\limits_0^\infty\ldots\int\limits_0^\infty\int\limits_0^\infty\prod_{\substack{k=1\\k\neq i}}^{n+1}\bar{F}_k^{(0)}(x_k+t)\;\bar{F}_i^{(0)}(t)d\bar{x}^{(i)}dt\right.$$
$$+\sum_{i=1}^{n}\prod_{k=1}^{i-1}\mathrm{E}\alpha_k^{(0)}\int\limits_0^\infty dx_{i+1}\int\limits_0^\infty dx_{i+2}\ldots\int\limits_0^\infty dx_{n+1}\int\limits_0^{h_i}\bar{F}_i^{(1,\varepsilon)}(t)\bar{F}_{i+1}^{(0)}(x_{i+1}+t)$$
$$\left.\ldots\,\bar{F}_{n+1}^{(0)}(x_{n+1}+t)dt\right]\approx\rho_0\left[\prod_{k=1}^{n+1}\mathrm{E}\alpha_k^{(0)}+\sum_{i=1}^{n}\prod_{\substack{k=1\\k\neq i}}^{n+1}\mathrm{E}\alpha_k^{(0)}\int\limits_0^{h_i}\bar{F}_i^{(1,\varepsilon)}(t)dt\right].$$

Thus, the stationary time to failure of the real system is approximately calculated by the formula

$$\mathrm{T}_+^{(h_1,\ldots,h_n)}\approx\left[\prod_{k=1}^{n+1}\mathrm{E}\alpha_k^{(0)}+\sum_{i=1}^{n}\prod_{\substack{k=1\\k\neq i}}^{n+1}\mathrm{E}\alpha_k^{(0)}\int\limits_0^{h_i}\bar{F}_i^{(1,\varepsilon)}(t)dt\right]\Bigg/\left[\prod_{k=1}^{n}\mathrm{E}\alpha_k^{(0)}\right.$$
$$+\sum_{i=1}^{n}\prod_{\substack{k=1\\k\neq i}}^{n}\mathrm{E}\alpha_k^{(0)}\int\limits_0^{h_i}\bar{F}_i^{(1,\varepsilon)}(t)dt+\sum_{i=1}^{n}\prod_{r=1}^{i-1}\mathrm{E}\alpha_k^{(0)}\bar{F}_i^{(1,\varepsilon)}(\sum_{k=i}^{n}h_k)$$
$$\left.\times\prod_{m=i+1}^{n+1}\int\limits_{\sum\limits_{l=i}^{m-1}h_l}^\infty\bar{F}_m^{(0)}(x_m)dx_m\right]. \tag{10}$$

Taking into account the smallness of the factors like $\int\limits_0^{h_i}\bar{F}_i^{(1,\varepsilon)}(t)dt$ due to condition (2) we can also write the next formula for $\mathrm{T}_+^{(h_1,\ldots,h_n)}$:

$$\mathrm{T}_+^{(h_1,\ldots,h_n)}\approx\frac{\prod\limits_{k=1}^{n+1}\mathrm{E}\alpha_k^{(0)}}{\sum\limits_{i=1}^{n}\prod\limits_{r=1}^{i-1}\mathrm{E}\alpha_k^{(0)}\bar{F}_i^{(1,\varepsilon)}(\sum\limits_{k=i}^{n}h_k)\prod\limits_{m=i+1}^{n+1}\int\limits_{\sum\limits_{l=i}^{m-1}h_l}^\infty\bar{F}_m^{(0)}(x_m)dx_m}. \tag{11}$$

Let us proceed to the obtaining of the average stationary renewal time, to do this it is necessary to determine the expression $\left(\rho,\mathrm{P}^{(r)}\bar{m}_0\right)$. Taking into account the formulas (8)–(9), we have

$$\left(\rho,\mathrm{P}^{(r)}\bar{m}_0\right)=\int\limits_{\mathrm{E}}\rho(dz)\int\limits_{\mathrm{E}_-}m(y)\mathrm{P}^{(r)}(z,dy)=\rho_0\left[\mathrm{E}\alpha_{n+1}^{(1)}\prod_{k=1}^{n}\mathrm{E}\alpha_k^{(0)}\right.$$

$$+\sum_{i=1}^{n}\underbrace{\int_0^\infty\dots\int_0^\infty}_{n}\prod_{k=1}^{i-1}\bar F_k^{(0)}(x_k)d\bar x^{(i)}\int_0^{h_i}f_{n+1}^{(0)}(x_{n+1}+t)\prod_{l=i+1}^{n}\bar F_l^{(0)}(x_l+t)dt\int_0^\infty\bar F_i^{(1,\varepsilon)}(t+y)$$

$$\times\bar F_{n+1}^{(1)}(y)dy+\sum_{i=1}^{n}\prod_{k=1}^{i-1}\mathrm{E}\alpha_k^{(0)}\int_{h_i}^\infty\bar F_{i+1}^{(0)}(x_{i+1})dx_{i+1}\int_{h_i+h_{i+1}}^\infty\bar F_{i+2}^{(0)}(x_{i+2})dx_{i+2}$$

$$\dots\int_{h_i+\dots+h_n}^\infty\bar F_{n+1}^{(0)}(x_{n+1})dx_{n+1}\int_0^\infty\bar F_i^{(1,\varepsilon)}\left(h_i+\dots+h_n+t\right)dt\Bigg]\approx\rho_0\Bigg[\mathrm{E}\alpha_{n+1}^{(1)}\prod_{k=1}^{n}\mathrm{E}\alpha_k^{(0)}$$

$$+\mathrm{E}\alpha_{n+1}^{(1)}\sum_{i=1}^{n}\prod_{\substack{k=1\\k\neq i}}^{n}\mathrm{E}\alpha_k^{(0)}\int_0^{h_i}\bar F_i^{(1,\varepsilon)}(t)dt$$

$$+\sum_{i=1}^{n}\prod_{r=1}^{i-1}\mathrm{E}\alpha_r^{(0)}\int_{\sum_{k=i}^{n}h_k}^\infty\bar F_i^{(1,\varepsilon)}(t)dt\prod_{m=i+1}^{n+1}\int_{\sum_{l=i}^{m-1}h_l}^\infty\bar F_m^{(0)}(x_m)dx_m\Bigg].$$

Let us obtain the next approximate formula for $\mathrm{T}_-^{(h_1,\dots,h_n)}$:

$$\mathrm{T}_-^{(h_1,\dots,h_n)}\approx[\mathrm{E}\,\alpha_{n+1}^{(1)}\prod_{i=1}^{n}\mathrm{E}\alpha_i^{(0)}+\mathrm{E}\alpha_{n+1}^{(1)}\sum_{i=1}^{n}\prod_{\substack{k=1,\\k\neq i}}^{n}\mathrm{E}\alpha_k^{(0)}\int_0^{h_i}\bar F_i^{(1,\varepsilon)}(t)dt$$

$$+\sum_{i=1}^{n}\prod_{r=1}^{i-1}\mathrm{E}\alpha_r^{(0)}\int_{\sum_{k=i}^{n}h_k}^\infty\bar F_i^{(1,\varepsilon)}(t)\,dt\prod_{m=i+1}^{n+1}\int_{\sum_{l=i}^{m-1}h_l}^\infty\bar F_m^{(0)}(x_m)dx_m\Bigg]\Bigg/\Bigg[\prod_{k=1}^{n}\mathrm{E}\alpha_k^{(0)}$$

$$+\sum_{i=1}^{n}\prod_{\substack{k=1,\\k\neq i}}^{n}\mathrm{E}\alpha_k^{(0)}\int_0^{h_i}\bar F_i^{(1,\varepsilon)}(t)dt$$

$$+\sum_{i=1}^{n}\bar F_i^{(1,\varepsilon)}(\sum_{k=i}^{n}h_k)\prod_{r=1}^{i-1}\mathrm{E}\alpha_r^{(0)}\prod_{m=i+1}^{n+1}\int_{\sum_{l=i}^{m-1}h_l}^\infty\bar F_m^{(0)}(x_m)dx_m\Bigg],\tag{12}$$

as well as

$$\mathrm{T}_-^{(h_1,\dots,h_n)}\approx\frac{\mathrm{E}\alpha_{n+1}^{(1)}\prod_{i=1}^{n}\mathrm{E}\alpha_i^{(0)}+\sum_{i=1}^{n}\prod_{r=1}^{i-1}\mathrm{E}\alpha_r^{(0)}\int_{\sum_{k=i}^{n}h_k}^\infty\bar F_i^{(1,\varepsilon)}(t)dt\prod_{m=i+1}^{n+1}\int_{\sum_{l=i}^{m-1}h_l}^\infty\bar F_m^{(0)}(x_m)dx_m}{\prod_{k=1}^{n}\mathrm{E}\alpha_k^{(0)}+\sum_{i=1}^{n}\bar F_i^{(1,\varepsilon)}(\sum_{k=i}^{n}h_k)\prod_{r=1}^{i-1}\mathrm{E}\alpha_r^{(0)}\prod_{m=i+1}^{n+1}\int_{\sum_{l=i}^{m-1}h_l}^\infty\bar F_m^{(0)}(x_m)dx_m}.\tag{13}$$

Using the formulas for $\mathrm{T}_{+}^{(h_1,\ldots,h_n)}$ and $\mathrm{T}_{-}^{(h_1,\ldots,h_n)}$ let us obtain the formulas for stationary availability factor. For example, based on the formulas (11), (13), we have:

$$\mathrm{K}_a^{(h_1,\ldots,h_n)} \approx \prod_{k=1}^{n} \mathrm{E}\alpha_k^{(0)} \Bigg/ \Bigg[\prod_{k=1}^{n+1} \mathrm{E}\alpha_k^{(0)} + \mathrm{E}\alpha_{n+1}^{(1)} \prod_{i=1}^{n} \mathrm{E}\alpha_i^{(0)} + \sum_{i=1}^{n} \prod_{r=1}^{i-1} \mathrm{E}\alpha_r^{(0)} \int\limits_{\sum_{k=i}^{n} h_k}^{\infty} \bar{F}_i^{(1,\varepsilon)}(t)\, dt \prod_{m=i+1}^{n+1} \int\limits_{\sum_{l=i}^{m-1} h_l}^{\infty} \bar{F}_m^{(0)}(x_m) dx_m \Bigg]. \tag{14}$$

The productivity $\mathrm{Pr}^{(h_1,\ldots,h_n)}$ of the real system is approximately defined by the next formula

$$\mathrm{Pr}^{(h_1,\ldots h_n)} = \mathrm{K}_a^{(h_1,\ldots,h_n)} \cdot c_{n+1},$$

where c_{n+1} is the productivity of the output device A_{n+1}.

To obtain approximately the probability of failure-free operation let us use the phase merging algorithm. Preliminary let us change the failure criterion for the real system: the system is in failure state if the output A_{n+1} device being in operable state can not produce the production due to emptness of the storage device B_n. The supporting system S_0, the class of the ergodic states E^0 and the supporting system EMC stationary distribution are stayed the same.

Let us calculate the expressions q and $\hat{m}$ comprised in the formula for the probability of failure-free operation

$$\mathrm{P}\left\{\varsigma^{(h_1,\ldots,h_n)} > t\right\} \approx \mathrm{e}^{-\frac{q}{\hat{m}}t},$$

where

$$q = \int\limits_{\mathrm{E}^0} \rho(dx) \int\limits_{\mathrm{E'}_1} \ldots \int\limits_{\mathrm{E'}_r} \mathrm{P}(x, dx_1)\ldots\mathrm{P}(x_{r-1}, dx_r)\mathrm{P}(x_r, \mathrm{E}_-)$$
$$= \rho_0 \sum_{i=1}^{n} \bar{F}_i^{(1,\varepsilon)}\Big(\sum_{k=i}^{n} h_k\Big) \prod_{r=1}^{i-1} \mathrm{E}\alpha_r^{(0)} \prod_{m=i+1}^{n+1} \int\limits_{\sum_{l=i}^{m-1} h_l}^{\infty} \bar{F}_m^{(0)}(x_m) dx_m.$$

Using the formulas (4), (5), let us obtain

$$\hat{m} = \int\limits_{\mathrm{E}^0} m(x)\rho(dx) = \rho_0 \Bigg[\mathrm{E}\alpha_{n+1}^{(1)} \int\limits_0^{\infty} \ldots \int\limits_0^{\infty} \prod_{k=1}^{n} \bar{F}_k^{(0)}(x_k) d\bar{x}^{(n+1)} + \sum_{i=1}^{n+1} \int\limits_0^{\infty} \ldots \int\limits_0^{\infty} d\bar{x}^{(i)} \int\limits_0^{\infty} \prod_{\substack{j=1,\\ j\neq i}}^{n+1} \bar{F}_j^{(0)}(x_j + t) \bar{F}_i^{(0)}(t) dt \Bigg]$$

$$= \rho_0 \prod_{k=1}^{n} \mathrm{E}\alpha_k^{(0)} \left[\mathrm{E}\alpha_{n+1}^{(0)} + \mathrm{E}\alpha_{n+1}^{(1)}\right].$$

Thus, based on phase merging algorithm and taking into account accepted additional failure criterion, we can approximately calculate the probability of failure-free operation for single-stream system using the formula

$$\mathrm{P}\left\{\varsigma^{(h_1,\ldots,h_n)} > t\right\} \approx \mathrm{e}^{-\Lambda(h_1,\ldots,h_n)t},$$

where the parameter $\Lambda(h_1,\ldots,h_n)$ is defined by the equation

$$\Lambda(h_1,\ldots,h_n) = \frac{\sum\limits_{i=1}^{n} \bar{F}_i^{(1,\varepsilon)}(\sum\limits_{k=i}^{n} h_k) \prod\limits_{r=1}^{i-1} \mathrm{E}\alpha_r^{(0)} \prod\limits_{m=i+1}^{n+1} \int\limits_{\sum\limits_{l=i}^{m-1} h_l}^{\infty} \bar{F}_m^{(0)}(x_m)dx_m}{\left[\mathrm{E}\alpha_{n+1}^{(0)} + \mathrm{E}\alpha_{n+1}^{(1)}\right] \prod\limits_{k=1}^{n} \mathrm{E}\alpha_k^{(0)}},$$

and the average time of the system's operation until the first failure is

$$\mathrm{E}\varsigma^{(h_1,\ldots,h_n)} \approx \frac{\left[\mathrm{E}\alpha_{n+1}^{(0)} + \mathrm{E}\alpha_{n+1}^{(1)}\right] \prod\limits_{k=1}^{n} \mathrm{E}\alpha_k^{(0)}}{\sum\limits_{i=1}^{n} \bar{F}_i^{(1,\varepsilon)}(\sum\limits_{k=i}^{n} h_k) \prod\limits_{r=1}^{i-1} \mathrm{E}\alpha_r^{(0)} \prod\limits_{m=i+1}^{n+1} \int\limits_{\sum\limits_{l=i}^{m-1} h_l}^{\infty} \bar{F}_m^{(0)}(x_m)dx_m}.$$

As an example let us consider the application of the formulas (11), (13), (14) in the case of the triple-phase system. For (n = 2) these formulas take form:

$$\mathrm{T}_+^{(h_1,h_2)} \approx \left[\mathrm{E}\alpha_1^{(0)}\mathrm{E}\alpha_2^{(0)}\mathrm{E}\alpha_3^{(0)}\right] \Big/ \left[\mathrm{E}\alpha_1^{(0)}\mathrm{E}\alpha_2^{(0)} + \mathrm{E}\alpha_1^{(0)}\bar{F}_2^{(1,\varepsilon)}(h_2) \times \int_{h_2}^{\infty} \bar{F}_3^{(0)}(t)dt + \bar{F}_1^{(1,\varepsilon)}(h_1+h_2) \int_{h_1}^{\infty} \bar{F}_2^{(0)}(t)dt \int_{h_1+h_2}^{\infty} F_3^{(0)}(x)dx\right],$$

$$\mathrm{T}_-^{(h_1,h_2)} \approx \left[\mathrm{E}\alpha_1^{(0)}\mathrm{E}\alpha_2^{(0)}\mathrm{E}\alpha_3^{(0)} + \mathrm{E}\alpha_1^{(0)} \int_{h_2}^{\infty} \bar{F}_1^{(1,\varepsilon)}(t)dt \int_{h_2}^{\infty} \bar{F}_3^{(0)}(x)dx + \int_{h_1+h_2}^{\infty} \bar{F}_1^{(1,\varepsilon)}(t)dt \int_{h_1}^{\infty} \bar{F}_2^{(0)}(x)dx \int_{h_1+h_2}^{\infty} \bar{F}_3^{(0)}(y)dy\right] \Big/ \left[\mathrm{E}\alpha_1^{(0)}\mathrm{E}\alpha_2^{(0)} + \mathrm{E}\alpha_1^{(0)}\bar{F}_2^{(1,\varepsilon)}(h_2) \int_{h_2}^{\infty} \bar{F}_3^{(0)}(t)dt + \bar{F}_1^{(1,\varepsilon)}(h_1+h_2) \int_{h_1}^{\infty} \bar{F}_2^{(0)}(t)dt \int_{h_1+h_2}^{\infty} F_3^{(0)}(x)dx\right],$$

$$K_a^{(h_1,h_2)} \approx \left[\mathrm{E}\alpha_1^{(0)}\mathrm{E}\alpha_2^{(0)}\mathrm{E}\alpha_3^{(0)}\right] \Big/ \left[\mathrm{E}\alpha_1^{(0)}\mathrm{E}\alpha_2^{(0)}\mathrm{E}\alpha_3^{(0)} + \mathrm{E}\alpha_1^{(0)}\mathrm{E}\alpha_2^{(0)}\mathrm{E}\alpha_3^{(1)} + \mathrm{E}\alpha_1^{(0)} \int_{h_2}^{\infty} \bar{F}_2^{(1,\varepsilon)}(t)dt \int_{h_2}^{\infty} \bar{F}_3^{(0)}(x)dx + \int_{h_1+h_2}^{\infty} \bar{F}_1^{(1,\varepsilon)}(t)dt \int_{h_1}^{\infty} \bar{F}_2^{(0)}(x)dx \int_{h_1+h_2}^{\infty} \bar{F}_3^{(0)}(y)dy$$

$$+\mathrm{E}\alpha_1^{(0)}\int\limits_{h_2}^{\infty}\bar{F}_2^{(1,\varepsilon)}(t)dt\int\limits_{h_2}^{\infty}\bar{F}_3^{(0)}(x)dx+\int\limits_{h_1+h_2}^{\infty}\bar{F}_1^{(1,\varepsilon)}(t)dt\int\limits_{h_1}^{\infty}\bar{F}_2^{(0)}(x)dx\int\limits_{h_1+h_2}^{\infty}\bar{F}_3^{(0)}(y)dy\Bigg].$$

Let the times to failure of the devices A_1, A_2, A_3 of the same productivity have third order Erlang distribution with the parameter $\lambda = 0,10$; the average time to failure of the devices is equal to 30 h. The renewal times of the devices has Veibull-Gnedenko distribution with scale parameter $\theta = 1$ and form parameter $\beta = 20$; the average renewal time for the devices is equal to 0,974h. The values of $\mathrm{T}_+^{(h_1,h_2)}$, $\mathrm{T}_-^{(h_1,h_2)}$, $\mathrm{K}_a^{(h_1,h_2)}$, calculated under the condition that $h_1+h_2=2$ are given in the Table 1. The data, given in the table, shows how the reliability characteristics of the triple-phase system change during redistribution of prescribed time reserve between storage devices.

Table 1. Reliability characteristics for the triple-phase system in the case of prescribed summary volume of the storage devices

Case	h_1	h_2	$\mathrm{T}_+^{(h_1,h_2)}$	$\mathrm{T}_-^{(h_1,h_2)}$	$\mathrm{K}_a^{(h_1,h_2)}$	$1-\mathrm{K}_a^{(h_1,h_2)}$
1	0	2.0	30	0.975	0.969	0.031
2	0.2	1.8	30	0.975	0.969	0.031
3	0.4	1.6	30	0.975	0.969	0.031
4	0.6	1.4	30	0.975	0.969	0.031
5	0.8	1.2	30	0.975	0.969	0.031
6	1.0	1.0	22.130	0.727	0.968	0.032
7	1.2	0.8	15.289	0.583	0.963	0.037
8	1.4	0.6	15.152	0.677	0.957	0.043
9	1.6	0.4	15.101	0.776	0.951	0.049
10	1.8	0.2	15.050	0.875	0.945	0.055
11	2.0	0	15	0.974	0.939	0.061

The Table 2 represents the results of four-phase information system stationary characteristics finding using the formulas (11), (13), (14) and using the simulation modeling. During the modeling it was assumed that all RV (operation and renewal times for all the devices) are distributed by the exponential law.

Table 2. Modeling results for the four-phase system

Initial data				
Four-phase system $E\alpha_1^{(0)} = E\alpha_2^{(0)} = E\alpha_3^{(0)} = E\alpha_4^{(0)} = 17\,h,\ \ E\alpha_1^{(1)} = E\alpha_2^{(1)} = 0.6\,h,$ $E\alpha_3^{(1)} = E\alpha_4^{(1)} = 0.8\,h,\ h_1 = h_2 = h_3 = h$, processing time - $0,2\,h$				
h	Analytic modeling results			Simulation modeling results
	$T_+^{(h_1,h_2,h_3)}, h$	$T_-^{(h_1,h_2,h_3)}, h$	$K_a^{(h_1,h_2,h_3)}$	$K_a^{(h_1,h_2,h_3)}$
0	4.250	0.700	0.859	0.879
0.1	6.518	0.736	0.899	0.929
0.2	8.693	0.736	0.919	0.932
0.3	10.507	0.780	0.931	0.923
0.4	11.929	0.790	0.938	0.933
0.5	13.029	0.795	0.943	0.927
0.6	13.885	0.797	0.946	0.937
0.7	14.556	0.799	0.948	0.940
0.8	15.084	0.799	0.950	0.939
0.9	15.501	0.800	0.951	0.935
1.0	15.830	0.800	0.952	0.941
1.1	16.089	0.800	0.953	0.943
1.2	16.292	0.800	0.953	0.945
1.3	16.451	0.800	0.954	0.946
1.4	16.575	0.800	0.954	0.947
1.5	16.671	0.800	0.954	0.945

3 Conclusion

In present paper the semi-Markov mode of a single-stream information system with intermediate storage devices is constructed. In information systems time reserve can be used for the switching of the system, renewal after failures, elimination of the malfunction and distortions. With the help of phase merging algorithm the stationary time to failure, the stationary renewal time and the stationary availability factor for the system considered. It is shown that occurence of intermediate storage devices affects sygnifically into information system reliability characteristics.

It is supposed to construct in the future the models of more complicated structure systems. The results of the present paper can be used for optimum choice of the storage devices capacity.

References

1. Cherkesov, G.N.: Reliability of Technical Systems with Time Redundancy. Sovietskoe Radio, Moscow (1974)
2. Cherkesov, G.N.: Reliability of Hardware-Software Complexes. Piter, Sanct-Petersburg (2005)
3. Kredentser, B.P.: Prediction of Reliability of Systems with Time Redundancy. Naukova Dumka, Kiev (1978)
4. Yao, D.D., Buzacot, J.A.: Models of flexible manufacturing systems with limited local buffers. Int. J. Prod. Res. **24**, 107–118 (1986)
5. Belyaev, Yu.K., Bogatyirev, V.A., Bolotin, V.V., et al.: Reliability of Technical Systems. Handbook. Radio i Svyaz, Moscow (1985). (Ushakov, I.A. (ed.))
6. Ushakov, I.A.: Probabilistic Reliability Models. Wiley, San Diego (2012)
7. Curry, G.L., Feldman, R.M.: Manufacturing Systems Modeling and Analysis, 2nd edn. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-16618-1
8. MacGregor, S.J., Tan, B. (eds.): Handbook of Stochastic Models and Analysis of Manufacturing System Operations. Springer, New York (2013). https://doi.org/10.1007/978-1-4614-6777-9
9. Koroluk, V.S., Turbin, A.F.: Markovian Restoration Processes in the Problems of System Reliability. Naukova Dumka, Kiev (1982)
10. Koroluk, V.S.: Stochastic System Models. Naukova Dumka, Kiev (1982)
11. Obzherin, Yu.E, Boyko, E.G.: Semi-Markov Models: Control of Restorable Systems with Latent Failures. Elsevier Academic Press, London (2015)
12. Kashtanov, V.A., Medvedev, A.I.: Reliability Theory of the Complex Systems (Theory and Practice). Fizmatlit, Moscow (2010)
13. Koroluk, V.S., Limnios, N.: Stochastic Systems in Merging Phase Space. World Scientific Imperial College Press, London (2005)
14. Korlat, A.N., Kuznetsov, V.N., Novikov, M.M., Turbin, A.F.: Semi-Markov Models of Recoverable Systems and Queuing Systems. Shtiinta, Chisinau (1991)
15. Grabskiy, F.: Semi-Markov Processes: Applications in System Reliability and Maintenance. Elsevier Science, Amsterdam (2014)
16. Sevastyanov, B.A.: Influence of storage bin capacity on the average standstill time of a production line. Theory Probab. Appl. **7**(4), 429–438 (1962)
17. Dimitrakosa, T.D., Kyriakidis, E.G.: A semi-Markov decision algorithm for the maintenance of a production system with buffer capacity and continuous repair times. Int. J. Prod. Econ. **111**, 752–762 (2008)
18. Prabhu, N.U., Pacheco, A.: A storage model for data communication systems. Theory Probab. Appl. **39**(4), 604–627 (1995)
19. Ushakov, I.A.: Using of the theorem on random strerams rarefaction to solve reliability problems for the system with buffers. Reliab. Qual. Control **10**, 27–33 (1986)

Cluster-Based Energy Consumption Forecasting in Smart Grids

Eugene Yu. Shchetinin(✉)

Financial University under the Government of the Russian Federation,
Leningradsky pr. 49, 111123 Moscow, Russia
riviera-molto@mail.ru
http://www.fa.ru/

Abstract. Clustering is a well-known machine learning algorithm which enables the determination of underlying groups in datasets. In electric power systems it has been traditionally utilized for different purposes like defining consumer individual profiles, tariff designs and improving load forecasting. A new age in power systems structure such as smart grids determined the wide investigations of applications and benefits of clustering methods for smart meter data analysis. This paper presents an improvement of energy consumption forecasting methods by performing cluster analysis. For clustering the centroid based method K-means with K-means centroids was used. Various forecasting methods were applied to find the most effective ones with clustering procedure application. Used smart meter data have an hourly measurements of energy consumption time series of russian central region customers. In our computer modeling investigations we have obtained significant improvement due to carrying out the cluster analysis for consumption forecasting.

Keywords: Energy consumption · Smart meter · Smart grids
Data mining · Forecasting · Cluster analysis · K-means

1 Introduction

The development of intelligent networks in manufacturing, finance, and services creates new opportunities for the development and application of effective methods of machine learning and data analysis. The installation of smart meters is usually considered as the starting point in the implementation of smart grids. Smart meters employ advanced metering, control, data storage, and communication technologies to offer a range of functions. The deployment of smart meters provides benefits to the end consumers (domestic and non-domestic), energy suppliers, and network operators by providing near real-time consumption information to the consumers that will help them to manage their energy use, save money, and reduce greenhouse gas emissions. At the same time, smart meters will benefit distribution network planning and operation, and demand management. In this regard, the smart metering data will enable more accurate demand forecasts, allow improved asset utilisation in distribution networks,

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 445–456, 2018.
https://doi.org/10.1007/978-3-319-99447-5_38

locate outages and shorten supply restoration times, and reduce the operational and maintenance costs of the networks. Smart technologies for collecting, recording and monitoring data on energy consumption create a huge amount of data of different nature for the energy suppliers and network operators to exploit. The data volume will vary according to the number of installed smart meters, the number of received smart meter messages, the message size (in bytes per message), and the frequency of recording the measurements – e.g., every 15 or 30 min. These data can be used for optimal network management, improving the accuracy of the forecasting load, detection of abnormal effects of power supply (peak load conditions), the formation of flexible price tariffs for different groups of consumers [1–3].

One of the most important issues in this area is to predict the power load consumption as accurately as possible. Consumption, as a rule, have a rather complicated stochastic structure, which are difficult for modeling and prediction for individual consumers. Therefore, the consumption data is aggregated (summed). Statistical, engineering and time-series methods [4–6] have been reported to analyse and extract the required information from the load profiles of customers. Additionally, statistical time-series and artificial intelligence (AI) methods have been applied to estimate and forecast the load in power networks. However, these methods can be costly and complex to implement and validate when large volumes of consumption measurements become available. Nevertheless, when different methods of aggregation are applied to the group of consumers having similar statistical characteristics of time series of power consumption, it is possible to count on considerable progress in the solution of objectives. One efficient approach to extract the necessary information from smart meter measurements is the employment of data mining techniques. Cluster analysis is one type of these techniques [7,8]. Clustering is the grouping of load profiles into a number of clusters such that profiles within the same cluster are similar to each other.

The main goal of this paper is to investigate the practical issues and possible benefits of combining clustering procedures with forecasting methods in order to improve their accuracy. We have used several known approaches to present time series and different forecasting methods such as Holt-Winters, ARIMA model, Support Vector Regression and some others. The results of cluster analysis can also be beneficial for finding the patterns in data [3,4] for classification of new customers. The paper is organized as follows: Sect. 2 contains some models for energy time series presentations that we used for clustering and forecasting. In Sect. 3 we described our clustering algorithm for classification and forecasting. Section 4 contains the review of our computer experiments and Sect. 5 presents conclusions.

2 Modeling of the Energy Consumption Time Series

The problems of application of clustering methods to the time series of electricity consumption are mainly in high dimension and high noise level of the data, which can be solved as mentioned above with the use of machine learning methods.

Papers dealing with aggregating of consumers usually use clustering primarily for mainly one purpose: immediate forecasts of time series [6,7]. These works, however, put a little focus on application of forecasting methods and methods for time series data mining. Shahzadeh et al. [6] explore the clustering of consumers for feature extraction from time series and its impact on the accuracy of forecast of energy consumption. K-means was used for clustering and neural network for forecasting. They used three different representation of time series: estimated regression coefficients, extractions of the averages of electricity consumption and the whole time series. The best results were achieved by the clustering with regression coefficients, which showed significant improvements in the accuracy of forecast with the help of clustering. Rodrigues [7] presented a hierarchical clustering method with optimization criterion of forecast error of ARIMA model. This method was compared against simple aggregation consumption forecast. Experiments showed that the positive impact of consumers' aggregation on forecast accuracy of certain methods (linear regression, multi-layer perceptron and support vector regression) depends not only on the number of clusters, but also on the size of the customer base. To evaluate this hypothesis, Monte-Carlo grouping of consumers and also forecasting methods Random Walk and Holt-Winters exponential smoothing were used [12]. McLoughlin et al. [4] presented dynamic clustering of consumers. With clustering approach, a large amount of mean daily profiles was created and then deeply analyzed. To link domestic load profiles with household characteristics, SOM clustering and multi-nominal logistic regression were used to perform analysis of dwelling, occupant and appliance characteristics [10].

In this paper we have focused our attention on the influence of clustering of consumers on accuracy of different forecasting methods and on the representations of time series, which are suitable for seasonal times series of electricity load. Based on investigated approaches we made the comparative analysis of two approaches of classification of consumers: with clustering application and without it (aggregation). Our methodology has four steps:

(1) to normalize the data and calculate the energy consumption model for each consumer. In the future, the study uses four different models based on the representation of time series, which serve as inputs to the clustering method.

(2) The second stage consists of calculating the optimal number of clusters for the given time series representation and the selected data learning window.

(3) The third stage is clustering and aggregation of consumption within clusters. For each cluster, the forecast model is trained and the forecast for the next period is run.

(4) The forecasts are aggregated and compared with the real consumption data. Next, we construct a forecast for day-ahead for the received representations of the clusters using the above-described prediction methods.

The process (1)–(4) is iteratively repeated as measurements from a next day become available. The training window of data is changed so that the new day is added to the training window and the oldest one is removed. The standard approach without clustering is just a summation of all measurements together.

2.1 Energy Consumption Time Series Modeling

The time series X is an ordered sequence of n real variables

$$X = (x_1, x_2, ..., x_n), \quad x_i \in R. \tag{1}$$

The main reason for time series presentation using is a significant decrease in the dimension of the analyzed data, respectively, reducing the required memory and the computational complexity. In our approach next four model-based representation methods were chosen: (a) Robust Linear Model (RLM), (b) Generalized Additive Model (GAM), (c) Holt-Winters Exponential Smoothing, and (d) Median Filter.

The first presentation is based on a robust linear model (RLM). Like other regression methods, it is aimed at modeling the dependent variable by independent variables

$$x_i = \beta_1 u_{i1} + \beta_2 u_{i2} + ... + \beta_s u_{is} + \varepsilon_i, \tag{2}$$

where $i = 1, .., n$, u_i - is energy consumption, $\beta_1, ..., \beta_s$ are the regression coefficients. Let's define the frequency of one season as s. $u_{i1}, .., u_{is}$ are independent binary variables, u_{ij}, $j = 1, 2, ..., s$, variable ε_i is a white noise with the normal distribution $N\left(0, \sigma^2\right)$. Obtaining an estimate of the vector $\beta_1, ..., \beta_s$ is calculated by using reiterated weighted least squares (IRLS). Extensions for RLM are generalized additive models (GAM) [9,11].

$$E\left(x_i\right) = \beta_0 + f_1\left(u_{i1}\right), \tag{3}$$

where f_1 is B-spline [17], $u_1 = (1, 2, ..., s, 1, ..., s, ...)$ is a vector $(day_j)_{j=1}^{d}$, $day = (1, ..., s)$, d is a number of days. Model parameters (3) can be evaluated by the weighted least squares (IRLS) iterative method [5].

Holt-Winters exponential smoothing (HW) was used as another method of representations based on the model. It is a method that is used mainly to forecast the seasonal time series and to smoothing time series from the noise [12]. Components of this model (with trend and seasonality) are:

$$L_i = \alpha\left(x_i - s_{t-s}\right) + (1 - \alpha)\left(L_{i-1} + b_{i-1}\right), \tag{4}$$

$$b_i = \beta\left(L_i - L_{i-1}\right) + (1 - \beta)\, b_i, \tag{5}$$

$$r_i = \gamma\left(x_i - L_{i-1}\right) + (1 - \gamma)\, r_{i-s}, \tag{6}$$

where L is smoothing component, b is trend component, r is seasonal component, α, β, γ are smoothing factors. Smoothing factors have been selected automatically, where the factors were optimized according to the average square error of the one-stepwise forecast. As HW representation (4)–(6) we have taken seasonal coefficients $(r_{n-s+1}, ..., r_n)$. Last presentation for time series model is a median filter as following

$$\widehat{x}_k = median\left(x_k, x_{k+s}, .., x_{k+s\times(d-1)}\right), \tag{7}$$

where $k = (1, ..., s)$, and $d-$ time series dimension. An example of application of presentation (7) is shown on Fig. 1.

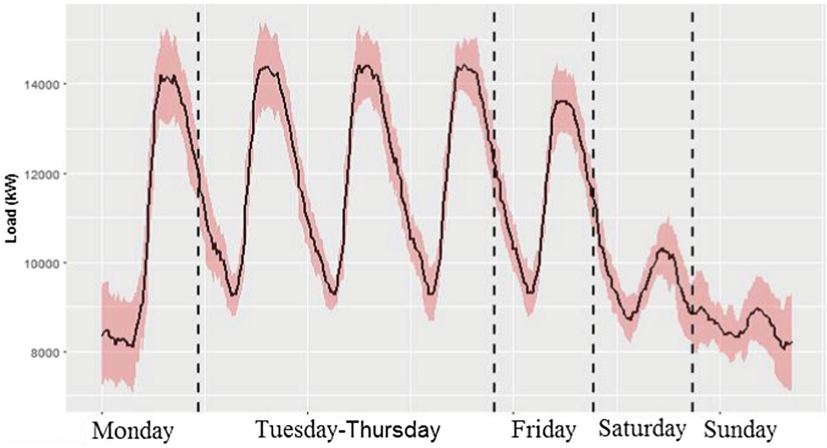


Fig. 1. Weekly median filter for energy consumption time series with absolute deviation.

2.2 Applications of Cluster Analysis for Smart Grids

As a common thing, utilities usually divided their customers in industrial, commercial and residential sectors based on some fixed information like voltage level, nominal demand etc. Based on this approach a set of customer class load profiles were defined and each user was assigned to one of these classes. However, this is still a fundamental problem, and the procedures for dealing with customers segmentation need to be revised greatly. Firstly, the consumption data of customers, those who have installed smart meters, are now accessible. Secondly, the time period of measurement is not restricted and usage information for some successive years is available. These two factors affect the dimensionality of data which is not comparable with previously used data sets. Finally, as the data is continuously recorded, it can have possible applications for real-time operation and management of power systems. All of these factors emphasize the use of new clustering methods for electricity consumption characterization. For classification consumers into groups (clusters), we used the centroid based clustering method K-means [9]. K-means is a method based on the mutual distances of objects, measured by Euclidean distance. The advantage over conventional K-means is based on carefully seeding of initial centroids, which improves the speed and accuracy of clustering. Before applying the K-means algorithm the optimal number of clusters k must be determined. For each representation of a data set, we have determined the optimal number of clusters to k using the internal validation rate Davies-Bouldin index [10]. The optimal number of clusters as been shown in our computer experiments ranged from 7 to 18. The results of this algorithm applied to energy consumption data [16] may be seen on Fig. 2. Computer code in R programming language, realized K-means algorithm, in Sect. 5 is reported.

Our algorithm works as follows. Let $d(x)$ denote the shortest Euclidean distance from a data point x to the closest centroid we have already chosen.

Step1. Choose an initial centroid K_1 randomly with uniform probability from set X.

Step 2. Choose the next center $K_i = \widehat{x} \in T$, selecting with some probability $P = \frac{d(\widehat{x})^2}{\sum_{x \in X} d(x)^2}$.

Step 3. Repeat previous step until we have chosen all K centers. Each object from data set is connected with a centroid that is closest to it. New centroids are then calculated.

Step 4. Last two steps are repeated until classification to clusters no longer changes. Euclidian distance measure is one of the best measures for comparison of time series of electricity load because of its stronger dependence on time. In each iteration of a batch processing, we have automatically determined the optimal number of clusters to K using the internal validation rate Davies-Bouldin index.

Computer code in R programming language, realized Euclidian function distance, is reported in Sect. 5 of this paper.

2.3 Energy Consumption Time Series Forecasting

We used mostly effective methods to improve forecasting energy consumption time series:

(1) Support Vector Regression (SVR), a method that works like simple linear regression but it tries to find a real regression function that best approximates output vector. SVR technique relies on kernel functions to construct the model. The commonly used kernel functions are: (a) Linear, (b) Polynomial, (c) Sigmoid and (d) Radial Basis. The selection of kernel function is a tricky and requires optimization techniques for the best selection. In the constructed SVR model, we used automated kernel selection. In our computer experiments the best forecasting results were shown by Radius Basis Function (RBF) kernel [8]

$$K(x_i, x_j) = exp\left(-\frac{(x_i - x_j)^2}{2\sigma^2}\right), \tag{8}$$

where x_i, x_j are energy consumption time series, σ^2 is variance.

(2) Seasonal decomposition of time series based on loses regression (STL) is a method, which decomposes seasonal time series into three parts: trend, seasonal component and remainder (noise) [11]. The seasonal component is found by loess (local regression) smoothing of the seasonal series, whereby smoothing can be effectively replaced by taking the mean. The seasonal values are removed, and the remainder is smoothed to find the trend. The

overall level is removed from the seasonal component and added to the trend component. For the final three time series any of the forecast methods is used separately, in our case either Holt-Winters exponential smoothing or ARIMA model.

(3) Random Forest (RF). First the Random Forests concept was proposed by Ho [4]. The method constructs the large number of decision trees at training time. Its output is the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

(4) Gradient Boosting Machine (GBM) is an efficient and scalable implementation of gradient boosting framework by Friedman [4]. The principle approach is to combine iteratively several simple models, called "weak learners", to obtain a "strong learner" with improved prediction accuracy. Paper [6] introduced a statistical point of view of boosting, connecting the boosting algorithm to the concepts of loss functions. It could extended the boosting to the regression by introducing the gradient boosting machines method (GBM). The GBM method can be seen as a numerical optimization algorithm that aims at finding an additive model that minimizes the loss function. Thus, the GBM algorithm iteratively adds at each step a new decision tree (i.e., "weak learner") that best reduces the loss function. More precisely, in regression, the algorithm starts by initializing the model by a first guess, which is usually a decision tree that maximally reduces the loss function (which is for regression the mean squared error), then at each step a new decision tree is fitted to the current residual and added to the previous model to update the residual. The algorithm continues to iterate until a maximum number of iterations, provided by the user, is reached. This process is so-called stage wise, meaning that at each new step the decision trees added to the model at prior steps are not modified. By fitting decision trees to the residuals the model is improved in the regions where it does not perform well.
We consider 4 seasonal variables to RF and GBM models for half-hourly and weekly periods in sinus and cosinus functions form $\frac{\left(\sin\left(2\pi\frac{day}{s}\right)+1\right)}{2}$ and $\frac{\left(\cos\left(2\pi\frac{day}{s}\right)+1\right)}{2}$ respectively, s is a period. For weekly periods $week$ is a vector $(s\ times\ week_j)_{j=1}^{d}$, $\frac{\left(\sin\left(2\pi\frac{week}{7}\right)+1\right)}{2}$ and $\frac{\left(\cos\left(2\pi\frac{week}{7}\right)+1\right)}{2}$, $week_j = (1, 2, ..., 7, 1, ..)$.

(5) Bagging (Bagg) predictors generate multiple versions of predictors and use them for determination an aggregated predictor. The aggregation is an average of all predictors. The bagging method gives substantial gains in accuracy, but the vital element is the instability of the prediction method. In the case that perturbing the learning set has significant influence on the constructed predictor, the bagging can improve accuracy.

(6) Regression Trees (R-Tree) are regression methods that consist of partitioning the input parameters space into distinct and non-overlapping regions following a set of if-then rules. The splitting rules identify regions that have the most homogeneous response to the predictor, and within each region a simple model, such as a constant, is fitted. The use of decision trees as a regression technique has several advantages, one of which is that the splitting rules represent an intuitive and very interpretable way to visualize the results. In addition, by their design, they can handle simultaneously numerical and categorical input parameters. They are robust to outliers and can efficiently deal with missing data in the input parameters space. The decision tree's hierarchical structure automatically models the interaction between the input parameters and naturally performs variable selection, e.g., if an input parameter is never used during the splitting procedure, then the prediction does not depend on this input parameter. Finally, decision trees algorithms are simple to implement and computationally efficient with a large amount of data [18].

The accuracy of the forecast of electricity consumption was measured by MAPE (Mean Absolute Percentage Error). MAPE is defined as follows:

$$MAPE = 100 \times \frac{1}{n} \sum_{i=1}^{n} \frac{|x_t - \overline{x}|}{x_t}, \tag{9}$$

where x_t is actual consumption, $\overline{x}$ - load forecast, n - length of time series.

3 Computer Experiments for Customer Energy Consumption

We performed the computer experiments to evaluate the profit of using clustering procedures on four time series representation methods for one day ahead forecast. Our testing data set contains measurements from customers of Central Russia Region [16]. Table 1 shows average daily MAPE forecast errors of 6 forecasting methods. Each forecasting method was evaluated on 5 datasets; 4 datasets are clustered with different representation methods ($Median, HW, GAM, RLM$) and aggregated electric load consumption (Sum). The following conclusions can be derived from Table 1. Optimized clustering of consumers significantly improves accuracy of forecast with forecasting methods $SVR, Bagging, GBM$. Despite this, clustering with STL+ARIMA, RF, R-Tree does not really improves accuracy of forecast. Three robust representation methods of time series Median, GAM and RLM performed best among all representations, while HW was the worst in most of the cases, because robust representations are stable and less fluctuate. The best result of all cases achieved by GBM with optimized clustering using GAM representation which mean daily MAPE error under 3,17% [14].

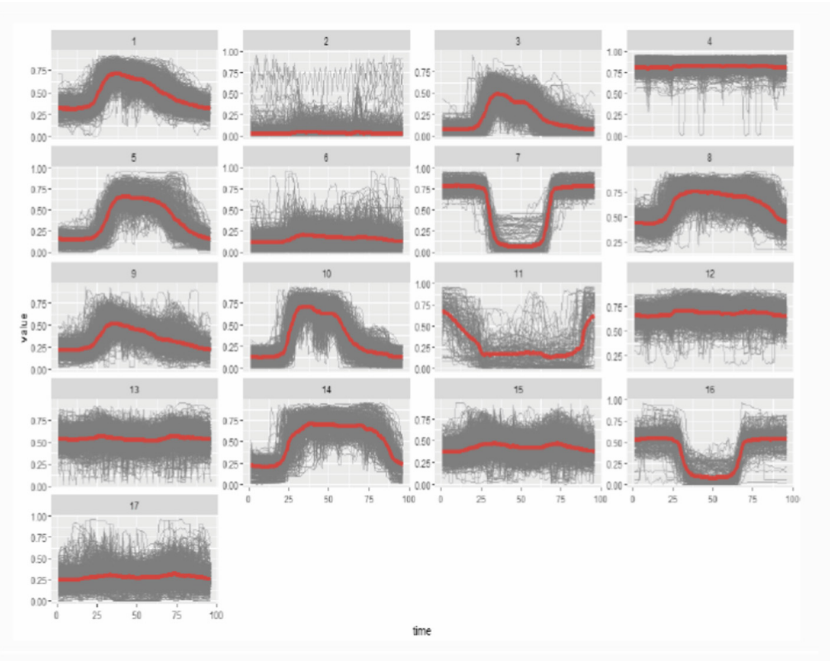


Fig. 2. Results of clustering of energy consumers: clusters and centroids.

Table 1. $MAPE(\%)$-error forecasting methods for aggregated load consumption. Repres.: model-based presentations of consumption time series. Meth.: forecasting methods applied with clustering.

Meth./Repres.	*Median*	*HW*	*GAM*	*RLM*	*SUM*
$STL + ARIMA$	4,873	4,947	4,423	4,674	4,56
SVR	4,073	4,072	4,42	4,216	4,163
Bagging	3,438	3,475	4, 23	3,34	4,13
GBM	4,036	4,036	3,282	4,321	4,241
$R - Forest$	4,479	4,476	4,36	4,62	4,61
$R - Tree$	4,419	4,476	4,33	4,26	4,98
Mean	4,541	4,739	4,56	4,723	5,27

4 Conclusions

Improving the accuracy of forecasts of electricity consumption is a key area in the development of intelligent energy grids. To implement this problem, we used machine learning methods, namely cluster analysis. The main purpose of this paper is to show that the application of the clustering procedure of consumers to the representation of time series of energy consumption can improve the

accuracy of their forecasts for energy consumption. Robust linear model, generalized additive model, exponential smoothing and median linear filter were used as such representations. In this paper we applied a modified K-means algorithm to more accurately select centroids and the Davis-Boldin index to evaluate clustering results. Numerical experiments have shown that the methods of forecasting such as $STL + ARIMA, SVR, RF, Bagging$ considered in the paper are more effective for improving forecast accuracy if used together with clustering. Prediction methods performed the best reliable representations of RLM, GAM, and median filter. The most accurate prediction result is obtained by GBM with the GAM presentation time series. Among the perspective applications of clustering for smart grids are benefits for tariff design, compilation smart demand response programs, improvement of load forecast, classifying new or non-metered customers and other tasks.

5 Program Code

R code for K-means algorithm

```
myKmeans <- function(x, centers, distFun, nItter=10) {
    clusterHistory <- vector(nItter, mode="list")
    centerHistory <- vector(nItter, mode="list")

    for(i in 1:nItter) {
        distsToCenters <- distFun(x, centers)
        clusters <- apply(distsToCenters, 1, which.min)
        centers <- apply(x, 2, tapply, clusters, mean)
        # Saving history
        clusterHistory[[i]] <- clusters
        centerHistory[[i]] <- centers
    }

    list(clusters=clusterHistory, centers=centerHistory)
}
```

R code for Euclidian distance between cluster centroids

```
myEuclid <- function(points1, points2) {
    distanceMatrix <- matrix(NA, nrow=dim(points1)[1], ncol=dim(points2)[1])
    for(i in 1:nrow(points2)) {
        distanceMatrix[,i] <- sqrt(rowSums(t(t(points1)-points2[i,])^2))
    }
    distanceMatrix
}
```

R code for modeling example of application K-means algorithm

```
mat <- matrix(rnorm(1000), ncol=2)
%initial centers of clusters definition
```

```
I-Centers <- mat[sample(nrow(mat), 7),]
%Resulted centers of clusters
K-Centers <- myKmeans(mat, I-Centers, myEuclid, 10)
%Plotting the results of K-means algorithm for  5 iterations
par(mfrow=c(2,2))
for(i in 1:5) {
plot(mat, col=theResult$clusters[[i]], main=paste("itteration:", i),
xlab="x", ylab="y")
points(theResult$centers[[i]], cex=3, pch=19,
col=1:nrow(theResult$centers[[i]]))
}
```

R code for GAM time series model presentation

```
%Generalized Additive Models
library(mgcv)
gamst <- proc.time()
z <- as.vector(log(ru.ext$rate$total))
x <- 1:nrow(ru.ext$rate$total)
y <- 1:ncol(ru.ext$rate$total)
xy <- expand.grid(x, y)
ru.gam <- gam(z~s(xy[,1],xy[,2], bs='ts', k=12^2))
gamen <- proc.time()
gamel <- gamen['elapsed'] - gamst['elapsed']
cat("Gam time passed:", gamel, "\n")
persp(matrix(fitted(ru.gam), nrow=length(x), ncol=length(y)))
persp(matrix(residuals(ru.gam), nrow=length(x), ncol=length(y)))
levelplot(matrix(residuals(ru.gam), nrow=length(x), ncol=length(y)))
wireframe(
matrix(fitted(ru.gam), nrow=52, ncol=52),
xlab = expression(a),
ylab = expression(y),
zlab = expression(m),
screen = list(z = 20, x = -70, y = 3)
)
```

References

1. Haben, S., Singleton, C., Grindrod, P.: Analysis and clustering of residential customers energy behavioral demand using smart meter data. IEEE Trans. Smart Grid **99**, 1–9 (2015)
2. Chicco, G., Napoli, R., Piglione, F.: Comparisons among clustering techniques for electricity customer classification. IEEE Trans. Power Sys. **21**, 933–940 (2013)
3. Gelling, C.W.: The Smart Grid: Enabling Energy Efficiency and Demand Response. The Fairmont Press Inc. (2009)
4. McLoughlin, F., Duffy, A., Conlon, M.: A clustering approach to domestic electricity load profile characterisation using smart metering data. Appl. Energy **141**, 190–199 (2015)
5. Aghabozorgi, S., Shirkhorshidi, A.: Time-series clustering: a decade review. Inf. Sys. **53**, 16–38 (2015)

6. Shahzadeh A., Khosravi A., Nahavandi S.: Improving load forecast accuracy by clustering consumers using smart meter data. In: International Joint Conference on Neural Networks (IJCNN), pp. 1–7 (2015)
7. Rodrigues, P., Gama, J., Pedroso, J.: Hierarchical clustering of time-series data streams. IEEE Trans. Knowl. Data Eng. **20**(5), 615–627 (2008)
8. Hsu, D.: Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data. Appl. Energy **160**, 153–163 (2015)
9. Andersen A.: Modern Methods for Robust Regression. SAGE Publications, Inc. (2008)
10. Wijaya T.K., Vasirani M., et al.: Cluster-based aggregate forecasting for residential electricity demand using smart meter data. In: 2015 IEEE International Conference on Big Data. IEEE Press, pp. 879–887 (2001)
11. Wood, S.: Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC (2006)
12. Hyndman, R.J., Koehler, A.B., Snyder, R.D., Grose, S.: A state space framework for automatic forecasting using exponential smoothing methods. Int. J. Forecast. **18**(3), 439–454 (2002)
13. Arthur D., Vassilvitskii S.: K-means++: the advantages of careful seeding. In: SODA 07 Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027–1035 (2007)
14. Hong, W.C.: Intelligent Energy Demand Forecasting. Springer, London (2013). https://doi.org/10.1007/978-1-4471-4968-2
15. Taylor, J.W.: Short-term electricity demand forecasting using double seasonal exponential smoothing. J. Oper. Res. Soc. **54**, 799–805 (2003)
16. http://br.so-ups.ru/Public/MainPageData/BR/GenConsum.aspx
17. Lyubin, P., Shchetinin, E.: Fast two-dimensional smoothing with discrete cosine transform. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds.) DCCN 2016. CCIS, vol. 678, pp. 646–656. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-51917-3_55
18. Liaw A.: Breiman and Cutler's Random Forests for Classification and Regression. CRAN (2015)

A Review of Metric Analysis Applications to the Problems of Interpolating, Filtering and Predicting the Values of Onevariable and Multivariable Functions

A. V. Kryanev[1,2], V. V. Ivanov[1,2], L. A. Sevastianov[2,3(✉)], and D. K. Udumyan[1,3,4]

[1] National Research Nuclear University "MEPhI", Kashirskoe Shosse 31, 115409 Moscow, Russia
[2] Joint Institute for Nuclear Research, Joliot-Curie st., 6, 141980 Dubna, Moscow Region, Russia
[3] Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, 117198 Moscow, Russian Federation
[4] University of Miami, 1320 S Dixie Hwy, Coral Gables, FL 33146, USA
AVKryanev@mephi.ru, ivanov@jinr.ru, sevastianov_la@rudn.university, mathudum@gmail.com

Abstract. At present, metric analysis schemes are developed to solve the problems of interpolation, smoothing, extrapolation of multivariable functions and their use for many applied problems [1–7]. In contrast to classical methods and schemes and a majority of other ones [8–20,23], the metric analysis, like artificial neuron networks, allows reconstructing the studied function values at each specified point of the definition domain separately. The individual position of this point with respect to the ones, where the values of the function are defined, is taken into account. Here we present a review of the published papers on the metric analysis used to solve the above problems, including those under the conditions of uncertainty of the defined values of the studied function. We present recommendations on using the metric analysis schemes and demonstrate the efficiency of the metric analysis methods and schemes.

Keywords: Multivariable function · Metric analysis
Interpolation · Smoothing · Extrapolation
Prediction of chaotic temporal series

1 Introduction

The problems of interpolating, smoothing, extrapolating multivariable functions, and predicting the values of temporal processes belong to the main urgent problems of mathematical analysis, relevant in many applied fields [1–7,21,22,24–27].

L. A. Sevastyanov—The publication has been prepared with the support of the "RUDN University Program 5-100".

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 457–468, 2018.
https://doi.org/10.1007/978-3-319-99447-5_39

As a rule, the widely known methods of interpolation, smoothing and extrapolation of multivariable functions are inefficient at sufficiently large number of arguments of the studied function. This is valid, e.g., for the classical methods and schemes, based on the representation of the desired function in the form of expansion in a certain set of basis functions, or for the use of multidimensional spline interpolations. At the same time, the metric analysis allows the reconstruction of the values of the studied function of any number of variables even under the conditions of a small number of points where the values of the function are known, which makes it impossible to use the knows methods and schemes, different from those of metric analysis.

2 Interpolation Of Multivariable Functions Using Metric Analysis

The interpolation schemes are related to the functional dependence

$$Y = F\left(X_1, ..., X_m\right) = F(\boldsymbol{X}), \tag{1}$$

where the unknown function $F\left(\boldsymbol{X}\right)$ is to be reconstructed either at one point $\boldsymbol{X}^*$, or at a set of specified points, basing on the known values of the function $Y_k, k = 1, ..., n$ at the fixed points $\boldsymbol{X}_k = \left(X_{k1}, ..., X_{km}\right)^T$ [1–3].

According to the main interpolation scheme based on the metric analysis, the interpolation values are found from the minimum of the metric uncertainty measure with respect to $\boldsymbol{z} = \left(z_1, ..., z_n\right)^T$ [1–6]

$$\sigma^2_{ND}(Y^*; \boldsymbol{z}) = (W(\boldsymbol{X}^*; \boldsymbol{X}_1; ...; \boldsymbol{X}_n)\boldsymbol{z}, \boldsymbol{z}), \tag{2}$$

where the interpolation value is defined by the equality

$$Y^* = \frac{(W^{-1}\boldsymbol{Y}, \boldsymbol{1})}{(W^{-1}\boldsymbol{1}, \boldsymbol{1})}, \tag{3}$$

and the metric uncertainty matrix is defined as

$$W = \begin{pmatrix} \rho^2(\boldsymbol{X}_1, \boldsymbol{X}^*)_{\boldsymbol{w}} & (\boldsymbol{X}_1, \boldsymbol{X}_2)_{\boldsymbol{w}} & ... & (\boldsymbol{X}_1, \boldsymbol{X}_n)_{\boldsymbol{w}} \\ (\boldsymbol{X}_2, \boldsymbol{X}_1)_{\boldsymbol{w}} & \rho^2(\boldsymbol{X}_2, \boldsymbol{X}^*)_{\boldsymbol{w}} & ... & (\boldsymbol{X}_2, \boldsymbol{X}_n)_{\boldsymbol{w}} \\ ... & ... & ... & ... \\ (\boldsymbol{X}_n, \boldsymbol{X}_1)_{\boldsymbol{w}} & (\boldsymbol{X}_n, \boldsymbol{X}_2)_{\boldsymbol{w}} & ... & \rho^2(\boldsymbol{X}_n, \boldsymbol{X}^*)_{\boldsymbol{w}} \end{pmatrix}, \tag{4}$$

where $\rho^2_{\boldsymbol{w}}\left(\boldsymbol{X}_i, \boldsymbol{X}^*\right) = \sum\limits_{k=1}^{m} w_k (X_{ik} - X^*_k)^2$,

$\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right)_{\boldsymbol{w}} = \sum\limits_{k=1}^{m} w_k \left(X_{ik} - X^*_k\right) \cdot \left(X_{jk} - X^*_k\right)$, $i, j = 1, ..., n$, w_k are the metric weights that determine the sensitivity of the function to the change of the arguments X_k, $k = 1, ..., m$ [1–3].

Like the interpolations schemes based on artificial neuron networks, the interpolation schemes based on the metric analysis execute the interpolation separately at each considered point of the definition domain.

Below we present numerical examples, in which functions of one and many variables are in-terpolated using the scheme of metric analysis.

Let us demonstrate the results of interpolation by particular examples in comparison with the interpolation by Lagrange polynomials and cubic splines. In these examples, to choose the inter-polation scheme in the subspace of zero eigenvalue we used Eq. (7), in which the "regularised" matrix W_α was taken in the form $W_\alpha = W + \alpha \cdot [diag(W_{1,1}, ..., W_{n,n}) + 0.5 \cdot di - ag(W_{1,2}, ..., W_{n-1,n}) + 0.5 \cdot diag(W_{2,1}, ..., W_{n,n-1})]$.

Example 1. Consider the function $y = |x|$ in the interval $[-1, 1]$. Figures 1, 2 and 3 show the results for $n = 7$, $n = 9$ and $n = 11$, respectively. The solid line is a plot of the function $y = |x|$ itself, the dotted line is the result of interpolation using the method of metric analysis, and the dashed line is obtained using the Lagrange polynomial interpolation. Black asterisks are the interpolation nodes. In Figs. 1, 2 and 3, one can see that the interpolation values corresponding to the scheme of metric analysis uniformly converge to the values of the interpolated functions, in spite of the presence of a salient point, where the derivative is discontinuous, whereas the Lagrange polyno-mial interpolation diverges.

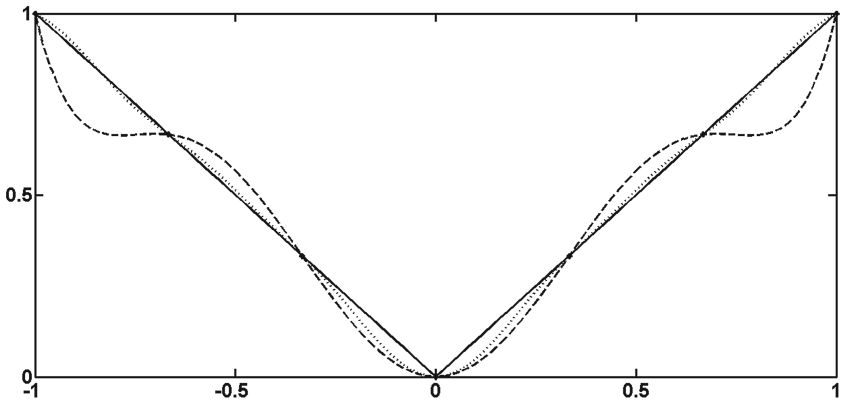


Fig. 1. The solid line plots the function $y = |x|$, the dotted line is the interpolation curve ob-tained using the metric analysis method, and the dashed line is a result of interpolation using the Lagrange polynomial for $n = 7$

Example 2. Consider the function $y(x) = 4 - \exp(x) \cdot \cos(2.1 \cdot \pi \cdot x)$ in the interval $[-1, 1]$. Figures 4, 5 and 6 present the interpolation values for $N = 100$ points uniformly distributed over the interval $[-1, 1]$ for $n = 6$, $n = 8$, and $n = 10$, respectively. The solid line corresponds to a plot of the function itself, the dotted line is obtained by interpolation using the method of metric analysis, and the dashed line is obtained by means of spline interpolation with zero boundary

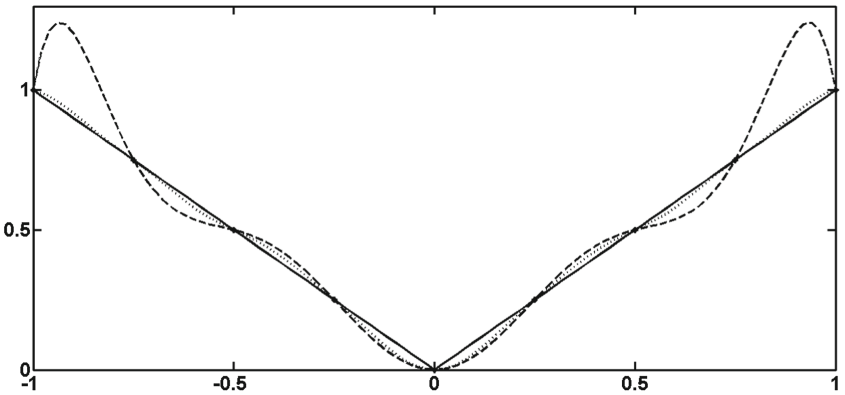


Fig. 2. The solid line plots the function $y = |x|$, the dotted line is the interpolation curve ob-tained using the metric analysis method, and the dashed line is a result of interpolation using the Lagrange polynomial for $n = 9$

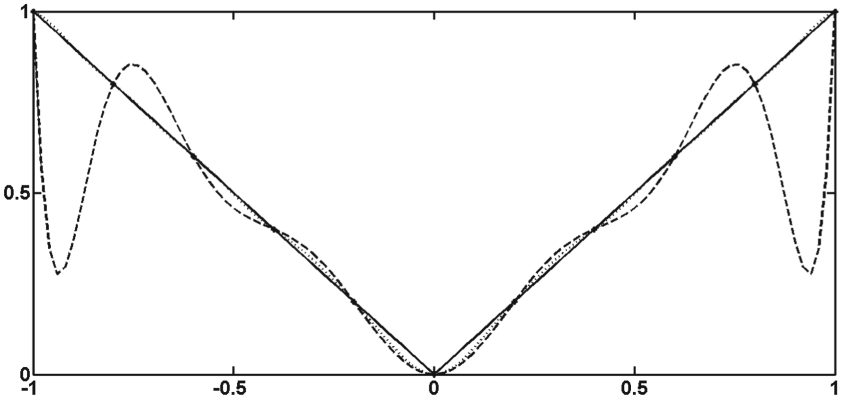


Fig. 3. The solid line plots the function $y = |x|$, the dotted line is the interpolation curve ob-tained using the metric analysis method, and the dashed line is a result of interpolation using the Lagrange polynomial for $n = 11$

conditions for the second derivative. The black asterisks show the interpolation nodes.

Example 3. Figures 7, 8, and 9 illustrate the numerical interpolation of the function of two variables $F(x, y) = 4{\cdot}\sin(2{\cdot}\pi{\cdot}x){\cdot}\cos(1.5{\cdot}\pi{\cdot}y){\cdot}(1{-}x^2){\cdot}y{\cdot}(1{-}y)$, $x \in [-1, 1]$, $y \in [0, 1]$, implemented using the modified metric interpolation scheme. Figure 7 shows the initial surface, the surface in Fig. 8 is drawn using only the

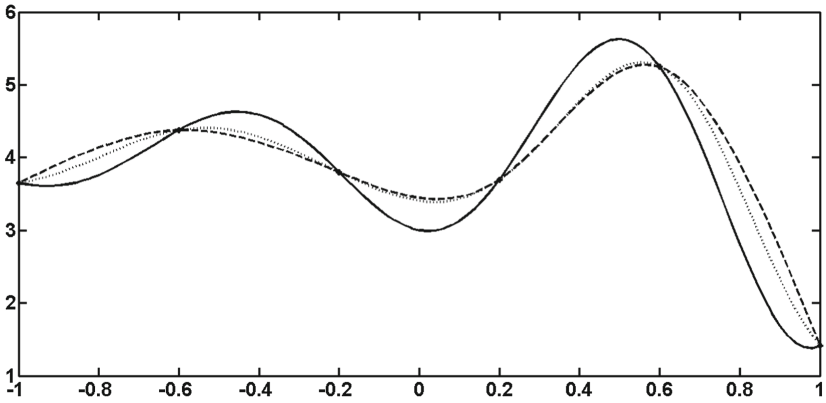


Fig. 4. The solid line plots the function $y(x) = 4 - \exp(x) \cdot \cos(2.1 \cdot \pi \cdot x)$, the dotted line is obtained by interpolation using the method of metric analysis, and the dashed line is a result of spline interpolation for $n = 6$

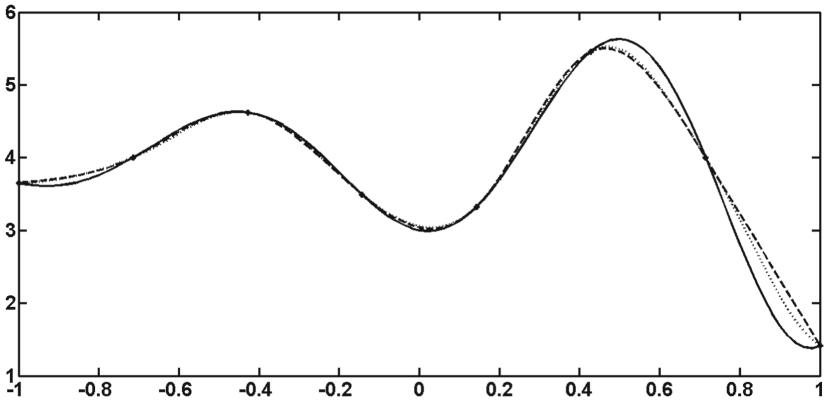


Fig. 5. The solid line plots the function $y(x) = 4 - \exp(x) \cdot \cos(2.1 \cdot \pi \cdot x)$, the dotted line is obtained by interpolation using the method of metric analysis, and the dashed line is a result of spline interpolation for $n = 8$

set of given values, and Fig. 9 shows the surface obtained by means of metric interpolation.

The use of the interpolation schemes in the solution of applied problems is presented in Refs. [1–3].

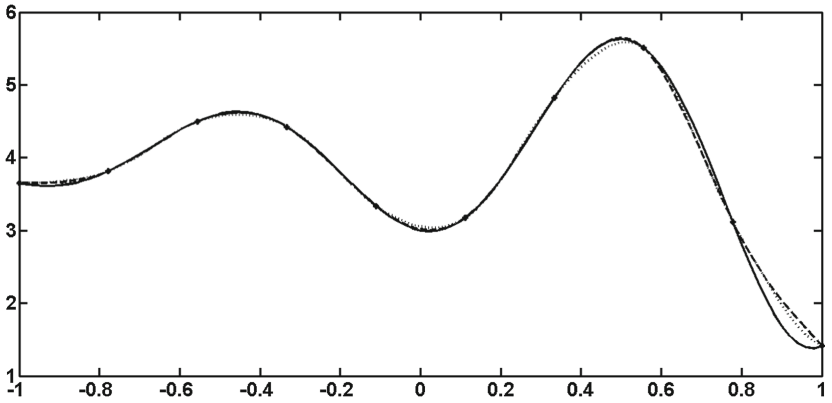


Fig. 6. The solid line plots the function $y(x) = 4 - \exp(x) \cdot \cos(2.1 \cdot \pi \cdot x)$, the dotted line is obtained by interpolation using the method of metric analysis, and the dashed line is a result of spline interpolation for $n = 10$

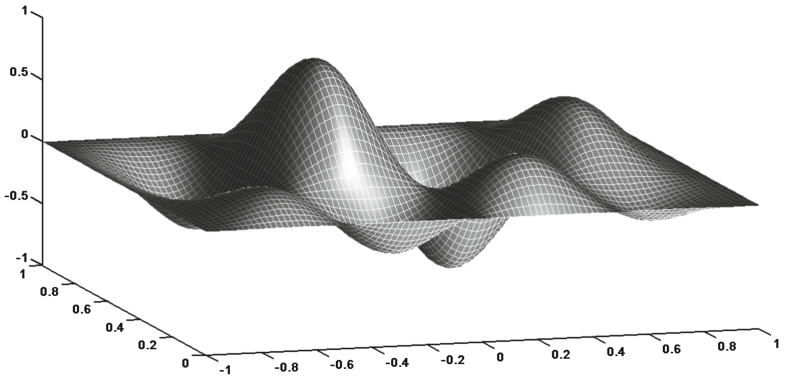


Fig. 7. Surface plot of the function $F(x, y) = 4{\cdot}\sin(2{\cdot}\pi{\cdot}x){\cdot}\cos(1.5{\cdot}\pi{\cdot}y){\cdot}(1{-}x^2){\cdot}y{\cdot}(1{-}y)$

3 Smoothing Multivariable Functions Using the Metric Analysis

Consider the problem of smoothing (selection of deterministic component) of the functional dependence $Y = F(X_1, \ \ldots \ , X_m) = F(\boldsymbol{X})$ in the presence of chaotic deflections from the exact values at the given points. The values of the function Y_k, $k = 1, \ \ldots, n$ are known with errors at the points $\boldsymbol{X}_k = (X_{k1}, \ \ldots, X_{km})^T$, $k = 1, \ \ldots, n$. Thus, we have the set of equalities

$$Y_k = Y_{kdet} + \varepsilon_k, \ k = 1, \ \ldots, n \tag{5}$$

where $\boldsymbol{Y}_{det} = (Y_{1det}, \ \ldots, Y_{n\,det})^T$ is the desired vector of deterministic components (estimates of the function values) at the points $\boldsymbol{X}_k = (X_{k1}, \ \ldots, X_{km})^T$,

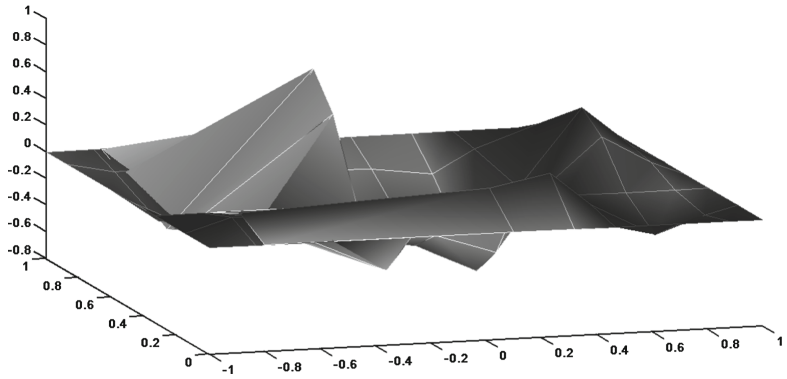


Fig. 8. The same function as in Fig. 7, plotted using a discrete set of given points

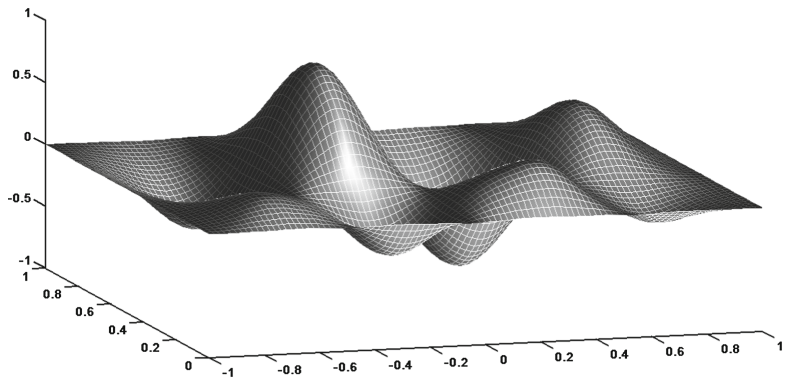


Fig. 9. The result of metric interpolation.

$k = 1, \ \ldots, n$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ \ldots \ , \varepsilon_n)^T$ is the vector of chaotic components. Below we assume that the matrix of metric uncertainty W can be singular. For any point $\boldsymbol{X}_k$ we search for the value $\boldsymbol{Y}_{det}$ of the deterministic component in the representation

$$Y_{k\,det} = \sum_{i=1}^{n} z_i \cdot Y_i = (\boldsymbol{z}, \boldsymbol{Y}), \tag{6}$$

where the vector $\boldsymbol{z}$ is a solution of the following problem of minimizing the total uncertainty:

$$\begin{aligned} &(W\,\boldsymbol{z}, \boldsymbol{z}) + \alpha \cdot (K_{\boldsymbol{Y}}\,\boldsymbol{z}, \boldsymbol{z}) \to \min, \\ &(\boldsymbol{z}, \boldsymbol{1}) = 1, \end{aligned} \tag{7}$$

$\alpha \geq 0$ is the smoothing parameter, $K_{\boldsymbol{Y}}$ is the covariance matrix of the vector of chaotic components $\boldsymbol{\varepsilon} = (\varepsilon_1, \ \ldots \ , \varepsilon_n)^T$ and the matrix of metric uncertainty $W$is calculated relative to the point $\boldsymbol{X}_k (\boldsymbol{X}^* = \boldsymbol{X}_k$, see Eq. (4)).

The expression $(W\,\boldsymbol{z},\boldsymbol{z})$ characterises the metric uncertainty, and the expression $(K_{\boldsymbol{Y}}\boldsymbol{z},\boldsymbol{z})$ characterises the stochastic uncertainty of the function value at the point $\boldsymbol{X}_k$. The problem (7) is solved using the Lagrange function, and the solution is expressed as:

$$Y_{k\alpha} = \left((W + \alpha \cdot K_{\boldsymbol{Y}})^{-1}\mathbf{1},\, \boldsymbol{Y}\right) / \left((W + \alpha \cdot K_{\boldsymbol{Y}})^{-1}\mathbf{1}, \mathbf{1}\right) \tag{8}$$

When $\alpha \to +\infty$ the smoothed value $Y_{k\alpha}$ for the point $\boldsymbol{X}_k$ is given by

$$Y_{k\alpha} = \left(K_{\boldsymbol{Y}}{}^{-1}\mathbf{1},\, \boldsymbol{Y}\right) / \left(K_{\boldsymbol{Y}}{}^{-1}\mathbf{1}, \mathbf{1}\right) \tag{9}$$

When $\alpha \to +0$, the smoothed value $Y_{k\alpha}$ for the point $\boldsymbol{X}_k$ is given by

$$Y_{k\alpha} = \frac{\left(W^{-1}\mathbf{1},\, \boldsymbol{Y}\right)}{\left(W^{-1}\mathbf{1}, \mathbf{1}\right)}. \tag{10}$$

The value of the deterministic component $Y_{k\,det}$ at the point $\boldsymbol{X}_k$ is expressed as

$$Y_{k\,det} = Y_{k\alpha_*}, \tag{11}$$

where α_* is found from the equation

$$\|\,\boldsymbol{Y} - \boldsymbol{Y}_\alpha\|^2 = n \cdot \sigma^2. \tag{12}$$

Here $\boldsymbol{Y}_\alpha = (Y_{1\alpha},\ \ldots\ , Y_{n\alpha})^T$, σ^2 is the variance of the chaotic components $\varepsilon_k,\ k = 1,\ \ldots, n$. The smoothing of multivariable functions by means of metric analysis was considered in Refs. [2,3,6,25,26].

4 Mathematical Formulae and References

Consider the function $y = f(t)$ of one variable t with the values $Y_1 = f(t_1), ..., Y_n = f(t_n)$ for $t_1 < ... < t_n \in [t_1, t_n]$. It is required to find the predicted value Y_{n+1} for t_{n+1}. Let us reduce the problem of finding the predicted value y_{n+1} to the problem of interpolating the multidimensional function using the nonlinear autoregression

$$\begin{aligned} &y(t_{m+1}) = y_{m+1} = F(y_1, ..., y_m) \\ &y(t_{m+2}) = y_{m+2} = F(y_2, ..., y_{m+1}) \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &y(t_N) = y_N = F(y_{N-m}, ..., y_{N-1}). \end{aligned} \tag{13}$$

Then the prediction of the function $y = f(t)$ of one variable t is reduced to the interpolation of the function of variables $Y = F(y_1, y_2, ..., y_m)$ with the values specified at points

$\boldsymbol{X}_1 = (Y_1, ..., Y_m)^T, \boldsymbol{X}_2 = (Y_2, ..., Y_{m+1})^T, ..., \boldsymbol{X}_{n-m} = (Y_{n-m}, ..., Y_{n-1})^T$. The predicted value $y_{n+1}$is determined as an interpolation value of the $m-$ dimensional function F at the point $\boldsymbol{X}^*$:

$$y_{n+1} = F(\boldsymbol{X}^*) = \frac{(W^{-1}\mathbf{1}, \boldsymbol{Y})}{(W^{-1}\mathbf{1}, \vec{1})}, \boldsymbol{X}^* = (Y_{n-m+1}, ..., Y_n)^T \tag{14}$$

where W^{-1} is the matrix inverse to the $(n-m) \times (n-m)$ matrix of metric uncertainty, $\boldsymbol{Y} = (Y_{m+1}, ..., Y_n)^T$ is the $(n-m)$-dimensional vector of the predicted temporal process values. The natural number m determines the dimension of the space of vectors $\boldsymbol{X}$ and its value is found as a solution of the extremum problem

$$m = \arg\min_m \| \boldsymbol{Y} - \boldsymbol{Y}_{for} \|, \tag{15}$$

where $\boldsymbol{Y}_{\text{for}}$ is the vector of the predicted values for the realised part of the values of the initial function $y = f(t)$, obtained using the above method.

Below we present numerical examples of predicting a temporal process by means of the schemes of metric analysis.

Example 4. Using the metric analysis we predicted the values of the function $Y(t)$, on which the additive noise $\varepsilon(t)$ was superposed $Y(t) = 2 \cdot \text{Sin}(\pi \cdot t/4) + 3 \cdot \text{Cos}(\pi \cdot t/3) + \varepsilon(t)$.

The scheme of nonlinear autoregression $Y(t_n) = f(Y(t_{n-1}), ..., Y(t_{n-m}))$ was used. The optimal value of $m = 17$ was determined by means of the least squares method on the realized part of the time series $Y(t_1), ..., Y(t_N)$.

Figure 10 presents the results of predicting 100 steps ahead. The error of the predicted values is equal to the noise level $\varepsilon(t)$.

Example 5. $y = e^t \sin(\omega t), \omega = 2.9; h = 0.1, n = 50, N_{ext} = 150$

The extrapolation result is presented in Fig. 11. In this example $\varepsilon = 10^{-3}$, $m_{opt} = 32$.

The prediction of temporal processes using the metric analysis was considered in Refs. [1–8].

5 Extrapolation of Multivariable Functions Using the Metric Analysis

The extrapolation scheme for the multivariable function (1) at the given point $\boldsymbol{X}^* = (X_1^*, \ldots, X_m^*)^T$consists of two stages.

At the first stage, in the cluster of realised values of the function $Y = F(\boldsymbol{X})$ the point$\boldsymbol{X}_0 = (X_{01}, \ldots, X_{0m})^T$is chosen. Then the points $\boldsymbol{X}_0$ and $\boldsymbol{X}^*$ are connected with the straight line segment

$$(1-s) \cdot \boldsymbol{X}_0 + s \cdot \boldsymbol{X}^*, \ \ 0 \le s \le 1, \tag{16}$$

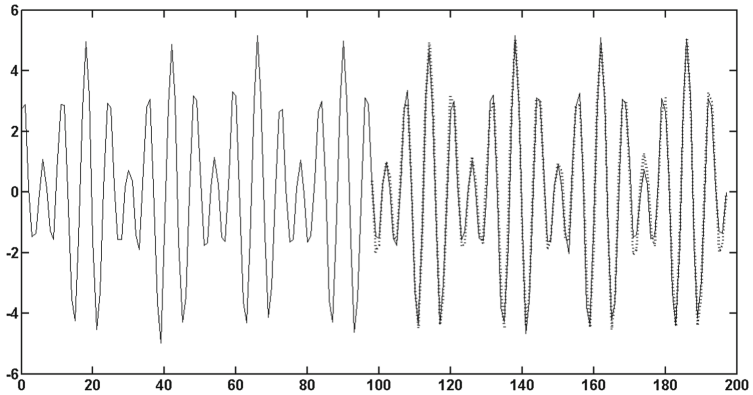


Fig. 10. Results of prediction of 100 time steps ahead, the level of prediction error is equal to the noise level $\varepsilon(t)$

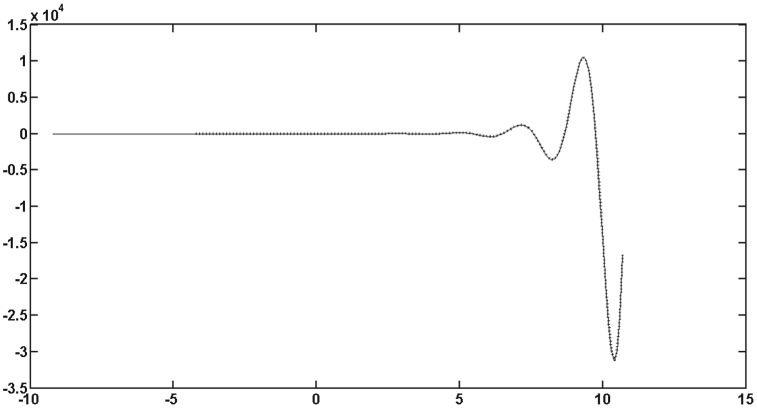


Fig. 11. The result of extrapolation of $y = e^t \sin(\omega t), \omega = 2.9; h = 0.1, n = 50, N_{ext} = 150$ for the extrapolation parameters $\varepsilon = 10^{-3}$, $m_{opt} = 32$

which is divided into L equal segments with the nodes

$$\boldsymbol{S}_k = (S_{k1}, \ \ldots, S_{km})^T, \ k = 1, \ \ldots, L, \quad \boldsymbol{S}_0 = \boldsymbol{X}_0, \ \ \boldsymbol{S}_L = \boldsymbol{X}^*. \tag{17}$$

At the points

$$\boldsymbol{S}_k = (S_{k1}, \ \ldots, S_{km})^T, \ k = 1, \ \ldots, l, \quad l < L, \tag{18}$$

the interpolation of the function (1), belonging to the cluster, is executed according to the scheme (3)–(4) using the set of known values of the function Y_k, $k = 1, ..., n$ at the points $\boldsymbol{X}_k = (X_{k1}, ..., X_{km})^T$. At the second stage, the interpolating values calculated at the first stage $\boldsymbol{Y} = (Y_1, \ \ldots, Y_l)^T$ at the points (18), are sequentially extrapolated to the rest of the nodes $\boldsymbol{S}_k = (S_{k1}, \ \ldots, S_{km})^T$, $k = l + 1, \ \ldots, L$, $\boldsymbol{S}_0 = \boldsymbol{X}_0$, $\boldsymbol{S}_L = \boldsymbol{X}^*$, where

$$f(s) = F\left((1 - s) \cdot \boldsymbol{X}_0 + s \cdot \boldsymbol{X}^*\right), \ \ 0 \le s \le 1. \tag{19}$$

Using the autoregression scheme (13)–(15) at the nodes

$$s = s_k, \;\; k = l+1, \;\; \ldots \;\; , L, \quad s_{k+1} = s_k + \Delta s, \;\; \Delta s = \frac{1}{L}, \tag{20}$$

we get the desired extrapolated value

$$Y_{ext} = F\left(\boldsymbol{X}^*\right) = f\left(s_L\right). \tag{21}$$

The extrapolation of multivariable functions using the metric analysis was considered in Ref. [27].

6 Conclusion

In the report, we present a review of methods and calculation schemes for interpolation, smoothing, and extrapolation of multivariable functions and prediction of values of a one-variable function, based on the metric analysis and developed during the last 13 years by the authors of the report in collaboration with the representatives of different research institutions. The metric analysis allows the solution of interpolation, smoothing and extrapolation problems for functions depending on a large number of variables (a few tens and more), considering the position of the point where the function value is sought for with respect to the points where the function values are known. Multiple examples of using the methods and calculation schemes for the solution of particular problems, including applied ones, demonstrate the efficiency of these methods and schemes.

References

1. Kryanev A. V., Lukin G. V.: Metric analysis for interpolation and forecasting for functions of many variables. Preprint MEPhI 003–2005, Moscow (2005)
2. Kryanev, A.V., Lukin, G.V., Udumyan, D.K.: Metric analysis and data processing. Science ed. Moscow (2012)
3. Kryanev, A.V., Lukin, G.V., Udumyan, D.K.: Metric analysis and applications. Numer. Methods Program. **10**, 408–414 (2009)
4. Kryanev, A.V., Udumyan, D.K., Lukin, G.V., Ivanov, V.V.: Metric analysis approach for interpolation and forecasting of time processes. Appl. Math. Sci. **8**(22), 1053–1060 (2014)
5. Kryanev, A.V., Udumyan, D.K.: Metric analysis, properties and applications as a tool for interpolation. Int. J. Math. Anal. **8**(45), 2221–2228 (2014)
6. Kryanev, A.V., Udumyan, D.K.: Metric analysis properties and applications as a tool for smoothing. Int. J. Math. Anal. **8**(47), 2337–2346 (2014)
7. Kryanev, A.V., Udumyan, D.K.: Metric analysis properties and applications as a tool for forecasting international. J. Math. Anal. **8**(60), 2971–2978 (2014)
8. Watson, G.A.: Approximation Theory and Numerical Methods. John Wiley, New York (1980)
9. Jazwinski, A.H.: Stochastic Processes and Filtering Theory. Academic Press, New York (1970)

10. Simonoff, J.S.: Smoothing Methods in Statistics, 2nd edn. Springer, New York (1998). https://doi.org/10.1007/978-1-4612-4026-6
11. Haykin, S.: Neural Networks: A Comprehensive Foundation. Prentice Hall, New York (1999)
12. De Boor, C.: A Practical Guide to Splines, Revised edn. Springer, New York (2001)
13. Zav'yalov, Y.S.: Smoothing by L-spline functions of many variables. Math. Notes Acad. Sci. USSR **15**(3), 212–217 (1974)
14. Sendov, B., Andreev, A.: Approximation and interpolation theory. Handb. Num. Anal. **3**, 223–462 (1994)
15. Söderström, T.: Discrete-time Stochastic Systems: Estimation and Control. Springer, London (2002). https://doi.org/10.1007/978-1-4471-0101-7
16. Simon, D.: Optimal State Estimation Kalman: H Infinity and Nonlinear Approaches. Wiley, New Jersey (2006)
17. Andreasen, M.M.: Non-linear DSGE models and the central difference Kalman filter. J. Appl. Econ. **28**(6), 929–955 (2013)
18. Strid, I., Walentin, K.: Block Kalman filtering for large-scale DSGE models. Comput. Econ. **33**(3), 277–304 (2009)
19. Chui, C.K.: An Introduction to Wavelets. Academic Press, New York (1992)
20. Antoniou, I., Ivanov, V.V., Ivanov, V.V., Zrelov, P.V.: Wavelet filtering of network traffic measurements. Phys. A: Stat. Mech. Appl. **324**, 733–753 (2003)
21. Samouylov, K.E., Abaev, P.O., Gaidamaka, Y.V., Pechinkin, A.V., Razumchik, R.V.: Analytical modeling of rate-based overload control with token bucket traffic shaping on client side. In: Proceedings of 29th European Conference on Modelling and Simulation, ECMS 2015, pp. 669–674 (2015)
22. Korolkova, A.V., Eferina, E.G., Laneev, E.B., Gudkova, I., Sevastianov, L.A., Kulyabov, D.S.: Stochastization of one-step processes in the occupations number representation. In: ECMS (2016)
23. Golyandina, N., Nekrutkin, V., Zhigljavsky, A.A.: Analysis of Time Series Structure: SSA and Related Techniques. Chapman & Hall/CRC (2001)
24. Kryanev, A.V., Udumyan, D.K., Kurchenkov, A.Y., Gagarinskiy, A.A.: Determination of power distribution in the VVER-440 core on the basis of data from in-core monitors by means of a metric analysis. Phys. Atomic Nuclei **77**(14), 1651–1655 (2014)
25. Ivanov, V.V., Klimanov, S.G., Kryanev, A.V., Lukin, G.V., Udumyan, D.K.: Forecasting of chaotic dynamic processes by means of allocation regular components. Comput. Math. Math. Phys. **55**(2), 340–347 (2015)
26. Kryanev, A.V., Ivanov, V.V., Romanova, A.O., Sevastyanov, L.A., Udumyan, D.K.: Separation of trend and chaotic components of time series and estimation of their characteristics by linear splines. Phys. Part. Nuclei Lett. **15**(2), 194–197 (2018)
27. Kryanev, A., Ivanov, V., Romanova, A., Sevastianov, L., Udumyan, D.: Extrapolation of functions of many variables by means of metric analysis. In: EPJ Web of Conferences, p. 173 (2018)

The Application of Helmholtz Decomposition Method to Investigation of Multicore Fibers and Their Application in Next-Generation Communications Systems

D. V. Divakov(✉), K. P. Lovetskiy, M. D. Malykh, and A. A. Tiutiunnik

Department of Applied Probability and Informatics,
Peoples' Friendship University of Russia (RUDN University),
6 Miklukho-Maklaya St., Moscow 117198, Russian Federation
{divakov_dv,lovetskiy_kp,malykh_md,tyutyunnik_aa}@rudn.university

Abstract. New optical multicore fibers use their spatial properties in the designs of next-generation systems. To investigate light propagation in such fiber waveguides we use Helmholtz decomposition method.

We consider a waveguide having the constant cross-section S with ideally conducting walls. We assume that the filling of waveguide does not change along its axis and is described by the piecewise continuous functions ϵ and μ defined on the waveguide cross section. We show that it is possible to make a substitution, which allows dealing only with continuous functions. Instead of discontinuous cross components of the electromagnetic field $\boldsymbol{E}$ and $\boldsymbol{H}$ we propose to use four potentials u_e, u_h and v_e, v_h. Generalizing the Thikhonov-Samarskii theorem, we have proved that any field in the waveguide allows such representation, if we consider the potentials u_e, u_h as elements of the Sobolev space $\overset{0}{W}{}_2^1(S)$ and the potentials v_e, v_h as elements of the Sobolev space $W_2^1(S)$.

If ϵ and μ are piecewise constant functions, then in terms of four potentials the Maxwell equations reduce to a pair of Helmholtz equations. This fact means that a few dielectric waveguides placed between ideally conducting walls can be described by a scalar boundary problem. This statement offers a new approach to the investigation of spectral properties of waveguides. First, we can prove the completeness of the system of the normal waves in closed waveguides using standard functional spaces. Second, we can propose a new technique for calculating the normal waves using standard finite elements.

A. A. Tiutiunnik—The publication has been prepared with the support of the "RUDN University Program 5-100" and funded by RFBR according to the research projects No. 18-07-00567 and No. 18-51-18005.

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 469–480, 2018.
https://doi.org/10.1007/978-3-319-99447-5_40

Keywords: Optical communications systems
Multicore fibers · Waveguides · Maxwell equations
Sobolev spaces · Normal modes
Connections between cloud data centers

1 Motivation

New optical fibers such as multicore, multimode and hollow core use their spatial properties in the designs of next-generation systems to solve future fiber capacity bottlenecks and other sticky problems [1,2].

In seeking to fully utilize its already-installed physical fiber-optic cable infrastructure, the optical communications industry has ingeniously and continuously managed to pack in more bandwidth. Yet current approaches may soon hit a barrier: the capacity threshold known as the Shannon Limit. Consequently, various new optical fiber designs are seeking to postpone the industry's arrival at this limit [3].

The need for more capacity is well illustrated by plans of multinational telecommunications companies. Where possible they aim to integrate fixed and mobile networks in order to maximize synergies. Therefore, they already use fiber to move enormous amounts of data between cell towers, and its ongoing network evolution will be a key factor in the path to 5G mobile networks. They are already building networks that are likely to bring some of the benefits of 5G sooner than 2020, when the industry expects 5G services to be rolled out [3].

Over the past 5 years, major progress has been made in developing new fibre technologies for high-capacity communications. Gradual improvements are being made in SSMFs in terms of effective area and loss reduction, new broadband amplifier options are emerging and steady progress is being made in the area of hollow core fibres. The primary advances though have been in the area of SDM, especially in the development of high-performance MCF, FMF and MCFMFs [4], as well as the associated components required to launch and amplify the individual spatial channels. As a result, various experiments have reported record per-fibre capacities and capacity length products. Initial demonstrations of switching/routing and associated networking have also now been performed. However, the research is still at the development stage and more intensive efforts will be required to show whether a significant reduction in costs can be achieved with a high level of reliability and performance on each channel. Since it is required to create a new technology that is competitive with existing single-mode fiber technologies. Photonic integration will be absolutely essential in realizing cost reduction through SDM and this work is really still at a very early stage [4].

However, the research is still very much in the exploratory phase with much further work needed to show whether significant cost-per-bit reductions can be achieved at levels of per-channel reliability and performance that are competitive with existing single-mode fibre technology.

In principle, each of the fiber cores in such a fiber can act as a separate waveguide, so that light can independently propagate through those cores. However,

there can be some coupling between the cores if the distance between two cores is so small that the corresponding mode fields have a significant spatial overlap. This means that light which is initially coupled into one core can eventually couple over to other cores; that effect is similar as in fused fiber couplers. For such a situation, one can compute so-called supermodes, i.e., approximate field configurations which are stationary despite the coupling. However, supermodes calculated in mode coupling theory for an idealized situation may not be the true modes of a fiber subject to random fluctuations in fabrication and/or due to operation conditions.

The simplest, correct model describing the propagation of waves along multicore optical fiber suggests placing optical fiber inside an ideally conducting casing separated from the optical fiber by a sufficient distance so that the fictitious walls do not have a noticeable effect on wave propagation along the optical fiber. The "optical fiber + casing" system is a classical object of the theory—a closed waveguide with ideally conducting walls and a complex filling described by piecewise constant functions ϵ and μ. Maxwell's equations without any seizures and simplifications. Modern problems in the theory of optical fibers, however, introduce their computational features into this classical model. In the old days it was usually assumed that the functions ϵ and μ vary along the cross-section very smoothly, on the contrary, we are interested in the case when the filling of the waveguide is described by rapidly varying functions, perhaps even having a "fine" structure.

To investigate light propagation in such waveguides we suggest to use Helmholtz decomposition methods [5]. For multicore fibers this method allows to split a boundary problem for Maxwell's equations into two "scalar" problem. We will show how to use this splitting for the theoretical and numerical analysis of boundary value problems of the mathematical theory of waveguide systems.

2 Introduction

Mathematical theory of the hollow waveguides was constructed by Tikhonov and Samarskii. In a hollow waveguide it is possible to introduce two scalar potentials, using which the Maxwell equations reduce to a pair of uncoupled wave equations, as it was proved in the classical papers by Tikhonov and Samarskii [6]. Such decomposition of the electromagnetic field is possible due to the cylindrical geometry of the waveguide that allows the application of the theory of Borgnis functions [7]. The most important consequence form the Tikhonov and Samarskii theorem is the completeness of the system of normal waves in a hollow waveguide, according to which any wave propagating through the waveguide can be presented as a superposition of TE and TM waves [8]. In 1990s this consequence extremely important for substantiating the partial radiation conditions and the incomplete Galerkin method [9], was generalized for the case of a waveguide, in which the filling varies over the transverse section, but is constant along the waveguide axis [10]. As a result, the theorem of field representation using potentials became shadowed by its consequence.

Due to this circumstance, the computational complexity of the spectral problems for hollow waveguides and for the waveguides filled with inhomogeneous matter differs in principle. In the first case, the resulting problems are scalar, and one can use the well-developed methods, equally applicable to acoustics and quantum mechanics. In the case of a waveguide filled with inhomogeneous matter, one has to solve numerically the full vector electrodynamic problem. Such problems possess the zero eigenvalue of infinite multiplicity, due to which their numerical solution requires nontrivial procedures hard for computer implementation, e.g., the method of mixed finite elements [11,12].

We should also note that for the problems of radiophysics the case of piecewise constant filling is of particular interest, since a waveguide with smoothly changing filling can be practically fabricated only by using thin homogeneous layers with finite but small difference of ϵ and μ between the adjacent layers. At the junction between different layers the transverse components of the vector fields $\boldsymbol{E}$ and $\boldsymbol{H}$ have discontinuities, which leads to additional difficulties in their approximation by continuous finite elements.

However, from the general considerations it is clear that the waveguide with transversely piecewise constant filling is nothing but a few waveguides with constant homogeneous filling coupled via their walls. Physically, the waveguide systems with the filling quickly varying across the section are fabricated by collecting a bundle of a few tens of homogeneous dielectric waveguides. To describe the field in each particular homogeneously filled waveguide it is quite enough to use two scalar equations, so it seems quite natural that these equations would also describe a composite waveguide system. A priori, it is not only clear how to describe the coupling between the waveguides in terms of boundary conditions.

In the present paper we return to the problem of presenting an arbitrary electromagnetic field in a waveguide with piecewise constant filling in its classical formulation. Usually, as in the case of a hollow waveguide, the introduction of potentials allows integrating some of the Maxwell equations and reducing the number of desired functions. It is well known that in a waveguide filled by inhomogeneous medium this is not the case. However, we believe that the main advantage of introducing potentials is dealing with continuous potentials instead of discontinuous field components. Form this point of view, the introduction of potentials can be considered as a change of variables providing a transition from discontinuous functions to continuous ones.

3 Notation

The subject of this article is a closed waveguide of constant cross-section S with piecewise-continuous filling ϵ and μ that does not vary along the axis of the waveguide. We denote the filling discontinuity line as Γ. We direct the Oz axis of the Cartesian coordinate system along the waveguide axis and take for brevity what

$$\boldsymbol{A}_\perp = (A_x, A_y, 0)^T \quad \text{and} \quad \nabla = (\partial_x, \partial_y, 0)^T, \quad \nabla' = (-\partial_y, \partial_x, 0)^T.$$

By the electromagnetic field in a closed waveguide $S \times Z \times T$ with filling ϵ, μ we mean vector fields $\boldsymbol{E}, \boldsymbol{H}$ whose components are defined on $(S - \Gamma) \times Z \times T$, provided that the narrowing of the fields $\boldsymbol{E}, \boldsymbol{H}$ and their partial derivatives with respect to z and t to the section S for any values of z and t are piecewise smooth functions satisfying

1. Maxwell's equations

$$\begin{cases} \operatorname{curl} \boldsymbol{E} = -\partial_t \dfrac{\mu}{c} \boldsymbol{H}, & \operatorname{div} \epsilon \boldsymbol{E} = 0, \\ \operatorname{curl} \boldsymbol{H} = +\partial_t \dfrac{\epsilon}{c} \boldsymbol{E}, & \operatorname{div} \mu \boldsymbol{H} = 0 \end{cases} \tag{1}$$

   inside the waveguide $S \times Z \times T$,
2. conditions for the ideal conductivity of the waveguide walls

$$\boldsymbol{E} \times \boldsymbol{n} = 0, \quad \boldsymbol{H} \cdot \boldsymbol{n} = 0 \tag{2}$$

   at regular points of the boundary $\partial S \times Z \times T$,
3. interface conditions

$$\begin{cases} [\boldsymbol{E} \times \boldsymbol{n}] = \boldsymbol{0}, & [\epsilon \boldsymbol{E} \cdot \boldsymbol{n}] = 0 \\ [\boldsymbol{H} \times \boldsymbol{n}] = \boldsymbol{0}, & [\mu \boldsymbol{H} \cdot \boldsymbol{n}] = 0 \end{cases} \tag{3}$$

   at regular points of the boundary of the filling discontinuity $\Gamma \times Z \times T$.

4 The Helmholtz Decomposition

The connection between the fields and the potentials is given in the following way

$$\boldsymbol{E}_\perp = \nabla u_e + \frac{1}{\epsilon} \nabla' v_e, \quad \boldsymbol{H}_\perp = \nabla v_h + \frac{1}{\mu} \nabla' u_h. \tag{4}$$

Each of these formulas is a two-dimensional analogue of the Helmholtz decomposition, which is well known in the theory of elasticity.

Note 1. In electrodynamics for the field $\boldsymbol{H}_\perp$ such potentials arose in the proof of the completeness of the system of normal modes as an auxiliary structure [13]. All four potentials were introduced in our works [5,14] for smooth filling without coefficients $\frac{1}{\epsilon}$ and $\frac{1}{\mu}$, important only for discontinuous case.

Theorem 1. *For any electromagnetic field $\boldsymbol{E}, \boldsymbol{H}$ in the waveguide, there are functions u_e, u_h of variables z, t with values in the Sobolev space $\overset{0}{W}{}_2^1(S)$ and the functions v_e, v_h of variables z, t with values in the Sobolev space $W_2^1(S)$ such that (4) holds. This representation is unique up to additive constants.*

The Theorem 1 means that when passing from the variables $\boldsymbol{E}, \boldsymbol{H}$ to the four potentials and two components E_z, H_z by the formulas (4) the solutions of Maxwell's equations are not lost. The conditions

$$u_e,\, u_h,\, E_z \in \overset{0}{W}{}_2^1(S) \quad \text{and} \quad v_e,\, v_h,\, H_z \in W_2^1(S)$$

replace the conditions on the filling discontinuities, as well as the boundary conditions. Thus, when investigating a multicore fiber described by discontinuous functions ϵ and μ, one can work with smooth potentials if leaving aside the subtle difference between belonging to the Sobolev space and the smoothness.

5 Splitting of the System of Maxwell's Equations in Waveguides with Piecewise Constant Filling

In the case of optical fibers, the filling of the waveguide is described by piecewise constant functions, which makes it possible to substantially simplify the Maxwell's equations in potentials. According to the Theorem 1, the electromagnetic field $\boldsymbol{E}, \boldsymbol{H}$ in such a waveguide can be represented as (4). In this case Maxwell's equations give that the potentials u_e, u_h and E_z are elements of $\overset{0}{W}{}_2^1(S)$ connected by equations

$$\begin{cases} \iint\limits_S \epsilon(\nabla u, \nabla u_e)dxdy = \partial_z \iint\limits_S \epsilon u E_z dxdy, \\ \iint\limits_S \dfrac{c}{\mu}(\nabla u, \nabla u_h)dxdy = -\partial_t \iint\limits_S \epsilon u E_z dxdy, \end{cases} \tag{5}$$

for any u from $C_0^\infty(S)$, where $E_z = \partial_z u_e + c^{-l}\partial_t u_h$, the potentials v_e, v_h and H_z are elements of $W_2^1(S)$ connected by equations

$$\begin{cases} \iint\limits_S \dfrac{c}{\epsilon}(\nabla v, \nabla v_e)dxdy = \partial_t \iint\limits_S \mu v H_z dxdy, \\ \iint\limits_S \mu(\nabla v, \nabla v_h)dxdy = \partial_z \iint\limits_S \mu v H_z dxdy, \end{cases} \tag{6}$$

for any v from $C^\infty(S)$, where $H_z = \partial_z v_h - c^{-1}\partial_t v_e$.

The Eqs. (5) and (6) can also be used to construct fields in a waveguide. If u_e, u_h and E_z from $\overset{0}{W}{}_2^1(S)$ satisfy the Eq. (5), and v_e, v_h and H_z from $W_2^1(S)$ satisfy the Eq. (6), then the field $\boldsymbol{E}, \boldsymbol{H}$, calculated from the formulas (4), satisfies Maxwell's equations in the generalized sense. Moreover, if this field has continuous partial derivatives of the first order in all variables outside the filling discontinuities, and discontinuities of the first kind on the filling discontinuities, then this field outside the filling discontinuities satisfies Maxwell's equations (1),

the interface conditions (3) on the filling discontinuities and ideal conductivity boundary conditions (2).

Since the system of Maxwell's equations splitted into two independent systems, the electromagnetic field $\boldsymbol{E}, \boldsymbol{H}$ in a waveguide whose filling is described by piecewise constant functions ϵ and μ is a superposition of TE- and TM-fields.

6 Monochromatic Fields in Waveguides with Piecewise Constant Filling

Let us apply the developed theory to the case of monochromatic fields, when the time dependence is described by the factor $e^{-i\omega t}$.

The monochromatic TM-field is described by the potentials

$$u_e = \tilde{u}_e e^{-i\omega t}, \quad u_h = \tilde{u}_h e^{-i\omega t},$$

which satisfy the equations

$$\begin{cases} \iint\limits_S \epsilon(\nabla u, \nabla \tilde{u}_e) dxdy = \partial_z^2 \iint\limits_S \epsilon u \tilde{u}_e dxdy - ik\partial_z \iint\limits_S \epsilon u \tilde{u}_h dxdy, \\ \iint\limits_S \frac{1}{\mu}(\nabla u, \nabla \tilde{u}_h) dxdy = ik\partial_z \iint\limits_S \epsilon u \tilde{u}_e dxdy + k^2 \iint\limits_S \epsilon u \tilde{u}_h dxdy, \end{cases} \tag{7}$$

for any $u \in \overset{0}{W}{}_2^1(S)$, here $k = \frac{\omega}{c}$. We rewrite this system of equations in operator form, using the standard technique of the Sobolev spaces theory [16].

Symmetric bilinear form

$$\iint\limits_S (\nabla u, \nabla \tilde{u}) k(x, y) dxdy$$

for any piecewise smooth k is bounded in the norm W_2^1, therefore there exists a bounded self-adjoint operator A_k such that this form is equal to $(u, A_k \tilde{u})$. Symmetric bilinear form

$$\iint\limits_S u \tilde{u} k(x, y) dxdy$$

for any piecewise smooth k is completely continuous in the norm W_2^1, therefore there exists a bounded self-adjoint operator B_k such that this form is equal to $(u, B_k \tilde{u})$. Therefore, the system (7) can be rewritten as follows

$$\begin{cases} A_\epsilon \tilde{u}_e = \partial_z^2 B_\epsilon \tilde{u}_e - i\omega \partial_z B_\epsilon \tilde{u}_h, \\ A_{\frac{1}{\mu}} \tilde{u}_h = i\omega \partial_z B_\epsilon \tilde{u}_e + \omega^2 B_\epsilon \tilde{u}_h. \end{cases} \tag{8}$$

Let us assume that the frequency ω of the field under consideration differs from the special frequencies (magnetic cutoff frequencies) at which the operator $A_{\frac{1}{\mu}} - \omega^2 B_\epsilon$ is not invertible. Eliminating $\hat{u}_h$ from this system, we obtain

$$A_\epsilon \tilde{u}_e = \partial_z^2 \left(B_\epsilon \tilde{u}_e + \omega^2 B_\epsilon (A_{\frac{1}{\mu}} - \omega^2 B_\epsilon)^{-1} B_\epsilon \right) \tilde{u}_e. \tag{9}$$

Therefore, a TM-field can be described as a solution of the Eq. (9), which plays the same role in the case of piecewise constant fillings as the Helmholtz equation in the case of hollow waveguides.

For linear differential equations, whose coefficients are compact operators, we can always write out the general solution by means of the root vectors system of the corresponding polynomial operator bundle [17]. In particular, the general solution of the Eq. (9) can be represented as a sum of TM-fields of the form

$$\boldsymbol{E}(x,y)e^{i\gamma z-i\omega t}, \quad \boldsymbol{H}(x,y)e^{i\gamma z-i\omega t},$$

each of which is a generalized solution of Maxwell's equations in the waveguide. Here the parameter γ can take purely real and purely imaginary values.

The object that arose under this consideration is well known in the theory of waveguides. A nontrivial field $\boldsymbol{E}, \boldsymbol{H}$ in the waveguide $S \times Z \times T$, which depends on z and t as $e^{i\gamma z-i\omega t}$, where ω, γ are constants, is called the normal mode of the waveguide, ω is called its frequency, and γ is the wave number. In this case complex values are allowed for γ.

The TE-modes are treated in a completely similar way.

Theorem 2. *Let the waveguide filling be described by piecewise constant functions ϵ and μ. At frequencies other than the cutoff frequency, any monochromatic electromagnetic field $\boldsymbol{E}, \boldsymbol{H}$ in the waveguide can be represented as a superposition of normal modes of the waveguide, and for the fields $\boldsymbol{E}, \boldsymbol{H}$ the corresponding series converge in the norm $L^2(S)$.*

Note 2. The completeness of the system of normal modes for waveguides with complex filling was established in the works of Delitsyn [10]. In these works, the original system of Maxwell's equations was reduced to a new one, which could be written down in compact but, unfortunately, non-self-adjoint operators. The theory developed by Keldysh [17] allowed us to establish the completeness of the system of eigenvectors and associated vectors. However, for non-self-adjoint operators, the completeness is not identical to the basis property, the latter was established by this method only for waveguides of circular cross-section [18]. In the proposed version of the theory of normal modes, based on the technique of four potentials, the search for normal modes reduces to a self-adjoint problem, which removes many subtle questions, including the question of basis property.

7 Calculation of Normal Modes

Theorem 2 reduces the calculations in waveguide problems to the determination of normal modes. The TM mode is an eigenfunction of the problem

$$\begin{cases} A_\epsilon \tilde{u}_e = -\gamma^2 B_\epsilon \tilde{u}_e + k\gamma B_\epsilon \tilde{u}_h, \\ A_{\frac{1}{\mu}} \tilde{u}_h = -k\gamma B_\epsilon \tilde{u}_e + k^2 B_\epsilon \tilde{u}_h \end{cases} \tag{10}$$

or

$$A_\epsilon \tilde{u}_e = -\gamma^2 \left(B_\epsilon + B_\epsilon \left(\frac{1}{k^2} A_{\frac{1}{\mu}} - B_\epsilon \right)^{-1} B_\epsilon \right) \tilde{u}_e. \tag{11}$$

From the definition of the operators A and B many properties of this problem can be derived in advance.

Theorem 3. *A waveguide with piecewise-constant filling possesses an infinite number of normal modes, only some of them being travelling waves; for the wavenumber the estimate $|\gamma| < k$ is valid.*

Note 3. Our calculations of the filled waveguide modes, based on the incomplete Galerkin method [14], led to wavenumber of higher-order modes having both real and imaginary parts, one of them being very small. Now it is clear that it was a purely numerical artefact.

To solve this problem it is natural to use the truncation method. Thus we change the operators $A_\epsilon, A_{\frac{1}{\mu}}$ and B_μ in eigenvalues problem (10) to the matrices. For example, we can use finite element space V_h instead of Sobolev space $\overset{0}{W}{}_2^1(S)$. Here h is the estimate for sides of triangles. In the case of a rapidly changing filling, as, for example, in multicore optical fibers, this value should be chosen substantially smaller than the characteristic linear dimension of the inhomogeneity. A standard algebraic eigenvalue problem for the operator bundle will result.

For applications the high frequency limit is especial interesting. We can't do limit transition $k \to \infty$ in (11), but we can do it after truncation of operators. The operator B_ϵ is compact and thus is not invertible. But the finite element method give us some regularization: the matrix B_ϵ has the inverse matrix. We can estimate norm of this inverse matrix using the last eigenvalue of B_ϵ that is

$$\|B_\epsilon^{-1}\| \simeq h^{-2}.$$

This circumstance allows to apply the Neumann formula in Eq. (11), thus

$$B_\epsilon \left(\frac{1}{k^2} A_{\frac{1}{\mu}} - B_\epsilon\right)^{-1} B_\epsilon = -B_\epsilon \left(B_\epsilon^{-1} + \frac{1}{k^2} B_\epsilon^{-1} A_{\frac{1}{\mu}} B_\epsilon^{-1} + \mathcal{O}\left(\frac{1}{k^4}\right)\right) B_\epsilon$$

and

$$A_\epsilon \tilde{u}_e = \frac{\gamma^2}{k^2}\left(A_{\frac{1}{\mu}} + \mathcal{O}\left(\frac{1}{k^2}\right)\right)\tilde{u}_e.$$

If we will designate eigenvalues of the operator bundle $A_\epsilon - \delta^2 A_{\frac{1}{\mu}}$ as δ_n then we can write the eigenvalues of (11) asymptotically as

$$\gamma_n = \delta_n k + \dots.$$

Thus eigenmodes have the form

$$\boldsymbol{E}(x,y)e^{ik\delta_n z - i\omega t}, \quad \boldsymbol{H}(x,y)e^{ik\delta_n z - i\omega t}$$

The condition of the application of Neumann formula consists in $h^{-2}k^{-2} \ll 1$, thus h has be much less then the wavelength. If ϵ and μ don't depend on x, y then $\frac{1}{\mu} A_\epsilon = \epsilon A_{\frac{1}{\mu}}$ and thus $\delta = \sqrt{\epsilon\mu}$. So the inhomogeneity can be used for control of the ratio γ/k in waveguide modes.

8 Results

In the present paper, the main theorems of the hollow waveguide theory are generalized over the case of waveguides with piecewise constant filling. Similar to the case of hollow waveguides, an arbitrary field can be presented as a sum of TE and TM fields. In the case of a hollow waveguide, the monochromatic TE and TM fields satisfy Helmholtz equation. Instead, for a filled waveguide the following equation arises

$$Au = \partial_z^2 Ku, \tag{12}$$

where A and K are bounded self-adjoint operators, A determines strictly positive defined quadratic form, and K is a completely continuous operator. This equation is by no means more complicated than the Helmholtz equation, so that the proposed method allows applying the numerical methods developed for scalar waveguides to the vector model.

9 Conclusion

Derived method for the analysis of wave propagation in waveguides filled with piecewise-constant media enriches the concept of splitting modes in coupled multicore waveguides. Also we showed how the splitting may be used for the theoretical and numerical analysis of boundary value problems of the mathematical theory of waveguide systems. This approach may have potential applications in mode-division multiplexing systems [19–21], and in other areas related to waveguide optics, such as optical phased arrays [22,23], beam combining [24,25] and fiber imaging systems [26,27], sensor fibers [28] and many other devices.

The results to date indicate that there is the potential to achieve a 10–100-fold improvement in per-fibre capacity, at least in the laboratory. Whether this will ever translate into commercial long-haul systems is yet to be seen, though there is already evidence that some of the technology may find use in the shorter term in short reach applications, for example in Data Centres, where achieving high spatial path densities is critical and the barrier to entry for new technologies is very much lower than for long-haul communication networks [4].

Applicability will depend on the volume of production and its competitive commercial availability. Anything that impacts the ease of installing and maintaining fiber lines efficiently on behalf of our customers will affect us. However, pushing spectral efficiency higher than is currently possible with 100 Gb/s QPSK (phase shift keying is one of the types of phase modulation) modulation coherent technology requires other options for moving beyond established G.652.D specification fiber designs. Long distance connections between cloud data centers in particular need to implement higher order modulation formats like 8QAM and 16QAM (quadrature amplitude modulation). These formats require higher optical signal-to-noise-ratio and are more sensitive to non-linear impairments from the 80 μm^2 G.652.D fiber core area [3].

References

1. Russell, P.St.J.: Photonic crystal fibers (review paper). Science **299**, 358–362 (2003)
2. Coffey, V.C.: Novel fibers use space to extend capacity limits. Photonics Spectra **47**, 52–55 (2013)
3. Extance, A.: Redefining the limits of optical fibre. Opt. Connect. **9**(Q2), 12–13 (2017)
4. Richardson, D.J.: New optical fibres for high-capacity optical communications. Phil. Trans. R. Soc. A **374**, 20140441 (2016)
5. Malykh, M.D., Sevastianov, L.A., Tiutiunnik, A.A., Nikolaev, N.E.: On the representation of electromagnetic fields in closed waveguides using four scalar potentials. J. Electromagn. Waves Appl. **32**(7), 886–898 (2017)
6. Samarskii, A.A., Tikhonov, A.N.: About representation of the field in a waveguide in the form of the sum of fields TE and TM (in Russian). J. Theor. Phys. **18**(7), 959–970 (1948)
7. Zhang, K., Li, D.: Electromagnetic Theory for Microwaves and Optoelectronics. Springer, Berlin (2007)
8. Chew, W.C.: Lectures on theory of microwave and optical waveguides (2012). http://wcchew.ece.illinois.edu
9. Sveshnikov, A.G.: A substantiation of a method for computing the propagation of electromagnetic oscillations in irregular waveguides. U.S.S.R Comput. Math. Math. Phys. **3**(2), 413–429 (1963)
10. Delitsyn, A.L.: On the completeness of the system of eigenvectors of electromagnetic waveguides. Comput. Math. Math. Phys. **51**(10), 1771–1776 (2011)
11. Delitsyn, A.L.: Application of the finite element method to the calculation of modes of dielectric waveguides. Comput. Math. Math. Phys. **39**(2), 298–304 (1999)
12. Lezar, E., Davidson, D.R.: Electromagnetic waveguide analysis. In: Logg, A., Mardal, K.A., Wells, G. (eds.) Automated Solution of Differential Equations by the Finite Element Method: The FEniCS Book. LNCSE, vol. 84, pp. 629–642. Springer, Berlin (2012). https://doi.org/10.1007/978-3-642-23099-8_34
13. Delitsyn, A.L.: An approach to the completeness of normal waves in a waveguide with magnetodielectric filling. Differ. Equ. **36**(5), 695–700 (2000)
14. Malykh, M.D., Sevastianov, L.A., Tiutiunnik, A.A., Nikolaev, N.E.: Diffraction of electromagnetic waves on a waveguide joint. In: EPJ Web of Conferences, vol. 173, p. 02014 (2018)
15. Duvaut, G., Lions, J.-L.: Les in quations en m canique et en physique. Dunod, Paris (1972)
16. Stummel, F.: Rand- und Eigenwertaufgaben in Sobolewschen Räumen. Springer, Berlin (1969). https://doi.org/10.1007/BFb0059060
17. Keldysh, M.V.: On the completeness of the eigenfunctions of some classes of non-selfadjoint linear operators. Russ. Math. Surv. **26**(4), 15–44 (1971)
18. Bogolyubov, A.N., Delitsyn, A.L., Malykh, M.D.: On the root vectors of a cylindrical waveguide. Comput. Math. Math. Phys. **41**(1), 121–124 (2001)
19. Xia, C., Bai, N., Ozdur, I., Zhou, X., Li, G.: Supermodes for optical transmission. Opt. Exp. **19**, 16653–16664 (2011)
20. Ryf, R., et al.: Space-division multiplexed transmission over 4200 km 3-core microstructured fiber. In: presented at the IEEE Optical Fiber Communication Conference, Los Angeles, paper PDP5C.2 (2012)

21. Arik, S.O., Kahn, J.M.: Coupled-core multi-core fibers for spatial multiplexing. IEEE Photon. Technol. Lett. **25**(21), 2054–2057 (2013)
22. Li, L., et al.: Phase locking and in-phase supermode selection in monolithic multi-core fiber lasers. Opt. Lett. **31**, 2577–2579 (2006)
23. Sun, J., Timurdogan, E., Yaacobi, A., Hosseini, E.S., Watts, M.R.: Large-scale nanophotonic phased array. Nature **493**, 195–199 (2013)
24. Bochove, E.J., Shakir, S.: Analysis of a spatial-filtering passive fiber laser beam combining system. IEEE J. Sel. Topics Quantum Electron. **15**(2), 320–327 (2009)
25. Corcoran, C.J., Durville, F.: Passive phasing in a coherent laser array. IEEE J. Sel. Topics Quantum Electron. **15**(2), 294–300 (2009)
26. Kim, D., et al.: Toward a miniature endomicroscope: pixelation-free and diffraction-limited imaging through a fiber bundle. Opt. Lett. **39**, 1921–1924 (2014)
27. Reichenbach, K.L., Xu, C.: Numerical analysis of light propagation in image fibers or coherent fiber bundles. Opt. Exp. **15**, 2151–2165 (2007)
28. Sensor fiber having a multicore optical waveguide including fiber Bragg gratings, US patient #20140029889A1. http://www.freepatentsonline.com/8123400.html

On-the-Fly Multiple Sources Data Analysis in AR-Based Decision Support Systems

Van Phu Tran, Maxim Shcherbakov(✉), and Van Cuong Sai

Volgograd State Technical University, Lenin Avenue 28, 400005 Volgograd, Russia
maxim.shcherbakov@vstu.ru
http://www.vstu.ru

Abstract. Augmented reality application requires both type of data analysis: image recognition or segmentation and appropriate data extracting related to semantic of image. Obtained information matched and provided for end user. The latency is a crucial point, so the fast approaches are mandatory to use. We suggest an approach for combining multiple sources analyses and positioning for augmented reality (AR)-based application. Distinctive features of the method are (1) the use of the coordinates of the observer and camera and the moving object to clarify the position; (2) identification of the object using image recognition and (3) processing of log data obtained from vehicles. The proposed method in this study can be applied in augmented reality-based decision support system which requires obtain and proceed data from multiple data sources. The proposed method was applied to the traffic analysis task based on video streaming and log data analysis where location of observer is the similar to location of camera.

Keywords: Augmented reality · Multiple source analysis · Positioning

1 Introduction

Augmented reality is used in many areas such as marketing, industry, construction, medicine, entertainment. AR technology is applied to help to analyse the current situation in situ and to support decision-making process [3]. AR-based decision making could be considered as a part of intelligent systems. Intelligent decision support systems that implement AR components leads to qualitative changes in the effectiveness of decision-making process. As an example, the problem of traffic situation can be considered in the concept of Smart Cities development. In this case, the decision support system captures the image of a vehicle passing by an observer, transfers data to a cloud computing system, receives additional information about the vehicle and presents it's to the observer. It

V. P. Tran—The reported study supported by RFBR research projects #16-37-60066 mol_a_dk.

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 481–492, 2018.
https://doi.org/10.1007/978-3-319-99447-5_41

allows understanding who is a driver, technical states of the vehicle and other useful information for the further decision.

When developing AR-applications as a component of intelligent decision support systems, the crucial issue is efficient synchronizing of heterogeneous data. Basically, a solution to the problem requires analyzing the image (static or video) and performs a search for the information relevant to these images. This problem relates to the tasks of processing heterogeneous data obtained from multiple data sources, the problem known as multiple sources analysis. The main criterion for evaluation quality of decisions is the delay in processing and obtaining relevant information. If an end user uses a mobile phone or a tablet, the data is transmitted to a server for further processing. Due to energy efficiency and memory issues, the performance of devices are not too high to make multiple sources data processing on site. In other hands, a server is a bottleneck in a server (or cloud) oriented architecture. The question here is, that is the efficient way to organize multiple sources data transmitting and processing in AR-based decision support system?

This paper presents a new method for multiple sources data analysis in AR-based decision support system. The method uses for getting additional information about a vehicle registered in the database. Distinctive features of the method are (1) the use of the coordinates of the observer and camera and the moving object to clarify the position; (2) identification of the object using image recognition and (3) processing of log data obtained from vehicles.

2 Background

A critical problem in management Decision Support Systems (DSS) is the analysis of heterogeneous data. In [17,19], the authors developed a method for merging and preprocessing sensory heterogeneous data to minimize the time required to execute real-time queries. To process large amounts of data, it is necessary to develop an effective method for storing and classifying the data obtained from various sources (sensors). In [20] an effective approach is presented, aimed at creating a small piece of data of various types with a fixed size of the volume at the stage of collecting raw data. This approach is suitable for systems that continuously receive events from various sources, and allows quick access to the database. In practice, this approach can lead to an increase latency, and therefore difficulties arise when applied in real-time systems in the case of high volume data arriving at high speed. In this case, distributed technologies can be used to collect and preprocess data, such as Apache Kafka, Flume, Spark Streaming [17].

In DSS, processing media data, like images, video, takes a long time, and therefore results in low system efficiency in general. In practice, use the OpenCV library with the cascade classifier model to recognize objects in images [6,11,21]. In 2016 a free open YOLO library was released, which can recognize objects in real time using the GPU [12]. This framework uses the Darknet model to extract image characteristics based on a Convolutional Neural Network (CNN) [4]. The

YOLOv3-tiny (the latest version) framework can process up to 220 images per second [22]. The YOLO outperformed OpenCV in terms of quality and speed. The advantage in the performance of the YOLO framework is primarily in the use of the GPU. For some cases, the results for OpenCV are comparable in quality with YOLO, in particular when considering the task of identifying vehicles in systems without a GPU. In addition, alternative frameworks have been developed that can be used to recognize objects with high accuracy, for example, the Mask R-CNN framework for recognition and semantic segmentation of images based on Tensoflow [7].

Nowadays, there are many libraries with different architectures based on the neural network where developed. e.g. Tensoflow [15], Theano [16], Caffe [2], Keras [8]. These libraries allow to develop various models of a neural networks, such as a convolutional neural network, for solving the problem of pattern recognition.

A Convolutional Neural Networks (CNN, or ConvNet) model is a special kind of multi-layer neural networks, designed to recognize visual patterns directly from pixel images with minimal preprocessing. There are various architectures of a convolutional neural network with high recognition accuracy, for instance: GoogLeNet [5], AlexNet [9]. LeNet [10], VGG Net [14], ResNet [13], ZFNet [23], etc. AlexNet has a parallel two CNN using two GPUs with cross-connections, GoogleNet has inception modules, ResNet has residual connections. Basically, training procedure is very time consuming. For example, GoogleNet might have more than 4 million hyperparameters. However, using modern CNN architecture, as the ResNet CNN architecture, allows to get error rate is only 3.57%, which is lower than the human readings on the ImageNet data set. The listed CNN architectures are suitable for the tasks of semantic image analysis. Using the power of graphics memory, GPUs based on these CNN architectures can be used in a real-time analysis of video streams that is suitable for the task considered in the research.

3 Proposed Method

3.1 Task Statement

Let, there is a set of vehicles V registered in the database SD, where SD is a static database containing general information about vehicles. For each vehicle in the set, the following data are stored: VIN code, ID, information about the owner of vehicle and optional road statistics (a number of accidents). Each vehicle sends a data package $DP_v(t)$ every discrete time period to dynamic database marked as DD [18]. Figure 1 shows the structure of transferred data.

There is an observer o which has its own location $(long_o, lat_o)$. The observer intents to get additional information about the passing vehicle. This information should be obtained from SD. The traffic camera formalized as $C = (long_c, lat_c)$ is installed in a certain location (a road section RS) with longitude $long_c$ and latitude lat_c. The camera is capturing video streams of part of roads with all traffic participant. Assume, that location of the camera is the same as observed

```
{ data
     [
     "VehicleID" : {String} ,
     "DriverID"  : {String} ,
     "eventStart": {Long int } ,
     "eventEnd"  : {Long int } ,
     "longitude" : {Double} ,
     "latitude"  : {Double} ,
     "velocity"  : {Double} ,
     "status"    : {String}
     ]
}
```

Fig. 1. A JSON-based scheme of transferred data, where uid - object ID, eventStart – time when event was started, eventEnd – time when event was terminated, longitude – longitude of the vehicle's location, latitude – latitude of the vehicle's location, velocity - the current velocity of the vehicle, status - a status of the vehicle (optional).

part of the road section. Video streaming VS_c is transmitted to the server for further analysis. Assume, that $long_o = long_c, lat_o = lat_c$.

Using video streams from camera VS_c, data of locations of observer, data gathered from vehicles, and vehicle data stored in the static database, it is necessary to provide relevant data about the recognized vehicle from SD as well as from DD to the observer.

Figure 2 represents an instance of transmitted data according to predefined scheme. *VehicleID* is an unique number of vehicle which is stored both in Static database and Dynamic database. We use the number to show the observer. *DriverID* is temporal data about driver (or owner) of the vehicle. *Timestamp* is Unix based time and values of parameters like *Speed*, *Latitude* and *Longitude* is captured at time declared for *Timestamp*.

```
{
"VehicleID":"0a955f61-65c3-44f9-8893-f49163225c05",
"DriverID":"123456",
"Speed":78.0,
"Timestamp":"1516661614478",
"Latitude":"45.07256",
"Longitude":"43.995274"
}
```

Fig. 2. An example of data transmitted from a vehicle to a server.

3.2 Description of the Method

The proposed method contains the next steps executed for the certain discrete time τ (time of start of the event).

We assume, that image from camera VS_c is transfered into server continuously, so the server splitting video stream into images every Δt. Also, data from vehicle DS_{lg} is transfered into server at time T_v, where $|T_c - T_o| < \epsilon$.

1. An observer o using AR-based system starts capturing the image of a road section at exact time τ (an event has started).
2. Send the data package DP about the start event to the server. The package contains data about observer (its location) and start-of-event time τ.
3. When data packages is received by server, do request video stream VS_c according to the location of observer, and the camera location.
4. Launch vehicle recognition procedure over images extracted from video streams VS_c in the time interval $[\tau - \epsilon; \tau + \epsilon]$.
5. If any vehicle is recognized, perform the following actions.
   (a) Store the timestamp of recognized vehicles as τ^*.
   (b) Create and send the request R_1 to DD. The request contains time of recognition τ^* and a pair $(long_c, lat_c)$.
   (c) Select from DD all vehicles which have the same location in the time interval $[\tau - \epsilon; \tau + \epsilon]$. The same location means that $L < d$, where d – is predefined threshold, and $L = R \cdot c$, where R – id the Earth radius. $c = 2 \cdot atan2(\sqrt{a}, \sqrt{1-a})$, and $a = sin^2((lat_v - lat_c)/2) + cos(lat_c) \cdot cos(lat_v) \cdot sin^2((long_v - long_c)/2)$.
   (d) If the vehicle is selected in previous step with ID^*, create and send the request R_2 to SD. The request contains the unique identificator of found vehicle ID^*.
   (e) Obtain and prepare response data package DP_r containing information about the vehicle with ID^*.
6. Send the data package DP_r to observer.

If there is no data in data bases DD or SD (responses of requests are empty), the DP_r contains the message like "No data available for the vehicle".

3.3 A Model for Data Storing

For a quick access in the lake of data, the structure of storing log files is developed according to a concept of data lake. Figure 3 shows the general structure of log data storage. The file name is the time t of the data log generation. Time t has a Unix timestamp format. Each file contains data d about the object examined.

When the log of the examined object is sent to the server, the log file is generated and indexed according to the indexing strategy, which is represented in the Fig. 3, where *IDSource* is the unique identifier of the considered object (source), *tstart* is the generation time of the *File_Start* file with the JSON format in which the data on the considered object.

Consider how we use the indexing strategy. Let's assume that we want to find out at what times in the interval $[t1, t2]$ the considered object with *IDSource2*

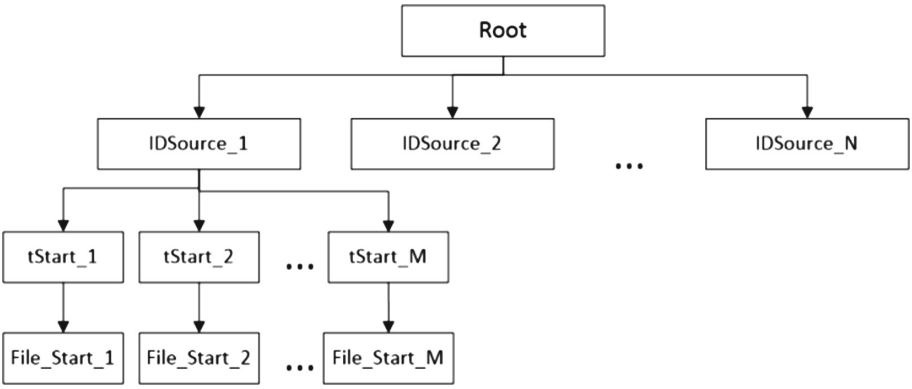


Fig. 3. Structure of stored data

has a speed that is greater than a given value of v. First we look for the path */theRootDirectoryToStoreObjectData/IDSource2*. Then parse all files */theRootDirectoryToStoreObjectData/IDSource2/FileName/*, whose name *FileName* satisfies the conditions:

$$t1 < String2Time(FileName) < t2 \tag{1}$$

As a result of parsing each file, we compare the speed value with a given value of v to add time t to the results list.

4 Use Case

4.1 Experimental Design

The section contains implementation details of the proposed method for the traffic analysis task.

The proposed method was used for traffic situation analysis with a prototype of AR-based information system. The experiment included a system with a camera which captures images from a part of the road. Video observes the part of the road with double-direction running. An observer used a tablet or mobile device for getting information about vehicle. The observer used its device to capture image of passing by vehicle. If the vehicle is recognized and registered in the database, the observer received additional information (e.g. VIN code) which was overlapped video images. To limit observation scene and synchronize two cameras (traffic camera and observer camera), the imaginary line was added to image. If a vehicle cross the line, the system reacts and transfer data. The line was highlighted to the observer.

As an example, we use ready-made videos obtained from a video camera on the road, and log-data about transport vehicles to determine what object is in an interesting area of the video camera or observer segment. As it was mentioned, the camera and the observer has nearly the same or the same location. Based

on this data and according to the task statement, the system provides the end user (observer) all the information associated with the recognized object.

We assume, that every vehicle in the video stream is registered both in static and dynamic data. Also, the simulation of data transferring from vehicle to dynamic database is on. In real life, we expect that there a number of vehicle are not equipped by data acquisition and transferring devises. It means the identification method return nothing to an observer.

4.2 System Architecture

Figure 4 shows the principle of the system which implements proposed method.

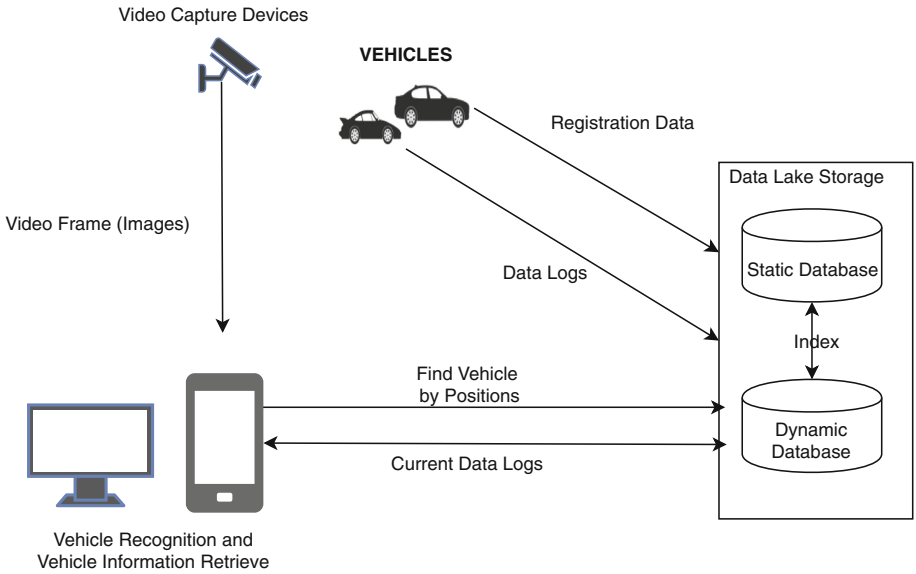


Fig. 4. The concept of system implementing proposed method.

The system consists of four main modules:

1. Data collection module. This module captures images that are generated by video cameras and sent to the heterogeneous data synchronization module.
2. Heterogeneous data synchronization module. The module processes received images from the data collection module and it performs an algorithm for localizing the object on the scene of the image. This module is based on OpenCV library. When an object is recognized in an interesting area of the image, the module specifies a request to the module for working with the data storage to find the information about the recognized object.

3. A module for working with the data storage. The module consists of methods for finding information about the recognized vehicle in the data storage (dynamic database), where the log data about objects is stored.
4. A data storage called Data lake, which is including Static database and Dynamic database.

4.3 Results and Discussion

All experiments were conducted on a laptop running the operating system Ubuntu x86_64 with an Intel (R) Core (TM) i5–2430M CPU with a frequency of 2.40 GHz.

The first scenario where all cars on the road are recognized (images of cars are framed by a green rectangle). As soon as the car crosses the red line (a selected section of a road), the identification method is performed by a system.

Figure 5 shows the results of the recognition of transport objects. Figure 6 shows information about the recognized object with ID *8389fa2f–52a2–4d1b–985c–2b7de1f6fa0b* retrieved from both databases *DS* and *DD*.

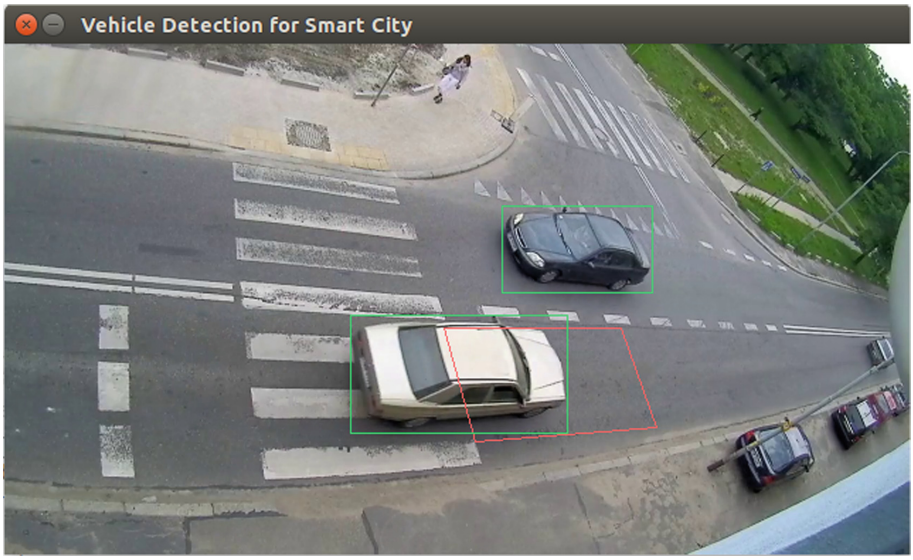


Fig. 5. Recognition of a transport object crossing an interesting area (red area) (Color figure online)

The second scenario shows a big vehicle (a bus). Information about this vehicle does not stored in database.

```
------------------------
Vehicle ID: 8389fa2f-52a2-4d1b-985c-2b7de1f6fa0b
Vehicle Speed: 57.91188681192046
Vehicle Location (latitude): 44.0000286909151
Vehicle Location (longitude): 46.00009722627896
------------------------
Current number of vehicle: 3
```

Fig. 6. Information about recognized vehicle with ID *8389fa2f-52a2-4d1b-985c-2b7de1f6fa0b* obtained from the data lake

Figure 7 shows the results of the recognition of transport objects. Figure 8 shows information that the recognized vehicle are not included in static or dynamic database.

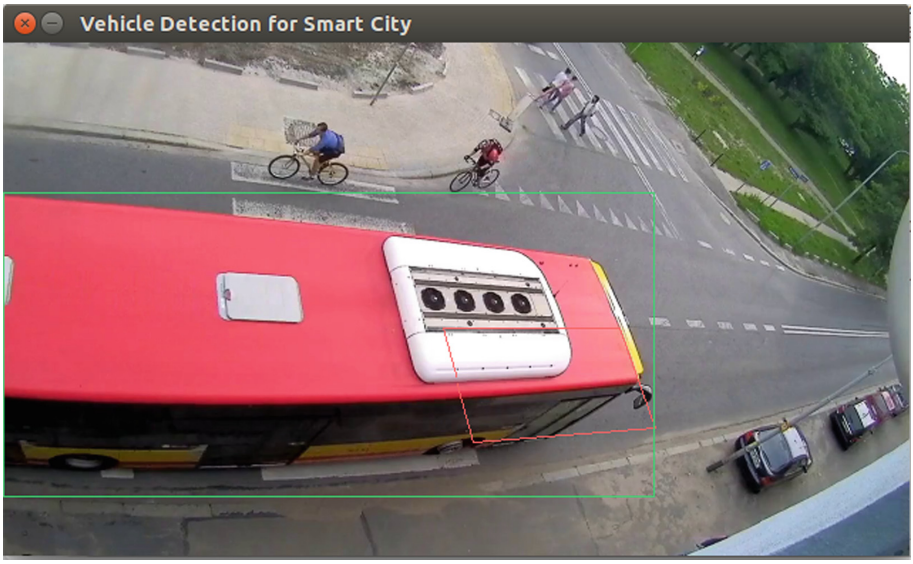


Fig. 7. Recognition of a bus crossing an interesting area (red area) (Color figure online)

Note, the large object hides other object, so it can be considered as a additional issue need to be solved in future.

In contrast with previous scenario, the latter one shows a small vehicle. This Note, the large object hides other object, so it can be considered as a additional issue need to be solved in future.

```
Current number of vehicle: 5
/home/kim/Desktop/logdata/22733645-ead0-414e-8370-e7293a19293c/1529273026332.txt
Eclapsed time:25 ms
No data available for the vehicle!
Current number of vehicle: 6
```

Fig. 8. Notification that there is no information about vehicle

Figure 9 shows the results of the recognition of small vehicle. Figure 10 shows information about the recognized object with ID *a9ff8a33-cda4-487b-a5a5-1827115507b2* obtained from both databases.

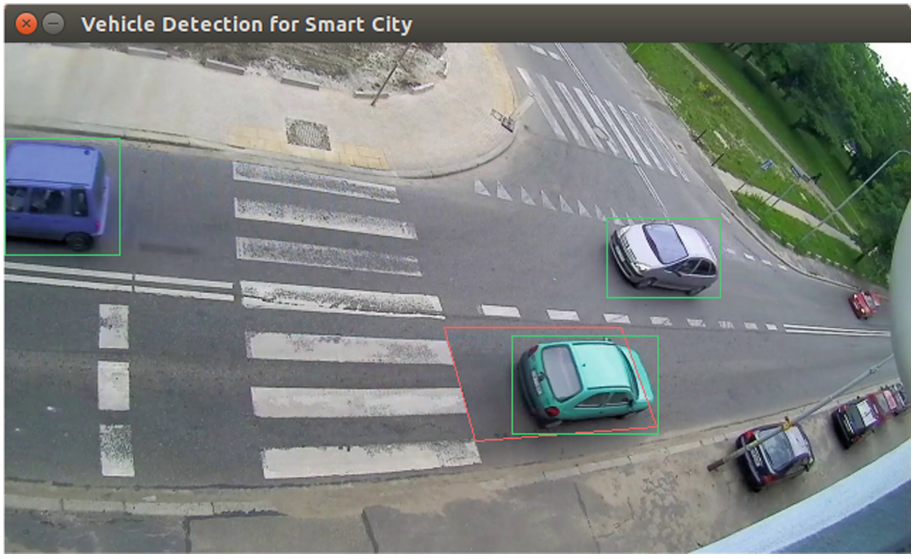


Fig. 9. Recognition of a transport object crossing an interesting area (red area) (Color figure online)

In this paper, we developed a method for recognizing moving transport objects based on the Background Subtraction method in the OpenCV 3.0 library. The processing time of one frame of video (image) is 35 ± 10 ms, so the system is suitable to work in real time. The volume of data log files in the dynamic database is up to 2 Mb, the processing time of the data log is 80 ± 20 ms. Note, that search time of vehicle data in *DD* depends on volume of log data. Technically, log data can be splitted into data chunks according to vehicle location and period of time.

```
------------------------
Vehicle ID: a9ff8a33-cda4-487b-a5a5-1827115507b2
Vehicle Speed: 23.58753950542249
Vehicle Location (latitude): 44.00004537958473
Vehicle Location (longitude): 46.00005788660227
------------------------
Current number of vehicle: 1
```

Fig. 10. Information about recognized vehicle with ID *a9ff8a33-cda4-487b-a5a5-1827115507b2* obtained from the data lake

5 Conclusion

We conclude, that the proposed method in this study can be applied in augmented reality-based decision support system which requires obtain and proceed data from multiple data sources. The proposed method was applied to the traffic analysis task based on video streaming and log data analysis.

In future work, we adapt the method for situation where a location of an observer differs from a location of a traffic camera. Also, deep learning techniques like convolutional neural networks (CNN) can be applied to vehicle recognition toward improving quality of recognition.

References

1. Alamri, A., Cha, J., Eid, M.: Evaluating the post-stroke patients progress using an augmented reality rehabilitation system. In: International Workshop on Medical Measurements and Applications, Cetraro, Italy, 29–30 May 2009, pp. 89–94 (2009)
2. Caffe is a deep learning framework. http://caffe.berkeleyvision.org/
3. Park, C.-S., Lee, D.-Y., Kwon, O.-S.: A framework for proactive construction defect management using BIM, augmented reality and ontology-based data collection template. Autom. Constr. **33**, 61–71 (2013)
4. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Computer Vision and Pattern Recognition. https://arxiv.org/abs/1311.2524
5. GoogleNet. https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/googlenet.html
6. Jalled, F., Voronkov, I.: Object detection using image processing. https://arxiv.org/pdf/1611.07791.pdf
7. He, K., Gkioxari, G., Dollar, P., Girshick R., (Facebook AI Research, FAIR): Mask R-CNN. https://arxiv.org/pdf/1703.06870.pdf
8. Keras: the Python deep learning library. https://keras.io/
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 25, pp. 1097–1105. Curran Associates, Inc. (2012). http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

10. LeCun, Y., Bottou, L., Bengio, Y.: Reading checks with graph transformer networks. In: International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 151–154. IEEE, Munich (1997)
11. OpenCV. https://opencv.org/
12. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement. https://pjreddie.com/media/files/papers/YOLOv3.pdf
13. ResNet in TensorFlow. https://github.com/ry/tensorflow-resnet
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
15. TensorFlow: an open source machine learning framework for everyone. https://www.tensorflow.org/
16. Theano. http://deeplearning.net/software/theano/
17. Tran, V.P., Shcherbakov, M., Nguyen, T.A.: Yet another method for heterogeneous data fusion and preprocessing in proactive decision support systems: distributed architecture approach. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds.) DCCN 2017. CCIS, vol. 700, pp. 319–330. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66836-9_27
18. Tran, V.P., Shcherbakov, M., Nguyen, T.A.: EVGEN: a framework for event generator in proactive system design. In: 2016 7th International Conference on Information, Intelligence, Systems Applications (IISA), pp. 1–5 (2016)
19. Tran, V.P., Shcherbakov, M.V., Nguyen, T.A., Skorobogatchenko, D.A.: A method for data acquisition and data fusion in intelligent proactive decision support systems. In: Neurocomputers: Development, Application, vol. 11, pp. 40–44 (2016). (in Russian)
20. Tyukov, A.P., Khrzhanovskaya, O., Sokolov, A.A., Shcherbakov, M.V., Kamaev, V.A.: Fast access to large timeseries datasets in SCADA systems. Res. J. Appl. Sci. **10**, 12–16 (2015)
21. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Conference on Computer Vision and Pattern Recognition. https://www.cs.cmu.edu/~efros/courses/LBMV07/Papers/viola-cvpr-01.pdf
22. YOLO: real-time object detection, 5 April 2018. https://pjreddie.com/darknet/yolo/
23. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53

Retrial Queue M/M/N with Impatient Customer in the Orbit

Elena Danilyuk(✉), Olga Vygoskaya, and Svetlana Moiseeva

National Research Tomsk State University, Lenina Avenue, 36, 634050 Tomsk, Russia
daniluc_elena@sibmail.com,osipovich.olga@bk.ru,smoiseeva@mail.ru

Abstract. In the paper, the retrial queueing system of M/M/N type with Poisson flow of events and impatient calls is considered. The delay time of calls in the orbit, the calls service time and the impatience time of calls in the system have exponential distribution. Asymptotic analysis method is proposed for the solving problem of finding distribution of the number of calls in the orbit under a system heavy load and long time patience of calls in the orbit condition. The theorem about the Gauss form of the asymptotic probability distribution of the number of calls in the orbit is formulated and proved. Numerical illustrations, results are also given.

Keywords: Finite-source retrial queueing system · Orbit
Asymptotic analysis · Impatient calls

1 Introduction

Queueing systems with repeated calls, or Retrial Queueing Systems, are mathematical models widely used for many real objects, systems and processes analysis and optimization, especially telecommunication systems, networks, mobile networks, call-centres, manufacturing, economics [1,3,4,8]. In these queueing systems unserved calls are not lost when there are not available service devices (servers are busy or broken). So, the customers that don't get a service repeat to occupy server after a random time.

There are many papers devoted to RQ-systems study. The main results and comprehensive description of retrial queues are contained in the books [5,6]. The first retrial model with impatience was considered by Cohen [2]. The M/M/1 retrial queue with impatient calls was studied by Falin [6].

Models with calls leaved RQ-system after failed attempt to get a service was considered by many scientists [13–18], etc. In these studies, an arriving call joints the orbit with some probability p and leaves the system with the probability 1-p when there are not available service devices at the time. Some authors name such customers as non-persistent or p-non-persistent customers.

We consider a different model which was not been investigated early. So, in present research impatient customer is a customer in the orbit that can repeat

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 493–504, 2018.
https://doi.org/10.1007/978-3-319-99447-5_42

an attempt to reach the server again or can leave the orbit after a random time without server recalling.

Classical retrial models consist of one server but real telecommunication systems are usually multiserver retrial queue [10,19–21]. In the proposed paper RQ-system consisting of N service devices is considered.

Asymptotic analysis method is widely applied for RQ-systems research. The method makes it possible to produce analytical result for different types of queueing systems and networks under given asymptotic condition. More information about the asymptotic analysis method is provided in [5–7,10,18,22], etc.

The general information about mathematical model of the retrial queueing system discussed in the paper and the problem statement are presented in the Sect. 2. In the Sect. 3 the detailed derivation of the model and the system of Kolmogorov equations for the stationary state probabilities are cited. The Sect. 4 consists of the decision of the problem under study by the asymptotic analysis method. As a result of the section the Theorem about stationary probability distribution of the calls number in the orbit for Retrial queueing system of M/M/N type with impatient calls in the orbit under a system heavy load and long time patience of calls in the orbit condition is formulated and proved. Some numerical results, graphs, that proved the theoretical results, are performed in the Sect. 5. Section 6 concludes the paper.

2 Mathematical Model

A retrial queueing system consisting of an infinite orbit and N servers is considered. The input flow is defined by the stationary Poisson process with parameter λ. The service times on every of the N servers are exponentially distributed with parameter μ. A customer which arrives into the system, when at least one of the N servers is free, instantly occupies this server. If all of the devices are busy, the call goes to the orbit, where it stays during a random time distributed exponentially with parameter σ. After the delay the customer makes an attempt to reach any server again. If it is free, the call occupies it, otherwise the call immediately joins the orbit. From the orbit calls (impatient calls) can leave the system after a random time distributed exponentially with parameter α.

The structure of the model is presented in Fig. 1.

The problem is to get stationary probability distribution of the number of calls in the orbit for the system under review.

3 Process of the System States: Stationary Distribution

Let us consider Markovian process $\{k(t), i(t)\}$ determined states of the Retrial queue M/M/N with impatient customer in the orbit where the random process $i(t)$ is the number of calls in the orbit at the moment t, $i(t) = 0, 1, 2, 3, \ldots$, the random process $k(t)$ defines device state at the moment t and takes one of the

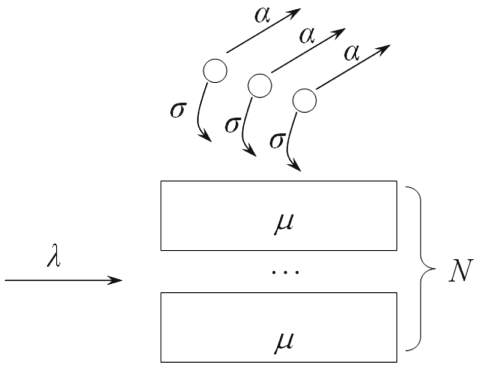


Fig. 1. Retrial queue M/M/N with impatient calls in the orbit

$(N+1)$ values

$$k(t)=\begin{cases}0, & \textit{if all servers are free at the moment } t;\\ 1, & \textit{if only one of } N \textit{ servers is busy at the moment } t;\\ 2, & \textit{if two of } N \textit{ servers are busy at the moment } t;\\ \dots & \\ N, & \textit{if all of } N \textit{ servers are busy at the moment } t.\end{cases}$$

Denote the probability that, at the moment t, the server (device) is in the state k, $k=0,1,\dots,N$, and there are i calls in the orbit, $i=0,1,2,\dots$, as $P\{k(t)=k,i(t)=i\}=P_k(i,t)$. We write the following system of equations

$$\begin{cases}\dfrac{\partial P_0(i,t)}{\partial t}=-(\lambda+i\sigma+i\alpha)P_0(i,t)+\mu P_1(i,t)+(i+1)\alpha P_0(i+1,t),\ k=0,\\ \dfrac{\partial P_k(i,t)}{\partial t}=-(\lambda+i\sigma+i\alpha+k\mu)P_k(i,t)+(i+1)\sigma P_{k-1}(i+1,t)\\ +\lambda P_{k-1}(i,t)+(i+1)\alpha P_k(i+1,t)+(k+1)\mu P_{k+1}(i,t),\ k=1,\dots,N-1,\\ \dfrac{\partial P_N(i,t)}{\partial t}=-(\lambda+i\alpha+N\mu)P_N(i,t)+(i+1)\sigma P_{N-1}(i+1,t)\\ +\lambda P_{N-1}(i,t)+(i+1)\alpha P_N(i+1,t)+\lambda P_N(i-1,t),\quad \textit{if}\quad k=N.\end{cases}\tag{1}$$

The system of Kolmogorov equations for the stationary state probabilities $\Pi_k(i)=\lim\limits_{t\to\infty}P(k,i,t)$ of the process $\{k(t),i(t)\}$ is written as follows

$$\begin{cases}-(\lambda+i\sigma+i\alpha)\Pi_0(i)+\mu\Pi_1(i)+(i+1)\alpha\Pi_0(i+1)=0,\quad \textit{if}\quad k=0,\\ -(\lambda+i\sigma+i\alpha+k\mu)\Pi_k(i)+\lambda\Pi_{k-1}(i)+(i+1)\sigma\Pi_{k-1}(i+1)\\ +(i+1)\alpha\Pi_k(i+1)+(k+1)\mu\Pi_{k+1}(i)=0,\quad \textit{if}\quad k=1,\dots,N-1,\\ -(\lambda+i\alpha+N\mu)\Pi_N(i)+\lambda\Pi_{N-1}(i)+(i+1)\sigma\Pi_{N-1}(i+1)\\ +(i+1)\alpha\Pi_N(i+1)+\lambda\Pi_N(i-1)=0,\quad \textit{if}\quad k=N.\end{cases}\tag{2}$$

We get in (2) the indefinite dimensional system of difference equations with variable coefficients. In common case it is not possible to produce the exact solution of this system. To find solution of (2), we will use the method of asymptotic

analysis under a system heavy load and long time patience of calls in the orbit condition.

4 Asymptotic Analysis Method

The method of asymptotic analysis in queueing theory is the method of research of the equations determining some characteristics of an queueing system under some limit (asymptotic) condition, which is specific for any model and solving problem.

We introduce the partial characteristic functions

$$H_k(u) = H(k,u) = \sum_{i=0}^{\infty} e^{jui} \Pi_k(i), \quad H_k(0) = H(k,0) = \sum_{i=0}^{\infty} \Pi_k(i) = R_k, \quad (3)$$

where $j = \sqrt{-1}$, $k = 0, 1, \ldots, N$, and R_k are stationary state probabilities of the process $k(t)$. It is obvious that $H(u) = \sum\limits_{k=0}^{N} H_k(u)$.

Using $H_k'(u) = j \sum\limits_{i=0}^{\infty} e^{jui} \Pi_k(i)$ and (3) we can write the system (2) as

$$\begin{cases} j\left(\sigma + \alpha\left(1 - e^{-ju}\right)\right) H_0'(u) + \mu H_1(u) - \lambda H_0(u) = 0, \quad if \quad k = 0, \\ j\left(\sigma + \alpha\left(1 - e^{-ju}\right)\right) H_k'(u) - \lambda H_k(u) - j\sigma e^{-ju} H_{k-1}'(u) - k\mu H_k(u) \\ +\lambda H_{k-1}(u) + (k+1)\mu H_{k+1}(u) = 0, \quad if \quad k = 1, \ldots, N-1, \\ j\alpha\left(1 - e^{-ju}\right) H_N'(u) + \lambda H_{N-1}(u) - j\sigma e^{-ju} H_{N-1}'(u) \\ -\left(N\mu + \lambda\left(1 - e^{ju}\right)\right) H_N(u) = 0, \quad if \quad k = N. \end{cases} \quad (4)$$

The system in (4) is the base system for analysis of Retrial queueing system of M/M/N type with impatient calls in the orbit under a system heavy load ($\lambda \gg \mu$) and long time patience of calls in the orbit ($\alpha \to 0$) condition.

Theorem 1. *Stationary probability distribution of the calls number in the orbit for Retrial queueing system of M/M/N type with impatient calls in the orbit under a system heavy load and long time patience of calls in the orbit condition can be approximated by the Gaussian distribution with mean and variance equal to* $\frac{\lambda - N\mu}{\alpha}$ *and* $\frac{\lambda}{\alpha}$ *respectively, where* λ *is the parameter of the Poisson input calls flow,* μ, σ, α *are the exponential distribution parameters, accordingly, of the calls service time, the calls delay time in the orbit, the calls leaving the system from the orbit.*

Proof. The Theorem 1 proving will carried out in two stages.

Stage 1. Let to denote

$\alpha = \varepsilon$, $u = \varepsilon w$, $H_0(u) = \varepsilon^N F_0(w,\varepsilon)$, $H_1(u) = \varepsilon^{N-1} F_1(w,\varepsilon)$, $H_k(u) = \varepsilon^{N-k} F_k(w,\varepsilon)$, $\ldots$, $H_{N-1}(u) = \varepsilon F_{N-1}(w,\varepsilon)$, $H_N(u) = F_N(w,\varepsilon)$, where $\varepsilon \to 0$ is infinitesimal.

Since $H_k'(u) = \varepsilon^{N-k-1}\dfrac{\partial F_k(w,\varepsilon)}{\partial w}$, $k = 0, 1, \ldots, N$, then the equations system (4) can be written as

$$\begin{cases} j\left(\sigma + \varepsilon\left(1 - e^{-j\varepsilon w}\right)\right)\varepsilon^{N-1}\dfrac{\partial F_0(w,\varepsilon)}{\partial w} + \mu\varepsilon^{N-1}F_1(w,\varepsilon) \\ -\lambda\varepsilon^N F_0(w,\varepsilon) = 0, \quad if \quad k = 0, \\ j\left(\sigma + \varepsilon\left(1 - e^{-j\varepsilon w}\right)\right)\varepsilon^{N-k-1}\dfrac{\partial F_k(w,\varepsilon)}{\partial w} - \varepsilon^{N-k}\left(\lambda + k\mu\right)F_k(w,\varepsilon) \\ -j\sigma e^{-j\varepsilon w}\varepsilon^{N-k}\dfrac{\partial F_{k-1}(w,\varepsilon)}{\partial w} + (k+1)\mu\varepsilon^{N-k-1}F_{k+1}(w,\varepsilon) \\ +\lambda\varepsilon^{N-k}F_{k-1}(w,\varepsilon) = 0, \quad if \quad k = 1, \ldots, N-1, \\ j\varepsilon\left(1 - e^{-j\varepsilon w}\right)\sum\limits_{k=0}^{N}\varepsilon^{N-k-1}\dfrac{\partial F_k(w,\varepsilon)}{\partial w} + \lambda e^{j\varepsilon w}\left(1 - e^{-j\varepsilon w}\right)F_N(w,\varepsilon) \\ +j\sigma\left(1 - e^{-j\varepsilon w}\right)\sum\limits_{k=0}^{N-1}\varepsilon^{N-k-1}\dfrac{\partial F_k(w,\varepsilon)}{\partial w} = 0, \quad if \quad k = N. \end{cases} \tag{5}$$

Let divide each equation of the system (5) by the ε to the minimum power and then we can obtain (6)

$$\begin{cases} j\left(\sigma + \varepsilon\left(1 - e^{-j\varepsilon w}\right)\right)\dfrac{\partial F_0(w,\varepsilon)}{\partial w} + \mu F_1(w,\varepsilon) \\ -\lambda\varepsilon F_0(w,\varepsilon) = 0, \quad if \quad k = 0, \\ j\left(\sigma + \varepsilon\left(1 - e^{-j\varepsilon w}\right)\right)\dfrac{\partial F_k(w,\varepsilon)}{\partial w} - \varepsilon\left(\lambda + k\mu\right)F_k(w,\varepsilon) \\ -j\sigma e^{-j\varepsilon w}\varepsilon\dfrac{\partial F_{k-1}(w,\varepsilon)}{\partial w} + (k+1)\mu F_{k+1}(w,\varepsilon) \\ +\lambda\varepsilon F_{k-1}(w,\varepsilon) = 0, \quad if \quad k = 1, \ldots, N-1, \\ j\sum\limits_{k=0}^{N}\varepsilon^{N-k}\dfrac{\partial F_k(w,\varepsilon)}{\partial w} + \lambda e^{j\varepsilon w}F_N(w,\varepsilon) \\ +j\sigma\sum\limits_{k=0}^{N-1}\varepsilon^{N-k-1}\dfrac{\partial F_k(w,\varepsilon)}{\partial w} = 0, \quad if \quad k = N. \end{cases} \tag{6}$$

The transformation of the (6) with $F_k(w) = \lim\limits_{\varepsilon\to 0} F_k(w,\varepsilon)$ and the expansion $e^{\pm j\varepsilon w} = 1 \pm j\varepsilon w + o(\varepsilon^2)$ under $\varepsilon \to 0$ leads to differential equations system for $F_k(w)$, $k = 0, 1, \ldots, N$,

$$\begin{cases} j\alpha F_0'(w) = -\mu F_1(w), \quad if \quad k = 0, \\ j\alpha F_k'(w) = -(k+1)\mu F_{k+1}(w), \quad if \quad k = 1, \ldots, N-1, \\ -j\alpha F_{N-1}'(w) = jF_N'(w) + \lambda F_N(w), \quad if \quad k = N, \end{cases} \tag{7}$$

where $F_k'(w) = dF_k(w)/dw$, $k = 0, 1, \ldots, N$.

Solving the (7) it is easy to obtain that $F_N(w) = R_N e^{jw(\lambda - N\mu)}$ where the constant R_N is defined above.

Pre-limit characteristic function $h(u)$ is approximately equal to

$$h(u) = \sum_{k=0}^{N} H_k(u) = F_N\left(\frac{u}{\varepsilon}\right) + o(\varepsilon) \approx F_N\left(\frac{u}{\varepsilon}\right).$$

So, the first-order asymptotic characteristic function $h^{(1)}(u)$ of the probability distribution of the number of calls in the orbit under the system heavy load and long time patience of calls in the orbit condition can be presented as

$$h^{(1)}(u) = F_N\left(\frac{u}{\varepsilon}\right) = R_N \exp\left\{(\lambda - N\mu)\frac{ju}{\varepsilon}\right\} = R_N \exp\left\{\frac{\lambda - N\mu}{\alpha} ju\right\}. \quad (8)$$

Stage 2.

In the base system of Eqs. (4) and (8) denoting

$$H_k(u) = R_N \exp\left\{\frac{\lambda - N\mu}{\alpha} ju\right\} H_k^{(2)}(u), \quad k = 0, 1, 2, \ldots, N, \quad (9)$$

and making some transformations with this system we get (10)

$$\begin{cases} j\left(\sigma + \alpha\left(1 - e^{-ju}\right)\right)\left(H_0^{(2)'}(u) + j\dfrac{\lambda - N\mu}{\alpha} H_0^{(2)}(u)\right) \\ +\mu H_1^{(2)}(u) - \lambda H_0^{(2)}(u) = 0, \quad if \quad k = 0, \\ j\left(\sigma + \alpha\left(1 - e^{-ju}\right)\right)\left(H_k^{(2)'}(u) + j\dfrac{\lambda - N\mu}{\alpha} H_k^{(2)}(u)\right) - \lambda H_k^{(2)}(u) \\ -j\sigma e^{-ju}\left(H_{k-1}^{(2)'}(u) + j\dfrac{\lambda - N\mu}{\alpha} H_{k-1}^{(2)}(u)\right) - k\mu H_k^{(2)}(u) \\ +\lambda H_{k-1}^{(2)}(u) + (k+1)\mu H_{k+1}^{(2)}(u) = 0, \quad if \quad k = 1, \ldots, N-1, \\ j\alpha\left(1 - e^{-ju}\right)\sum\limits_{k=0}^{N}\left[H_k^{(2)'}(u) + j\dfrac{\lambda - N\mu}{\alpha} H_k^{(2)}(u)\right] - \lambda\left(1 - e^{ju}\right) H_N^{(2)}(u) \\ +j\sigma\left(1 - e^{-ju}\right)\sum\limits_{k=0}^{N-1}\left[H_k^{(2)'}(u) + j\dfrac{\lambda - N\mu}{\alpha} H_k^{(2)}(u)\right] = 0, \quad if \quad k = N. \end{cases} \quad (10)$$

Let $\alpha = \varepsilon^2$, $u = \varepsilon w$, $H_0^{(2)}(u) = \varepsilon^{2N} F_0(w, \varepsilon)$, $H_1^{(2)}(u) = \varepsilon^{2(N-1)} F_1(w, \varepsilon)$, $H_k^{(2)}(u) = \varepsilon^{2(N-k)} F_k(w, \varepsilon)$, $H_N^{(2)}(u) = F_N(w, \varepsilon)$, where $\varepsilon \to 0$ is infinitesimal.

Taking into account $H_0^{(2)'}(u) = \varepsilon^{2N-1} F_0'(w, \varepsilon)$, $H_1^{(2)'}(u) = \varepsilon^{2N-3} F_1'(w, \varepsilon)$, $H_k^{(2)'}(u) = \varepsilon^{2N-2k-1} F_k'(w, \varepsilon)$, $H_N^{(2)'}(u) = \dfrac{1}{\varepsilon} F_N'(w, \varepsilon)$, we get the system (10) in the form below

$$\begin{cases} j\left(\sigma + \varepsilon^2 j\varepsilon w\right)\left[\varepsilon^{2N-1} F_0'(w, \varepsilon) + j(\lambda - N\mu)\varepsilon^{2N-2} F_0(w, \varepsilon)\right] \\ +\mu\varepsilon^{2(N-1)} F_1(w, \varepsilon) - \lambda\varepsilon^{2N} F_0(w, \varepsilon) = 0, \quad if \quad k = 0, \\ j\left(\sigma + \varepsilon^2 j\varepsilon w\right)\left[\varepsilon^{2(N-k)-1} F_k'(w, \varepsilon) + j(\lambda - N\mu)\varepsilon^{2(N-k)-2} F_k(w, \varepsilon)\right] \\ -j\sigma\left(1 - j\varepsilon w\right)\left[\varepsilon^{2(N-k)+1} F_{k-1}'(w, \varepsilon) + j(\lambda - N\mu)\varepsilon^{2(N-k)} F_{k-1}(w, \varepsilon)\right] \\ -\lambda\varepsilon^{2(N-k)} F_k(w, \varepsilon) - k\mu\varepsilon^{2(N-k)} F_k(w, \varepsilon) + \lambda\varepsilon^{2(N-k+1)} F_{k-1}(w, \varepsilon) \\ +(k+1)\mu\varepsilon^{2(N-k-1)} F_{k+1}(w, \varepsilon) = 0, \quad if \quad k = 1, \ldots, N-1, \\ -\varepsilon^2 w\left[\varepsilon^{2N} F_0'(w, \varepsilon) + \cdots + F_N'(w, \varepsilon)\right] + \lambda j\varepsilon w F_N(w, \varepsilon) \\ -j\varepsilon w(\lambda - N\mu)\left[\varepsilon^{2N} F_0(w, \varepsilon) + \cdots + F_N(w, \varepsilon)\right] \\ -\sigma\varepsilon w\left[\varepsilon^{2N-1} F_0'(w, \varepsilon) + \cdots + \varepsilon F_{N-1}'(w, \varepsilon)\right] \\ -j\sigma\varepsilon w(\lambda - N\mu)\left[\varepsilon^{2N-2} F_0(w, \varepsilon) + \cdots + F_{N-1}(w, \varepsilon)\right] = 0, \ if \quad k = N. \end{cases} \quad (11)$$

Let us divide each equation of the system (11) by the ε to the minimum power and then we can obtain (12) by a limiting process $\varepsilon \to 0$ in (11)

$$\begin{cases} \sigma(\lambda - N\mu)F_0(w) - \mu F_1(w) = 0, \quad if \quad k = 0, \\ \sigma(\lambda - N\mu)F_k(w) - (k+1)\mu F_{k+1}(w) = 0, \quad if \quad k = 1, \ldots, N-1, \\ \sigma(\lambda - N\mu)F_{N-1}(w) - N\mu F_N(w) = 0, \quad if \quad k = N, \end{cases} \tag{12}$$

where $F_k(w) = \lim\limits_{\varepsilon \to 0} F_k(w, \varepsilon)$.

The solving of equations system (11) has the following form

$$F_k(w, \varepsilon) = F_k(w) + j\varepsilon w f_k(w) + o(\varepsilon^2). \tag{13}$$

Using (12), (13) and $F_k'(w, \varepsilon) = F_k'(w) + j\varepsilon f_k(w) + j\varepsilon w f_k'(w)$ we can write (11) as

$$\begin{cases} \sigma F_0'(w) - \sigma(\lambda - N\mu)w f_0(w) + \mu w f_1(w) = 0, \quad if \quad k = 0, \\ \sigma F_k'(w) - \sigma(\lambda - N\mu)w f_k(w) + (k+1)\mu w f_{k+1}(w) = 0, \ if\, k = 1, ..., N-1, \\ \sigma F_{N-1}'(w) - \sigma(\lambda - N\mu)w f_{N-1}(w) + N\mu w f_N(w) \\ +\lambda w F_N(w) + F_N'(w) = 0, \quad if \quad k = N. \end{cases} \tag{14}$$

It is easy to get the solution of the system (14) as $F_N(w) = R_N \exp\{-\lambda w^2/2\}$.

Based on (9) pre-limit characteristic function $h(u)$ has form

$$h(u) = \sum_{k=0}^{N} H_k(u) = R_N \exp\left\{\frac{\lambda - N\mu}{\alpha} ju\right\} \sum_{k=0}^{N} H_k^{(2)}(u) =$$

$$= R_N \exp\left\{\frac{\lambda - N\mu}{\alpha} ju\right\} H_N^{(2)}(u) + o(\varepsilon^2) \approx R_N \exp\left\{\frac{\lambda - N\mu}{\alpha} ju\right\} H_N^{(2)}(u).$$

Turning back to expressions $H_N^{(2)}(u) = F_N(w, \varepsilon) \approx F_N(w) = R_N \exp\{-\lambda w^2/2\}$, we finally obtain the second order asymptotic characteristic function $h^{(2)}(u)$ of the probability distribution of the number of calls in the orbit under the system heavy load and long time patience of calls in the orbit condition

$$h^{(2)}(u) = R_N^2 \exp\left\{\frac{\lambda - N\mu}{\alpha} ju + \frac{\lambda}{\alpha}\frac{(ju)^2}{2}\right\}. \tag{15}$$

The Theorem 1 is proved.

5 Numerical Results

In this section, some numerical examples are presented. It demonstrate the applicability area of the asymptotic results depending on parameters of the Retrial queueing system of M/M/N type with impatient customer in the orbit. Let the system parameters be $\mu = 1$, $N = 2$ and $N = 5$. So, we compare asymptotic and exact distributions for different values of parameters λ and α using the Kolmogorov distance between respective cumulative distribution functions

$$\Delta = \max_{0 \le i < \infty} \left| \sum_{\nu=0}^{i} D_\nu - \sum_{\nu=0}^{i} P_\nu \right|$$

where D_ν and P_ν are an exact and an asymptotic probability distributions respectively.

In Figs. 2 and 3 there are examples of comparison of the asymptotic and the exact distribution densities.

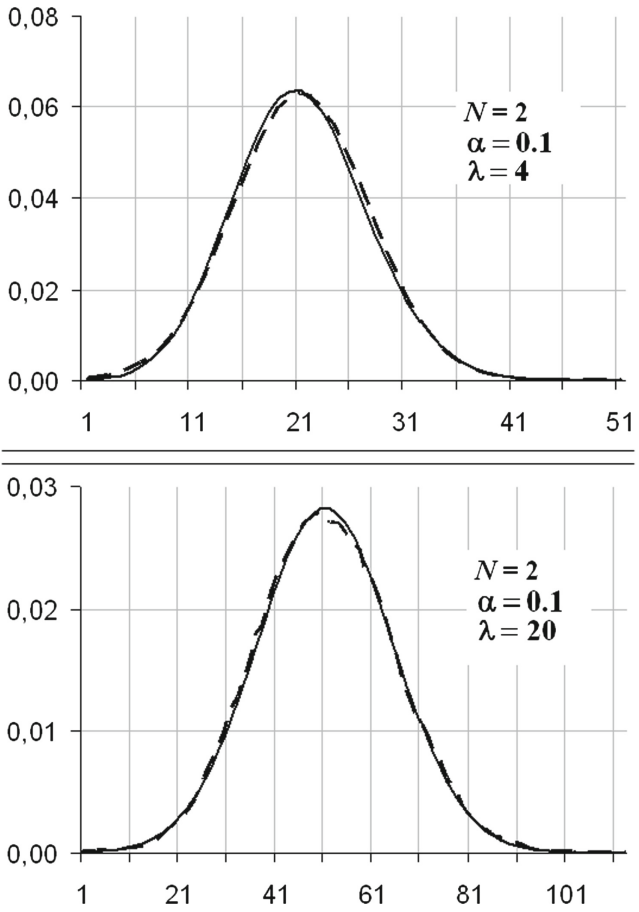


Fig. 2. Comparisons of the asymptotic (dashed line) and the exact (solid line) probability densities

Values of the Kolmogorov distance for these examples are presented in Tables 1 and 2. If we suppose the Kolmogorov distance equal to 0.05 and less as acceptable accuracy of a result, we can find parameters values in which the approximation (15) can be applied. Figures 2 and 3 show that increasing of the parameter λ when parameter α and number of the servers are fixed leads to reduction of the Kolmogorov distances between asymptotic and exact distributions. Figure 4 shows that decreasing of the parameter α when parameter λ and

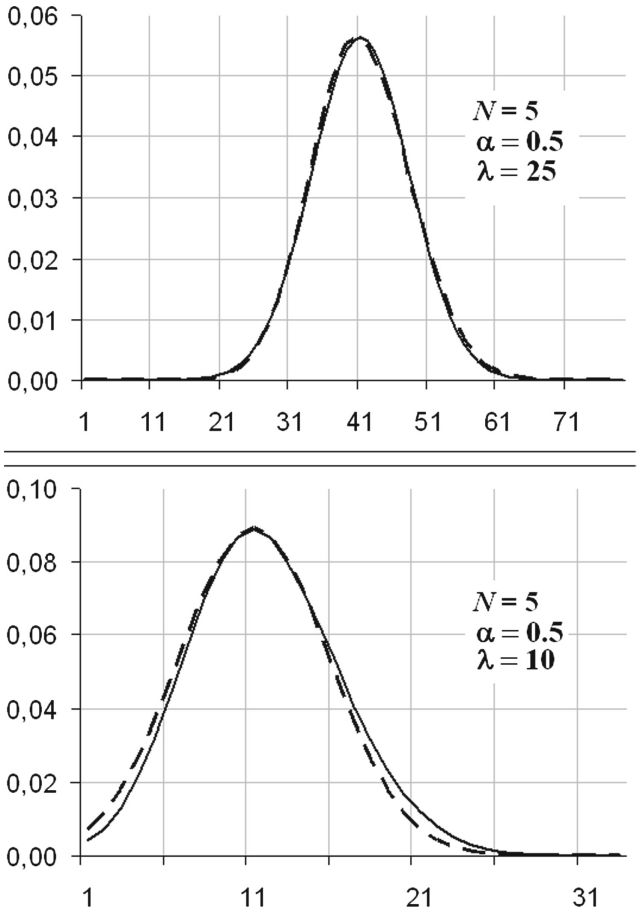


Fig. 3. Comparisons of the asymptotic (dashed line) and the exact (solid line) probability densities

Table 1. Kolmogorov distances between asymptotic and exact distributions if $N = 2$, $\mu = 1$.

	$\alpha = 2$	$\alpha = 1$	$\alpha = 0.5$	$\alpha = 0.1$
$\lambda = 2$	0.204	0.214	0.199	0.126
$\lambda = 4$	0.146	0.088	0.035	0.013
$\lambda = 10$	0.041	0.015	0.011	0.008
$\lambda = 20$	0.018	0.017	0.014	0.007

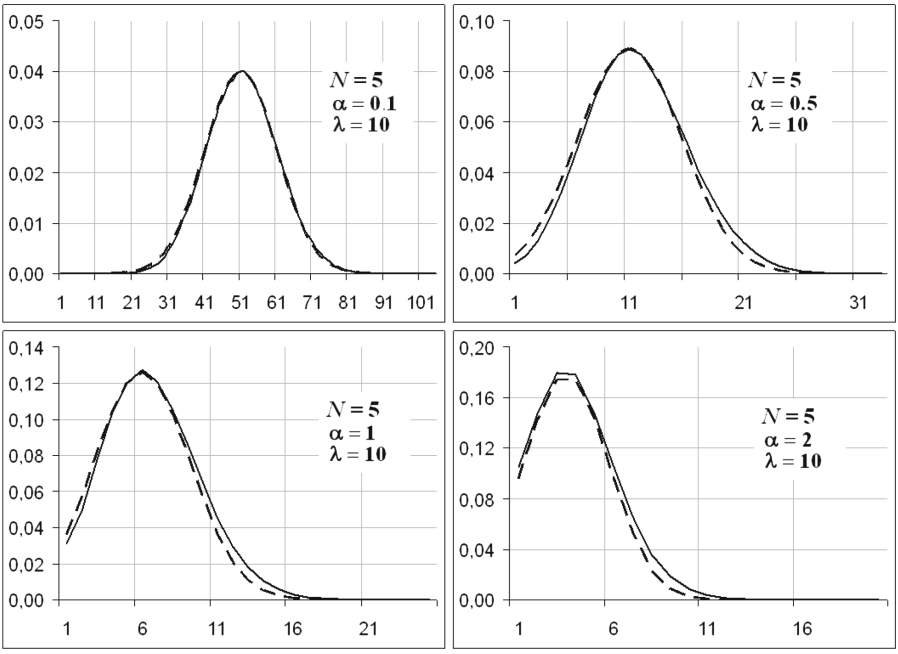


Fig. 4. Comparisons of the asymptotic (dashed line) and the exact (solid line) probability densities

Table 2. Kolmogorov distances between asymptotic and exact distributions if $N = 5$, $\mu = 1$.

	$\alpha = 2$	$\alpha = 1$	$\alpha = 0.5$	$\alpha = 0.1$
$\lambda = 5$	0.095	0.156	0.180	0.262
$\lambda = 10$	0.038	0.043	0.039	0.015
$\lambda = 25$	0.016	0.010	0.007	0.006
$\lambda = 30$	0.012	0.007	0.009	0.005

number of the servers are fixed leads to reduction of the Kolmogorov distances between asymptotic and exact distributions.

6 Conclusion

In the present paper, multiserver retrial queueing system of M/M/N type with impatient customer in the orbit is considered. It is proved that the probability distribution of the calls number in the orbit can be approximated by the Gaussian distribution under the system heavy load and long time patience of calls in the orbit condition with obtained parameters. Numerical results allow to draw a conclusion about an applicability area of the asymptotic result.

References

1. Wilkinson, R.I.: Theories for toll traffic engineering in the USA. Bell Syst. Tech. J. **35**(2), 421–507 (1956)
2. Cohen, J.W.: Basic problems of telephone traffic and the influence of repeated calls. Philips Telecommun. Rev. **18**(2), 49–100 (1957)
3. Gosztony, G.: Repeated call attempts and their effect on traffic engineering. Budavox Telecommun. Rev. **2**, 16–26 (1976)
4. Elldin, A., Lind, G.: Elementary Telephone Traffic Theory. Ericsson Public Telecommunications, Stockholm (1971)
5. Artalejo, J.R., Gomez-Corral, A.: Retrial Queueing Systems. A Computational Approach. Springer, Stockholm (2008). https://doi.org/10.1007/978-3-540-78725-9
6. Falin, G.I., Templeton, J.G.C.: Retrial Queues. Chapman & Hall, London (1997)
7. Artalejo, J.R., Falin, G.I.: Standard and retrial queueing systems: a comparative analysis. Revista Matematica Complutense **15**, 101–129 (2002)
8. Roszik, J., Sztrik, J., Kim, C.: Retrial queues in the performance modelling of cellular mobile networks using MOSEL. Int. J. Simul. **6**, 38–47 (2005)
9. Aguir, S., Karaesmen, F., Askin, O.Z., Chauvet, F.: The impact of retrials on call center performance. OR Spektrum **26**, 353–376 (2004)
10. Nazarov, A., Sztrik, J., Kvach, A.: Comparative analysis of methods of residual and elapsed service time in the study of the closed retrial queuing system M/GI/1//N with collision of the customers and unreliable server. In: Dudin, A., Nazarov, A., Kirpichnikov, A. (eds.) ITMM 2017. CCIS, vol. 800, pp. 97–110. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68069-9_8
11. Dudin, A., Deepak, T.G., Joshua, V.C., Krishnamoorthy, A., Vishnevsky, V.: On a $BMAP/G/1$ retrial system with two types of search of customers from the orbit. In: Dudin, A., Nazarov, A., Kirpichnikov, A. (eds.) ITMM 2017. CCIS, vol. 800, pp. 1–12. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68069-9_1
12. Dudin, A.N., Klimenok, V.I.: Queueing system BMAP/G/1 with repeated calls. Math. Comput. Model. **30**(3–4), 115–128 (1999)
13. Yang, T., Posner, M., Templeton, J.: The M/G/1 retrial queue with non-persistent customers. Queueing Syst. **7**(2), 209–218 (1990)
14. Krishnamoorthy, A., Deepak, T.G., Joshua, V.C.: An M/G/1 retrial queue with non-persistent customers and orbital search. Stochast. Anal. Appl. **23**, 975–997 (2005)
15. Kim, J.: Retrial queueing system with collision and impatience. Commun. Korean Math. Soc. **4**, 647–653 (2010)
16. Martin, M., Artalejo, J.: Analysis of an M/G/1 queue with two types of impatient units. Adv. Appl. Probab. **27**, 647–653 (1995)
17. Kumar, M., Arumuganathan, R.: Performance analysis of single server retrial queue with general retrial time, impatient subscribers, two phases of service and Bernoulli schedule. Tamkang J. Sci. Eng. **13**(2), 135–143 (2010)
18. Fedorova, E., Voytikov, K.: Retrial queue M/G/1 with impatient calls under heavy load condition. In: Dudin, A., Nazarov, A., Kirpichnikov, A. (eds.) ITMM 2017. CCIS, vol. 800, pp. 347–357. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68069-9_28
19. Berczes, T., Sztrik, J., Toth, A., Nazarov, A.: Performance modeling of finite-source retrial queueing systems with collisions and non-reliable server using MOSEL. In:

Dudin, A., Nazarov, A. (eds.) Information Technologies and Mathematical Modelling. Queueing Theory and Applications. ITMM 2017, CCIS, vol. 800, pp. 248–258, Springer, Cham (2017).https://doi.org/10.1007/978-3-319-68069-9
20. Artalejo, J.R., Pozo, M.: Numerical calculation of the stationary distribution of the main multiserver retrial queue. Ann. Oper. Res. **116**, 41–56 (2002)
21. Neuts, M.F., Rao, B.M.: Numerical investigation of a multiserver retrial model. Queueing Syst. **7**(2), 169–189 (1990)
22. Borovkov, A.A.: Asymptotic Methods in Queueing Theory. Wiley, New York (1984)

On a Problem of Base Stations Optimal Placement in Wireless Networks with Linear Topology

Roman Ivanov[1(✉)], Oleg Pershin[1,2], Andrey Larionov[1], and Vladimir Vishnevsky[1]

[1] V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, 65 Profsoyuznaya street, Moscow 117997, Russia
iromcorp@gmail.com, larioandr@gmail.com, vishn@inbox.ru
[2] Gubkin Russian State University of Oil and Gas, Leninsky Pr., 65, Moscow, Russia
pershino@mail.ru

Abstract. A backbone network of road side units (RSUs) is an essential part of modern intelligent transportation systems, where the roadside units are used to collect and distribute information among the mobile users. Connecting each RSU to the data center or Internet is not always feasible due to location and cost constraints. When the RSUs are located close enough to each other, they can be connected via a multihop wireless network where they play a role of the wireless routers. In this paper we consider a partial case when the roadside units are connected in a wireless network with linear topology. Each RSU is equipped with IEEE 802.11 access point that is used by the mobile users to send their data via the network, and with a relay equipment that allows the RSU to connect to the neighbouring RSUs. Each station type is defined by the coverage radius of the access point, connection distance of the relay links and the station price. We also assume that the road has several possible discrete locations where the stations can be deployed. The problem is to find out which stations should be deployed to maximize the overall coverage while providing the given solution cost. First, we formulate the problem in combinatorial form and use this formulation to prove NP-hardness of the problem. Then we define an integer linear program that can be used to find the optimal solution using a well-known software like GLPK or CPLEX.

Keywords: Wireless network · Optimisation problem
Linear programming · Optimal placement · Linear topology

The publication was supported in part by Russian Foundation for Basic Research (RFBR) according to the research project No.18-57-00002.

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 505–513, 2018.
https://doi.org/10.1007/978-3-319-99447-5_43

1 Introduction

Nowadays the functionality of highways deeply relies on an up-to-date data transmission infrastructure for heterogeneous traffic, e.g. voice, video, background traffic. This issue is specifically vital for those countries that have spacious territory. Such infrastructure enables a real-time monitoring for highway parameters via the wireless transmission of data collected from sensors and vehicles, ensuring of proper road safety with the utilization of speed cameras, providing Internet access and the related services.

The infrastructure consits of multiple base stations (roadside unit, RSU) that connects to each other into a network [8,9]. These RSUs can be used either to collect data from sensors, speed cameras or vehicles or to support a vehicular ad-hoc network (VANET). Though RSUs can be connected with optic-fibre or satellite channels, deploying such network is a costly project and it is not always realizable in practice. An inexpensive alternative is to take advantage of wireless connection lines. There is a number of technologies that can be used for this purpose. In this paper we will focus on the scenario assuming the usage of IEEE 802.11-2016 standard [1].

In this paper we focus on the RSUs networks with linear topology. This is an appropriate case for the highway scenario (let us notice that in urban areas it is possible to create more complicated topologies in case of deploying such networks over the street lines). Each RSU connects to exactly two neighbours, the first and the last are connected to the gateways. Each base station provides access to the end users (human or machine). We will refer to the area where end-users are able to connect to the specific station as a coverage area of this station. It is vital to maximize a total coverage area while deploying RSUs in order to increase the VANET connectivity, as well as to minimize transmission delays and the time when the users remains unconnected. But for most cases we are restricted by a total cost of the network deployment (a budget restriction). Thus an optimal placement problem arises out: being given a set of stations specified by the cost, connection range and coverage area, it is required to maximize an overall network coverage and not to exceed a budget granted.

The paper is organized as follows. In Sect. 2 a brief review of existing researches devoted to the problem are given. The optimal placement problem is stated in Sect. 3 in a combinatorial form to help us prove its NP-hardness in Sect. 4. In Sect. 5 a linear program representing the optimal placement problem is defined.

2 Related Work

The numerous papers are devoted to study the problem of deploying RSUs networks. Brahim et. al. describe the problem of station deploying maximizing coverage area while restricted by a total cost [2]. The input data for the problem are the set of places to deploy the stations and preliminarily collected the statistics of mobile and stationary users traffic. Close problem was solved by Cavalcante

et. al. [3]. The authors suggested the model of maximum coverage with delays restrictions and analysed it with a genetic algorithm.

Liu et al. [6] formulated RSU placement problem maximizing the probability of the average connectivity in VANETs as an combinatorial optimisation problem and solved using an expansion and colouration algorithm (ECA). The solution considered road traffic characteristics. The authors presented computation and simulation results in an actual urban area.

In the paper [7], the IEEE 802.11p/WAVE standard considered and an analytical model for the network that takes into account communication delay in a highway scenario was presented. Lee et al. [4] suggested an analyser of the duration of the connection links between vehicles and RSU based on telematic data. The correct choice of RSU placement scheme aiming at maximizing network coverage are presented and the authors illustrate a model effectiveness on an example for the network in South Korea [5].

Xie et al. [11] researched an information dissemination in VANETs. An RSU placement problem in grid road networks without vehicles trajectories knowing are presented. For estimating unknown trajectories the authors presented a probabilistic model. Wu et al. [10] gave a description of the network with single-hop connections if a vehicle is within line-of-sight area and multi-hop connection if vehicle is outside. In the paper a scenario of highway with large number of lanes are considered; for both connection types a placement problem is stated as an integer linear program maximizing a network throughput. In these statements the interference effects, vehicle velocities and vehicle distribution are taken into account. To support an analytical results, NS-2 simulation was presented.

3 The Placement Problem in the Combinatorial Form

The base station placement problem is to put the stations, specified by its maximum radio-relay link distance, a coverage radius and a cost, into determined places in the way to maximize the overall coverage while being restricted by the total cost (a budget limit). So let us give a mathematical formulation of the problem.

Suppose we have a line segment α of length L with the ends in the points a_0 and a_{n+1}. That segment models a long transport route. Inside of the segment $\alpha = [a_0, a_{n+1}]$ a finite set of arranged points $A = \{a_i\}_{i=1}^{n}$, $a_{i+1} > a_i$ is given; that points correspond to the set of vacant places where the stations can be deployed. Each point a_i is defined with its one-dimensional coordinate l_i. Let us also denote a set of station types as $S = \{s_j\}_{j=1}^{m}$. Each station has three attributes described above: a coverage radius r_j, a connection link distance R_j and the station cost c_j. Then we can define a station type $s_j \in S$ as a set of parameters: $s_j = \{r_j, R_j, c_j\}$ (Fig. 1).

There is a special station type s_0 which is a gateway; stations of that type are already placed at the ends a_0 and a_{n+1} of the segment α. For that stations $r_0 = c_0 = 0$ and $R_0 = \max\limits_{1 \le i \le n} R_i$. Defining the gateways in this way we gurantee

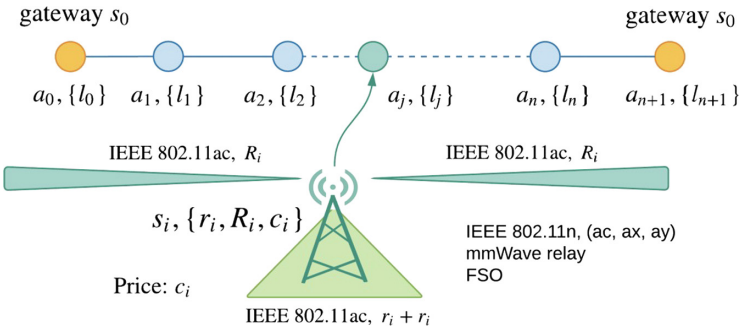


Fig. 1. An optimal station placement problem parameters including places and stations description.

that other stations placed inside the segment will be able to connect to them, and also that these gateways will affect neigher the network coverage, nor cost.

Let us define C as a *budget limit* as a maximum cost of all deployed stations. A *stations placement* is a sequence of pairs increasing by the value of l_i:

$$P = \{(a_i,\ s_j) \mid a_i \in A,\ s_j \in S,\ i \neq 0,\ i \neq n+1\}.$$

Let us denote a set of all available stations as U. Each element $(a_i, s_j) \in P$ corresponds to a station $u \in U$ and all such elements form a set $U_P \subset U$. We will denote a position and type of station u as $a(u) \equiv a_i$ and $s(u) \equiv s_j$ respectively. The placement is called *feasible*, if the following three constraints are fulfilled:

1. **left connectivity**: $\forall\ (a_i, s_j)$, $1 \leq i \leq n$, either $\exists (a_k, s_q) : l_k < l_i$ and $l_i - l_k \leq \min\{R_j, R_q\}$, or $l_i - l_0 \leq R_j$;
2. **right connectivity**: $\forall\ (a_i, s_j)$, $1 \leq i \leq n$, either $\exists (a_t, s_g) :\ l_t > l_i$ and $l_t - l_i \leq \min\{R_j, R_g\}$, or $l_{n+1} - l_i \leq R_j$;
3. **budget limit**: $\sum\limits_{(a_i,s_j)\in P} c_j \leq C$.

The first and the second requirements guarantee that all stations are connected in a chain and this chain ends in the gateways. It is also assumed that a set of available stations U of all types is large enough to implement every feasible placement (e.g. it contains n stations of each type $s_j \in S$). Let us denote the set of all permissible placements as G.

Each placement P can be set into correspondence with an overall coverage value $f(P)$. This function is defined as the length of disjoint union of segments τ, $\tau \subset \alpha$ such that each segment is included in the coverage only if it is contained in the coverage area of some station from placement P, and each point $a \in \alpha$ belongs to a signle segment τ.

Now we can formulate the optimal placement problem as an extremal discrete problem in the combinatorial form:

Problem 1. It is required to find a permissible placement P^*, such that

$$P^* = \underset{P \in G}{\operatorname{argmax}} f(P)$$

4 The Proof of NP-hardness

The problem described is NP-hard. In this section we proof this statement via showing that the recognition problem corresponding to the partial case of the initial placement problem is NP-complete ([8], p. 85).

To formulate the partial case let us introduce additional restrictions on the parameters of the problem 1:

1. **guaranteed connectivity**: each station is able to connect to any other station or a gateway, i.e. the link range of each station surpasses the length of the segment where the stations are placed:

$$\forall s_j : R_j \geq L.$$

2. **absence of intersections**: in any placement the stations coverage areas are not intersecting pairwise:

$$\forall u_i, u_j \in U \text{ and } \forall a_k, a_g \in A : |l_k - l_g| \geq r_i + r_j.$$

3. **vacant places availability**: the number of possible places where stations can be deployed is essentially greater than the number of stations in any feasible placement:

$$\forall P = \{(s_i, a_j)\} \in G \quad |A| > |P|.$$

While these conditions are valid the following statement holds:

Lemma 1. *for any feasible placement $P = \{(s, a)\} \in G$ the value of the overall coverage $F = f(P)$ equals to the sum of the coverages of the stations from $U_P \subset U$, i.e. $f(P) = f(U_P) = \sum_{(a_i, s_j) \in P} r_j$, and doesn't depend on the locations $\{a_i\}$.*

Proof. According to the vacant places availability constraint, all stations from any feasible placement P can be deployed on the set of locations A. According to the guaranteed connectivity constraint, stations from U_P are connected and due to the absence of intersections in the coverage area, the total coverage area $F = f(U_P)$ remains the same for any actual deployment on the vacant places A.

Lemma 1 allows us to define a placement using a subset of stations U_P and don't think about their locations. Now we are going to formulate the extremal problem for the partial case considering 1:

Problem 2. Assume the set of stations $U = \{u\}$ is given and, accordingly, each station is specified by two parameters: the coverage radius $r(u)$ and its cost $c(u)$. Let the budget C is given as well. It is required to find such subset of stations U^* that

$$U^* = \underset{U' \subseteq U}{\operatorname{argmax}} f(U'),$$

where

$$f(U') = \sum_{u \in U'} r(u)$$

with condition

$$\sum_{u \in U'} c(u) \leq C.$$

Then a relevant recognition problem have a form:

Problem 3. The parameters U, C and a number K are given. Does such subset $U' \subset U$ exist, that

$$\sum_{u \in U'} c(u) \leq C,$$

$$\sum_{u \in U'} r(u) \geq K?$$

The problem 3 is exactly a knapsack problem, that is known to be NP-complete. This proves the key statement:

Theorem 1. *The optimal station placement problem 1 is NP-hard.*

5 A Linear Program Model

Now let us define the problem 1 as an integer linear program. Suppose the number of vacant places is n and a number of station types is m. Next we define the following variables:

- x_{ij} for $i = \overline{1..n}, j = \overline{1..m}$, where $x_{ij} = 1$ if the station of type s_j has been deployed in the location a_i, and $x_{ij} = 0$ otherwise;
- $y_i^+ \geq 0$ and $y_i^- \geq 0$ for i=$\overline{1..n}$ and $y_0^+ = y_0^- = y_{n+1}^+ = y_{n+1}^- = 0$. These variables define the value of coverage for the station placed in a_i (due to possible coverage area intersections we cannot simply use parameter r_j determined by the station type s_j as the coverage area will be considered twice at each intersection).

5.1 Stations Placement Without Connectivity Constraint

Let us define a target function to be maximized:

$$f = \sum_{i=1}^{n} (y_i^+ + y_i^-) \to \max \tag{1}$$

Then let us describe the conditions. The coverage value has to be less or equal than the station coverage radius if the station is deployed at the location a_i and equals to 0 if the station is not deployed:

$$y_i^+ \le \sum_{i=1}^{m} x_{ij} r_j, \; i = \overline{1..n} \tag{2a}$$

$$y_i^- \le \sum_{i=1}^{m} x_{ij} r_j, \; i = \overline{1..n} \tag{2b}$$

The total coverage between any two points a_i and a_k has no to be greater the distance between these points:

$$\forall i = \overline{1..n} \quad y_i^+ - y_k^- \le l_k - l_i, \quad k = \overline{i+1..n+1} \tag{3a}$$

$$\forall i = \overline{1..n} \quad y_i^- - y_k^+ \le l_i - l_k, \quad k = \overline{i-1..0}. \tag{3b}$$

The condition given above excludes the case when the coverage is calculated twice due to the coverage areas intersection. The cost function and budget restriction are

$$\sum_{i=1}^{n} \sum_{j=1}^{m} c_j x_{ij} \le C. \tag{4}$$

5.2 Connectivity Constraint

Here we define a connectivity condition that allows us to complete the problem description as an integer linear program. According to problem 1, the station placed in a_i has to be connected to, at least, one station to the left and one station to the right, considering gateway stations of type s_0. The connection link is assumed to be simplex. Let us define the variables e_i, $i = \overline{1,n}$ and set $e_i = 1$ if a station of any type is deployed in a_i, and $e_i = 0$ otherwise; $e_0 = e_{n+1} = 1$ since the gateway stations are already placed at the ends of the segment α.

One station at one place constraint: at each location at most one station can be placed:

$$e_i = \sum_{j=1}^{m} x_{ij}, \quad i = \overline{1..n}. \tag{5}$$

Let us also define the following variables:

- $z_{ijk}, i = \overline{1..n}, j = \overline{1..m}, k = \overline{1..n}, k \neq i$: $z_{ijk} = 1$ if at a point a_i the station of type s_j is deployed and connected to the station deployed in a_k, and $z_{ijk} = 0$ otherwise;
- $z_{ij0} = 1$ if at point a_i the station of type s_j is deployed and connected to the gateway station at a_0, $z_{ij0} = 0$ otherwise;
- $z_{ij,n+1} = 1$ if at point a_i the station of type s_j is deployed and connected to the gateway station at a_{n+1}, $z_{ij,n+1} = 0$ otherwise.

Making use of these variables we describe below a set of conditions derived from the connectivity requirements and radio-relay link symmetry.

Connectivity 1: if the locations a_i and a_j are connected, then stations must be placed at these locations:

$$z_{ijk} \leq e_i \quad \forall i, j, k \tag{6a}$$

$$z_{ijk} \leq e_k \quad \forall i, j, k \tag{6b}$$

Connectivity 2: the station placed in a_i (a_k) has to be connected for at least one station to the right from a_k (a_i), $k > i$ ($k < i$):

$$\sum_{k=i+1}^{n+1} \sum_{j=0}^{m} z_{ijk} \geq e_i, \tag{7a}$$

$$\sum_{k=0}^{i-1} \sum_{j=0}^{m} z_{ijk} \geq e_i, \tag{7b}$$

Link symmetry: if a station u_i has established a connection to a station u_k then the station u_k should also be connected to the station u_i:

$$\sum_{j=0}^{m} z_{ijk} = \sum_{j=0}^{m} z_{kji} \quad \forall i \neq k \tag{8}$$

Connectivity 3: connection radius of a station of type s_j deployed in a_i should be as large as the distance to the location a_k, where the connected station is placed:

$$\forall i \quad z_{ijk} \left(R_j - (a_i - a_k)\right) \geq 0, \quad j = \overline{1..m},\ k = \overline{0..i-1} \tag{9a}$$

$$\forall i \quad z_{ijk} \left(R_j - (a_k - a_i)\right) \geq 0, \quad j = \overline{1..m},\ k = \overline{i+1..n+1} \tag{9b}$$

Let us note that if all coordinates l_i are integer numbers, than the optimal values of variables y_i^+ and y_i^- are also integer and the problem (1)–(9) is an integer linear program.

6 Conclusion

In this paper the optimal base stations placement in wireless networks with linear topology is considered. The problem is shown to be NP-hard and its linear programming model is presented. In this paper we assume radio-relay to be used to connect RSUs to each other and consider the additional condition for this case (**link symmetry**); it also affects conditions **left-** and **right-connectivity** described in the Sect. 3. If IEEE 802.11-based connections are considered the condition **link-symmetry** should be skipped.

References

1. IEEE std 802.11-2016 - IEEE standard for information technology-telecommunications and information exchange between systems local and metropolitan area networks-specific requirements (2016). https://standards.ieee.org/findstds/standard/802.11-2016.html
2. Brahim, M., Drira, W., Filali, F.: Roadside units placement within city scaled area in vehicular ad-hoc networks. In: 3rd International Conference on Connected Vehicles and Expo (ICCYE 2014) (2014)
3. Cavalcante, E., Aquino, A., Pappa, G., Loureiro, A.: Roadside unit deployment for information dissemination in a VANET: an evolutionary approach. In: 14th Annual Conference Companion on Genetic and Evolutionary Computation (aECCO 2012), pp. 27–34 (2012)
4. Lee, J.: Design of a network coverage analyzer for roadside-to-vehicle telematics networks. In: 9th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD 2008) (2008)
5. Lee, J., Kim, C.M.: A roadside unit placement scheme for vehicular telematics networks. In: Kim, T., Adeli, H. (eds.) ACN/AST/ISA/UCMA -2010. LNCS, vol. 6059, pp. 196–202. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13577-4_17
6. Liu, H., Ding, S., Yang, L., Yang, T.: A connectivity-based strategy for roadside units placement in vehicular ad hoc networks. Int. J. Hybrid Inf. Technol. **7**, 91–108 (2014)
7. Reis, A., Sargento, S., Neves, F., Tonguz, O.: Deploying roadside units in sparse vehicular networks: what really works and what does not. IEEE Trans. Veh. Technol. **63**, 2794–2806 (2014)
8. Vishnevsky, V., Portnoy, S., Shakhnovich, I.: WiMAX Encyclopedia. The Way to 4G. Technosphera, Moscow (2009). (in Russian)
9. Vishnevsky, V., Semenova, O.: Polling Systems: Theory and Applications for Broadband Wireless Networks. Academic Publishing, London (2012)
10. Wu, T., Liao, W., Chang, C.: A cost-effective strategy for road-side unit placement in vehicular networks. IEEE Trans. Commun. **60**, 2295–2303 (2012)
11. Xie, B., Xia, G., Chen, Y., Xu, M.: Roadside infrastructure placement for information dissemination in urban ITS based on a probabilistic model. In: Hsu, C.-H., Li, X., Shi, X., Zheng, R. (eds.) NPC 2013. LNCS, vol. 8147, pp. 322–331. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40820-5_27

Analysis of the Possibilities of Using the Means of Tropospheric cm-Wave Radio Communication with a Time Division Duplex in Telecommunication Systems

V. G. Anisimov[1], V. N. Perelomov[1], L. O. Myrova[1], and D. A. Aminev[2(✉)]

[1] JSC "Moscow order of the red banner of labor scientific research institute of radio engineering (MNIRTI)", Bolshoi Trehsvyatitel'skii lane 2/1, Moscow, Russia
astra@mnirti.ru

[2] V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, 65, Profsoyuznaya Street, Moscow 117997, Russia
aminev.d.a@ya.ru

Abstract. The need for development of promising tropospheric stations requires the definition of optimal principles for their design and construction. We consider the features of packet transmission with adaptation of operating frequencies in the construction of promising tropospheric stations with a time division duplex, and study the peculiarities of the proposed technical solution. A variant of the basic tactical and technical requirements for prospective types of tropospheric communication stations of various control units of the RF Armed Forces was proposed. The use of systems with a time division duplex when constructing the tactical control tropospheric stations is justified.

Keywords: Frequency adaptation · Time division duplex
Over-horizon communication · Intersymbol interference
Multipath propogation · Modular design · Modulation
Optimal frequency · Antijam · Frequency division duplex · Power gain

1 Introduction

Analysis of the application of existing tropospheric stations in the special-purpose communication systems reveals their shortcomings, determined by the meeting of the Council of Chief Designers of the RF Armed Forces Communication System for radio relay and tropospheric communication. These shortcomings include:

– low throughput, impossibility of providing the relaying by one station in 512 and 2048 kbit/s modes;

This work has been partially financially supported by the Russian Foundation for Basic Research (grant No.18-57-00002).

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 514–524, 2018.
https://doi.org/10.1007/978-3-319-99447-5_44

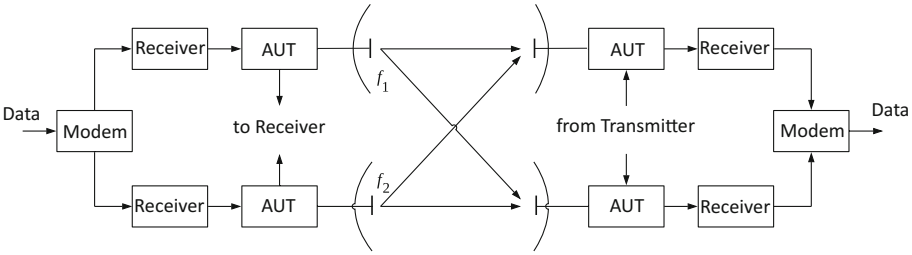


Fig. 1. Structural diagram of a tropospheric line with diverted reception

- insufficient communication range at the speed of 2048 kbit/s, low power potential;
- operation in the range of 4.4 . . . 5.0 GHz with a fixed frequency separation, low jam resistence, a large level of lateral and rear lobes;
- absence of Ethernet 10/100 Base-T, 100 Base-FX;
- large mass-dimensional characteristics and power consumption.

Structural design of tropospheric communication means is determined by the anti-fast fading method [6,7]. In traditional tropospheric communication systems, a diversity reception is used to combat fast fading, which consists in creating several uncorrelated parallel transmission paths with further combining of reception signals. As a rule, spatial diversity is used (two antennas in reception and transmission paths) in combination with frequency diversity [6,7]. The frequencies of the radiated signals f_1 and f_2 are spread by an amount exceeding the radius of frequency correlation. On the receiver's side, space-frequency diversity is realized with an equivalent multiplicity factor n_E equal to four (Fig. 1).

The required reserve for fast (interference) fading (L_{FF}) under tropospheric communication depends on the equivalent diversity multiplicity (n_E) and has the values given in Table 1.

Table 1.

n_E	1	2	3	4	6	8	12
L_{FF}, dB	20,0	10,5	7,0	4,5	2,5	1,7	1,0

To reduce the fade margin to a value of 1 dB or less, the diversity multiplicity factor n_E should be equal to 12–16. Therefore, in addition to the duration of the information package, a signal of the form of a time-frequency matrix with dimension of 3×3 or 4×4, is generated. This allows us to achieve the required magnitude of the equivalent diversity and the minimum margin for fast fading.

The use of two sets of antennas and receivers for spatial separation makes tropospheric communication equipment cumbersome and more expensive.

Thus it seems relevant to justify the methods for design and construction of promising tropospheric stations unified to provide communication using various mechanisms of propagation of radio waves, reduced mass-size characteristics, energy consumption and increased mobility.

2 Principle of Construction of Tropospheric Communication Systems with a Time Division Duplex. Advantages and Disadvantages

The idea of using adaptation to the conditions of distribution on the tropospheric interval exists since 60s of the 20th century. The implementation was hampered by the need to organize a reverse control channel, obsolescence of information on the optimal frequency (OCH), the complexity of the equipment and the possibility of failures in the operation of the radio link with the loss of information of the reverse control channel.

In JSC "MNIRTI", a method of transmission with frequency adaptation without a reverse channel was proposed and implemented using the principle of symmetry of the communication channel with respect to the forward and reverse directions of transmission [1–4].

The absence of a return channel allows to increase the communication reliability [8,9] and to reduce the time of aging of the optimal frequency, since the time interval from the moment of choosing the optimum frequency in the receiver to the moment of its use for transmission is shortened.

In a system with adaptation, the N (independent) uncorrelated frequencies $f_1 \dots f_N$ of the tropospheric channel are periodically probed in the working frequency band. At reception by probing pulses, the optimum frequency (OF) is determined, at which the transmission coefficient of the tropospheric channel was maximum. At the chosen frequency, the transmitter transmits the next packet of information until the next sounding cycle, preceded by its preamble - to transmit data on the nominal value of the selected OF. The structure of the time cycle of a system with frequency adaptation is shown in Fig. 2.

It is known [1] that with probability $p \geq 0.99$, the time correlation interval of fast fading of the signal τ_0 exceeds 40 ms. The loss of noise immunity due to aging of the optimum frequency will be negligible if the duration of the T_P packet is chosen from the T_P condition $T_P/\tau_0 \leq 0.1$ [3]. Thus, the duration of the packet (the permissible frequency obsolescence time) will be equal to $T_P = 0.1 \times \tau_0 = 4$ ms, and the cycle time T_C, taking into account the propagation time of the radio signal over an interval of 300 km, will be 10 ms.

The transmission (reception) packet includes a preamble and an information packet. The preamble provides the correspondent with information on the number of the used frequency for tuning the receiver, and also the probing frequencies for determining the optimum frequency by the correspondent for the next cycle.

The information package includes a sync word, the encrypted user information, interval control channel information (IC), and service channel (SC) pulses.

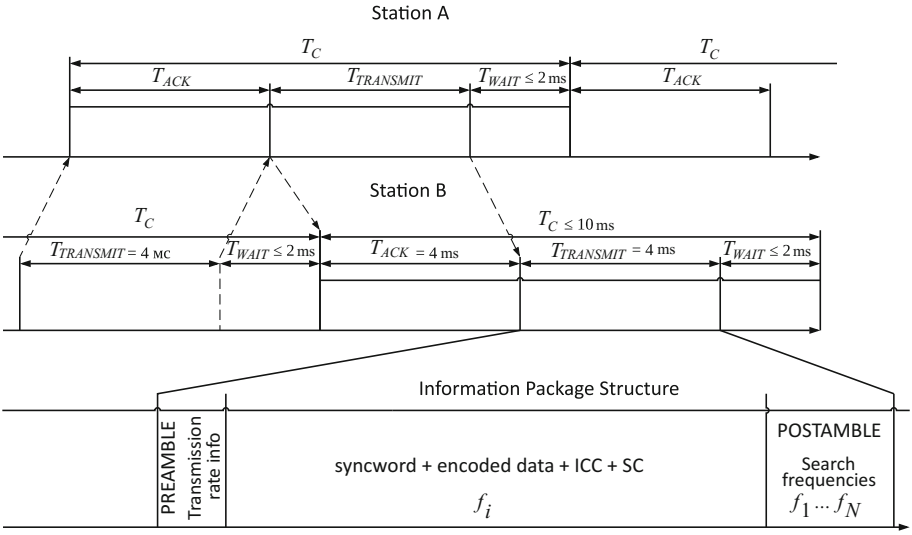


Fig. 2. The structure of the time cycle of a system with frequency adaptation

Let us give the fundamental differences between the time division duplex (TDD) method and the frequency division duplex (FDD) method.

1. Energy efficiency. Transmission at the optimum frequency allows obtaining an energy gain of 3 . . . 5 dB in comparison with the multifrequency signal (MFS) method, depending on the number of uncorrelated frequencies N. The results of comparative analysis of noise immunity for different diversity multiplicities in a system with multifrequency signals and the number of used frequencies N in the system with frequency adaptation (FA) [5] are shown in Fig. 3.

It follows from Fig. 3 that the system with FA exceeds the noise-immunity system with MFS (for a probability of an error of $p_{err} = 10^{-4}$ when $N = 4$, the gain is 2.6 dB, when $N = 8$–3.5 dB, when $N = 16$–4.5 dB).

The choice of the optimum frequency, i.e., the frequency adaptation to the propagation conditions in the tropospheric channel, makes it possible to transmit the signal at the instant of time at a frequency that is the best for propagation conditions. Transmission power is used with maximum efficiency without loss of fading.

The quantitative estimates of the gain were determined repeatedly by the calculation method [2,4] as well as by direct experimental measurements at the tropospheric station “Ladya”. The gain is about 3.5 dB.

Under the FDD, the transmit power is divided into several frequencies, the propagation conditions are different, which leads to loss of signal power at those frequencies at which the signal propagates with fading.

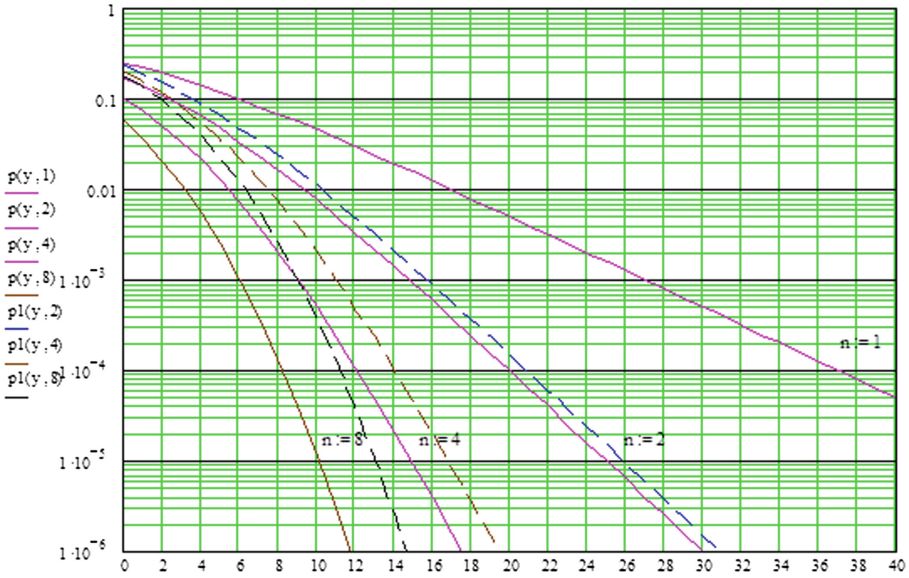


Fig. 3. Noise immunity of the system with FA and MFS

2. Frequency efficiency. Under the FDD for operation, the frequency bands of 150 . . . 200 MHz are used at the bottom and at the top of the operating range 4.4 . . . 5.0 GHz with an unused guard interval (Fig. 4).

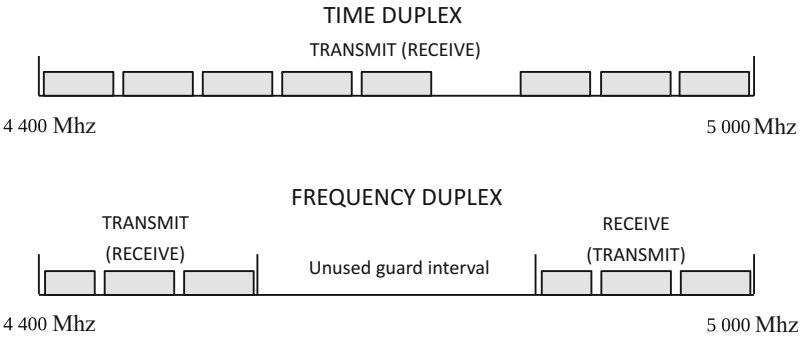


Fig. 4. Use of the frequency band under TDD and FDD

Under TDD and when a spacing between frequencies $\Delta f = 4$ MHz (where fading of the signal at neighboring frequencies is practically independent), using 8 frequencies, the total bandwidth required for operation is 32 MHz $-f_0 \pm 16$ MHz. The TDD provides the possibility of communication in any 35 MHz band located within the frequency range 4.4 . . . 5.0 GHz.

The TDD is the only possible option for those departments that have a relatively narrow frequency band, where there is no possibility of forming the FDD.

3. Design features. In systems with a frequency division duplex, the separation of transmission and reception paths requires bulky and complex waveguide filters (duplexors).

When working with a time division duplex in the common frequency band, the transmitter and the receiver are alternately connected to the antenna, which excludes the use of a duplexer and allows to reduce the mass-dimension parameters of the transceiver and the antenna post.

4. When packetizing information, the transmission rate in a packet is higher in the packetizing factor times (3 times) than the initial information rate, which requires an increase in the transmit power as compared with the case of transmission of the signal without packetization.

However, considering that the transmitter operates in a pulsed mode without emitting a signal in a pause, the average power consumption in the power circuit will practically not change. Therefore, the gain in noise immunity, which gives an adaptive system with the choice of the optimal frequency, will be realized in reducing the overall power consumption of the station.

5. An increase in the transmission rate when packetizing by a factor of 3 leads to a decrease in the sensitivity of the receiving device by $10lg3 = 4.8$ dB [2,4]. The loss in sensitivity is compensated by the reduction in losses in the absent duplexer and the energy gain in accordance with Fig. 3.

3 Justification of Application Fields of Systems with Time Division Duplex

Structuring, energy parameters, mass-dimensional parameters of advanced tropospheric communication facilities are determined by a command unit of the RF Armed Forces. The requirements in the types of communication services of various command units of the Armed Forces of the Russian Federation make it possible to assess the main tactical and technical requirements for prospective types of tropospheric communication stations of various control links listed in Table 2, where the following acronyms are used: TCU—tactical command unit, OTCU—operational-tactical command unit, OCU—operational command unit, OSCU—operational-strategic command unit.

Similar requirements for promising tropospheric stations (TRS) are made by the Federal Guard Service (FGS) and the Federal Security Service (FSS) of the Russian Federation. For the mobile units of the Center for Special Communication and Information of the FSO, the most acceptable variant of a central TRS performance can be a transportable one-pack one with a capacity of 64 kbit/s (for a length of 150 km) and up to 2,048 kbit/s (for a length of 90 km), consisting of 3–4 packages weighing not more than 100 kg for use in the complex hardware and command post vehicles.

Table 2.

Characteristics	Type of tropospheric station		
	Type 1 TCU,OTCU	Type 2 OTCU,OCU	Type 3 OSCU
Range, GHz	4,4 . . . 5,0		
Number of antennas × diameter, m	1 × 1,5	100	500
Number of transmitters × Power, W	1 × 100	2 × 500	2 × 1000
Throughput, Mbps	2/0,5/0,064	8/2/0,5	34/8/2/0,5
Range of communication, km	70/100/150	150/180/210	110/180/210/240
Power consumption, kW	2	8	16
Wheel base	APC	KamAZ	2 KamAZ cars

The FSS assumes the use of light tropospheric stations in a stationary version for communication in the territorial units of the border service of the North-Western and Far Eastern regions.

Peculiarities of providing tropospheric communication in a tactical command unit are the following:

- the need to provide communication from unprepared areas, in conditions of cross-cut terrain, forest vegetation;
- increased intelligence and noise immunity;
- limitations of the location and volume of equipment installed on the armored base and the command post vehicle, taking into account the available means of communication and armament;
- minimum power consumption;
- ease of equipment management;
- automation of pointing antennas on the correspondent and establishing communication;
- ensuring the biological safety of personnel from microwave radiation.

Analysis of the data given in Table 2, as well as the peculiarities of the TDD method, justifies the expediency of its implementation in the development of promising tropospheric stations of a tactical control link.

Providing communication from unprepared areas using TRS of a tactical command unit (TCU), a wide range of required communication ranges, which can be from 40 to 120 km, make use of various mechanisms of distribution - tropospheric, diffraction line of sight. At the same time, the range can be exchanged for throughput.

The time duplex method has advantages in over-the-horizon communication stations, which can work both in the diffraction zone and in the tropospheric zone. For multipath characteristics, these zones are significantly different, since there is only one beam in the diffraction propagation. This allows to operate at a single frequency on such routes, without occupying a redundant band [2,4].

Adaptation of the frequency and power of transmission under the packet transmission method with time duplex increases the intelligence and noise immunity, since the carrier frequency continuously changes randomly in the 32 MHz band with a cycle of 10 ms [2,4].

The weight of the antenna station with an antenna diameter of 1.5 m, a transceiver of 100 W and a pivoting device is 65 kg, and the internal equipment does not exceed 15 kg. When using a range of 10.7 . . . 11.7 GHz for the TCU TRS, the diameter of the antenna can be reduced to 70 . . . 75 cm while maintaining the same energy. Reduction of the antenna and the weight of the antenna post will facilitate the placement of equipment on the transport base.

Power consumption at a transmitter power of 100 W does not exceed 350 W, which does not create a significant additional burden on the standard power supplies.

The control is carried out in a programmatic manner. The operator's software provides visual control of the equipment status, excludes the possibility of incorrect actions by personnel, requires minimal time for mastering.

Thus, the basic operational and technical capabilities required for the TRS of TCU are best implemented when using the packet transfer method with adaptation of operating frequencies to the propagation conditions in the tropospheric channel.

4 The Results of Creation of Tropospheric Communication Means with a Time Division Duplex and Further Areas of Work

The main directions for the development of promising tropospheric communications include:

- development and implementation of a unified modem that provides work in tropospheric and radio relay modes;
- development and application of a line of solid-state transmitters with a power from 100 to 1000 W and active antenna arrays to increase the energy potential and communication range;
- reduction of weight and size characteristics and increase of mobility;
- Support for Ethernet 10/100Base-T, 100 Base-FX

In the means of tropospheric communication, the microwave path, including a transceiver with an antenna, is the most expensive equipment. At the same time, the power amplifier of the transceiver, as a rule, in the solid state version, lacks the hardware reliability.

To reduce costs and improve reliability, the microwave path is expedient to realize on the basis of unified antenna transmit/receive modules (ATRM) with an output power of 50 W. The TDD principle and the absence of a duplexer make it possible to implement the ATRM in the form of small-size planar printed antennas with distributed amplification.

ATRMs will have reliable protection against the effects of climatic and mechanical factors with radio-transparent shelter.

The combination of several ATRMs into the common antenna array (2×2, 3×3, 4×4) makes it possible to produce transmitting devices with various radiated power for promising tropospheric stations, depending on the requirements of specific consumers. This will replace the powerful microwave transceiver with a set of low-power transmitters, with in-phase addition of their power in the air.

To fulfill the requirements of electromagnetic safety of personnel, the variability of deployment of antenna posts on towers, roofs of buildings, own support, a universal operational system has been technically implemented for removing antenna posts 200–400 m from the equipment room.

To improve performance and reduce the number of connecting cables, the transmission of intermediate frequency signals, remote control and signaling, power supply of the antenna post is carried out in a single coaxial microwave cable based on FDD.

The implementation of the aforementioned tasks will allow to:

- provide the possibility of developing promising tropospheric stations of various types (light, medium and heavy) with variable energy provided by changing the number of ATRMs taking into account the requirements of specific consumers;
- unify equipment for various types of promising tropospheric stations, significantly reduce development and serial production costs;
- significantly improve the reliability of microwave paths based on active antenna arrays, since the failure of one of the modules will not lead to communication interruption, but only a decrease in the signal level;
- reduce the terms of repair and reduce the cost of it with the inclusion in spare-voltage modules of ATRM in spare parts;
- reduce operating costs and ensure the possibility of long-term operation of tropospheric stations without maintenance, in a manner similar to radio-relay stations of line of sight;
- ensure import substitution due to the lack of the need for the use of sharply scarce powerful foreign radio components, and to enhance the technological security of the state.

Taking into account the implementation of the proposed solutions, the appearance of an antenna post with a concentrated and a distributed amplification is shown in Fig. 5.

a) with a parabolic antenna and concentrated gain

b) with an active antenna array and distributed gain

Fig. 5. Appearance of the antenna post

5 Conclusions

1. Currently, there is a wide area of applications of tropospheric communication means, both in networks of special and commercial purposes:
   - in special-purpose networks, the advantage of tropospheric means over satellite ones is their higher survivability in armed conflicts and/or counter-terrorist activities;
   - the use of tropospheric stations is also possible when deploying communication links in high northern latitudes, where the use of satellite communication through geostationary satellites is fundamentally impossible;
   - in commercial networks, the use of tropospheric means can in some cases be economically more feasible than the use of satellite communications. Due to the greater length of the intervals, the over-horizon link lines have the advantage over the line-of- sight links when organizing communications in hard-to-reach, mountainous and sparsely populated areas;
2. It seems practically expedient to create an over-the-horizon digital communication station of the centimeter wave band, which allows to transmit information both on the line of sight intervals and the diffraction and long-distance tropospheric propagation intervals. At the same time, the equipment can have admissible dimensions and cost, should not consume a lot of electricity, should not require the installation of antennas on high masts.

3. The method of packet transmission with a time division duplex is promising and should be used as a basis for the development of promising tropospheric TCU stations. Features of the packet transfer method with a TDD make it possible to implement microwave paths (a transceiver with an antenna system) based on active antenna arrays.

On the basis of the ATRM modules, it will be possible to create various high-potential and highly reliable transmit/receive devices in the form of active antenna solutions for TRRS, the radiated power of which can be increased by the acquisition of a different number of modules.

References

1. Matskov, A.A., Serov, V.V., Chernobelsky, L.I.: Prospects for the use of over-horizon communication lines. Telecommunications, No. 8 (2006). (in Russian)
2. Serov, V.V.: Features of propagation of radio waves in over-horizon radio communication systems. Telecommunications, No. 1 (2009). (in Russian)
3. A method for transmitting and receiving information by packets and a device for its implementation. Patent of FSUE MNIRTI No. 2411651 for the invention of 2008141382/09
4. Serov, V.V.: Adaptive transmission system for high-speed signals in a multi-path channel with fading. Telecommunications, No. 5 (2010). (in Russian)
5. Kliot, E.I., Kozlov, D.G.: Investigation of noise immunity of tropospheric radio line with frequency adaptation. Radio Engineering, No. 11 (1994). (in Russian)
6. Kalinin, A.I., Cherenkova, E.L.: Propagation of radio waves and the operation of radio links. Communication (1971). (in Russian)
7. Borodich, S.V. (ed.): Handbook on microwave relay communication. Radio and Communication (1981). (in Russian)
8. Aminev, D., Zhurkov, A., Polesskiy, S., Kulygin, V., Kozyrev, D.: Comparative analysis of reliability prediction models for a distributed radio direction finding telecommunication system. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds.) DCCN 2016. CCIS, vol. 678, pp. 194–209. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-51917-3_18
9. Aminev, D.A., Kozyrev, D.V., Zhurkov, A.P., Romanov, A.Y., Romanova, I.I.: Method of automated control of distributed radio direction finding system. In: 2017 Dynamics of Systems, Mechanisms and Machines (Dynamics), Omsk, Russia, pp. 1–9 (2017). https://doi.org/10.1109/Dynamics.2017.8239426

The Recognition of the Output Function of a Finite Automaton with Random Input

S. Yu. Melnikov(✉) and K. E. Samouylov

Peoples' Friendship University of Russia (RUDN University),
6 Miklukho-Maklaya St, Moscow 117198, Russia
melnikov@linfotech.ru, ksam@sci.pfu.edu.ru

Abstract. The output function recognition of a binary automaton with random Bernoulli input by the character frequency in the output sequence is considered. On the set of output functions, an equivalence relation is introduced, the classes of which consist of functions that are indistinguishable in the scheme. The problem of recognizing the equivalence class of the output function is reduced to the integer optimization problem. Three partial classes of automata close to shift registers are considered.

Keywords: Automaton with random input · Statistical equivalence
Shift register

1 Introduction

The recognition of the type of malfunction of a discrete device is a kind of recognition of a finite state machine. A number of studies solve the problems of discrete devices' testing and diagnosing by using the random sequences as an input and analyzing the output statistics [1,2]. An example of this statistics is the frequency of symbols in output sequence. This paper presents the recognition method for output function of binary machine with random input, which is based on the calculation of relative frequency of symbol occurrence in the output sequence. The method is developed for the following machine types: shift register, generalized shift register, shift register with internal XOR.

2 Finite Moore Machine Probability Function

Let $A = (X, Y, Q, h, f)$ be strongly connected finite Moore machine with $X = Y = \{0, 1\}$ as input and output alphabets, $Q = \{q_1, q_2, ..., q_r\}$ as set of states, $h : Q \times X \to Q$ as transition function, $f : Q \to Y$ as output function. Let $\mathbf{q}^{(0)} = \left(q_1^{(0)}, q_2^{(0)}, ..., q_r^{(0)}\right)$, $\sum\limits_{i=1}^{r} q_i^{(0)} = 1, \quad q_i^{(0)} \geq 0, \ i = 1, 2, ..., r$, - initial probability

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 525–531, 2018.
https://doi.org/10.1007/978-3-319-99447-5_45

distribution, and the input of the automaton is a sequence of independent binary random variables $x^{(i)}$, $i = 1, 2, ...$, with $P\left(x^{(i)} = 1\right) = p$, $P\left(x^{(i)} = 0\right) = 1 - p$, $0 < p < 1$. Under these conditions, the Markov chain with the set of states $Q = \{q_1, q_2, ..., q_r\}$, matrix of transition probabilities $M = (m_{ij})$, $i, j = 1, 2, ..., r$,

$$m_{ij} = \begin{cases} 0, \ h(q_i, 0) \neq q_j, \ h(q_i, 1) \neq q_j, \\ p, \ h(q_i, 0) \neq q_j, \ h(q_i, 1) = q_j, \\ 1 - p, \ h(q_i, 0) = q_j, \ h(q_i, 1) \neq q_j, \\ 1, \ h(q_i, 0) = h(q_i, 1) = q_j \end{cases}$$

and the initial distribution $\mathbf{q}^{(0)}$ is irreducible. The stationary distribution of the probabilities of chain states is given by a single stochastic vector $\boldsymbol{\pi} = (\pi_1, ..., \pi_r)$, $\boldsymbol{\pi} M = \boldsymbol{\pi}$. The probability function of the automaton for the character "1" is $\Phi(p) = \sum_{i=1}^{r} \pi_i f(q_i)$ [2].

3 Recognition of the Output Function of an Automaton by the Value of Its Probability Function

Let $\mathrm{A} = \left\{ A = (X, Y, Q, h, f), \ f \in F_r^{(2)} \right\}$ be the class of the automata described above, which output functions belong to the set $F_r^{(2)}$ of all possible output functions. We consider the problem of recognizing the unknown output function f of an automaton $A \in \mathrm{A}$ by the value of its probability function $\Phi_f(p)$. We define two functions f_1 and f_2 as statistically equivalent if $\Phi_{f_1}(p) = \Phi_{f_2}(p)$ for all $p \in (0, 1)$. The introduced relation splits $F_r^{(2)}$ into disjoint classes. The equivalent output functions are indistinguishable by the value $\Phi_f(p)$, and the output function can be specified only to within an equivalence class.

Theorem 1. *For all $p \in (0, 1)$, except for some finite set $\Omega \subset (0, 1)$, the value $\Phi_f(p)$ corresponds to a single equivalence class of the function f.*

Let $\mathbf{f} = (f(q), \ q \in Q)^{\mathrm{T}}$ denote column-vector of the table assignment of the function f. By construction $\mathbf{f} \in V_r$, where V_r is the space of r-dimensional binary vectors. The probability function is represented as a scalar product $\Phi_f(p) = \boldsymbol{\pi}\mathbf{f}$. Let t denote the dimension of the linear space generated by functions $\pi_i(p)$, $1 \leq t \leq r$, over the field of real numbers. If $\mathbf{d} = (d_1(p), d_2(p), ..., d_t(p))$ is basis of this space, then for some real matrix B of size $t \times r$ the equality $\boldsymbol{\pi} = \mathbf{d}B$ is true.

By choosing an appropriate basis and by re-numbering the set of states, we reduce the matrix B to the form:

$$B = \begin{pmatrix} b_1^1 \, b_2^1 \cdots b_{s_1}^1 & * * \cdots * & \vdots & * * \cdots * \\ 0 \, 0 \cdots 0 & b_1^2 \, b_2^2 \cdots b_{s_2}^2 & \vdots & * * \cdots * \\ \cdots & \cdots & \ddots & \cdots \\ 0 \, 0 \cdots 0 & 0 \, 0 \cdots 0 & \vdots & b_1^t \, b_2^t \cdots b_{s_t}^t \end{pmatrix}.$$

We will assume that the selected basis $\mathbf{d} = (d_1(p), d_2(p), ..., d_t(p))$ corresponds to the form of the matrix B presented above, and the used renumbering corresponds to the partition of the set Q into t blocks: $Q = \bigcup_{i=1}^{t} Q_i$, $|Q_i| = s_i$, $\sum_{i=1}^{t} s_i = r$, $Q_i \cap Q_j = \emptyset$, $i \neq j$, $i, j = 1, 2, ..., r$. Since the statistical equivalence of the output functions f_1 and f_2 is equivalent to the vector equality $B\mathbf{f}_1 = B\mathbf{f}_2$, we can identify the equivalence class of a function f with a vector $B\mathbf{f}$.

Theorem 2. *For $p \in (0,1) \setminus \Omega$ the equation $\mathbf{d}\boldsymbol{\mu} = \Phi_f(p)$ under restrictions on unknowns $\boldsymbol{\mu} = (\mu_1, \mu_2, ...\mu_t)^{\mathrm{T}} \in \{B\mathbf{f},\ \mathbf{f} \in V_r\}$ has a unique solution $\boldsymbol{\mu}_0 = B\mathbf{f}$, corresponding to the equivalence class of the function f.*

4 Output Function Recognition by Symbol Statistics

Let the sequence $x^{(1)}, x^{(2)}, ..., x^{(N)}$ be the machine A input and the output be the sequence $y^{(1)}, y^{(2)}, ..., y^{(N)}$. Let $Y_N = \frac{1}{N} \sum_{k=1}^{N} y^{(k)}$. Let us estimate the length N which is enough to recognize the statistical equivalence class by the statistics Y_N with the reliability level δ, $0 < \delta < 1$. Let us use normal approximation. According to the central limit theorem the random value $\frac{Y_N - N\Phi_f(p)}{\sqrt{Nd_f}}$ converges in distribution to a standard normal distribution. The method of the limiting variance d_f calculation by fundamental matrix is described in [3]. For given $p \notin \Omega$ we denote $\varepsilon_0 = \min_{\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2} |\mathbf{d}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)|$, where the minimum is calculated over not equal $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \{B\mathbf{f},\ \mathbf{f} \in V_r\}$, $D = \max d_f$. According to the equivalence relation condition $\varepsilon_0 > 0$. Taking into account that $P\left\{|Y_N - \Phi_f(p)| < \frac{\varepsilon_0}{2}\right\} \approx 2\Phi\left(\frac{\varepsilon_0}{\sqrt{Nd_f}}\right)$, where Φ is the standard normal distribution function, we will find the length N_0 as the minimum integer N with

$$\frac{\delta}{2} > \Phi\left(\frac{\varepsilon_0}{\sqrt{ND}}\right). \tag{1}$$

Theorem 3. *Let for $N \geq N_0$, $p \notin \Omega$, the vector $\boldsymbol{\mu}_0$ be the solution of the minimization problem*

$$\begin{cases} |Y_N - \mathbf{d}\boldsymbol{\mu}| \to \min \\ \boldsymbol{\mu} \in \{B\boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \in V_r\} \end{cases}. \tag{2}$$

Then with the probability not less than δ the statistical equivalence class of the output function f is $\boldsymbol{\mu}_0$.

5 The Shift Register Case

Let F_n be the set of all Boolean functions of n arguments, $n = 1, 2, ...$ Let $A_f = (X = \{0,1\},\ V_n,\ Y = \{0,1\}, h, f)$ be the Moore machine (n-bit shift register)

with the states V_n, the transition function

$$h\left((a_1,...,a_n),x\right)=(a_2,...,a_n,x),$$

where $x,a_i\in\{0,1\}$, $i=1,2,...,n$, the output function $f(x_1,x_2,...,x_n)$. The machine A_f probability function is polynomial

$$\Phi_f(p)=\sum_{j=0}^{n}s_jp^j(1-p)^{n-j},$$

where $s_k=\sum\limits_{(x_1,x_2,...,x_n):\sum x_i=k}f(x_1,x_2,...,x_n)$, $k=0,1,...,n$.

Let us choose $d_i(p)=p^{i-1}(1-p)^{n-i+1}$, $i=1,2,...,n+1$ as basis functions. The partition of the set Q into classes Q_i corresponds to a partition V_n into classes of vectors of fixed weight. We arrange the vectors V_n in an ascending order of weight, and in the groups of vectors of the constant weight lexicographically. We have

$$B=\begin{pmatrix}1&&&&&&\\&1&1&1&\cdots&&\\&&\cdots&&1&1&1\\&&&&&&1\end{pmatrix}.$$

There are exactly $\binom{n}{i-1}$ ones in the i-th row of the matrix B (only non-zero elements are indicated). The statistical equivalence of functions f and g is equivalent to the relations

$$\sum_{(x_1,x_2,...,x_n):\sum x_i=k}f(x_1,x_2,...,x_n)=\sum_{(x_1,x_2,...,x_n):\sum x_i=k}g(x_1,x_2,...,x_n),k=0,1,...,n.$$

The equivalence class of the function f is given by the vector $(\mu_0,\mu_1,...,\mu_n)$, where $\mu_k=\sum\limits_{(x_1,x_2,...,x_n):\sum x_i=k}f(x_1,x_2,...,x_n)$, $0\le\mu_k\le\binom{n}{k}$, $0\le k\le n$. The set Ω is the union of the roots of the family of nonzero polynomials with integer coefficients $\sum\limits_{i=0}^{n}c_ip^i(1-p)^{n-i}$, where $-\binom{n}{i}\le c_i\le\binom{n}{i}$, $i=0,1,...,n$, $0<p<1$.

Theorem 4. *For $p\in(0,1)\setminus\Omega$ the equation $\Phi_f(p)=\sum\limits_{j=0}^{n}\mu_jp^j(1-p)^{n-j}$ with the constraints $0\le\mu_j\le\binom{n}{j}$, $j=0,...,n$, has the unique integer solution $\left(\mu_0^{(0)},\mu_1^{(0)},...,\mu_n^{(0)}\right)$, which corresponds to equivalence class of function f.*

Theorem 5. *Let N_0 be defined by (1). For $N > N_0$, $p \notin \Omega$, let the vector $\left(\mu_0^{(0)}, \mu_1^{(0)}, ..., \mu_n^{(0)}\right)$ be the solution of the integer minimization problem*

$$\begin{cases} \left| Y_N - \sum\limits_{i=0}^{n} \mu_j p^j (1-p)^{n-j} \right| \to \min \\ 0 \le \mu_j \le \binom{n}{j}, \ \mu_j \in \mathrm{Z} \end{cases}.$$

Then with the probability not less than δ, the statistical equivalence class of the truth output function f corresponds with the vector $\left(\mu_0^{(0)}, \mu_1^{(0)}, ..., \mu_n^{(0)}\right)$.

6 The Generalized Shift Register Case

Generalized shift registers are defined in [4]. Transition graphs of these registers are generalized de Bruijn graphs. A binary generalized register of the order r, $r = 1, 2, ...$, is a Moore automaton $A_f^{(r)} = (X, Y, Q, h, f)$, where the input and output alphabets are $X = Y = \{0, 1\}$, the set of states is $Q = \{0, 1, ..., r-1\}$, the transition function is defined by the rule

$$h(q, \varepsilon) = (2q + \varepsilon) \bmod r,$$

where $q \in Q$, $\varepsilon = 0, 1$, the output function is some mapping $f : Q \to \{0, 1\}$. Let $r = s2^k$, s be odd, $k \ge 0$. Let $b(q)$ be the number of ones in the binary notation of the number $q \bmod 2^k$, $0 \le q \le r-1$.

Theorem 6. *The machine $A_f^{(r)}$ probability function is*

$$\Phi_{A_f^{(r)}}(p) = \frac{1}{s} \sum_{i=0}^{k} \left\| f/_{S_i} \right\| p^i (1-p)^{k-i},$$

where $S_i = \{q \,|\, b(q) = i\}$, $0 \le i \le k$.

The statistical equivalence of output functions f and g is equivalent to the relations

$$\sum_{q \in S_i} f(q) = \sum_{q \in S_i} g(q), i = 0, 1, ..., k.$$

The equivalence class of a function f is given by the vector $(\mu_0, \mu_1, ..., \mu_k)$, where $\mu_i = \sum\limits_{q \in S_i} f(q)$, $0 \le \mu_i \le \binom{k}{i}$, $0 \le i \le k$. The set Ω is the union of the sets of the roots of a family of nonzero polynomials with integer coefficients $\sum\limits_{i=0}^{k} c_i p^i (1-p)^{k-i}$, where $-\binom{k}{i} \le c_i \le \binom{k}{i}$, $i = 0, 1, ..., k$, $0 < p < 1$.

The problem (2) has the form:

$$\begin{cases} \left| Y_N - \sum\limits_{i=0}^{k} \mu_j p^j (1-p)^{k-j} \right| \to \min \\ 0 \le \mu_j \le \binom{k}{j}, \ \mu_j \in \mathrm{Z} \end{cases}.$$

7 The Shift Register with Internal XOR Case

A binary shift register with internal XOR is a Moore automaton $A_f^{\oplus}$ where the input and output alphabets are $X = Y = \{0, 1\}$, the set of states is V_n, $n \geq 1$, the transition function is defined by the rule

$$h\left((a_1, a_2, ..., a_n), a_0\right) = \left(a_0 \oplus a_1, a_1 \oplus a_2, ..., a_{n-1} \oplus a_n\right),$$

where $a_0 \in \{0, 1\}$, $\oplus$ is the summation modulo 2, and the output function is $f(x_1, x_2, ..., x_n) \in F_n$. The implementation of similar registers in the autonomous case is considered in [5].

Theorem 7. *The machine $A_f^{\oplus}$ probability function is*

$$P_{A_f^{\oplus}}(p) = \frac{1}{2^n} \sum_{(x_1, x_2, ..., x_n) \in V_n} f(x_1, x_2, ..., x_n).$$

The statistical equivalence of output functions f and g is equivalent to the relation

$$\sum_{(x_1, x_2, ..., x_n) \in V_n} f(x_1, x_2, ..., x_n) = \sum_{(x_1, x_2, ..., x_n) \in V_n} g(x_1, x_2, ..., x_n).$$

The equivalence class of a function f is given by the scalar

$$\mu_0 = \sum_{(x_1, x_2, ..., x_n) \in V_n} f(x_1, x_2, ..., x_n),$$

where $0 \leq \mu_0 \leq 2^n$. The set Ω is empty. The problem (2) has the form:

$$\begin{cases} \left|Y_N - \frac{\mu}{2^n}\right| \to \min \\ 0 \leq \mu \leq 2^n, \ \mu \in \mathrm{Z} \end{cases}.$$

8 Conclusion

The equivalence relation for output functions of finite binary Moore machine with the random Bernoulli input is studied. The relation takes place when the limits of relative frequency of occurrence of one in output sequences are equal. The problem of output function equivalence class recognition by the value of relative frequency of symbol occurrence in output sequence is formulated as the linear form modulus discrete minimization problem. The problem is formulated as the integer linear form modulus minimization problem with linear constraints for several machine classes.

References

1. Frenkel, S.: Probabilistic model of control-flow altering based malicious attacks. Hardware and Software: Verification and Testing. LNCS, vol. 10629, pp. 249–252. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70389-3_22
2. Barashko, A.S.: Modelirovaniye i testirovaniye diskretnykh ustroystv. In: Skobtsov, Y.A., Speranskiy, D.V., 288 p. Naukova dumka, Kiev (1992). (in Russian)
3. Kemeny, J.G., Snell, J.: Finite Markov Chains, p. 226. Springer, New York (1976)
4. Melnikov, S.Yu.: Statistical properties of generalized binary shift registers. Dokl. Tomsk. Un-ta Sist. Upr. i Radioel. (TUSUR) **20**(1), 93–95 (2017). (in Russian)
5. Dharma Teja, B., Swetha, V., Lokesh, D., Prasad, K.V.R.L.: Design and analysis of a 128 bit linear feedback shift register using VHDL. Int. J. Adv. Res. Sci. Eng. Technol. (IJAET) **3**(2), 1442–1446 (2016)

Issues in the Software Implementation of Stochastic Numerical Runge–Kutta

Migran N. Gevorkyan[1], Anastasiya V. Demidova[1], Anna V. Korolkova[1], and Dmitry S. Kulyabov[1,2(✉)]

[1] Department of Applied Probability and Informatics, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya str., Moscow 117198, Russia
{gevorkyan_mn,demidova_av,korolkova_av,kulyabov_ds}@rudn.university
[2] Laboratory of Information Technologies, Joint Institute for Nuclear Research, 6 Joliot-Curie, Dubna, Moscow region 141980, Russia

Abstract. This paper discusses the application of stochastic Runge-Kutta-like numerical methods with weak and strong convergences for systems of stochastic differential equations in Itô form. At the beginning a brief overview of available publications about stochastic numerical methods and information from the theory of stochastic differential equations are given. Then the difficulties that arise when trying to implement stochastic numerical methods and motivate to use source code generation are described. We discuss some implementation details, such as program languages (Python, Julia) and libraries (Jinja2, Numpy). Also the link to the repository with source code is provided in the article.

Keywords: Stochastic differential equations
Stochastic numerical methods · Automatic code generation
Python language · Julia language · Template engine

1 Introduction

The article [16] describes the Python [30] implementation of stochastic numerical Runge-Kutta type methods. This implementations heavily relies on NumPy and SciPy [18] libraries. We chose Pyhon language because of it's simplicity and development speed. NumPy's capability to work with multidimensional arrays as tensors (functions `tensor_dot` and `einsum`) was also very helpful. However, the performance was low, and not so much because of Python slowness, as because we used the large number of nested loops (up to seven). In this paper, we consider an alternative approach of stochastic numerical methods implementation, based on automatic code generation.

This article is divided into three sections. The first section provides the overview of the main sources and presents information from the theory of stochastic differential equations (SDE) and methods for their numerical solution. The second section presents stochastic numerical schemes for scalar SDE

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 532–546, 2018.
https://doi.org/10.1007/978-3-319-99447-5_46

with strong convergence and for SDE systems with strong and weak convergence. In addition to the general schemes, several coefficient tables are provided. This allows to implement a specific numerical method. Finally, the third section explains the use of code generation for stochastic numerical methods and describes some details of the generator we have implemented (we use Jinja2 [1] template engine).

2 Background Overview

In this section, we give a brief overview of the available publications on stochastic Runge-Kutta methods. We study multistage numerical schemes without partial derivatives from the drift vector $\mathbf{f}(t, \mathbf{X})$ and the diffusion matrix $\mathbf{G}(t, \mathbf{X})$, so we don't consider Milstein methods [25–27]).

First, who used a stochastic Brownian process for mathematical modeling was a French mathematician, a student of Henri Poincare—Bachelier (1870–1946) in 1900 in the work [3].

The book by Kloeden and Platen [19] is classical work on numerical methods for SDE. The book provides a brief introduction to the theory of stochastic Ito and Stratonovich differential equations and their applications. The last two thirds of the book are devoted to the presentation of numerical methods in the sense of strict and weak approximations, including a number of Runge-Kutta methods.

The dissertation by Rößler [32] is a consistent report of stochastic numerical Runge-Kutta-like methods. The author considers the approximation of Ito and Stratonovich SDE systems in a weak sense for the scalar and multidimensional Wiener process. After a brief review of the previous works, the author develops the stochastic equivalent of labelled trees theory (labelled trees are used to derive the order conditions in the case of deterministic Runge-Kutta methods, see, for example, [13,17]).

Rossler considers weakly convergent stochastic Runge-Kutta-like methods for Ito and Stratonovich SDE systems for both the scalar and the multidimensional Wiener process. In the third and the fifth part of the dissertation the specific implementation of the explicit stochastic numerical methods for weak convergence was described.

Further results of Rosler studies were described in articles [14,15] in collaboration with Debrabant. In the preprint [15] authors continued classification of stochastic methods, Runge-Kutta method with a weak convergence. Several concrete realizations and results of numerical experiments were obtained. In the another preprint [31], they gave tables for fourth stage and strong order convergence methods $p = 3.0$.

Euler–Maruyama method described by Maruyama in the paper [23] may be considered as the first stochastic Runge-Kutta-like method. The first systematic study of stochastic numerical Runge–Kutta-like methods of strong order of convergence $p = 1.0$ is given by Rümelin [33] and Platen in his thesis [29].

Great contribution was made by Burrage and Burrage in a series of articles [7–10,12]. In these papers, they not only studied methods of strong order $p = 1.5$, but also extended the theory of labeled trees to the stochastic case.

The article of Soheili and Namjoo [34] presented the three methods with strong convergence $p = 1.0$ and the numerical comparison with the method from the book [19].

Some of the first methods with weak convergence are given in the book [19]. Further development they received in article by Komori and Mitsui [20] and in [22]. In the article [35] two three-stage methods, the weak convergence of the $p = 2.0$, as well as numerical experiments were introduced.

In view of the extreme complexity of further improving the order of accuracy of stochastic numerical schemes, the modern studies are devoted to obtaining numerical schemes for special SDE cases. It is possible to point out some of such studies about stochastic symplectic Runge–Kutta-like methods [11,21,37] and stochastic analogues of the Rosenbrock method [2].

3 Stochastic Wiener Process and Software Generation of Its Trajectories

The stochastic process $W(t), t \geqslant 0$ is called scalar *Wiener process* if the following conditions are true [19,28]:

- $\mathrm{P}\{W(0) = 0\} = 1$, or in other words, $W(0) = 0$ is almost certain;
- $W(t)$ is process with independent increments, i.e. $\{\Delta W_i\}_0^{N-1}$ are independent random variables: $\Delta W_I = W(t_{I+1}) - W(t_I)$ and $0 \leqslant t_0 < t_1 < t_2 < \ldots < t_N \leqslant T$;
- $\Delta W_i = W(t_{I+1}) - W(t_I) \sim \mathcal{N}(0, t_{I+1} - t_I)$ where $0 \leqslant t_{I+1} < t_I < t$, $I = 0, 1, \ldots, N-1$.

The symbol $\Delta W_i \sim \mathcal{N}(0, \Delta t_i)$ denotes that ΔW_i is normally distributed random variable with expected value $\mathbb{E}[\Delta W_i] = \mu = 0$ and variance $\mathbb{D}[\Delta W_i] = \sigma^2 = \Delta t_i$.

The Wiener process is a model of *Brownian motion* (random walk). If we consider the process $W(t)$ in time points $0 = t_0 < t_1 < t_2 < \ldots < t_{N-1} < t_N$ when it experiences random additive changes, then directly from the definition of Wiener process follows:

$$W(t_1) = W(t_0) + \Delta W_0, W(t_2) = W(t_1) + \Delta W_1, \ldots, W(t_N) = W(t_{N-1}) + \Delta W_{N-1},$$

where $\Delta W_i \sim \mathcal{N}(0, \Delta t_i)$, $\forall i = 0, \ldots, N-1$.

The multidimensional Wiener process $\mathbf{W}(t)\colon \Omega \times [t_0, T] \to \mathbb{R}^m$ is defined as a random process composed of jointly independent one-dimensional Wiener processes $W^1(t), \ldots, W^m(t)$. Increments of ΔW_I^α, $\forall \alpha = 1, \ldots, m$ are jointly independent normally distributed random variables. On the other hand, the vector ΔW_I^α can be represented as a multidimensional normally distributed random variable with the expectation vector $\mu = \mathbf{0}$ and the diagonal covariance matrix.

In the case of a multidimensional stochastic process one has to generate m sequences of n normally distributed random variables should be generated.

3.1 Program Generation of Wiener Process

To simulate a one-dimensional Wiener process, it is necessary to generate N normally distributed random numbers $\varepsilon_1, \ldots, \varepsilon_N$ and to construct their cumulative sums. The result will be the simulated *sample path* of the Wiener process $W(t)$ or—using a different terminology—concrete implementation of the Wiener process.

In the case of a multidimensional random process, n sequences of m normally distributed random variables should be generated (that is, n arrays, each of m elements).

We implemented Wiener process generator in `Python` [30] and `Julia` [5]. To generate an array of numbers distributed according to the standard normal distribution in the case of `Python`, we used the function `random.normal` from the `NumPy` [18] library and, in the case of Julia, the built-in `randn` function. Both functions give qualitative pseudorandom sequences, since their work uses generators of uniformly distributed pseudorandom numbers based on an algorithm called Mersenne's vortex [24] (Mersenne Twister), and generators of pseudorandom normally distributed numbers use the Box–Mueller transformation [4,6].

To generate the Wiener process in `Python` one should use the `WienerProcess` class. The following code gives an example of this class usage.

```
import stochastic

N = 100
T = (0.0, 10.0)
W = stochastic.WienerProcess(N=N, time_interval=T)

print("Step size: ", W.dt)
print("Time points: ", W.t)
print("Process iterations: ", W.dx)
print("Wiener Process trajectory: ", W.x)
print("Intervals numbers: ", len(W.dx))
print("Points number: ", len(W.x))
```

The class constructor does not have any required arguments. By default, a process is generated on a time interval [0, 1], which is divided into 1000 parts (`N=1000`). Thus, by default, a path consisting of 1001 points with step `dt` equal to 0.001.

In the case of `Julia`, the Wiener process generator is implemented as the composite data type `struct`

```
"""Stochastic Wiener process"""
struct WienerProcess <: AbstractStochasticProcess
    "Number of process steps"
    N::Int64
    "Time interval starting point"
    t_0::Float64
```

```
    "Time interval end point"
    t_N::Float64
    "Step size"
    dt::Float64
    "Time points"
    T::Vector{Float64}
    "Winer process values"
    X::Vector{Float64}
    "Winer process increments dX ~ N(0, dt)"
    dX::Vector{Float64}
end
```

With following contractors

```
WienerProcess(N::Int64, t_0::Float64, t_N::Float64)
WienerProcess(N::Int64, dt::Float64)
WienerProcess(N::Int64)
WienerProcess()
```

3.2 Calculation and Approximation of Multiple Ito Integrals of Special Form

Here we will not go into the general theory of multiple stochastic Ito integrals, a reader can refer to the book [19] for additional information. Here we consider multiple special integrals, which are included in the stochastic numerical schemes.

In General, for the construction of numerical schemes with order of convergence greater than $p = \frac{1}{2}$, it is necessary to calculate single, double and triple Ito integrals of the following form:

$$I^{\alpha}(t_n, t_{n+1}) = I^{\alpha}(h_n) = \int\limits_{t_n}^{t_{n+1}} \mathrm{d}W^{\alpha}(\tau),$$

$$I^{\alpha\beta}(t_n, t_{n+1}) = I^{\alpha\beta}(h_n) = \int\limits_{t_n}^{t_{n+1}} \int\limits_{t_n}^{\tau_1} \mathrm{d}W^{\alpha}(\tau_2)\mathrm{d}W^{\beta}(\tau_1),$$

$$I^{\alpha\beta\gamma}(t_n, t_{n+1}) = I^{\alpha\beta\gamma}(h_n) = \int\limits_{t_n}^{t_{n+1}} \int\limits_{t_n}^{\tau_1} \int\limits_{t_n}^{\tau_2} \mathrm{d}W^{\alpha}(\tau_3)\mathrm{d}W^{\beta}(\tau_2)\mathrm{d}W^{\gamma}(\tau_1),$$

where $\alpha, \beta, \gamma = 0 \ldots, m$ and $W^{\alpha}, \alpha = 1, \ldots, m$ are components of multidimensional Wiener process. In the case of $\alpha, \beta, \gamma = 0$, the increment of $\mathrm{d}W^0(\tau)$ is assumed to be $\mathrm{d}\tau$.

The problem is to get analytical formulas for these integrals with $\Delta W_n^I = W^I(t_{n+1}) - W^I(t_n)$ in them. Despite its apparent simplicity, this is not achievable

for all possible combinations of indices. Let us consider in the beginning those cases when it is possible to obtain an analytical expression, and then turn to those cases when it is necessary to use an approximating formulas.

In the case of a single integral, the problem is trivial and the analytic expression can be obtained for any index α:

$$I^0(h_n) = \Delta t_n = h_n, \;\; I^\alpha(h_n) = \Delta W_n^\alpha, \; \alpha = 1, \ldots, m.$$

In the case of a double integral $I^{\alpha\beta}(h_n)$, the exact formula takes place only at $\alpha = \beta$:

$$I^{00}(h_n) = \frac{1}{2}\Delta t_n = \frac{1}{2}h_n^2, \;\; I^{\alpha\alpha}(h_n) = \frac{1}{2}\left((\Delta W_n^\alpha)^2 - \Delta t_n\right), \; \alpha = 1, \ldots, m,$$

in other cases, when $\alpha \neq \beta$ Express $I^{\alpha\beta}(h_n)$ by increments of ΔW_n^α and Δt_n in the final form is not possible, so we can only use numerical approximation.

For the mixed case $I^{0\alpha}$ and $I^{\alpha 0}$ in [32], simple formulas of the following form are given:

$$I^{0\alpha}(h_n) = \frac{1}{2}h_n\left(I^\alpha(h_n) - \frac{1}{\sqrt{3}}\zeta^\alpha(h_n)\right),$$
$$I^{\alpha 0}(h_n) = \frac{1}{2}h_n\left(I^\alpha(h_n) + \frac{1}{\sqrt{3}}\zeta^\alpha(h_n)\right),$$

where $\zeta_n^\alpha \sim \mathcal{N}(0, h_n)$ are multidimensional normal distributed random variables.

For the General case $\alpha, \beta = 1, \ldots, m$, the book [19] provides the following formulas for approximating the double Ito integral $I^{\alpha\beta}$:

$$I^{\alpha\beta}(h_n) = \frac{\Delta W_n^\alpha \Delta W_n^\beta - h_n\delta^{\alpha\beta}}{2} + A^{\alpha\beta}(h_n),$$
$$A^{\alpha\beta}(h_n) = \frac{h}{2\pi}\sum_{k=1}^{\infty}\frac{1}{k}\left[V_k^\alpha\left(U_k^\beta + \sqrt{\tfrac{2}{h_n}}\Delta W_n^\beta\right) - V_k^\beta\left(U_k^\alpha + \sqrt{\tfrac{2}{h_n}}\Delta W_n^\alpha\right)\right],$$

where $V_k^\alpha \sim \mathcal{N}(0,1)$, $U_k^\alpha \sim \mathcal{N}(0,1)$, $\alpha = 1, \ldots, m$; $k = 1, \ldots, \infty$; $n = 1, \ldots, N$ is numerical schema number. From the formulas it is seen that in the case $\alpha = \beta$, we get the final expression for the $I^{\alpha\beta}$, which we mentioned above. In the case of $\alpha \neq \beta$, one has to sum the infinite series $a^{\alpha\beta}$. This algorithm gives an approximation error of order $O(h^2/n)$, where n is number of left terms of an infinite series a^{ij}.

In the article [36] a matrix form of approximating formulas is introduced. Let $\mathbf{1}_{m\times m}$, $\mathbf{0}_{m\times m}$ be the unit and zero matrices $m \times m$, then

$$\mathbf{I}(h_n) = \frac{\Delta\mathbf{W}_n\Delta\mathbf{W}_n^T - h_n\mathbf{1}_{m\times m}}{2} + \mathbf{A}(h_n),$$
$$\mathbf{A}(h_n) = \frac{h}{2\pi}\sum_{k=1}^{\infty}\frac{1}{k}\left(\mathbf{V}_k(\mathbf{U}_k + \sqrt{2/h_n}\Delta\mathbf{W}_n)^T - (\mathbf{U}_k + \sqrt{2/h_n}\Delta\mathbf{W}_n)\mathbf{V}_k^T\right),$$

where $\Delta\mathbf{W}_n, \mathbf{V}_k, \mathbf{U}_k$ are independent normally distributed multidimensional random variables:

$$\Delta\mathbf{W}_n = (\Delta W_n^1, \Delta W_n^2, \ldots, \Delta W_n^m)^T \sim \mathcal{N}(\mathbf{0}_{m\times m}, h_n\mathbf{1}_{m\times m}),$$
$$\mathbf{V}_k = (V_k^1, V_k^2, \ldots, V_k^m)^T \sim \mathcal{N}(\mathbf{0}_{m\times m}, \mathbf{1}_{m\times m}),$$
$$\mathbf{U}_k = (U_k^1, U_k^2, \ldots, U_k^m)^T \sim \mathcal{N}(\mathbf{0}_{m\times m}, \mathbf{1}_{m\times m}).$$

If the programming language supports vectored operations with multidimensional arrays, these formulas can provide a benefit to the performance of the program.

Finally, consider a triple integral. In the only numerical scheme in which it occurs, it is necessary to be able to calculate only the case of identical indexes $\alpha = \beta = \gamma$. For this case, [32] gives the following formula:

$$I^{\alpha\alpha\alpha}(h_n) = \frac{1}{6}\left((I^{\alpha}(h_n))^3 - 3I^0(h_n)I^{\alpha}(h_n)\right) = \frac{1}{6}\left((\Delta W_n^{\alpha})^3 - 3h_n\Delta W_n^{\alpha}\right).$$

3.3 Strong and Weak Convergence of the Approximating Function

Before proceeding to the formulation of numerical schemes, it is necessary to determine the criterion of accuracy of approximation of the simulated process $\mathbf{x}(t)$ by the grid function $\mathbf{x}_n$. Two criteria are used: *weak* and *strong* convergence.

The sequence of approximating functions $\{\mathbf{x}_n\}_1^N$ converges with order p to the exact solution $\mathbf{x}(t)$ of SDE in moment T in *strong sense* if constant $C > 0$ exists and $\delta_0 > 0$ such as $\forall h \in (0, \delta_0$ and following condition is fulfilled:

$$(\|\mathbf{x}(T) - \mathbf{x}_N\|) \leqslant Ch^p.$$

The sequence of approximating functions $\{\mathbf{x}_n\}_1^N$ converges with order p to the exact solution $\mathbf{x}(t)$ of SDE in moment T in *weak sense* if constant $C_F > 0$ exists and $\delta_0 > 0$ such as $\forall h \in (0, \delta_0]$ and the following condition is fulfilled:

$$|\mathbb{E}\left[F(\mathbf{x}(T))\right] - \mathbb{E}\left[F(\mathbf{x}_N)\right]| \leqslant C_F h^p.$$

Here $F \in C_{\mathrm{P}}^{2(p+1)}(\mathbb{R}, \mathbb{R}^d)$ is a continuous differentiable functional with polynomial growth.

If the $\mathbf{G}$ matrix is zero, then the strong convergence condition is equivalent to the deterministic case, but the order of strong convergence is not necessarily a natural number and can take fractional-rational values.

It is important to note that the choice of the convergence type depends on the problem one has to solve. Increasing the order of strict convergence leads to more accurate approximation of the trajectories of $\mathbf{x}(t)$. If one wants to calculate, for example, the moment of a random process $\mathbf{x}(t)$ or a generalized functional of the form $\mathbb{E}[F(\mathbf{x}(t))]$, one should increase the order of weak convergence.

4 Stochastic Runge–Kutta-like Numerical Methods

4.1 Euler–Maruyama Numerical Method

The simplest numerical method for solving scalar equations and systems of SDEs is the Euler–Maruyama method, named in honor of Gisiro Maruyama, who extended the classical Euler method for ODEs to the case of equation [23]. The method is easily The each step requires only corresponding to this step increment ΔW_n^{β}. The method has a strong order $(p_d, p_s) = (1.0, 0.5)$. The value p_d denotes the deterministic accuracy order, when the method is used for the equation with $G(t, x^{\alpha}(t)) \equiv 0$. The value p_s denotes the order of the stochastic part approximation.

4.2 Weak Stochastic Runge–Kutta-like Method with Order 1.5 for a Scalar Wiener Process

In the case of a scalar SDE, the drift vector $f^{\alpha}(t, x^{\gamma})$ and the diffusion matrix $G^{\alpha}_{\beta}(t, x^{\gamma})$ become $f(t, x)$ and $g(t, x)$ scalar functions, and the driving Wiener process W^{β}_{t} is the scalar W_t. For scalar SDE it is possible to construct a numerical scheme with strong convergence $p = 1.5$. In the above numerical scheme, the Wiener stochastic process is present in implicit way. It is "hidden" inside the stochastic Ito integrals: $I^{10}(h_n)$, $I^{1}(h_n)$, $I^{11}(h_n)$, $I^{111}(h_n)$.

4.3 Stochastic Runge–Kutta Method with Strong Order $p = 1.0$ for Vector Wiener Process

For SDE system with a multidimensional Wiener process, one can construct a stochastic numerical Runge-Kutta scheme of strong order $p_s = 1.0$ by using single and double Ito integrals [31].

Methods `SRK1Wm` and `SRK2Wm` have strong order $(p_d, p_s) = (1.0, 1.0)$ and $(p_d, p_s) = (2.0, 1.0)$.

4.4 Stochastic Runge–Kutta Method with Weak Order $p = 2.0$ for the Vector Wiener Process

Numerical methods with weak convergence are good for approximation of the distribution characteristics of stochastic process $x^{\alpha}(t)$. The weak numerical method does not need information about the trajectory of driving Wiener process W^{α}_{n} and random increments for these methods can be generated on another probability space. From the paper [15] we get two Butcher tables.

5 Analysis of Implementation Difficulties of Stochastic Runge–Kutta Numerical Methods

As can be seen from the formulas, stochastic Runge-Kutta methods are much more complicated than their classical analogues. In addition to the cumbersome formulas, we can highlight the following factors that complicate the implementation of stochastic methods in software, as well as their application to the numerical solution of SDEs.

- When choosing a particular method, it is necessary to consider what type of convergence is necessary to provide for this problem, as well as which of the stochastic equations should be solved—in Ito or Stratonovich form. This increases the number of algorithms one has to implement.
- For methods with strong convergence greater than one at each step it is necessary to solve the resource-intensive problem of stochastic integrals approximation.

– In the numerical scheme, there are not only matrices and vectors, but also tensors (four-dimensional arrays), it is necessary to perform a convolution operation on several indexes. The implementation of convolution via summation by using normal cycles results in a significant performance drop.
– Weak methods requires the Monte Carlo simulation and, therefore, a large number of repeated computations for the numerical solution. Since the Monte Carlo method converges approximately as $1/\sqrt{N}$, where N—number of calculations, to achieve an accuracy of at least 10^{-3}, it is necessary to perform minimum 10^6 tests.

The most significant performance drop occurs when implementing a universal algorithm, that is, a program that can make a calculation using an arbitrary coefficient table. In this case, we have to use a large number of nested loops in order to organize the summation. The presence of double sums in the schemes as well as complex combination of indices in the multipliers under the sign of these sums complicates the implementation even more and the number of nested cycles increases to six. In addition to these specific features, we mention a few reasons for the performance drop, which also take place in case of deterministic numerical methods. The obvious way to store the coefficients of the methods is to use arrays. However, in explicit methods, that we consider, the matrix is lower-diagonal and storing it as a two-dimensional array results in more than half of the allocated memory being spent on storing zeros.

While examine the source codes of popular routines that implement classical explicit embedded Runge–Kutta methods, one may find that these programs use a set of named constants rather than arrays to store the coefficients of the method. It is also caused by the fact that the operations with scalar variables in most programming languages are faster than operations on arrays.

We wish to preserve the requirement of code flexibility and at the same time to increase the speed of calculations and reduce the memory consumption. That led us to automatic code generation from one template.

In addition to performance gains, automatic code generation allows you to add or modify all functions at once by editing only one template. This allows both to reduce the number of errors and to generate different variants of functions for different purposes.

6 Automatic Code Generation

For code generation we use Python 3 language. The program is open source and available on bitbucket repository by URL bitbucket.org/mngev/sde_num_generation. The repository contains module `stochastic`. This module implements Wiener stochastic process and the numerical methods we considered in this paper. Most part of the module's code are generated by scripts from `generator` directory.

For the code generation, we used Jinja2 [1] template engine. This library was originally developed to generate HTML pages, but it has a very flexible syntax and can be used as a universal tool for generating text files of any kind, including

source codes in any programming languages. In addition to Jinja2, we also used NumPy library to work with arrays and speed-up some calculations.

In addition to the two external libraries listed above, the standard `fraction` module was used. It allows to specify the coefficients of the method as rational fractions, and then convert them to float type with the desired order of accuracy. Also we use `typing` module to annotate the types of function arguments (Python 3.5 and above feature).

Templates are files with Python source code with insertions of Jinja2 specific commands. Information about the coefficients of the methods is stored separately, in a structured form of JSON format. This makes it easy to add new methods and modify old ones by editing JSON files. Currently we use methods with coefficients presented in [14,15,19].

Python itself is used as the language for already generated functions with the active use of NumPy library, which allows to get acceptable performance. However, the generated code can be easily reformatted to match the syntax of any other programming language. We plan to modify the program to generate code in Julia language (julialang.org). This language was introduced in 2012 and initially focused on scientific computing. Currently, he is intensively developing and gaining popularity. To date, the current version is 0.6.2. Julia provides performance comparable to C++ and Fortran, but it is a dynamic language with interactive command line (REPL) capability similar to IPython and can be integrated into an interactive Jupyter environment.

The current version of the library exceeds the one described by the authors in [16]. The use of auto-generation made it possible not to use nested loops, which reduced the number of memory allocations, and greatly simplified the code.

6.1 Realisation of Automatic Code Generation

To study the calculation errors and the efficiency of different stochastic numerical methods, it is necessary to have a universal implementation of such methods. The universality means the possibility to use any stochastic method with a desired strong or weak error by setting its coefficient table. With direct transfer of mathematical formulas to the program code, one need to use about five nested cycles, which extremely reduces performance, since such code does not take into account a large number of zeros in the coefficient tables and arithmetic operations on zero components are still performed, although this is an extra waste of processor time.

One way to achieve versatility and acceptable performance is to generate code for a numerical method step. This approach minimizes the number of arithmetic operations and saves memory, since the zero coefficients of the method do not have to be stored.

We implemented a code generator for the three stochastic numerical methods mentioned above:

– scalar method with strong convergence $p_s = 1.5$,

– vector method with strong convergence $p_s = 1.0$,
– vector method with weak convergence of $p_s = 2.0$.

We use Python to implement the code generator and Jinja2 [1] template engine. This template engine was originally created to generate HTML code, but its syntax is universal and allows you to generate text of any kind without reference to any programming or markup language.

Information about the coefficients of each particular method is stored as a JSON file of the following structure:

```
{
  "name": "method's name (the future name of the function)",
  "description": "method's short description",
  "stage": 4,
  "det_order": "2.0",
  "stoch_order": "1.5",
  "A0": [...],
  "B0": [...],
  "A1": [...],
  "B1": [
    ["0", "0", "0", "0"],
    ["1/2", "0", "0", "0"],
    ["-1", "0", "0", "0"],
    ["-5", "3", "1/2", "0"]
  ],
  "c0": ["0", "3/4", "0", "0"],
  "c1": ["0", "1/4", "1", "1/4"],
  "a": ["1/3", "2/3", "0", "0"],
  "b1": ["-1", "4/3", "2/3", "0"],
  "b2": ["-1", "4/3", "-1/3", "0"],
  "b3": ["2", "-4/3", "-2/3", "0"],
  "b4": ["-2", "5/3", "-2/3", "1"]
}
```

The parameter `stage` is the number of method's stages, `det_order` is the error order of the deterministic part (p_d), `stoch_order` is the error order of the stochastic part (p_s), `name` is the name of the method, which will then be used to create the name of the generated function, so it should be written in one word without spaces. All other parameters are the coefficients of the method. In this case, we give the coefficients of the scalar method with strong convergence $p_s = 1.5$, omitting the coefficients $\mathbf{a}_0$, $\mathbf{a}_1$ and $\mathbf{B}_0$ to save text space. It is necessary to note that the values of the coefficients can be specified in the form of rational fractions, for which they should be presented as JSON strings and enclosed in double quotes.

For internal representation of stochastic numerical methods we created three Python classes: `ScalarMethod`, `StrongVectorMethod` and `WeakVectorMethod`. The implementation of these classes is contained in the file `coefficients_`

`table.py`. The constructors of these classes read the JSON file and, based on them, create objects, which can later be used for code generation. The Fraction class from the Python standard library is used to represent rational coefficients. Each class has a method that generates a coefficient table in LaTeX format.

The file `stoch_rk_generator.py` is a script which handles the `jinja2` templates and, based on them, generates a code of python functions. For vector stochastic methods, a code is generated for dimensions up to 6. Functions are named based on the information specified in `JSON` files, such as `strong_srk1w2`, `strong_srk2w5`, `weak_srk2w6`, and so on.

In addition to the code in `Python`, LaTeX formulas are generated. It allows one to check the correctness of the generator. For example, we give below the formula generated automatically based on the data from `JSON` file for Runge–Kutta method `strong_srk1w2` with stages $s = 3$, and 2 dimensioned Wiener process. Nonzero coefficients of the method are as follows:

$$A^2_{01} = 1,\ A^2_{11} = 1,\ A^3_{11} = 1,\ B^2_{11} = 1,\ B^3_{11} = -1,$$
$$a_1 = 1/2,\ a_2 = 1/2,\ c^2_0 = 1,\ c^2_1 = 1,\ c^3_1 = 1,\ b^1_1 = 1,\ b^2_2 = 1/2,\ b^2_3 = -1/2.$$

7 Parallel SDE Integration with Weak Numerical Methods

Stochastic numerical methods with strong convergence are well suited for computing a specific trajectory of SDE solution. If we are not interested in a specific trajectory, but in some probabilistic characteristics (distribution of a random process, mathematical expectation, variance, etc.), then we should use numerical methods with weak convergence.

In the case of numerical methods with weak convergence, we have to use Monte Carlo method. It means that we should solve our SDE system multiple times and each time with different trajectory. The error of the Monte Carlo method depends on the number of trials N as $\sqrt{N}$, so to achieve the accuracy of 10^{-3} we need 10^6 trials. However, since the trajectories of the Wiener process are independent, the SDE for each specific trajectory can be solved independently in parallel mode.

We have implemented a script in Python, which allows to find solutions of SDE for N different trajectories in parallel mode by spawning a given number of processes. For processes spawning we use `multiprocessing` module. The following features of the Cpython interpreter should be noted.

- Because of the global interpreter lock (GIL), it is not possible to use threads for the Monte Carlo method. The standard `threading` module is only suitable for asynchronous tasks.
- When using processes, you should reinitialize the random number generator with new seed for each process separately, because otherwise all generated processes will generate the same sequence of random numbers.

The source code of the implemented script is located in the tests directory. It is based on two functions.

– Function `calculation` performs the necessary calculations for a given number of trajectories. As arguments, the function takes the drift vector, the diffusion matrix, the required number of simulations, the initializing value for the random generator, the initial value of the SDU solution, the number of steps of the Wiener process, the time interval at which it is necessary to carry out integration, the dimension of the Wiener process and optionally the function for testing the obtained solution for adequacy.
– Function `run_parallel` distributes the Monte Carlo tests equally between processes, creates a pool of processes, and runs them. Each process performs the function `calculation`.

When carrying out a large number of tests, the storage of all the resulting trajectories requires a significant amount of RAM. Therefore, it is more reasonable to immediately decide what probabilistic characteristics we need and calculate them using on-line algorithms. For example, to calculate the average trajectory, we use the following formula

$$\bar{\mathbf{x}}_n = \bar{\mathbf{x}}_{n-1} + \frac{\mathbf{x}_n - \bar{\mathbf{x}}_{n-1}}{n}.$$

This formula allows you to update the mean values of all path steps $\bar{\mathbf{x}}_n$ based on the previous mean values $\bar{\mathbf{x}}_{n-1}$ and the current value $\mathbf{x}_n$. As a result, each process must store only one array of constant length, which saves memory.

8 Conclusion

Stochastic numerical schemes with convergence order higher than 0.5 are considered. It is shown that such methods are much more complicated than equivalent numerical methods for systems of ordinary differential equations. Their specifics makes efficient software implementation of such methods not a trivial task. We discuss an approach based on automatic generation of code, which allows to obtain an efficient implementation of the methods and gives the possibility to use any table of coefficients. We also give a short description of our program and provide the url link to the repository with the source code.

Acknowledgments. The work is partially supported by Russian Foundation for Basic Research (RFBR) grants No 16-07-00556. Also the publication was prepared with the support of the “RUDN University Program 5-100”.

References

1. Jinja2 official site. http://jinja.pocoo.org
2. Amiri, S., Hosseini, S.M.: Stochastic Runge-Kutta rosenbrock type methods for SDE systems. Appl. Numer. Math. **115**, 1–15 (2017). https://doi.org/10.1016/j.apnum.2016.11.010
3. Bachelier, L.: Théorie de la spéculation. Ann. Sci. l’École Norm. Supér. **3**(17), 21–86 (1900)

4. Bell, J.R.: Algorithm 334: normal random deviates. Commun. ACM **11**(7), 498 (1968). https://doi.org/10.1016/j.apnum.2016.11.010
5. Bezanson, J., Edelman, A., Karpinski, S., Shah, V.B.: Julia: a fresh approach to numerical computing. SIAM Rev. **59**, 65–98 (2017)
6. Box, G.E.P., Muller, M.E.: A note on the generation of random normal deviates. Ann. Math. Stat. **29**(2), 610–611 (1958). https://doi.org/10.1214/aoms/1177706645
7. Burrage, K., Burrage, P.M.: High strong order explicit Runge-Kutta methods for stochastic ordinary differential equations. Appl. Numer. Math. **22**, 81–101 (1996)
8. Burrage, K., Burrage, P.M.: General order conditions for stochastic Runge-Kutta methods for both commuting and non-commuting stochastic ordinary differential equation systems. Appl. Numer. Math. **28**, 161–177 (1998)
9. Burrage, K., Burrage, P.M.: Order conditions of stochastic Runge-Kutta methods by B-series. SIAM J. Numer. Anal. **38**, 1626–1646 (2000)
10. Burrage, K., Burrage, P.M., Belward, J.A.: A bound on the maximum strong order of stochastic Runge-Kutta methods for stochastic ordinary differential equations. BIT **37**, 771–780 (1997)
11. Burrage, K., Burrage, P.M.: Low rank Runge-kutta methods, symplecticity and stochastic hamiltonian problems with additive noise. J. Computat. Appl. Math. **236**(16), 3920–3930 (2012). https://doi.org/10.1016/j.cam.2012.03.007
12. Burrage, P.M.: Runge-Kutta methods for stochastic differential equations. Ph.D. thesis, University of Qeensland, Australia (1999)
13. Butcher, J.: Numerical Methods for Ordinary Differential Equations, 2nd edn. Wiley, New Zealand (2003)
14. Debrabant, K., Rößler, A.: Continuous weak approximation for stochastic differential equations. J. Comput. Appl. Math. **214**, 259–273 (2008)
15. Debrabant, K., Rößler, A.: Classification of stochastic Runge-Kutta methods for the weak approximation of stochastic differential equations, March 2013. arXiv:1303.4510v1
16. Gevorkyan, M.N., Velieva, T.R., Korolkova, A.V., Kulyabov, D.S., Sevastyanov, L.A.: Stochastic Runge–Kutta software package for stochastic differential equations. In: Zamojski, W., Mazurkiewicz, J., Sugier, J., Walkowiak, T., Kacprzyk, J. (eds.) Dependability Engineering and Complex Systems. AISC, vol. 470, pp. 169–179. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39639-2_15
17. Hairer, E., Nørsett, S.P., Wanner, G.: Solving Ordinary Differential Equations I, 2nd edn. Springer, Berlin (2008). https://doi.org/10.1007/978-3-662-12607-3
18. Jones, E., Oliphant, T., Peterson, P., et al.: SciPy: open source scientific tools for Python (2001). http://www.scipy.org/. Accessed 08 10 2017
19. Kloeden, P.E., Platen, E.: Numerical Solution of Stochastic Differential Equations, 2nd edn. Springer, Heidelberg (1995). https://doi.org/10.1007/978-3-662-12616-5
20. Komori, Y., Mitsuri, T.: Stable ROW-type weak scheme for stochastic differential equations. RIMS Kokyuroku (932), 29–45 (1995)
21. Ma, Q., Ding, X.: Stochastic symplectic partitioned Runge-Kutta methods for stochastic hamiltonian systems with multiplicative noise. Appl. Math. Comput. **252**, 520–534 (2015). https://doi.org/10.1016/j.amc.2014.12.045
22. Mackevičius, V.: Second-order weak approximations for stratonovich stochastic differential equations. Lith. Math. J. **34**(2), 183–200 (1994). https://doi.org/10.1007/BF02333416
23. Maruyama, G.: Continuous Markov processes and stochastic equations. Rend. Circ. Mat. **4**, 48–90 (1955)

24. Matsumoto, M., Nishimura, T.: Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Trans. Model. Comput. Simul. **8**(1), 3–30 (1998). https://doi.org/10.1145/272991.272995
25. Milstein, G.N.: Approximate integration of stochastic differential equations. Theory Probab. Appl. **19**, 557–562 (1974)
26. Milstein, G.N.: A method of second-order accuracy integration of stochastic differential equations. Theory Probab. Appl. **23**, 396–401 (1979)
27. Milstein, G.N.: Weak approximation of solutions of systems of stochastic differential equations. Theory Probab. Appl. **30**, 750–766 (1986)
28. Øksendal, B.: Stochastic Differential Equations. An Introduction with Applications, 6th edn. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-642-14394-6
29. Platen, E.: Beiträge zur zeitdiskreten Approximation von Itoprozessen. Ph.D. thesis, Akad. der Wiss., Berlin (1984)
30. Rossum, G.: Python reference manual. Technical report, Amsterdam, The Netherlands (1995). https://docs.python.org/3/
31. Rößler, A.: Strong and weak approximation methods for stochastic differential equations—some recent developments. In: Devroye, L., Karasözen, B., Kohler, M., Korn, R. (eds.) Recent Developments in Applied Probability and Statistics, pp. 127–153. Physica-Verlag HD, Heidelberg (2010). https://doi.org/10.1007/978-3-7908-2598-5_6
32. Rößler, A.: Runge-Kutta methods for the numerical solution of stochastic differential equations. Ph.D. thesis, Technischen Universität Darmstadt, Darmstadt, februar 2003
33. Rümelin, W.: Numerical treatment of stochastic differential equations. SIAM J. Numer. Anal. **19**(3), 604–613 (1982)
34. Soheili, A.R., Namjoo, M.: Strong approximation of stochastic differential equations with Runge-Kutta methods. World J. Model. Simul. **4**(2), 83–93 (2008)
35. Tocino, A., Ardanuy, R.: Runge-Kutta methods for numerical solution of stochastic differential equations. J. Comput. Appl. Math. **138**, 219–241 (2002)
36. Wiktorsson, M.: Joint characteristic function and simultaneous simulation of iterated Itô integrals for multiple independent Brownian motions. Ann. Appl. Probab. **11**(2), 470–487 (2001)
37. Zhou, W., Zhang, J., Hong, J., Song, S.: Stochastic symplectic Runge-Kutta methods for the strong approximation of hamiltonian systems with additive noise. J. Comput. Appl. Math. **325**, 134–148 (2017). https://doi.org/10.1016/j.cam.2017.04.050

Automatic Recognition of a Weakly Identified Animal Activity State Based on Data Transformation of 3D Acceleration Sensor

Valentin Sturm[1], Julia Mayer[1], Dmitry Efrosinin[2,3(✉)], Leonie Roland[4], Michael Iwersen[4], Marc Drillich[4], and Wolfgang Auer[5]

[1] Linz Center of Mechatronics Gmbh, Altenbergerstraße 69, 4040 Linz, Austria
{valentin.sturm,julia.mayer}@lcm.at
[2] Johannes Kepler University, Altenbergerstraße, 69, 4040 Linz, Austria
dmitry.efrosinin@jku.at
[3] Institute of Control Sciences, RAS, Profsoyuznaya st., 65, 117997 Moscow, Russia
[4] University of Veterinary Medicine Vienna, Veterinaerplatz, 1, 1210 Vienna, Austria
[5] Smartbow GmbH, Jutogasse, 3, Weibern 4675, Austria
https://www.lcm.at, http://www.jku.at, http://www.ipu.ru

Abstract. Smartbow ear-attached motion active sensor with a 3d accelerometer is used for animal activity tracking. Such technology is required to understand the welfare, nutrition scheme and management strategies for breeding cattle. The ear-tag with integrated sensor has no fixed location and orientation that leads to necessity to use the orientation independent features by solving a time series classification problem. In this paper we propose an accelerometer data transformation techniques based on Euler angle rotation and signal projection and show their equivalence relative to a reference coordinate system. The main aim is to increase a recognition accuracy for the weakly-identified states or actions. The previous research for the fitting of the calves has demonstrated certain difficulties by recognition of some rare states and actions, e.g. milk intake. The results show that an average area under the ROC-curve of 0.740 is achieved with improvement of 0.252 over classifications without data transformation.

Keywords: Activity recognition · Accelerometer
Data transformation · Machine learning

1 Introduction

The wireless data transmission has become an integral part of our daily activities. It has found widespread use in different applications particularly in radio finding systems [1], location tracking systems [11] and activity recognition [2].

V. Sturm—This work has been supported by the "LCM – K2 Center for Symbiotic Mechatronics" within the framework of the Austrian COMET-K2 program.

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 547–560, 2018.
https://doi.org/10.1007/978-3-319-99447-5_47

The animal activity recognition based on behavioral monitoring systems with wireless data transmission from acceleration sensors is a rapidly growing area of smart technologies in agriculture. The animal activity characteristics offer potential for health and performance care of the farm animals in particular of dairy calves. Different technologies have been already implemented. Ear-attached accelerometer which is based on radio-frequency identification, quantifies ear movements and estimates feeding and rumination was used in [10]. The authors have confirmed that acceleration technology is a promising tool to measure feeding behavior in beef cattle. Another behavioral monitoring system based on accelerometer technology and ear temperature measurements with application for fitting the calves was discussed in [6]. The SensOor ear attached 3D-accelerometer has been used in [5] as element of measurement tool for rumination, eating and inactivity times. The authors have noticed that ear placement and environmental conditions are critical for successful activity recognition. The modified ear-tag attached to the tail was used in [7] to illustrate the potential in recognizing the parturition. Smartbow activity recognition system was used by authors in [9] who have already proposed a number of algorithms for two-class and multi-class time series classification using stochastic and chaos-theoretical approaches for evaluation of feature vectors. While the first steps in animal activity recognition have been already performed, the problem of rare states recognition in unbalanced data sets is still actual. The ear tag fasted to the animal is equipped with an active 3d accelerometer transmitting the data sets of accelerations along all axes to the central server. Such device has no fixed location and orientation that leads to necessity to use the orientation independent features by solving a classification problem. In this paper we specifically focus on recognition of such rare action as milk intake by eliminating as far as possible the influence of the ear-tag orientation. Data collected from the acceleration sensor according to the local coordinate system is converted using Euler angle rotation and signal projection. This type of transformation can be more useful for fitting physical movements of the ear-tag that in turn leads to increasing of a recognition accuracy for the weakly identified action. The transformed data sets are used for feature extraction for action of interest and subsequent classification by different types of classifiers.

The paper is organized as follows. In Sect. 2, we shortly discuss the materials and methods of data generation and introduce the methodology for data transformation. The activity recognition algorithm together with performance metrics is discussed in Sect. 3. Experiments and numerical results are presented and discussed in Sect. 4. Section 5 summarizes the main results and proposes some ideas for future research.

2 Data Transformation

The data sets of accelerations are transmitted from ear-tags via Smartbow radio wall points to the central server. We consider 15 acceleration data sets of fifteen calves with length of 4 h each. The acceleration data generated during head

moving consists of the recordings along three axes x, y, z and the acceleration magnitude, thus for every point in time i we have four dimensional recordings of the form

$$(x_i, y_i, z_i, \sqrt{x_i^2 + y_i^2 + z_i^2})^T,$$

evaluated with a frequency of 10 per second. The video observations were implemented to get data needed for learn stage in the machine-learning algorithms. From available data sets we select the time intervals with a relative rare event such as milk intake which constitutes only about 4.5% of the total observation time and had a very low sensitivity in realized tests obtained on the basis of a multi-class recognition.

The acceleration sensor measures the values for three axes according to the local coordinate system which is specified by the sensor orientation at certain point of time. The acceleration signals synchronously change with positions and spacial orientations of the ear tag fasted to the head of animal. Due to the sensor fixation in the ear of the calves, the directional recordings are not directly comparable between calves and also between different times. Therefore we want to seek remedies to tackle this problem. The real observations show that during milk drinking the calf's head behaves mostly in horizontal plane. The ear-tag coordinate system is defined according to its default orientation, see Fig. 1. In the global coordinate system we assume that vertical Z-axis captures the upward and downward motions while X- and Y-axis capture forward and backward motions as well as motions to the left and to the right.

Further we expect that appropriate data transformation with respect to the global coordinate system and calculation of corresponding orientation-independent features will improve the quality of recognition of a weakly identifying states like milk intake. We define two transformations based on projection and rotation and compare them with the original time series. Moreover we show that these two transformations are equivalent. The basic idea is that we estimate the direction of gravity in short intervals and use this information to filter out the gravitational acceleration. The sensor data transformed with respect to a global coordinate system becomes then more independent of the local ear tag orientation and therefore is getting more comparing to a raw stream accelerometer readings.

Fig. 1. Ear-tag coordinate system

We analyse three different time series which are explained in more detail below,

1. acceleration magnitude
$$\sqrt{x_i^2 + y_i^2 + z_i^2}$$
2. horizontal magnitude after projection on direction of gravity
$$||h_i||$$
3. horizontal magnitude after rotation
$$\sqrt{x_{i,rot}^2 + y_{i,rot}^2}$$

2.1 Horizontal Projection

The horizontal projection was used in [12] in human activity recognition problem. Here we describe the main principles of such transformation. Assume we have given a time series of the length n

$$(x_i, y_i, z_i)^T, i = 1, \ldots, n,$$

where x_i, y_i and z_i denotes acceleration in x, y and z direction at time i respectively. The projected 2D (Vertical) $\boldsymbol{p_i}$ is defined by

$$\boldsymbol{p_i} = \frac{\boldsymbol{d_i} \cdot \boldsymbol{g}}{\boldsymbol{g} \cdot \boldsymbol{g}} \cdot \boldsymbol{g},$$

where

$$\boldsymbol{g} = (\bar{x}, \bar{y}, \bar{z})^T$$

is an estimate for the direction of gravity using $\bar{x}$, $\bar{y}$ and $\bar{z}$, the means of x-direction, y-direction and z-direction over the considered interval

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i \quad \bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i \quad \bar{z} := \frac{1}{n}\sum_{i=1}^{n} z_i.$$

Defining

$$\boldsymbol{d_i} = (x_i - \bar{x}, y_i - \bar{y}, z_i - \bar{z})^T,$$

we calculate the projected 2D (Horizontal) $\boldsymbol{h_i}$ by

$$\boldsymbol{h_i} = \boldsymbol{d_i} - \boldsymbol{p_i}.$$

and finally the *magnitude of the horizontal projection* as

$$\|\boldsymbol{h_i}\| = \|\boldsymbol{d_i} - \boldsymbol{p_i}\|$$

In Figs. 2 and 3 we see an exemplary data set before and after the projection transformation.

2.2 Rotation

To rotate a vector with an angle α in x-, y- or z- direction we multiply it with the rotation matrices M_x, M_y or M_z

$$M_x(\alpha) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & -\sin(\alpha) \\ 0 & \sin(\alpha) & \cos(\alpha) \end{pmatrix},$$

$$M_y(\alpha) = \begin{pmatrix} \cos(\alpha) & 0 & \sin(\alpha) \\ 0 & 1 & 0 \\ -\sin(\alpha) & 0 & cos(\alpha) \end{pmatrix},$$

$$M_z(\alpha) = \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) & 0 \\ \sin(\alpha) & \cos(\alpha) & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

According to [8,12] the mean of accelerometer readings, computed over a long time period gives an estimate of the gravity $\boldsymbol{g}$ hence we assume exactly as above in our section on projection that

$$\boldsymbol{g} = (\bar{x}, \bar{y}, \bar{z})^T$$

In this approach we want to find a transformation, i.e. a rotation, to transform the vector of gravitation to point along the negative z-axis. The rotation along three axes with respective angles α, β, γ can be calculated using a composition of matrix-multiplications

$$\boldsymbol{a} = M_z M_y M_x a$$

where we want to find a rotation such that

$$M_z M_y M_x \begin{pmatrix} \bar{x} \\ \bar{y} \\ \bar{z} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -\sqrt{\bar{x}^2 + \bar{y}^2 + \bar{z}^2} \end{pmatrix} \tag{1}$$

We assume angle γ to be zero, which is equivalent to setting M_z to be the identity matrix which is equivalent to omitting it. Moreover, since we are not actually interested in the angles and to avoid calculations with inverse trigonometric functions, we use the well known algebraic substitutions

$$\sin\alpha = \frac{2a}{1+a^2}; \qquad \cos\alpha = \frac{1-a^2}{1+a^2}$$

$$\sin\beta = \frac{2b}{1+b^2}; \qquad \cos\beta = \frac{1-b^2}{1+b^2}$$

and solve the equation which is equivalent to Eq. (1)

$$\begin{pmatrix} \frac{1-b^2}{1+b^2} & 0 & \frac{2b}{1+b^2} \\ 0 & 1 & 0 \\ \frac{-2b}{1+b^2} & 0 & \frac{1-b^2}{1+b^2} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1-a^2}{1+a^2} & \frac{-2a}{1+a^2} \\ 0 & \frac{2a}{1+a^2} & \frac{1-a^2}{1+a^2} \end{pmatrix} \begin{pmatrix} \bar{x} \\ \bar{y} \\ \bar{z} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -\sqrt{\bar{x}^2 + \bar{y}^2 + \bar{z}^2} \end{pmatrix}$$

which gives us 2 solutions for a and b and more importantly, two solutions for $M := M_z M_y M_x$:

$$M_1 = \frac{1}{\sqrt{\bar{y}^2+\bar{z}^2}\sqrt{\bar{x}^2+\bar{y}^2+\bar{z}^2}} \begin{pmatrix} \bar{y}^2+\bar{z}^2 & -\bar{x}\bar{y} & -\bar{x}\bar{z} \\ 0 & -\bar{z}\sqrt{\bar{x}^2+\bar{y}^2+\bar{z}^2} & \bar{y}\sqrt{\bar{x}^2+\bar{y}^2+\bar{z}^2} \\ -\bar{x}\sqrt{\bar{y}^2+\bar{z}^2} & -\bar{y}\sqrt{\bar{y}^2+\bar{z}^2} & -\bar{z}\sqrt{\bar{y}^2+\bar{z}^2} \end{pmatrix}$$

$$M_2 = \frac{1}{\sqrt{\bar{y}^2+\bar{z}^2}\sqrt{\bar{x}^2+\bar{y}^2+\bar{z}^2}} \begin{pmatrix} -\bar{y}^2-\bar{z}^2 & \bar{x}\bar{y} & \bar{x}\bar{z} \\ 0 & \bar{z}\sqrt{\bar{x}^2+\bar{y}^2+\bar{z}^2} & -\bar{y}\sqrt{\bar{x}^2+\bar{y}^2+\bar{z}^2} \\ -\bar{x}\sqrt{\bar{y}^2+\bar{z}^2} & -\bar{y}\sqrt{\bar{y}^2+\bar{z}^2} & -\bar{z}\sqrt{\bar{y}^2+\bar{z}^2} \end{pmatrix}$$

Using one of the two solutions, we can rotate the three axes in every segment of our time series. Therefore we get two new time series:

$$\begin{pmatrix} x_{i,rot,j} \\ y_{i,rot,j} \\ z_{i,rot,j} \end{pmatrix} = M_j \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix}, j = 1, 2, i = 1, \ldots, n$$

In Fig. 2 we see this rotation applied to an example dataset.

2.3 Equivalence of Transformations

Proposition 1. *The magnitudes* $\|h\|$ *and* $\sqrt{x_{rot}^2 + y_{rot}^2}$ *satisfy the equality*

$$\|h\|^2 = x_{rot}^2 + y_{rot}^2. \tag{2}$$

Proof. First we inspect the two time series which are constructed by taking either M_1 or M_2. Since the only difference in the first two rows of the matrix is switched signs, it's trivial to conclude that the following holds $x_{rot,1}^2 + y_{rot,1}^2 = x_{rot,2}^2 + y_{rot,2}^2$.

The calculations are straightforward but very lengthy, so we present them in a shortened form. Basically we can factor out a term of $\|h\|$ namely $\bar{x}^2+\bar{y}^2+\bar{z}^2$ and also a factor in $\sqrt{x_{rot}^2 + y_{rot}^2}$, namely $\bar{y}^2+\bar{z}^2$, which leads us to see that both expressions are equal. To simplify calculations we multiply two transformations by one factor

$$\begin{aligned}
&(\bar{y}^2+\bar{z}^2)(\bar{x}^2+\bar{y}^2+\bar{z}^2)\|h\|^2 = \\
&(\bar{y}^2+\bar{z}^2)(\bar{x}^2+\bar{y}^2+\bar{z}^2) \left\| \frac{1}{\bar{x}^2+\bar{y}^2+\bar{z}^2} \begin{pmatrix} x(\bar{y}^2+\bar{z}^2) - \bar{x}(y\bar{y}+z\bar{z}) \\ y(\bar{x}^2+\bar{z}^2) - \bar{y}(x\bar{x}+z\bar{z}) \\ z(\bar{x}^2+\bar{y}^2) - \bar{z}(x\bar{x}+y\bar{y}) \end{pmatrix} \right\|^2 \\
&= \frac{\bar{y}^2+\bar{z}^2}{\bar{x}^2+\bar{y}^2+\bar{z}^2} \Big(x^2(\bar{y}^2+\bar{z}^2)^2 + \bar{x}^2(y\bar{y}+z\bar{z})^2 - 2x\bar{x}(\bar{y}^2+\bar{z}^2)(y\bar{y}+z\bar{z}) \\
&+ y^2(\bar{x}^2+\bar{z}^2)^2 + \bar{y}^2(x\bar{x}+z\bar{z})^2 - 2y\bar{y}(\bar{x}^2+\bar{z}^2)(x\bar{x}+z\bar{z}) \\
&+ z^2(\bar{x}^2+\bar{y}^2)^2 + \bar{z}^2(x\bar{x}+y\bar{y})^2 - 2z\bar{z}(\bar{x}^2+\bar{y}^2)(x\bar{x}+y\bar{y}) \Big) \\
&= (\bar{y}^2+\bar{z}^2)\Big(\bar{x}(y^2+z^2) + \bar{y}(x^2+z^2) + \bar{z}(x^2+y^2) - 2(x\bar{x}y\bar{y} + x\bar{x}z\bar{z} + y\bar{y}z\bar{z}) \Big).
\end{aligned} \tag{3}$$

Next we transform $(\bar{y}^2+\bar{z}^2)(\bar{x}^2+\bar{y}^2+\bar{z}^2)(x_{rot}^2+y_{rot}^2)$:

$$
\begin{aligned}
&(\bar{y}^2+\bar{z}^2)(\bar{x}^2+\bar{y}^2+\bar{z}^2) \qquad (4)\\
&\times\left[\left(\frac{-x\bar{y}^2-x\bar{z}^2+\bar{x}y\bar{y}+\bar{x}z\bar{z}}{\sqrt{\bar{y}^2+\bar{z}^2}\sqrt{\bar{x}^2+\bar{y}^2+\bar{z}^2}}\right)^2+\left(\frac{(y\bar{z}-\bar{y}z)\sqrt{\bar{x}^2+\bar{y}^2+\bar{z}^2}}{\sqrt{\bar{y}^2+\bar{z}^2}\sqrt{\bar{x}^2+\bar{y}^2+\bar{z}^2}}\right)^2\right]\\
&=\bar{y}^2x^2\bar{y}^2+\bar{z}^2x^2\bar{z}^2+\bar{y}^2\bar{x}^2y^2+\bar{z}^2\bar{x}^2z^2\\
&+2x^2\bar{y}^2\bar{z}^2-2\bar{y}^2x\bar{x}y\bar{y}-2\bar{y}^2x\bar{x}z\bar{z}-2\bar{z}^2x\bar{x}y\bar{y}-2\bar{z}^2x\bar{x}z\bar{z}-2\bar{x}^2y\bar{y}z\bar{z}\\
&+\bar{z}^2\bar{x}^2y^2+\bar{y}^2\bar{y}^2z^2+\bar{z}^2y^2\bar{z}^2+\bar{y}^2\bar{x}^2z^2+\bar{y}^2\bar{y}^2z^2+\bar{z}^2\bar{y}^2z^2\\
&-2\bar{x}^2y\bar{y}z\bar{z}-2\bar{y}^2y\bar{y}z\bar{z}-2\bar{z}^2y\bar{y}z\bar{z}\\
&=(\bar{y}^2+\bar{z}^2)\Big(\bar{x}(y^2+z^2)+\bar{y}(x^2+z^2)+\bar{z}(x^2+y^2)-2(x\bar{x}y\bar{y}+x\bar{x}z\bar{z}+y\bar{y}z\bar{z})\Big).
\end{aligned}
$$

Comparing (3) and (4) we can conclude equivalence.

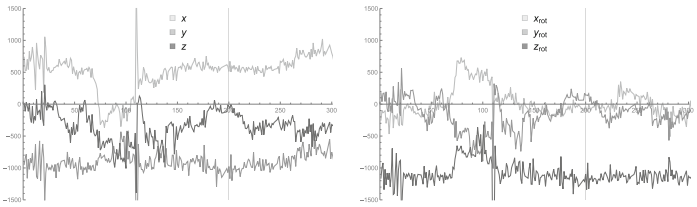


Fig. 2. Example time series: original (left) and rotated (right)

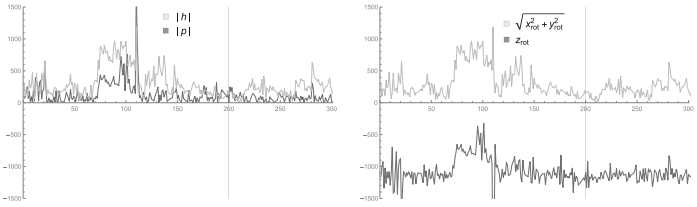


Fig. 3. Transformed time series: projected (left) and horizontal magnitude after rotation (right)

3 Performance Metrics in Activity Recognition

In [4,14] the accuracy, sensitivity and specificity are defined: A classification model (or classifier) is a mapping from instances to predicted classes. Given a classifier and an instance, there are four possible outcomes. If the condition is positive and it is classified as positive, it is counted as *true positive (TP)*; if it is

classified as negative, it is counted as a *false negative (FN)*. If the condition is negative and it is classified as negative, it is counted as *true negative (TN)*; if it is classified as positive, it is counted as a *false positive (FP)*. A confusion matrix consists of these four elements, a schematic depiction can be seen in Fig. 4. The numbers along the major diagonal (TP and TN) represent the correct classifications. The number beside the major diagonal describe the errors - Type I error (FP) and Type II error (FN). The performance of the classifier can be measured by calculating different metrics (e.g. accuracy, specificity and sensitivity).

Outcome of the diagnostic test	Condition (e.g. Disease) As determined by the Standard of Truth		
	Positive	Negative	Row Total
Positive	TP	FP	TP+FP (Total number of subjects with positive test)
Negative	FN	TN	FN + TN (Total number of subjects with negative test)
Column total	TP+FN (Total number of subjects with given condition)	FP+TN (Total number of subjects without given condition)	N = TP+TN+FP+FN (Total number of subjects in study)

Fig. 4. Confusion matrix ([14])

Sensitivity, specificity and accuracy are described in terms of TP, TN, FN, FP.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{5}$$

The accuracy describes the proportion of the right classified values (TP and TN) to all considered cases.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{6}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{7}$$

The sensitivity and specificity describe the proportion of the TP (and TN) and positive (and negative) condition.

3.1 Receiver Operating Characteristics (ROC) Analysis

In [14] the Receiver Operating Characteristics Analysis are defined as follows. For a given diagnostic test, the true positive rate (TPR) against false positive rate (FPR) can be measured, where

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \text{Sensitivity}, \tag{8}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} = 1 - \text{Specificity}. \tag{9}$$

All possible combinations of TPR and FPR compose a ROC space. One TPR and one FPR together determine a single point in the ROC space. The position of a point in the ROC space shows the tradeoff between sensitivity and specificity, i.e. the increase in sensitivity is accompanied by a decrease in specificity. Thus the location of the point in the ROC space depicts whether the diagnostic classification is good or not. In an ideal situation, a point determined by both TPR and FPR yields a coordinate $(0, 1)$ - 100% sensitivity and 100% specificity (It is also called perfect classification). This point is the upper left corner of the ROC space. Theoretically, a random guess would give a point along this diagonal. A good classification is if the point predicted by a diagnostic test fall into the area above the diagonal (otherwise a bad prediction).

Different possible cut-points of a diagnostic test determine a curve in ROC space, which is also called ROC curve. ROC curve is often plotted by using TPR against FPR for different cut-points of a diagnostic test, starting from coordinate $(0, 0)$ and ending at coordinate $(1, 1)$. The interpretation of ROC curve is similar to a single point in the ROC space. The closer the point on the ROC curve to the ideal coordinate, the more accurate the test is. The closer the points on the ROC curve to the diagonal, the less accurate the test is. Figure 4 illustrates the ROC curve and the ROC space.

The area under the ROC curve provides a way to measure the accuracy of a diagnostic test. The larger the area, the more accurate the diagnostic test is. AUC ("Area under curve") is defined by

$$\text{AUC} = \int_0^1 \text{ROC}(t)dt. \tag{10}$$

Since we have detailed information on the behaviour of the individual calves we can define the following limitations: We only consider standing states inside the 4 h interval of each calf, since we assume that these drinking events only occur during standing and we distinguish between two classes - Calf is drinking milk and Calf is not drinking milk.

3.2 Procedure

The features we use for classification are summarized in the table below. Assume that we have given a time series $\mathbf{x} := (x_1, \ldots, x_n)$ and corresponding observations arranging in an ascending order $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$ (Fig. 5).

The activity recognition algorithm can be represented as follows:
This algorithm consists of the following main steps:

1. Perform the rotation/projection in intervals of 1 min length.

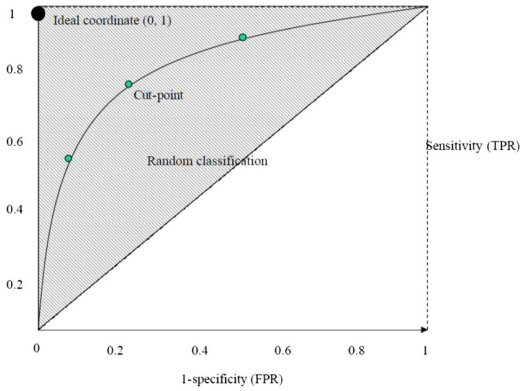


Fig. 5. ROC curve and ROC space ([14])

Table 1. Table of used features

Name	Formula
Mean, $f_1(\mathbf{x})$	$\frac{1}{n}\sum_{i=1}^{n} x_i$
Median, $f_2(\mathbf{x})$	$x^*_{\lfloor \frac{n+1}{2} \rfloor}$
First quartile, $f_3(\mathbf{x})$	$x^*_{\lfloor \frac{n+1}{4} \rfloor}$
Third quartile, $f_4(\mathbf{x})$	$x^*_{\lfloor \frac{3(n+1)}{4} \rfloor}$
Mean Deviation, $f_5(\mathbf{x})$	$\frac{1}{n}\sum_{i=1}^{n} \lvert x_i - f_1(\mathbf{x})\rvert$
MAD $f_6(\mathbf{x})$	$f_2((\lvert x_1 - f_2(\mathbf{x})\rvert, \ldots, \lvert x_n - f_2(\mathbf{x})\rvert))$
Periodogram-Sum $f_7(\mathbf{x})$	$\sum_{k=0}^{n} \left\lvert \sum_{j=0}^{n-1} x_j e^{-\frac{2\pi i}{n} kj} \right\rvert^2$

2. Calculate different features in each one-second-interval for both the original series (magnitude of the acceleration) and the transformed horizontal magnitude.
3. Split the data according to a 15-fold cross-validation: Fourteen calves are used for training and one for validation.
4. Build 4 different algorithms (Naive Bayes, Logistic Regression, Decision Tree and Nearest Neighbors) for both type of considered acceleration data using our current training set and use them on the current validation set.
5. Repeat procedure fifteen times.

Data: Acceleration datasets $\mathbf{x}_1, \ldots, \mathbf{x}_{15}$ with labels representing drinking state
Calculate rotation/projection in 1-minute intervals ($\hat{=}$ 600 data-points)
$\rightarrow \hat{\mathbf{x}}_1 \ldots, \hat{\mathbf{x}}_{15}$;
and vector magnitude $\rightarrow \tilde{\mathbf{x}}_1 \ldots, \tilde{\mathbf{x}}_{15}$;
Split up $\{\hat{\mathbf{x}}_1 \ldots, \hat{\mathbf{x}}_{15}\}$and $\{\tilde{\mathbf{x}}_1 \ldots, \tilde{\mathbf{x}}_{15}\}$ into intervals of 1 second length ($\hat{=}$ 10 data-points):
for $k := 1, k \leq 15$ **do**
  $M_k := \{\hat{\mathbf{x}}_{k,1}, , \ldots \hat{\mathbf{x}}_{k,l_k}\}$
  $T_k := \{\tilde{\mathbf{x}}_{k,1}, , \ldots \tilde{\mathbf{x}}_{k,l_k}\}$
end
Calculate features and combine them with their respective class label $c \in \{0, 1\}$ (1:= Drinking milk, 0:= Not drinking milk):
for $k := 1, k \leq 15$ **do**
  $f_{1,k} =$
  $\{(f_1(M_{k,1}), \ldots, f_7(M_{k,1}), c_{k,1}), \ldots, (f_1(M_{k,l_k}), \ldots, f_7(M_{k,l_k}), c_{k,l_k})\}$;
  $f_{2,k} = \{(f_1(T_{k,1}), \ldots, f_7(T_{k,1}), c_{k,1}), \ldots, (f_1(T_{k,l_k}), \ldots, f_7(T_{k,l_k}), c_{k,l_k})\}$;
end
Train Models and assess performance by using on validation set:
for $k := 1, k \leq 15$ **do**
  for $j := 1, j \leq 2$ **do**
    $C_{j,k,1} = NaiveBayes(f_{j,1}, \ldots, f_{j,k-1}, f_{j,k+1}, \ldots, f_{j,15})$;
    $C_{j,k,2} = LogisticRegression(f_{j,1}, \ldots, f_{j,k-1}, f_{j,k+1}, \ldots, f_{j,15})$;
    $C_{j,k,3} = DecisionTree(f_{j,1}, \ldots, f_{j,k-1}, f_{j,k+1}, \ldots, f_{j,15})$;
    $C_{j,k,4} = NearestNeighbors(f_{j,1}, \ldots, f_{j,k-1}, f_{j,k+1}, \ldots, f_{j,15})$;
    $j++$;
  end
  for $j := 1, j \leq 2$ **do**
    for $i := 1, i \leq 4$ **do**
      $Performance_{j,k,i} := C_{j,k,i}(f_{j,k})$
      $i++$;
    end
    $j++$;
  end
  $k++$;
end

4 Results

In our results, the different classifiers assign a probability to each class and the ROC is constructed by defining different percentage thresholds for assigning to one of the two classes, the results are depicted in Fig. 6. Moreover we present some classical performance measures which are calculated by adding up the 15 confusion matrices from our cross validation in Table 2.

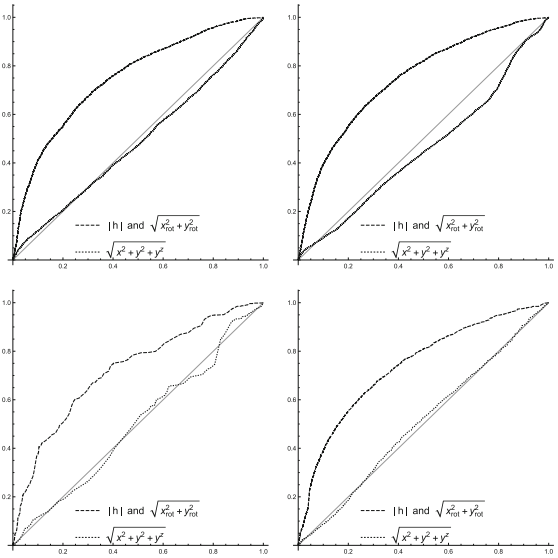


Fig. 6. Comparison of different Receiver operating characteristic curves to compare the quality of classification: Naive Bayes (top left), Logistic Regression (top right), Decison tree (bottom left) and Nearest Neighbors (bottom right)

As we can see from both Fig. 6 and the AUC values in Table 2 is that the discriminative power of the features from the transformed time series are superior to the ones calculated from the original series. Moreover we can conclude that the usage of classical performance measures should be done with caution, since the sometimes fail to represent the difference in quality.

Table 2. Performance measures for different classifiers and our two considered approaches, higher values in each row are highlighted in gray

Measure	Classifier	Proj/trans	Magnitude
AUC (10)	Naive Bayes	0.750	0.486
	Logistic Regression	0.747	0.458
	Decision Tree	0.722	0.498
	Nearest Neighbors	0.739	0.507
Sensitivity (6)	Naive Bayes	0.454	0.026
	Logistic Regression	0	0
	Decision Tree	0.01	0
	Nearest Neighbors	0	0
Specificity (7)	Naive Bayes	0.877	0.988
	Logistic Regression	1	1
	Decision Tree	0.997	0.999
	Nearest Neighbors	1	1
Accuracy (5)	Naive Bayes	0.858	0.944
	Logistic Regression	0.955	0.955
	Decision Tree	0.953	0.954
	Nearest Neighbors	0.955	0.955
Cohens κ [3]	Naive Bayes	0.167	0.020
	Logistic Regression	0	0
	Decision Tree	0.019	-0.002
	Nearest Neighbors	0	0
Youdens Index $\mathcal{J}$[13]	Naive Bayes	0.331	0.013
	Logistic Regression	0	0
	Decision Tree	0.011	-0.001
	Nearest Neighbors	0	0

5 Conclusion

In this paper we focus on a orientation-independent animal behavioural recognition of a weakly identified rare action of milk intake. The raw accelerometer data from ear-tag coordinate system were transformed into the global coordinate system. The numerical results show that the features evaluated with transformed data sets are more appropriate to fit the rare actions and states. We expect that such technology can be applied in multi-class recognition problem as part of an ensemble method in machine learning where a combination of different models is used.

Acknowledgements. This work has been supported by the COMET-K2 "Center for Symbiotic Mechatronics" of the Linz Center of Mechatronics (LCM) funded by the Austrian federal government and the federal state of Upper Austria.

References

1. Aminev, D.A., Kozyrev, D.V., Zhurkov, A.P., Romanov, A.Y., Romanova, I.I.: Method of automated control of distributed radio direction finding system. In: Dynamics of Systems, Mechanisms and Machines (Dynamics), Omsk, Russia, pp. 1–9 (2017)
2. Ann, O.C., Theng, L.B.: Human activity recognition: a review. In: 2014 IEEE International Conference on Control System, Computing and Engineering (ICCSCE 2014), Batu Ferringhi, pp. 389–393 (2014)
3. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**, 37–46 (1960)
4. Fawcett, T.: An introduction to ROC analysis. Pattern Recognit. Lett. **27**(8), 861–874 (2006)
5. Hill, T.M., Suarez-Mena, F.X., Hu, W., Dennis, T.S., Schlotterbeck, R.L., Timms, L.L.: Technical note: evaluation of an ear-attached movement sensor to record rumination, eating, and activity behaviors in 1-month-old calves. Prof. Anim. Sci. **33**, 743–747 (2017)
6. Hodson, M., Timms, L.: Use of an Ear Tag Based Behavioral and Temperature Monitor (Cow ManageR) on Dairy Calves (Preliminary Report). Animal Industry Report: AS 663, ASL R3164 (2017). https://lib.dr.iastate.edu/ans_air/vol663/iss1/37
7. Krieger, S., Sattlecker, G., Kickinger, F., Auer, W., Drillich, M., Iwersen, M.: Prediction of calving in dairy cows using a tail-mounted tri-axial accelerometer: a pilot study. Biosyst. Eng. (2017). https://doi.org/10.1016/j.biosystemseng.2017.11.010
8. Lu, H., Yang, J., Liu, Z., Lane, N.D., Choudhury, T., Campbell, A.T.: The jigsaw continuous sensing engine for mobile phone applications. In: SenSys, pp. 71–84 (2010)
9. Sturm, V., et al.: A chaos theoretic approach to animal activity recognition. J. Math. Sci. (2017, to appear). XXXIV International Seminar on Stability Problems for Stochastic Models, Debrecen, Hungary
10. Wolfger, B., Timsit, E., Pajor, E.A., Cook, N., Barkema, H.W., Orsel, K.: Technical note: accuracy of an ear tag-attached accelerometer to monitor rumination and feeding behavior in feedlot cattle. J. Anim. Sci. **93**(6), 3164–3168 (2015)
11. Vishnevsky, V., Kozyrev, D., Rykov, V.: New generation of safety systems for automobile traffic control using RFID technology and broadband wireless communication. In: Vishnevsky, V., Kozyrev, D., Larionov, A. (eds.) DCCN 2013. CCIS, vol. 279, pp. 145–153. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-05209-0_13
12. Yan, Z., Chakraborty, D., Misra, A., Jeung, H., Aberer, K.: Semantic Acitivity Classification Using Locomotive Signatures from Mobile Phones (2012). https://infoscience.epfl.ch/record/174016/files/2012_semActi.pdf
13. Youden, J.W.: Index for rating diagnostic tests. Cancer **3**, 32–35 (1950)
14. Zhu, W., Zeng, N., Wang, N.: Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS® Implementations (2010). https://www.lexjansen.com/nesug/nesug10/hl/hl07.pdf

Principles of Construction of Mobile and Stationary Tethered High-Altitude Unmanned Telecommunication Platforms of Long-Term Operation

V. M. Vishnevsky[1(✉)], D. V. Efrosinin[1,2], and A. Krishnamoorthy[3]

[1] V.A. Trapeznikov Institute of Control Sciences of RAS,
Profsoyuznaya st., 65, 117997 Moscow, Russia
vishn@inbox.ru
[2] Johannes Kepler University, Altenbergerstrasse, 69,
4040 Linz, Austria
dmitry.efrosinin@jku.at
[3] Department of Mathematics, CMS College,
Kottayam 686001, India
achyuthacusat@gmail.com

Abstract. This work considers the principles of designing of the new generation of tethered high-altitude telecommunication platforms. A set of theoretical and engineering problems for the development of technology and a system for ground-to-aircraft transmission of high-power energy based on the development of the principle of resonance high-frequency energy transfer by N. Tesla is formulated; of the development of a highly reliable unmanned vehicle for long-term use; of the development of a local navigation system that provides high positioning accuracy and increased noise immunity compared to satellite navigation systems, etc.

Keywords: Tethered high-altitude unmanned telecommunication platforms · Autonomous unmanned aerial vehicles
High-power energy transmission · Multi-rotor facility

1 Introduction

At present, high-altitude telecommunication platforms implemented on autonomous unmanned aerial vehicles and underwater robots have been widely developed. The main disadvantage of the autonomous unmanned vehicles

A. Krishnamoorthy—The work has been carried out with the partial financial support from the Russian Science Foundation and the DST (India) (grant No. 16-49-02021) within the joint research project of the V.A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences and the CMS College of Kottayam, India.

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 561–569, 2018.
https://doi.org/10.1007/978-3-319-99447-5_48

(UAVs) is a limited time of operation due to the small battery resource of UAVs equipped with electric motors or the fuel reserve for internal combustion engines. In this regard, these UAVs can not be effectively used in systems that require long-term operation, for example, in the safety management systems and counter-terrorist surveillance systems, protecting critical facilities (nuclear power plants, airports, extended bridges and sections of the border, etc.) from the terrorist threats. Long-term operation can be provided by tethered high-altitude unmanned platforms in which the power supply of engines and payload equipment is provided from the ground-based energy sources.

The tethered high-altitude platforms fall in between satellite systems and terrestrial systems, whose equipment (cellular base stations, radio relay and radar equipment, etc.) is located on high-altitude structures. Compared with expensive satellite systems, tethered high-altitude platforms are highly economical, and outperform terrestrial telecommunication systems in the vast area of telecommunications and video coverage. Considering the extensive practical application of tethered unmanned high-altitude platforms in both civilian and defense industries, intensive development engineering and development of such platforms is being carried out in the research centers of the advanced countries of the world. The main direction of research is the creation of high-power energy transmission systems through thin copper cables and unmanned vehicles with high reliability and long operating time without lowering to the ground.

Currently, numerous projects are known for creating tethered high-altitude platforms with low altitude and small (1–3 kW) power consumption of propulsion systems (for example, the development of AeroVironment (USA) http://avia.pro/blog/aerovironment-tether-eye-tehnicheskie-harakteristiki-foto). Similar products are supplied to the international market by the Chinese company Beijing Dagong Technology (https://dagongtech.en.alibaba.com), the German company Copting (www.copting.de), the French company Elistair (https://elistair.com), etc. However, these systems do not provide a rise on significant payload heights of any significant weight, as most of energy transmitted from the ground to the aircraft is spent on the operation of the propulsion system and the retention of the cable-rope.

The development considered in this article has significant competitive advantages, having much better basic characteristics [1–5]. The main advantage is the possibility of remote transmission of energy up to 10 kW by copper wires of small section (small weight) from the ground to the board for powering electric motors and equipment of high-altitude rotorcraft. The new technology of energy transfer will enable the platform to be raised to a height of up to 200 m with a payload of up to 15 kg and a long operation period limited only by the reliability characteristics of the unmanned vehicle. It should be noted that this technology of transferring energy through small-section wires can also be effectively used in the creation of deep-water robots. The originality of the technology is confirmed by patent No. 2572822 “Method of remote wire power supply of facilities” dated December 16, 2015. Another advantage of the proposed project is that the Kevlar strength-power-communications cable includes not only copper

cables but also an optical fiber providing transmission of large amounts of data from the board to the ground and vice versa. This allows to install only the necessary antenna equipment on board of the high-altitude platform, leaving on the ground the most dimensional and heavy parts of the base station of cellular network of 4th generation (LTE), radar systems, radio link equipment and video surveillance equipment. It is also planned to develop a multifunctional multi-rotor unmanned vehicle with heavy payload and long working life, as well as a local navigation system based on tethered high-altitude platforms, which provides high positioning accuracy and increased noise immunity compared to satellite navigation systems.

Recently, there was information about the beginning of new developments in the field of tethered high-altitude platforms. In 2017 it was announced that the largest US telecommunications company AT&T is planning to implement a project to create "LTE-towers" based on tethered multicopters (https://3dnews.ru/947864). In accordance with the project, the base stations of cellular networks LTE (Long Term Evolution), installed on board of a tethered high-altitude platform, must provide services to mobile users (cell phones, gadgets, etc.) on the territory up to 100 km^2. The Defense Advanced Research Projects Agency (DARPA) has announced a tender for the creation of a monitoring system for mobile objects using tethered high-altitude rotorcraft platforms of long-term operation (http://www.darpa.mil/news-events/2016-09-13).

These projects indicate the relevance of the described development and additional extensive areas of its application.

2 The Architecture of Tethered High-Altitude Telecommunication Platforms

The architecture of a tethered high-altitude platform includes the following main components.

1. Unmanned multi-rotor equipment of large carrying capacity and long operating time, designed for lifting and holding the telecommunication payload and video surveillance equipment at a height of up to 200 m.
2. Ground-to-board energy transmission system of high-power (up to 10 kW), providing power supply to propulsion systems of unmanned multi-rotor facility and to payload equipment.
3. Control and stabilization system of the high-altitude platform, including a local navigation subsystem with ground-based radio beacons, providing increased positioning accuracy and noise immunity in the absence of signals from satellite navigation systems.
4. On-board payload equipment including: base station of the cellular network of the fourth generation (LTE); radar and radio relay equipment; equipment for video surveillance and environmental monitoring, etc.
5. Cable-rope on Kevlar base, including copper wires of small cross-section (0.5 mm^2) for transmission of high-voltage (up to 2000 V), high-frequency

(up to 100 kHz) signals and optical fiber for digital information transmission with a speed of up to 10 Gbit/s.

6. Ground control complex, which includes an AC voltage converter 380/2000 V, a system for diagnostics of the parameters of the high-altitude platform and an intelligent wrench with a microprocessor unit to control the cable-rope tension during lifting, descending and wind loads. In mobile configuration, the ground control center is located on a mobile platform with an electric generator installed on it, the output power of which is not less than 15 kW.

The scheme for the transfer of high-power energy, providing the possibility of developing an unmanned multi-rotor facility with a large take-off weight (up to 100 kg) is shown in Fig. 1.

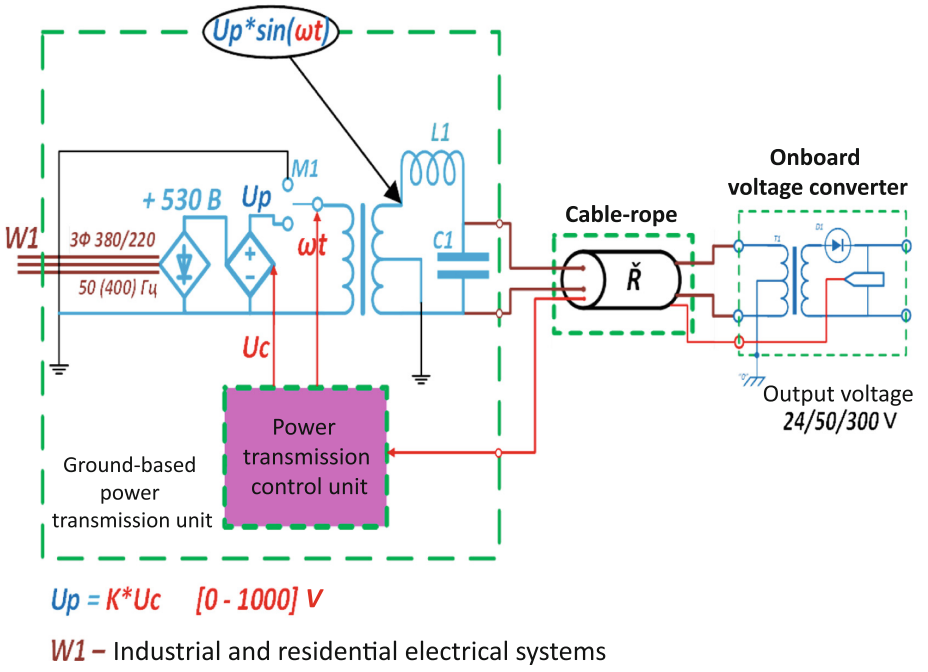


Fig. 1. High-power energy transfer scheme

The input voltage of 380/220 V 50 Hz is delivered to the rectifying 3-phase bridge, the output of which generates a constant voltage of 530 V. The rectified voltage is applied to the input of direct-voltage transducer controlled by a feedback signal, at the output of which a constant voltage is formed in the range 0–1000 V, depending on the feedback signal, which is formed by the deviation of the output voltage from the nominal value of the on-board converter. Thus, the output voltage of the energy transmission system is stabilized. Further, this controlled DC voltage is applied to the input of the electronic commutation

bridge, which transmits its input voltage to the output in alternating forward and reverse polarities.

A unique scheme is introduced for coordination of resistances in power circuits using the LC gyrator (L1, C1), which is presented by a resonance circuit, induced by voltage drops at the output of the electronic commuting bridge. The switching frequency of the commuting bridge coincides with the natural frequency of the gyrator and is 150–200 kHz. The ratios of the inductance and capacity values of gyrator are selected so that the output resistance of the gyrator is equal to the wave resistance of the wire line, over which energy is transferred from the ground-based transducer to the high-altitude platform. With this type of coordination, an increase in the voltage directly applied to the line up to values of the order of 2000 V and suppression of the current harmonics arising from the non-linearity of the rectifying part of the on-board converter circuit occurs.

Wired connection line is connected to the input transformer of the on-board voltage converter of the high-altitude platform. The ratio of the turns of the primary and secondary windings of the transformer is chosen in such a way that to provide the equality of the wave resistance of the wire line and the input resistance of the rectifying part of the on-board converter circuit near the maximum levels of energy consumption. Capacities for flattening the ripple of the output voltage are connected to the output of the rectifying circuit of the on-board converter. Such scheme for building an on-board converter provides a theoretically possible limit to the reliability of the device. The output voltage of the on-board converter on a thin wire channel is transmitted to the ground block at the input of the feedback signal analysis circuit.

The above diagram illustrates the proposed methodology for the transmission of high-power energy. Unlike traditional low-frequency approaches, a resonance frequency method of energy transfer by high-voltage (up to 2000 V), high-frequency (up to 100 kHz) signal is proposed, which allows to sharply reduce the weight and size of ground and on-board voltage converters. The sharp decrease in the weight of the on-board converter and the connecting cable-rope, which is fundamental for creating a tethered high-altitude platform, is one of the main advantages of the proposed approach.

3 Complex of Problems Under Consideration

Development of technology and methods of engineering of a new generation of tethered high-altitude platforms require the use and development of approaches and methods from various fields of science, including: a resonance theory of energy transfer; reliability theory and queuing theory; the theory of optimal control and mathematical programming; theoretical foundations of electrical engineering and broadband wireless communication, etc. Designing and manufacturing of experimental samples of a tethered high-altitude platform for long-term operation will require the use of the latest developments in the field of electronics, materials science, wireless radio and optical communications, aircraft construction, etc., as well as solving the following set of new scientific and engineering tasks.

1. Development of a complex of models, methods and algorithms for assessing reliability based on multidimensional Markov random processes for comparative analysis and selection of options for the optimal construction of a high-altitude platform for long-term operation.
2. Development of analytical and simulation models of a high-power energy transmission system to define optimal parameters for high-frequency sources and high-power energy receivers; solution of the problem of controlling high voltage output of the source when the load in the on-board receiver changes using digital feedback; minimization of losses in resonance converters of resistances between the energy source, the wave-making resistance of hardwired communication medium and the on-board receiver, and minimization of the magnitude of electromagnetic radiation during the transmission of energy by a high-frequency signal.
3. Calculation of characteristics and selection of parameters for a brushless electric motor of systems and the whole architecture of high-altitude rotary-wing module providing lifting and holding at a given height of the telecommunications platform for a long time of operation.
4. Development of an architecture of a ground control complex and a connecting cable-rope of a high-altitude platform, which provides high-speed transmission of multimedia information over an optical fiber channel and transfer of high-power energy from the ground to the aircraft.
5. Development of a new set of models of adaptive stochastic polling to minimize interference and disturbances, defining optimal modes of operation of on-board telecommunications equipment (LTE cellular network base stations, radar and video control) in interaction with mobile and stationary users.
6. Development of a mathematical model of the dynamics of a tethered platform with a complex cable-rope loading in turbulent atmosphere, the study of which is the basis for determining the parameters of the autopilot and the aerodynamic characteristics of the high-altitude rotor-craft.
7. Development of principles of construction, mathematical models and algorithms of a high-altitude platform control system with a backup stabilization system that does not use navigation satellite systems (GLONASS/GPS).
8. Development of a diagnostics system of the parameters of a high-altitude unmanned module (vibration, temperature, voltage and current) transmitted on-line via the wireless communication channel to the ground control system, to automatically monitor the output of these parameters beyond acceptable threshold values.

To study a new set of models for estimating the stationary and non-stationary reliability characteristics of multi-rotor unmanned high-altitude platforms, including finding distribution function of the time before the first system failure and the moments of this random variable, it is proposed to use an apparatus of multidimensional Markov processes, the block structure of matrix of the transition intensity of which allows to apply matrix-analytic methods of Newts for finding the solution. The solution of the problem of finding the reliability function of large-scale systems is based on the application of asymptotic

approaches and methods of simulation. Methods and algorithms of stochastic controllable processes are supposed to be used to solve the problem of choosing the optimal moments for carrying out preventive maintenance of an unmanned multi-rotor vehicle.

The technology and transmission system of high power energy is based on the method of converting the industrial voltage 220/380V with frequency of 50 Hz to a high-frequency signal (up to 200 kHz) of high voltage (up to 2000 V), followed by a reverse conversion of a high voltage to a constant voltage (24V, 48V) necessary power supply of engines and high-altitude platform equipment. The specified technology based on the evolution of N. Tesla principle of high-frequency resonance energy transmission and on original circuit designs, using high-frequency electronic devices and switching elements of high power and speed, allows to significantly reduce the size and weight of the on-board and ground transformers and inductors, which is essential when creating tethered high-altitude platforms. A set of analytical and simulation models under development, including the analysis of the Heaviside equations' solutions for the case of a long link, allows to choose the optimal parameters of an energy transfer system, and exclude the effect of reflected waves, which can lead to breakdown of the insulation and reduced line efficiency.

The development of a local control and stabilization system that provides small deviations of the altitude platform from the hovering point in the absence or weakening of GPS/GLONASS signals involves the installation of ground beacons capable of recording the time of receipt of a signal from an active sensor aboard an unmanned module. It is planned to use deterministic and probabilistic methods of object localization. In the first case, hyperbolic trilateration will be used, where the position of the object is calculated by the difference in the time of receiving the signals. In the second case, it is assumed that probabilistic localization methods are based on hidden Markov chains. As hidden states, the coordinates of the object are used, and the observed states will be described by differences in the time of signal acquisition. The Viterbi dynamic programming algorithm is planned to be used to find the list of hidden states. Baum-Welsh algorithm will be used to reconfigure (learning) the parameters of the hidden Markov chain, until the point of distribution function vector of the observed random variables stabilization.

The development of a complex of statistical models of the system for diagnosing the operability of a tethered unmanned module is planned on the basis of machine learning methods used to solve problems related to the automatic search for breakpoints, anomalies, and classification of segments in piecewise-stationary time series. During the actual testing, as well as with the help of simulation, it is planned to obtain samples of the observed performance indicators of the high-altitude unmanned module, such as the vibration parameters obtained with a sensor with a three-dimensional accelerometer, temperature, amperage and voltage.

A set of stochastic features (metrics), including, for example, dispersion, asymmetry and kurtosis coefficients, maximum and minimum values of param-

eters at a certain time interval, spectral density, etc., and metric, dynamic and topological invariants of attractors of dynamic models obtained from the corresponding time series transformed with the help of band and low-frequency filters, as well as wavelet transforms for removing noise in data will be proposed based on these indicators. It is planned to conduct an empirical evaluation of the threshold values of these characteristics, indicating a decrease in the working capacity of the module. The proposed stochastic and deterministic metrics based on observable parameters are supposed to be used as input data for the learning process and further binary metric classification in real time mode, carried out with the help of various models of machine learning such as neural networks, support vector method, hidden Markov processes, etc., as well as the convolutional neural network, which is part of the technology of in-depth training. The parameters of sensitivity and specificity will be used as statistical indicators of the effectiveness of the diagnostic test. In the study of adaptive polling algorithms that minimize interference and disturbances of on-board telecommunications equipment, approaches and methods for calculating the performance characteristics of $BMAP/G/n$ queuing systems with correlated input flows will be used and developed.

To determine the parameters of the autopilot and aerodynamic characteristics of an unmanned aerial vehicle for long-term operation, a mathematical model of the dynamics of a tethered platform in the conditions of a turbulent atmosphere will be developed and investigated. Solving the system of differential equations describing the functioning of the tethered platform in the turbulent atmosphere will allow to determine the required values of the ground-board power depending on the lift height and the magnitude of the wind loads.

When implementing the experimental model of a tethered high-altitude platform, the latest developments in the field of aircraft construction, broadband wireless communications, materials science and electronics will be used.

4 Conclusions

The article gives a brief overview of the current state and development prospects of tethered high-altitude unmanned telecommunication platforms. The architecture, a complex of scientific and technical problems and principles of construction of a new generation of such high-altitude platforms are described.

References

1. Vishnevsky, V., Tereschenko, B., Tumchenok, D., Shirvanyan, A.: Optimal method for uplink transfer of power and the design of high-voltage cable for tethered high-altitude unmanned telecommunication platforms. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds.) DCCN 2017. CCIS, vol. 700, pp. 240–247. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66836-9_20
2. Vishnevsky, V.M., Andronov, A.N.: Estimating the throughput of wireless hybrid systems operating in a semi-Markov stochastic environment. Autom. Remote Control **78**(12), 2154–2165 (2017). https://doi.org/10.1134/S0005117917120049

3. Andreev, M.A., Vishnevsky, V.M., Gregoriev, F.N., Tershchenko, B.N.: Identifying the dynamics of azimuth angle change of the antenna of a high-altitude telecommunication platform. In: Distributed Computer and Communication Networks. Theory and Applications (DCCN-2008), Sofia, Bulgaria, pp. 120–129 (2008)
4. Dudin, A.N., Vishnevsky, V.M., Sinyugina, Y.V.: Analysis of the BMAP/G/1 queue with gated service and adaptive vacations duration. Telecommun. Syst. J. **61**(3), 403–415 (2016). https://doi.org/10.1007/s10958-016-2814-1
5. Vishnevsky, V.M., Tereshchenko, B.N.: Development and research of high-altitude tethered telecommunications platforms. In: Telekommunikatsii and Transport, no. 7, pp. 20–24 (2013). (in Russian)

Reliability of Two Communication Channels in a Random Environment

A. M. Andronov and V. M. Vishnevsky(✉)

Transport and Telecommunication Institute, Lomonosov Str., 1, Riga 1019, Latvia
lora@mailbox.riga.lv,vishn@inbox.ru

Abstract. Considered system consists of two renewable channels that connected in parallel. The system operates in a random environment having k states. The functioning of both components are described by two continuous time alternating processes. The sojourn time in the state 0 (work state) of both channels has exponential distribution with parameters $\mu_{1,i}$ and $\mu_{2,i}$ if the random environment has state i. The sojourn times in the state 1 (failed state) have general absolute continuous distributions. These sojourn times are independent and doesn't depend on the random environment state too. The system is working at time t if at least one channel is working. The system reliability on given time interval is calculated for the known initial states of the components.

Keywords: Continuous-time Markov chain · Recurrent event
Renewal equation · System reliability

1 Introduction

The reliability of double redundant renewable system was a subject of the consideration a long time ago [5,7,11,12]. This problem acquires a special interest if it is supposed that the system operates in a random environment. Usually the random environment is described as continuous-time finite Markov chain [1–3,8–10]. In this paper we consider a reliability system, which consists from two channels connected in parallel, each of which has its own repair facility. The system operates in a random environment having k states. In this case the functioning of both channels can be described by continuous time alternating processes $X_1(t)$ and $X_2(t)$. It is supposed that the sojourn time in the state 0 (work state) of both channels has exponential distribution with parameters $\mu_{1,i}$ and $\mu_{2,i}$ if the random environment has state i. The sojourn times in the state 1 (failed state) have general absolute continuous distributions with probability density functions (p.d.f.) $\alpha_1(t)$ and $\alpha_2(t)$. These sojourn times are independent and doesn't depend on the random environment state too. The system is working at time t if at least one channel is working. We wish calculate system reliability on interval (0, t). The paper is organized as follows. The random environment and the channels are described in Sects. 2 and 3. The Sect. 4 is devoted to the channel reliability if the renewal absents. To take into consideration the renewals,

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 570–576, 2018.
https://doi.org/10.1007/978-3-319-99447-5_49

recurrent events are used (Sect. 5). The reliability function of the system is presented in Sect. 6. Section 7 contains a numerical example. The paper ends with conclusion.

2 Random Environment

We suppose that considered system operates in a random environment. The last is presented [10] as a continuous-time homogeneous irreducible Markov chain $J(t), t \geq 0$, with finite state set $N = 1, 2, \ldots, k$. Let $\lambda_{i,j}$ be the known transition rate from state i to state $j(\lambda_{i,j} = 0)$ and $\lambda = (\lambda_{i,j})$ be a $k \times k$ matrix, $\Lambda = diag(\sum\limits_{j} \lambda_{i,j})$ be a diagonal matrix, $P_{i,j}(t) = PY(t) = j|Y(0) = i$ be the transition probability of Markov chain $Y(t)$, and let $P(t) = (P_{i,j}(t))_{k\times k}$ denote the corresponding matrix. If all eigenvalues of matrix $A = \lambda - \Lambda$ are different then probabilities $P(t) = (P_{i,j}(t))_{k\times k}$ can be represented simply. Let ν_η and $\chi_\eta, \eta = 1, \ldots, k$, be the eigenvalue and the corresponding unit norm's eigenvector of $A, \chi = (\chi_1, \ldots, \chi_k)$ be the matrix of eigenvectors, $\overline{\chi} = \chi^{-1} = (\overline{\chi}_1^T, \ldots, \overline{\chi}_k^T)^T$ be the corresponding inverse matrix (here $\overline{\chi}_\eta$ is the $\eta - th$ row of $\overline{Z}$. Then [4,10]

$$P(t) = exp(tA) = \chi diag(exp(\nu_1 t), \ldots, exp(\nu_k t))\chi^{-1} = \sum_{\eta=1}^{k} \chi_\eta exp(\gamma_\eta t)\overline{\chi}_\eta. \quad (1)$$

It is known that for the considered Markov chain one eigenvalue equals 0 (we give him number 1), another eigenvalues (with numbers $2, \ldots, k$) are negative.

3 Channels

We consider two channels that are connected in parallel. The first and the second channels are described by continuous time alternative processes $X_1(t)$ and $X_2(t)$ correspondingly. These processes are independent, if a traectory of the random environment is fixed. The sojourn time in the state 0 (work state) of both processes has exponential distribution with parameters $\mu_{1,i}$ and $\mu_{2,i}$, if the random environment has state i. The sojourn time in the state 1 (failed state) has nonnegative distribution density $\alpha_1(t)$ and $\alpha_w(t)$ for $X_1(t)$ and $X_2(t)$, independently from state of the random environment. Let $\overline{A}_\nu(t) = 1 - \int\limits_0^t \alpha_\nu(\varsigma)d\tau$. All sojourn times are independent.

The system is worked at time t if at least one channel is worked. Then the integrated state of the system $Z(t) \in 0, 1$ can be presented as $Z(t) = X_2(t) \wedge X_2(t)$. We wish calculate system reliability on interval $(0, t)$ if initially the random environment has state $i \in 1, \ldots, k$:

$$R_i(t) = R(t|Y(0) = i) = PZ(t) = 0 : \tau \in (0, t)|Y(0) = i, X_1(0) = 0, X_2(0) = 0 \quad (2)$$

4 The Reliability Without Renewals

At time we consider one channel $\eta = 1, 2$ and suppose that renewal of the failed channel absent. What is probability $\widetilde{P}_{i,j}^{(\eta)(t)}$ that channel η will be worked during time t and final state of the random environment is j if the initial state of the random environment is i?

In this case the following difference with respect to upper considered continuous-time homogeneous irreducible Markov chain $J(t)$ takes place: we have absorbing chain with an absorbing state. The transition rate for the η-th component from state $i = 1, \ldots, k$ to the absorbing state equals $\mu_{\eta,i}$. Now instead of Λ and $A = \lambda - \Lambda$ we have $\widetilde{\Lambda} = diag(\sum\limits_{i} \lambda_{i,j} + \mu_{\eta,i}$ and $\widetilde{A} = \lambda - \widetilde{\Lambda}$. Analogously to (2) we have

$$\widetilde{P}^{\eta}(t) = (P_{i,j}^{(\eta)}(t))_{k\times k} = exp(t\widetilde{A}) = \widetilde{\chi} diag(exp(\widetilde{\nu}_1 t), \ldots, exp(\widetilde{\nu}_k t))\widetilde{\chi}^{-1} = \\ = \sum_{\eta=1}^{k} \widetilde{\chi}_\eta exp(\widetilde{\gamma}_\eta t)\widetilde{\overline{\chi}}_\eta, \tag{3}$$

where the tilda means that it is related to the generator $\widetilde{A}$

5 Recurrent Events

We call *recurrent event of the i-th kind* a time moment t, when process $W(t) = (Y(t), X_1(t), X_2(t))$ goes in the state $(i, 0, 0)$ from any another state.

Let $f_{i,j}(t)$ be be the probability density of a time between two neighbouring *recurrent events of the i-th and j-th kinds*, so that another recurrent events and system failure absent between them). Then

$$f_{i,j}(t) = \lambda_{i,j} exp)(-(\lambda_i + \mu_{1,i} + \mu_{2,i})t) \\ + \sum_{\eta=1}^{2} \int_0^t \mu_\eta exp(-(\lambda_i + \mu_{1,i} + \mu_{2,i})\tau)\alpha_\eta(t-\tau)\widetilde{P}_{i,j}^{3-\eta}(t-\tau)d\tau, t \geq 0. \tag{4}$$

6 Reliability Function

We have the following equation for the reliability function:

$$R_i(t) = exp(-(\lambda_i + \mu_{1,i} + \mu_{2,i})t) + \sum_{j=1}^{k} \int_0^t f_{i,j}(\tau) R_j(t-\tau)\, d\tau \\ + \sum_{\nu=1}^{2} \int_0^t \mu_{\nu,i} exp(-(\lambda_i + \mu_{1,i} + \mu_{2,i})\tau)\overline{A}_\nu(t-\tau) \sum_{j=1}^{k} \widetilde{P}_{j=1}^{(3-\nu)}(t-\tau)d\tau, t \geq 0. \tag{5}$$

Let us introduce the following notation: $f(t) = (f_{i,j}(t))$ is $k \times k$-matrix, $R(t) = (R_1(t), \ldots, R_k(t))^T$ and $H(t) = (H_1(t), \ldots, H_k(t))T$ are column vectors, where

$$H_i(t) = exp(-(\lambda_i + \mu_{1,i} + \mu_{2,i})t)$$
$$+ \sum_{\nu=1}^{2} \int_0^t \mu\nu, iexp(-(\lambda_i + \mu_{1,i} + \mu_{2,i})\tau)\overline{A}_\nu(t-\tau) \sum_{j=1}^{k} \widetilde{P}_{i,j}^{(3-\nu)}(t-\tau)d\tau, t \geq 0. \quad (6)$$

Then the matrix form of 5 is the following:

$$R(t) = H(t) + \int_0^t f(\tau)R(t-\tau)d\tau, t \geq 0. \quad (7)$$

The solution of this matrix renewal equation is such [6]:

$$R(t) = H(t) + \int_0^t u(\tau)H(t-\tau)d\tau, t \geq 0, \quad (8)$$

where $u(t)$ is *the renewal matrix-function*:

$$u(t) = \sum_{\eta=2}^{\infty} f^\eta, t \geq 0, f^{(1)}(t) = f(t), f^{\eta+1}(t) = \int_0^t f(\tau)f^\eta(t-\tau)d\tau, t = 1, 2, \ldots \quad (9)$$

7 Numerical Example

Our numerical example has the following initial data. The matrix of the transition rates of Markov chain $J(t)$ is the following:

$$\lambda = \begin{pmatrix} 0 & 0.2 & 0.3 \\ 0.4 & 0 & 0.2 \\ 0.2 & 0.2 & 0 \end{pmatrix}$$

The parameter matrix for the exponential distributions of a time till the failure

$$\mu = \begin{pmatrix} 0.4 & 0.2 & 0.7 \\ 0.3 & 0.0 & 0.4 \end{pmatrix}$$

The sojourn time in the failed state has: uniform distribution with parameters $a = 0$ and $b = 4$ for the first component and Weibull distribution with parameter $\beta = 1.5$ and $c = 2$ for the second component. The corresponding probability densities and distribution functions are the following:

$$\alpha_1(t) = \begin{cases} \frac{1}{b-a}, a < t < b, \\ 0, otherwise, \end{cases}$$

$$A_1(t) = \begin{cases} 0, t < a, \\ \frac{t-a}{b-a}, a \le t < b, \\ 1, t > b, \end{cases}$$

$$\alpha_2(t) = \begin{cases} 0, t < 0, \\ \frac{c}{\beta}(\frac{t}{\beta}^{c-1} exp(-(\frac{t}{\beta})^c), t \ge 0, \end{cases}$$

$$A_2(t) = \begin{cases} 0, t < 0, \\ 1 - exp(-(\frac{t}{\beta})^c), t \ge 0. \end{cases}$$

Further calculation results are presented. The probability densities $f_{i,j(t)}$ are calculated according to 4. To simplify next calculations, an approximation of these functions is performed. Functions $\widetilde{f}_{i,j}(t)$ for $i \ne j$ are approximated as

$$\widetilde{f}_{i,j}(t) = exp\Big(-\sum_{\eta=0}^{5} \beta_\eta(i,j) t^{\eta/2}\Big), t \ge 0. \tag{10}$$

Values of approximation coefficients are presented in Table 1.

Table 1.

ij	$\eta = 0$	$\eta = 1$	$\eta = 2$	$\eta = 3$	$\eta = 4$	$\eta = 5$
12	−1.609	0.073	−1.570	0.523	−0.162	0.021
13	−1.203	0.079	−1.588	0.538	−0.164	0.021
21	−0.917	0.081	−1.047	0.262	−0.088	0.010
23	−1.611	0.91	−1.061	0.249	−0.061	0.003
31	−1.611	0.509	−3.317	2.019	−0.614	0.067
32	−1.610	0.336	−2.742	1.415	−0.397	0.041

Functions $f_{i,i}(t)$ are approximated by the density of Gamma distribution:

$$\widetilde{f}_{i,j}(t) = \frac{1}{\Gamma(c_i)} v_i^{c_i} t^{c_i - 1} exp(-v_i t), t \ge 0. \tag{11}$$

Values of approximation coefficients are the following: $c_1 = 2.121, v_1 = 1.235; c_2 = 2.203, v_2 = 1.085; c_3 = 2.045, v_3 = 1.228$.

The probability density of the time till recurrent event 4 and its approximation $fApp(t)_i$ are presented in Fig. 1.

Considered approximations are used for the calculation of the renewal matrix-function 8. This function is presented by matrix for any fixed time moment $t = \Delta\xi, \xi = 0, 1, \ldots$, where $\Delta > 0$ is mesh width.

The graphics of functions 6 are presented in Fig. 2.

The last figure presents the reliability function 8. The graphic $Reliab(t, 0.2, 3)_i$ corresponds to the reliability 8 if the i-th state of the random environment takes place initially, the time t is considered with mesh width $\Delta = 0.2$, and the number of the summands in the infinite sum 9 equals 3 (Fig. 3).

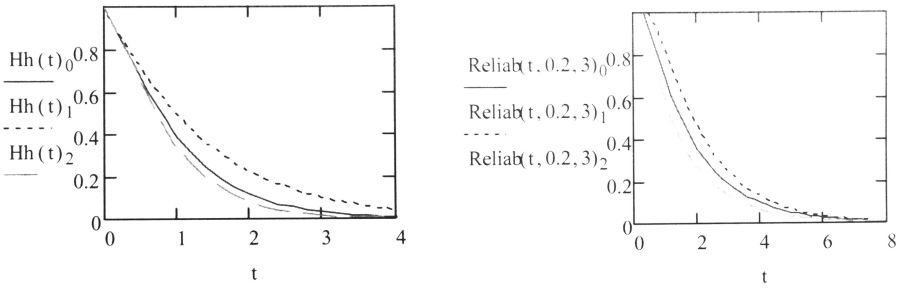


Fig. 1. Graphics of density 4 and its approximation fApp(t) for i = 0

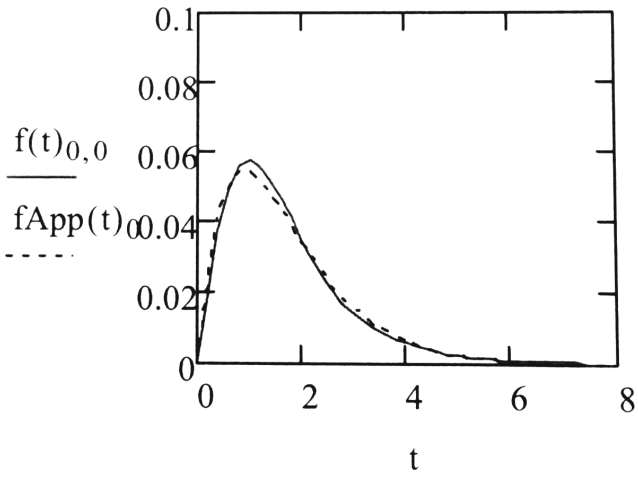


Fig. 2. Graphics of functions (5.2)

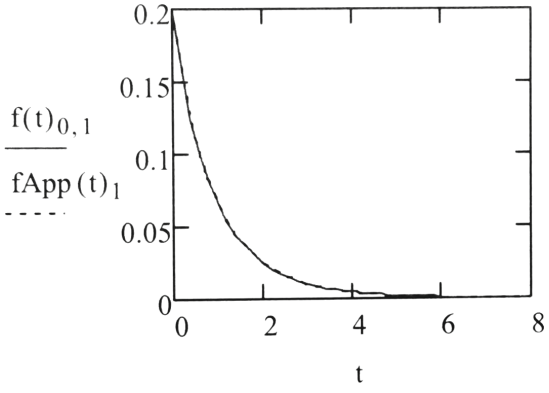


Fig. 3. Graphics of reliability functions (5.4)

8 Conclusions

Considered approach is used for the analysis and optimization of the hybrid communication channels [13]. The authors intent to continue current investigation in two directions in future. Firstly, to consider a case of three parallel channels. Secondly, to reject a supposition about an independence of sojourn times in the state 1 on the random environment state.

References

1. Andronov, A.M.: Markov-modulated birth-death processes. Autom. Control Comput. Sci. **45**(3), 123–132 (2011)
2. Andronov, A.M., Vishnevsky, V.M.: Algorithm of state stationary probability computing for continuous-time finite Markov chain modulated by semi-Markov process. In: Vishnevsky, V., Kozyrev, D. (eds.) DCCN 2015. CCIS, vol. 601, pp. 167–176. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30843-2_18
3. Andronov, A., Gertsbakh, I.B.: Signatures in Markov-modulated processes. Stochast. Models **30**, 1–15 (2014)
4. Bellmar, R.: Introduction to Matrix Analysis. McGraw Hill Book Company, New York (1969)
5. Efrosinin, D., Rykov, V.: Sensitivity analysis of reliability characteristics to the shape of the life and repair time distributions. In: Dudin, A., Nazarov, A., Yakupov, R., Gortsev, A. (eds.) ITMM 2014. CCIS, vol. 487, pp. 101–112. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13671-4_13
6. Feller, W.: An Introduction to Probability Theory and its Applications, vol. II. Wiley, New York, London, Sydney, Toronto (1971)
7. Gnedenko, B.V., Belyaev, Yu.K., Solovyev, A.D.: Mathematical Methods of Reliability. Academic Press, New York (1969)
8. Kim, C.S., Dudin, A., Klimenok, V., Khramova, V.: Erlang loss queueing system with butch arrivals operating in a random environment. Comput. Oper. Res. **36**(3), 674–697 (2009)
9. Kim, C.S., Klimenok, V., Mushko, V., Dudin, A.: The BMAP/PH/N retrial queueing system operatinh on Markovian random environment. Comput. Oper. Res. **37**(7), 1228–1237 (2010)
10. Pacheco, A., Tang, L.C., Prabhu, N.U.: Markov-Modulated Processes and Semiregenerative Phenomena. World Scientific, New Jersey, London (2009)
11. Rykov, V.: Multidimensional alternative processes reliability models. In: Dudin, A., Klimenok, V., Tsarenkov, G., Dudin, S. (eds.) BWWQT 2013. CCIS, vol. 356, pp. 147–156. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-35980-4_17
12. Sranavasan, S.K., Gopalan, M.N.: Probabilistic analysis of a two-unit system with a warm standby and a single repair facility. Oper. Res. **21**(3), 748–754 (1973)
13. Vishnevsky, V.M., Semenova, O.V.: Modeling and analysis of a hybrid communication channels based on free-space optical and radio-frequency. Autom. Remote Control **72**, 345–352 (2013)

Self Rising Tri Layers MLP for Time Series Forecasting

T. D. Balabanov(✉), I. I. Blagoev, and K. I. Dineva

Institute of Information and Communication Technologies,
Bulgarian Academy of Sciences, acad. Georgi Bonchev Str, block 2,
office 514, 1113 Sofia, Bulgaria
todorb@iinf.bas.bg, {i.blagoev,k.dineva}@iit.bas.bg

Abstract. Time series forecasting is an attractive and heavily researched area. A very popular approach in this field is the usage of artificial neural networks. Some artificial neural network are oriented to deep learning as training algorithm. In this study instead of hidden layers number extension the size of input layer of tri layers multilayer perceptron is extended. The network starts with 1-1-1 topology. The input layer rise to n, according the size of input time series. In parallel hidden layer goes to m by application of pruning algorithm. Achieved topology n-m-1 is trained with classical backpropagation of the error.

Keywords: Data mining · Time series forecasting
Artificial neural networks

1 Introduction

There are numerous techniques applied in the field of times series forecasting [1]. Artificial neural networks are one technique which is very successfully applied for such forecasting. Time series are values measured in the time by keeping strict order of the measurements (time-value pairs) [2]. In most cases measurement interval is fixed, but expectations are also possible. The idea in time series is that values are not independent in time. Values are related in such way that future values are dependent from the past values. Forecasting problem is defined as - by knowing past values to do a prediction for the future values. In order such forecasting to be successful prediction model construction is needed. Artificial neural networks are one of the proven models in time series forecasting. Initially artificial neural networks were inspired by biological neural systems. First appearance of artificial neural networks was in the middle of 20th century [3]. The most commonly used artificial neural networks are oriented weighted graphs. Nodes of the graph are called neurons. The links between the neurons have weights and these weights are the core of the information presented in the network. Artificial neural networks are working in two common modes - training

This work was supported by private funding of Velbazhd Software LLC.

V. M. Vishnevskiy and D. V. Kozyrev (Eds.): DCCN 2018, CCIS 919, pp. 577–584, 2018.
https://doi.org/10.1007/978-3-319-99447-5_50

and operation. The training mode is executed as an optimization task in which weights in the network should be modified in such way in which the network will learn the training patterns best. There are a lot of training algorithms developed during last four decades, but the most popular one is the backpropagation of the error. Backpropagation of the error is an exact numerical method and it is the prefered training method in this study. The idea is a minimization of the total neural network error achieved during processing of all training examples. The gradient of the total error is used for weights updating as direction of the update and the magnitude of the update. The way in which links between neurons are organized is common for the artificial neural network topology. There are a lot of different topologies widely investigated in the literature, like generalized nets [4] or deep learning neural networks. When the time series are too noise the input information can be filtered with Kalman filter for example [5].

In this study the main idea used into deep learning neural networks is reverted and instead of hidden layers number rising the size of the input and the hidden layers are extended during the neural network training. Extension of the input layer is related with the fact that each time series rise by appearance of a new measurement. The the goal of the training is the size of the input layer to get as big as the size of the full time series.

The paper is organized as follows: Sect. 1 introduces the problem; Sect. 2 presents a model and optimization approach; Sect. 3 gives experiment details; Sect. 4 concludes and some further ideas for research are pointed.

2 Model Proposition

Conditionally the time series is divided to past and future. The values supplied into the input of the artificial neural network are called lag and they are subset of the past values nearest to the future values. The values obtained in the output of the artificial neural network are the prediction and they are compared with the subset of the future values called lead. As artificial neural network base, for the model proposed, multilayer perceptron is used with input, one hidden and output layers.

In the proposed model set of artificial neural sub-networks is used and sub-networks are merged into a general artificial neural network. The smallest artificial neural sub-network has 1-1-1 topology (Fig. 1-left). The network is trained with examples which input has only single value. The goal in the model is a forecast for only one value ahead in time. That is why all sub-networks have only single output. Figure 1-left shows only 3 intermediate examples of the training. All 29 input values are supplied as training examples to resilient backpropagation training. Training stops on certain epsilon level for total neural network error change.

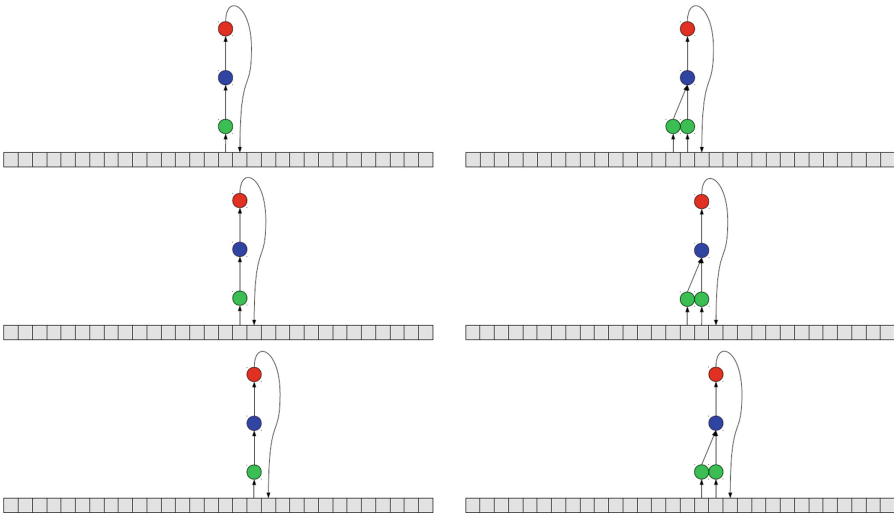


Fig. 1. Training of artificial neural sub-networks with 1-1-1 topology (left) and 2-1-1 topology (right).

After training of 1-1-1 topology the weights values of the first sub-network are loaded into second sub-network with 2-1-1 topology (Fig. 1-right). It is obvious that one of the weights would not be loaded, because it is not presented in the first sub-network. This weight has the value from the previous training of the biggest sub-network. Time series is reorganized to supply two values for input and to expect one forecast value in the output. For the second sub-network there are 28 input examples and single output is expected. Training is the same as with the first sub-network - resilient backpropagation training. As with the first sub-network, training stops on certain epsilon level for total neural network error change.

Third sub-network has 3-2-1 topology. The size of the hidden layer is automatically selected by incremental pruning algorithm implemented in Encog Machine Learning Framework [6]. Figure 2-left shows two neurons in the hidden layer, but it is only illustrative the real size of the hidden layer is estimated by the algorithm. Training algorithm and stop criteria are the same as with the previous sub-networks.

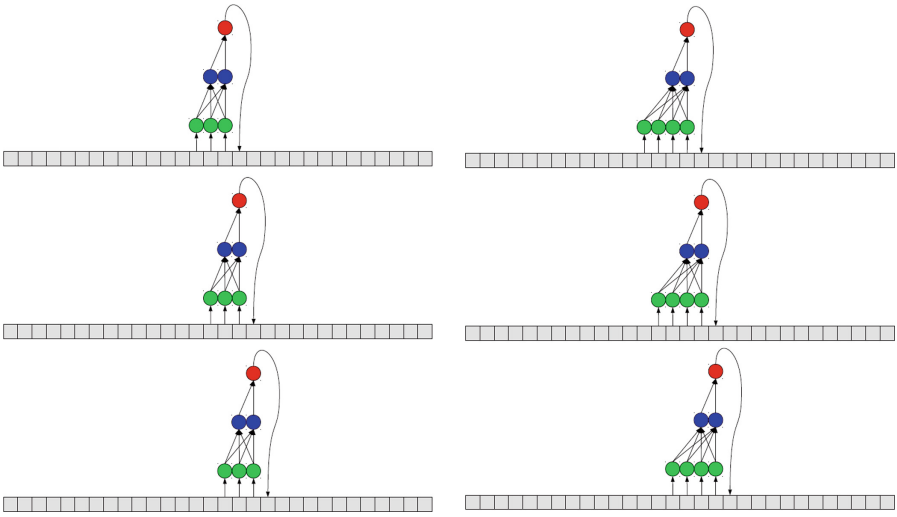


Fig. 2. Training of artificial neural sub-networks with 3-2-1 topology (left) and 4-2-1 topology (right).

Fourth sub-network has 4-2-1 topology and once again the size of the hidden layer is only illustrative (Fig. 2-right). The real size of the hidden layer is estimated by incremental pruning algorithm. Training examples are one less than with the previous sub-network, because the input size is bigger by one. Training and stopping criteria are the same as with the previous sub-networks. Figures 1 and 2 show only initial 4 sub-networks. In the model implementation many more sub-networks are involved. Sub-networks typologies are formed by addition of a single neuron in the input layer and adjustment of the hidden layer size with incremental pruning algorithm. The final goal is to reach n-m-1 topology (Fig. 3), which covers all known time series values. Most of the links between the input and the hidden layers in Fig. 3 are missed for better visualization, but both layer are fully-connected in the model implementation.

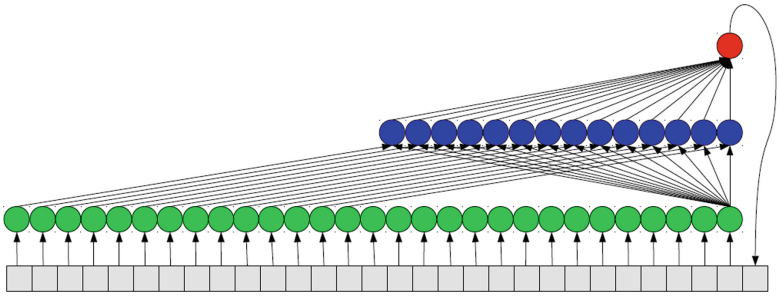


Fig. 3. Training of artificial neural sub-network with n-m-1 topology. Some of the links between input and hidden layer are not visualized for better appearance.

After the training of the biggest sub-network hierarchical procedure goes back in a loop to the smallest sub-network. Values of the weights from the biggest sub-network, which correspond to the links in the smaller sub-network, are taken and are loaded in the smallest sub-network. In a similar manner weights are taken from the biggest sub-network for the other sub-networks in combination with the weights of the previous smaller sub-network weights. For example, the sub-network with 4-2-1 topology will take some of its weights from 3-2-1 sub-networks, but links which are not presented into the smaller sub-network will be taken from the biggest sub-network.

The common idea behind the proposed model is the incremental training of racing in size artificial neural networks. Such training is inspired by the natural neural systems where biological cells grow in number and form connections between each other. The common problem in artificial neural networks training is the size of the network. By splitting the biggest network in many smaller networks speed up of the training process is achieved. It is very well known in the time series forecasting field that the oldest measurements have the smallest impact for the forecast. The proposed model takes this fact into account and the oldest measurements are taken in the biggest sub-network, but they have relatively smaller impact in the final forecast. The proposed model has higher degree of self-adaptation, because when a new value in the time series appears the size of the artificial neural network grows, which means that training phase and operation phase are simultaneous.

3 Experiments and Results

All experiments are done as Java program where the artificial neural networks are implemented with the API provided by Encog Machine Learning Framework [6]. All input neurons do not have activation function, because their task is only to supply the input signals into the hidden layer. The neurons in the hidden and the output layer are used with hyperbolic tangent function. Hyperbolic tangent is preferred instead of the sigmoid function, because it has symmetry against

X axis. This symmetry helps for the training speed up when backpropagation training is used, because the output values of the neurons have positive and negative values. With the sigmoid function output of the neurons is only positive and negative signals (if they are needed) can be achieved only by negative weights.

Fig. 4. Currencies values for two months on daily basis - EUR/USD currency pair.

As input data for the experiments FOREX financial time series are used (Figs. 4 and 5). Date are taken for daily trading of two months for EUR/USD and USD/JPY currencies pairs. Time series values are scaled in the range of −0.99 to +0.99 with MinMax scaling rule. The output of the artificial neural network is re-scaled back to the original rage with the same rule, but used in the opposite direction.

The results of the experiments are still in the range of the statistical error, which comes from the complexity of the financial processes and the high-frequency noise inside the data.

Fig. 5. Currencies values for two months on daily basis - USD/JPY currency pair.

4 Conclusion

The proposed model for self rising tri layers MLP for time series forecasting is a promising approach for artificial neural networks training speed-up. The rising size of the input layer involve a maximum information available in the time series, but the proposed procedure for artificial neural network training takes in account that older values should be less informative. As further research it will be interesting such self-rising training to be implemented az parallel computing solution.

References

1. Atanasova T., Barova, M.: Exploratory analysis of time series for hypothesize feature values. In: Proceedings of International Scientific Conference UniTech17, Gabrovo, Bulgaria, vol. 2, pp. 399–403 (2017). ISSN 1313-230X
2. Balabanov, T., Zankinski, I., Dobrinkova, N.: Time series prediction by artificial neural networks and differential evolution in distributed environment. In: Lirkov, I., Margenov, S., Waśniewski, J. (eds.) LSSC 2011. LNCS, vol. 7116, pp. 198–205. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-29843-1_22. ISBN 978-3-642-29842-4

3. Balabanov, T.: Long short term memory in MPL pair. In: Proceedings of the International Scientific Conference UniTech17, Gabrovo, Bulgaria, vol. 2, pp. 375–379 (2017). ISSN 1313-230X
4. Tashev, T., Hristov, H.: Modeling of synthesis of information processes with generalized nets. In: Drinov, M. (ed.) Cybernetics and Information Technologies, vol. 2, pp. 92–104. Academic Publishing House, Sofia (2003)
5. Alexandrov, A.: Ad-hoc Kalman filter based fusion algorithm for real-time wireless sensor data integration. Flexible Query Answering Systems 2015. AISC, vol. 400, pp. 151–159. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-26154-6_12. ISBN 978-3-319-26153-9
6. Heaton, J.: Encog Machine Learning Framework. Heaton Research, Inc. http://www.heatonresearch.com/encog/

Author Index