Российская академия наук (РАН)

Институт проблем управления им. В.А. Трапезникова
Российской академии наук (ИПУ РАН)

Российский университет дружбы народов (РУДН)

Институт информационных и телекоммуникационных технологий
Болгарской академии наук (София, Болгария)

Национальный исследовательский Томский государственный университет
(НИ ТГУ)

Научно-производственное объединение
«Информационные и сетевые технологии» («ИНСЕТ»)

РАСПРЕДЕЛЕННЫЕ КОМПЬЮТЕРНЫЕ И ТЕЛЕКОММУНИКАЦИОННЫЕ СЕТИ: УПРАВЛЕНИЕ, ВЫЧИСЛЕНИЕ, СВЯЗЬ (DCCN-2023)

МАТЕРИАЛЫ XXVI МЕЖДУНАРОДНОЙ НАУЧНОЙ КОНФЕРЕНЦИИ
(25–29 СЕНТЯБРЯ 2023 г., МОСКВА, РОССИЯ)

Под общей редакцией д.т.н. В.М. Вишневского, д.т.н. К.Е. Самуйлова

Москва
ИПУ РАН
2023

Russian Academy of Sciences (RAS)
V.A. Trapeznikov Institute of Control Sciences of RAS (ICS RAS)
Peoples' Friendship University of Russia (RUDN University)
Institute of Information and Communication Technologies of Bulgarian Academy of Sciences (Sofia, Bulgaria)
National Research Tomsk State University (NR TSU)
Research and development company
"Information and networking technologies"

DISTRIBUTED COMPUTER AND COMMUNICATION NETWORKS: CONTROL, COMPUTATION, COMMUNICATIONS (DCCN-2023)

PROCEEDINGS OF THE XXVI INTERNATIONAL SCIENTIFIC CONFERENCE
(September 25–29, 2023, Moscow, Russia)

Under the general editorship of D.Sc. V.M. Vishnevskiy, D.Sc. K.E. Samouylov

MOSCOW
ISC RAS
2023

УДК 004.7:004.4].001:621.391:007
ББК 32.973.202:32.968
Р 24

Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь (DCCN-2023) = Distributed computer and communication networks: control, computation, communications (DCCN-2023) **:** материалы XXVI Междунар. научн. конфер, 25–29 сент. 2023 г., Москва / под общ. ред. В.М. Вишневского, К.Е. Самуйлова; Ин-т проблем упр. им. В.А. Трапезникова Рос. акад. наук. – Электрон. текстовые дан. (1 файл: 26,1 Мб). – Москва : ИПУ РАН, 2023. – 1 электрон. опт. диск (CD-R). – Мин. систем. требования: Pentium 4; 1,3 ГГц и выше; Windows 7/8; Acrobat Reader 4.0 и выше. – Загл. с экрана. – ISBN 978-5-91450-269-7. – № госрегистрации 0322303901. – Текст: электронный.

В научном электронном издании представлены материалы XXVI Международной научной конференции «Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь» по следующим направлениям:

- Алгоритмы и протоколы телекоммуникационных сетей
- Управление в компьютерных и инфокоммуникационных системах
- Анализ производительности, оценка QoS / QoE и эффективность сетей
- Аналитическое и имитационное моделирование коммуникационных систем последующих поколений
- Эволюция беспроводных сетей в направлении 5G;
- Технологии сантиметрового и миллиметрового диапазона радиоволн;
- RFID-технологии и их приложения;
- Интернет вещей и туманные вычисления
- Системы облачного вычисления, распределенные и параллельные системы
- Анализ больших данных
- Вероятностные и статистические модели в информационных системах
- Теория массового обслуживания, теория надежности и их приложения
- Высотные беспилотные платформы и летательные аппараты: управление, передача данных, приложения

В материалах научной конференции DCCN-2023, подготовленных к выпуску к.ф.-м.н. Козыревым Д.В., обсуждены перспективы развития и сотрудничества в этой сфере.

Сборник материалов конференции предназначен для научных работников и специалистов в области управления крупномасштабными системами.

Утверждено к изданию Программным комитетом конференции

Текст воспроизводится в виде, представленном авторами

ISBN 978-5-91450-269-7

Содержание / Contents

UDC: 519.234.6

Information Spreading in Evolving Networks with node and edge deletion *

Natalia M. Markovich[1] and Maksim S. Ryzhov[1]

[1]V.A. Trapeznikov Institute of Control Sciences, Russian Academy of Sciences ,
Profsoyuznaya Str. 65, 117997 Moscow , Russia
markovic@ipu.rssi.ru, nat.markovich@gmail.com,maksim.ryzhov@frtk.ru

Abstract

A preferential attachment (PA) has been suggested to model network evolution and to explain conjectured power-law node degree distributions in real-world networks. In Markovich, Ryzhov (2022a,b), the schemes of the linear PA proposed in Wan et al. (2020) for the network evolution were suggested for information spreading. The PA and the well-known algorithm SPREAD proposed in Mosk-Aoyama, Shah (2006) were compared regarding the minimum number of evolution steps K^* required to spread a single message among a fixed number of nodes in non-homogeneous directed networks. This comparison was done in Markovich, Ryzhov (2022a,b) without node and edge deletion during the evolution. The objective of the current study is to investigate the impact of the PA parameters on spreading of a single message to a fixed number of nodes in the graph when an existing node or edge is uniformly deleted at each step of the PA evolution. The results are provided for simulated graphs.

Keywords: evolving network, information spreading, linear preferential attachment, SPREAD algorithm, node and edge deletion

1. Introduction

Let $G_t = (V_t, E_t)$ be the graph with sets of nodes V_t and edges E_t generated by the attachment of a new node and a new edge at time t. The preferential attachment (PA) has been suggested to model network evolution and to explain conjectured power-law degree distributions in real-world networks [1]. Such networks contain rare giant nodes with a large number of links (node degrees). The node degrees grow fast at each evolution step since new nodes prefer to be attached to valuable nodes. Here, we focus on the rate of the spreading of one message among a fixed number n of nodes in the direct networks. The information spreading arises particularly for

*The authors were supported by the Russian Science Foundation RSF, project number 22-21-00177.

distributed computation [2]. In [3], [4] the linear PA $\alpha-$, $\beta-$, $\gamma-$ schemes proposed in [5] to model the network evolution were suggested for the information spreading. A message to be spread among n nodes is assumed to be in the disposal of some node set $S_t \subset V_t$ at time t. A new node v_t that is appended to the network at a time t may get the message from one of the nodes in S_t (let say, node i) if v_t follows the node i, i.e. a new edge is created such that it is directed from v_t to $i \in S_t$. This case corresponds to the $\alpha-$scheme where the directed edge is created from a "new" node to an "old" one. For example, if a customer follows a Web page, then he(she) can upload the content of the page. Peer-to-peer networks provide another example when a follower tends to attach to a node with a large number of peers to upload a required peer. Another way to spread an information is to create a new edge between a pair of existing nodes i and j. If the node i has the message, i.e. $i \in S_t$ and the node j has not for the time t, and if there is an edge $j \to i$, then node i shares the message with node j. We assume that the message cannot be spread to the node j if the edge $i \to j$ is created at the evolution step. Both existing nodes may have not the message and then no sharing is happened. The $\beta-$scheme corresponds to the creation of the edge with any direction between two "old" nodes.

The PA spreading has been compared with a well-known algorithm SPREAD proposed in [2] by the minimum number of evolution steps K^* required to spread a message among n nodes in directed networks [3], [4]. The idea of the original SPREAD is to share all messages nodes have among all nodes in an undirected network. In [4] the modified SPREAD has been applied to spread a single message among n nodes in directed networks. The latter comparison was done in [3], [4] without nodes and edges deletion at each step of the evolution. Less attention in the literature related to the PA is devoted to the node and edge deletion, e.g. [5], [6], [7]. In [6] the node deletion is interpreted in the context of agent systems. Then the collapsing agent that loses its incoming connections (consumption links) leads to the breaking of some production links of its neighbors, and thus also leading to their collapse. In [5] the node deletion is provided in the evolving graph at the final time of the evolution but not permanently at each step when a new node is appended. In [7] the tail index of the PageRank and the Max-linear model used as node influence characteristics is estimated nonparametrically for graphs generated by the PA $\alpha-$, $\beta-$, $\gamma-$ schemes where the existing node or edge is deleted.

Our objective is to extend the results in [3], [4] by investigating the impact of the PA parameters α, β, γ on the spreading of one message to n nodes in the graph, when a node or an edge is uniformly deleted at each step of the evolution. We aim to compare the spreading rate of the linear PA schemes and of the SPREAD algorithm for different values of α, β, γ.

The message may be lost in the network due to the deletion of nodes with the message. Such situation is impossible for classical SPREAD where nodes share all messages what they have with each other. If an edge is deleted, then the message cannot be lost as in the case of node deletion, but it may not be transmitted further.

The paper is organized as follows. In Section 2, related works regarding the linear PA schemes (Section 2.1) and the spreading information (Section 2.2) are described. In Section 3, our main results concerning the spreading with the node or edge deletion for simulated graphs are presented. We end with conclusions.

2. Related works

2.1. Preferential attachment. The linear PA schemes [1], [5] start with an initial directed graph $G(k_0)$ with at least one node and k_0 edges and construct a growing sequence of directed random graphs $G(k) = (V(k), E(k))$ for evolution steps $k \geq 1$. A graph $G(k)$ is produced from $G(k-1)$ by adding a directed edge. Let us denote the number of nodes at step k as $N(k)$, and in- and out-degree of node w in the graph $G(k)$ with k edges as $I_k(w)$ and $O_k(w)$. The edge creation proposed in [1], [5] is activated by flipping a 3-sided coin with probabilities α, β and γ such that $\alpha + \beta + \gamma = 1$. The independent identically distributed (i.i.d.) trinomial r.v.s with values 1, 2 and 3 and the corresponding probabilities α, β and γ are generated to select schemes. Let $\Delta_{in}, \Delta_{out}$ be other PA parameters.

- By the α-scheme, one appends to $G(k-1)$ a new node $v \in V(k) \setminus V(k-1)$ and a new edge $(v \to w)$ to an existing node $w \in V(k-1)$ with probability α. The node w is chosen with probability depending on its in-degree in $G(k-1)$

$$P(\textit{choose } w \in V(k-1)) \;=\; \frac{I_{k-1}(w) + \Delta_{in}}{k - 1 + \Delta_{in} N(k-1)}.$$

- By the β-scheme, one adds a new edge $(v \to w)$ to $E(k-1)$ with probability β, where the existing nodes $v, w \in V(k-1)$ are chosen independently of the nodes of $G(k-1)$ with probabilities

$$P(\textit{choose } (v,w)) \;=\; \frac{O_{k-1}(v) + \Delta_{out}}{k - 1 + \Delta_{out} N(k-1)} \cdot \frac{I_{k-1}(w) + \Delta_{in}}{k - 1 + \Delta_{in} N(k-1)}.$$

- By the γ-scheme, one adds a new node $v \in V(k) \setminus V(k-1)$ and an edge $(w \to v)$ with probability γ. The existing node $w \in V(k-1)$ is chosen with probability

$$P(\textit{choose } w \in V(k-1)) \;=\; \frac{O_{k-1}(w) + \Delta_{out}}{k - 1 + \Delta_{out} N(k-1)}.$$

Note that $N(k) = N(k-1)$ for the β-schema and $N(k) = N(k-1) + 1$ for the others. These scenarios realize a 'rich-get-richer' mechanism, when a node with a large degree can likely increase it with a high probability. As mentioned in [5], the PA schemes allow creating multiple edges between two nodes and self loops.

2.2. Information spreading. Let us describe the idea of the information spreading first by the SPREAD method and then by the PA. We assume that all nodes in the network have asynchronized clocks. Let $k \geq 0$ denote the index of a tick, on which at most one node can receive messages by communicating with another node. To this end, on a clock tick one of n nodes (let say a node i) of the graph is chosen uniformly. Then the node i chooses a node j uniformly among its neighbors with probability $P_{ij} = 1/d_{\max}$, where $d_{\max} = \max_{i \in V} d_i$, d_i is the degree of node i. The usage $P_{ij} = 1/d_i$ as in [8] allows us to avoid the necessity to know the maximal node degree in the network.

In Algorithm 1 by [4], the SPREAD algorithm proposed in [2] for undirected graphs has been modified for directed graphs assuming that a single message of an initial graph G_0 can be spread to a part of the rest nodes. The node i may share the message with the node j if there is a directed edge from j to i. We use $P_{ij} = 1/I_i$, where I_i is a node in-degree in contrast to [4] where out-degree O_i was used instead of I_i. Such a mechanism simulates the message spreading when users search and collect data from pages in Internet and can share it further with other users through their web-pages. The next node j is proposed to select uniformly among nodes $V \backslash S(k)$ without the message at the clock tick k. $S(k)$ denotes the set of nodes that have the message at the end of the clock tick k. The linear PA can also be used for spreading using the directed edge $(j \to i)$ from the new node j to the old node i or between two existing nodes i and j [4]. The edge $(j \to i)$ can be created by the $\alpha-$ or $\beta-$ schemes. As in [4] we compare the SPREAD algorithm and the PA schemes by the minimal number of clock ticks or evolution steps required to disseminate the message from G_0 to n nodes with probability not less than $1 - \delta$, namely,

$$K^* \quad = \quad K^*(n, \delta) = inf\{0 < k \leq K' : Pr(\|S(k)\| = n) > 1 - \delta\}, \quad \delta \in (0, 1). \quad (1)$$

$\|S(k)\|$ is a cardinality of the set. The number of steps is bounded by K'. If $K^* \leq K'$ holds, then $S(K^*) = n$ is likely hold for a sufficiently small δ. If $S(K^*) < n$ holds, then K' evolution steps are likely not enough to disseminate the message to n nodes.

3. The PA and the SPREAD for simulated directed graphs

We compare a spreading ability of the PA and the SPREAD for simulated directed graphs and three deletion strategies. The first strategy 'without node and edge deletion' has been studied in [3], [4]. Here, we focus on two strategies of uniform node or edge deletion at each evolution step when a new node is appended.

Our experiment is the following. We generate 100 graphs by the PA schemes starting with an initial graph G_0 up to step $K^*(n, \delta) \leq K'$ with $K' = 2.5 \cdot 10^5$ and $\delta = 0.01$ in (1). We aim to spread a message from G_0 to $n = 100$ other nodes. The

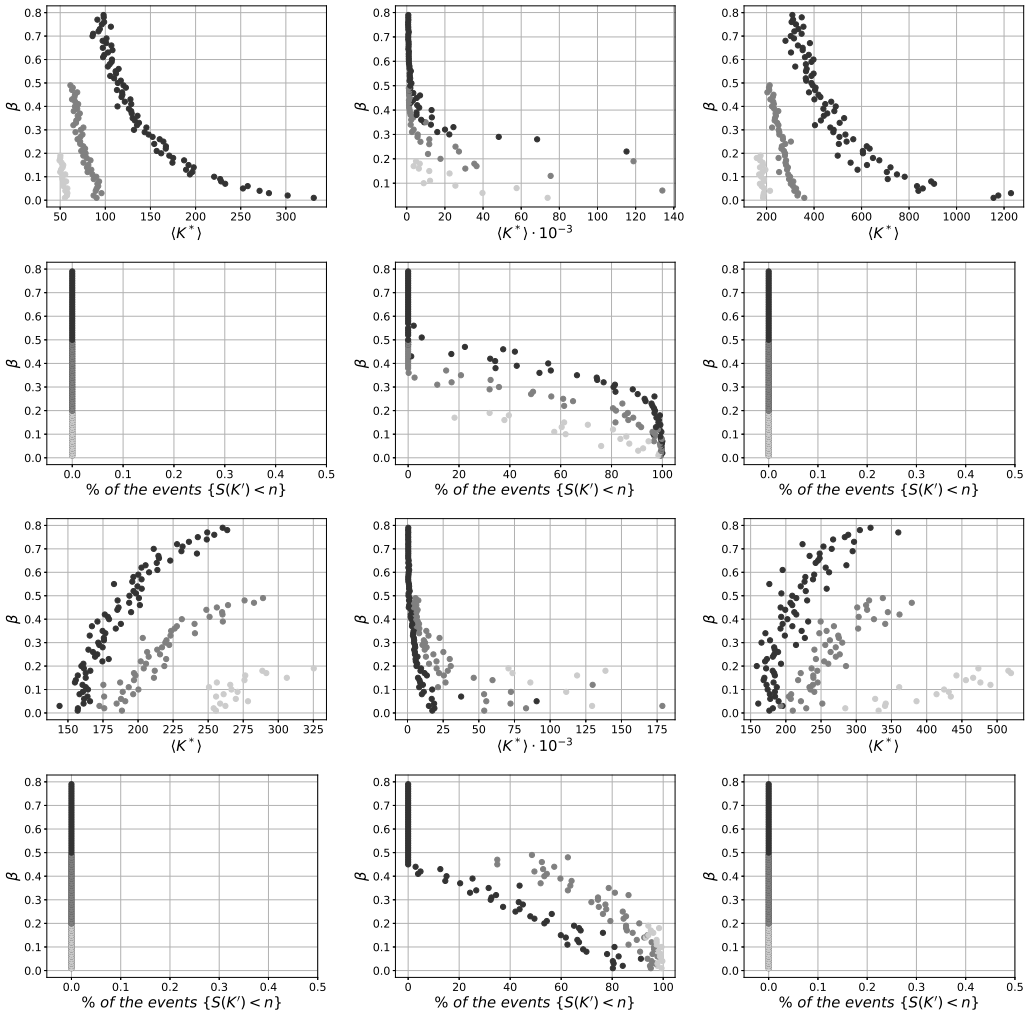


Figure 1. The PA parameter β against the average number of steps $\langle K^* \rangle$ over 100 graphs and against the proportion of the events $\{S(K') < n\}$ for the PA schemes without node and edge deletion (left column); with node deletion (middle column); with edge deletion (right column) for spreading by the PA schemes (two top lines) and by the SPREAD algorithm (two bottom lines). The dark, light dark and grey points correspond to $\alpha \in \{0.8, 0.5, 0.2\}$, respectively.

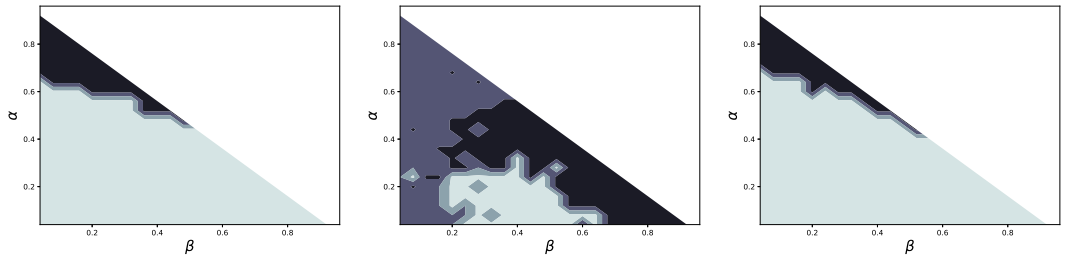
probability in (1) is approximated by a proportion of the event $\{\|S(k)\| = n\}$ for a given k over 100 graphs. A triangle of connected nodes has been used as G_0 for the PA evolution without node and edge deletion in [4]. If we use the triangle as G_0

for the evolution with node deletion at each step, then the evolving graph always contains the same 3 nodes as G_0 and a random number of edges. The number of edges depends on the frequency of using the $\beta-$scheme. The node is deleted with its edges. If an edge is deleted at each step k, then the graph may contain the same edges 3 as G_0 and k nodes. The number of isolated nodes grows up. All nodes having the message can be quickly lost during evolution with the node deletion, or the transmission of the message may become difficult for the case of the edge deletion. To give more opportunities for the message spreading in case of node or edge deletion, we take G_0 sufficiently large. G_0 is generated by 10^3 evolution steps by the PA schemes with parameters $(\alpha, \beta, \gamma, \Delta_{in}, \Delta_{out}) = (0.4, 0.2, 0.4, 1, 1)$. Starting with G_0, 100 graphs are evolved by the PA for each set of parameters α, β, γ which all are taken in the interval $[0.01, 0.99]$ with step 0.01 as far as $\Delta_{in} = \Delta_{out} = 1$. The comparison of the PA and the SPREAD algorithms is shown in Fig. 1. The PA schemes spread the information faster for large values of α and small values of β, that implies the spreading among newly appended nodes mostly, see Fig. 1 (top line). In contrast, the SPREAD is faster for small α's and small β's apart of the case with node deletion; Fig. 1 (third line). In the latter case, the SPREAD is faster if $\alpha + \beta$ is close to one, i.e. there are a few edges directed from the existing nodes to newly appended ones. The impact of α and β on the proportion of events $\{S(K') < n\}$, i.e. when the message cannot be delivered to n nodes by K' steps, is shown in Fig. 1 (second and fourth lines). The PA and the SPREAD without node and edge deletion and with edge deletion lead to the full spreading of the message among n nodes independently of α and β. The case of the node deletion is different. The PA may spread the message to all n nodes for any α if $\beta \geq 0.5$. For $\beta < 0.5$ the message will be delivered to a part of n nodes. Similar conclusions can be done for the SPREAD.

Fig.2 shows the impact of α and β on K^*_{PA}/K^*_{SPREAD}, where K^*_{PA} and K^*_{SPREAD} denote the minimum number of steps required for the PA and SPREAD algorithms, respectively, to spread a single message to $n = 100$ new nodes. The options $\alpha + \beta > 1$ were not considered due to condition $\alpha + \beta + \gamma = 1$. One can see areas where the PA is faster spreader than the SPREAD, i.e. $K^*_{PA} < K^*_{SPREAD}$, and vice versa. Fig. 2a and 2c look similar that is in agreement with Fig. 1.

4. Conclusions

We study the linear PA schemes as the tool to spread one message from an initial graph to n rest nodes of the network. We compare the PA and the SPREAD algorithm on directed simulated graphs generated by the PA $\alpha-$, $\beta-$, $\gamma-$schemes with different sets of the PA parameters. The graph evolution with the uniform node or edge deletion and without node and edge deletion is considered. One may conclude that the PA $\alpha-$ and $\beta-$ schemes may be the faster spreader than the SPREAD



(a) (b) (c)

Figure 2. The dependence of K^*_{PA}/K^*_{SPREAD} averaged over 100 graphs for the PA parameters α and β and the evolution without node and edge deletion (Fig. 2a), with uniform node deletion (Fig. 2b) and with uniform edge deletion (Fig. 2c). The dark area indicates $K^*_{PA} < K^*_{SPREAD}$, the grey area - $K^*_{PA} > K^*_{SPREAD}$. In Fig. 2b, the light dark area indicates that the message is missed for both PA and SPREAD algorithms.

algorithm for large values of α and for the evolution without node and edge deletion or with edge deletion. The PA may be faster than the SPREAD for smaller α in the case of node deletion. For node deletion, the message may be lost, and both spreading algorithms will not be effective. The extended version will contain an application to real temporal graphs. The future work may concern to finding the probability to lose the message due to the node deletion.

REFERENCES

1. Bollobás B., Borgs C., Chayes J., Riordan O. Directed Scale-Free Graphs // Society for Industrial and Applied Mathematics, USA, SODA '03. 2003. P. 132-139.
2. Mosk-Aoyama D., Shah D. Computing separable functions via gossip // In Proceedings of the 25th ACM symposium on Principles of distributed computing (PODC '06), ACM, New York, USA. 2006. P. 113-122.
3. Markovich N.M., Ryzhov M.S. Information Spreading with Application to Non-homogeneous Evolving Networks // CCIS. 2022a. V. 1552. P. 284-292.
4. Markovich N.M., Ryzhov M.S. Information Spreading and Evolution of Non-Homogeneous Networks // Adv Syst Sci Appl. 2022b. V. 2, P. 21-33.
5. Wan P., Wang T., Davis R. A., Resnick S.I. Are extreme value estimation methods useful for network data? // Extremes. 2020. V. 23 P. 171-195.
6. da Cruz J.P., Lind P.G. The bounds of heavy-tailed return distributions in evolving complex networks // *Physics Letters A*. 2013. V. 377 P. 189-194.

7. Markovich N.M., Ryzhov M.S., Vaičiulis M. Tail Index Estimation of PageRanks in Evolving Random Graphs // Mathematics. 2022. V. 10 I. 16 N. 3026.
8. Censor-Hillel, K., Shachnai, H. Partial Information Spreading with Application to Distributed Maximum Coverage // In Proceedings of the 29th ACM symposium on Principles of distributed computing (PODC '10), ACM, New York, USA. 2010. P. 161-170.

UDC: 51.74

On Comparison of Waiting Buffer Scheduling Methods in a Resource Loss System

V.A. Stepanov[1], A.V. Daraseliya[1], E.S. Sopin[1,2]

[1]Peoples' Friendship University of Russia (RUDN University), Moscow, Russian Federation

[2]Institute of Informatics Problems, Federal Research Center Computer Science and Control of Russian Academy of Sciences, Moscow, Russian Federation

{daraselia-av,sopin-es}@rudn.ru

Abstract

Fifth-generation (5G) networks are expected to revolutionize wireless communication by enabling faster and more reliable data transfer rates. However, the use of millimeter-wave frequencies in 5G networks introduces unique challenges in data transmission due to phenomena such as dynamic signal blockage, which can result in sudden and intense resource demands for sessions. This paper presents a mathematical model of a 5G base station using a resource allocation system with waiting and non-homogeneous resource requirements for sessions. The study aims to determine the most efficient way to select sessions from the waiting buffer to increase system performance and minimize the probability of blocking data sessions and reduce waiting time in the waiting buffer. As a result of the research, it was found, that the method that prioritized sessions with the highest resource requirements and took them from the waiting buffer until the system reached its limit, demonstrates the best system performance under certain parameters.

Keywords: 5G networks, millimeter-wave frequencies, dynamic signal blockage, data session continuity, resource allocation, queueing theory.

1. Introduction

The emergence of 5G technology has brought about significant advancements in the field of data transmission. With its high speed, low latency, and massive capacity, 5G promises to revolutionize the way we communicate and consume data. However, the implementation of 5G also poses unique challenges that must be addressed for the technology to reach its full potential. One of these challenges is the dynamic blocking

The research was funded by the Russian Science Foundation, project No.22-79-10128, `https://rscf.ru/en/project/22-79-10128/`.

of signals, which causes a sudden increase in resource demands for a particular session, leading to resource congestion and blocking of other sessions.

To address this challenge, several techniques have been proposed, including the use of Channel Quality Indicator (CQI) and Modulation and Coding Scheme (MCS) in 5G networks [1]. CQI and MCS are used to optimize data transmission by selecting the appropriate modulation and coding scheme based on the channel conditions. This optimization can significantly improve the efficiency of data transmission and reduce the likelihood of blocking.

In several studies [2, 3, 4], resource loss systems (LS) with finite resources and random requirements have been used as one of the tools for modeling 5G cellular networks. However, the authors did not take into account the possibility of storing sessions in the waiting buffer. Our previous research [5] has been conducted on the analysis of resource LSs with waiting buffer, for which the equilibrium equations for the stationary probabilities of the system, as well as, the loss probability of the system, average waiting time, average number of sessions, and average resource requirements of blocked sessions were analytically derived.

In the current paper, we extend this work [5] by considering the specific case of 5G networks and analyzing different methods for selecting sessions from the waiting buffer to minimize the probability of blocking and waiting time. The simulation model is implemented in Python using methods from [6], and the parameters of the system are defined by the number of servers, the number of available resource units, and the waiting buffer size. We compare six different methods of selecting sessions from the waiting buffer and evaluate their performance based on several metrics.

2. Model description

We consider a resource LS with finite resources R, limited number of servers N and waiting buffer size V, see Fig. 1. Sessions arrive according to the Poisson law with rate λ. Serving process of each session requires a server and a random number r_i of resources, $1 \leq i \leq N$, $0 < r_i \leq R$. If there are not available servers in the system, the sessions are placed into a waiting buffer. In case there is space available in the waiting buffer, the sessions are held until the system has sufficient available resources and available servers to handle them. Upon entering service, sessions are serviced with a service rate μ, and once their service is complete, they exit the system. Sessions that cannot be accepted into the waiting buffer due to lack of buffer space are lost from the system.

The performance of the system is evaluated in terms of various metrics, including the blocking probability, the average number of customers in service, the average number of occupied resources, the average number of customers in the buffer, and

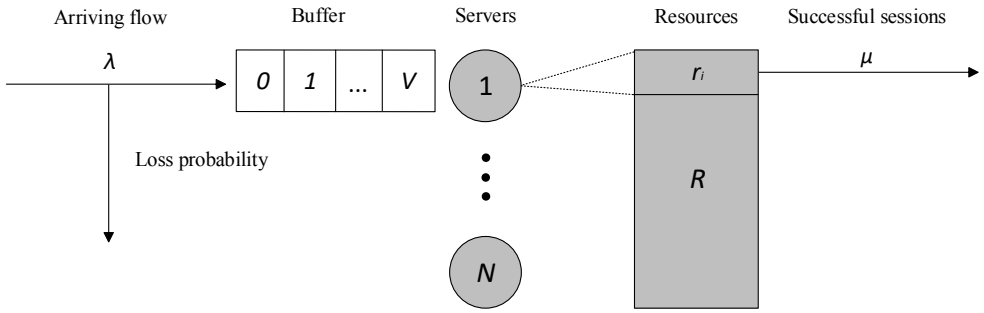


Fig. 1. Resource LS with the waiting buffer

the average waiting time in the buffer. These metrics of interest are calculated based on the chosen method of selecting sessions from the buffer for service.

Additionally, six different methods for selecting sessions from the buffer for service are studied:

1. service is provided to sessions in the buffer until the system runs out of resources or the buffer is full;
2. sessions are sorted in ascending order by resource requirement, and the ones with the lowest requirement are serviced first until the system runs out of resources or the buffer is full;
3. sessions are sorted in descending order by resource requirement, and the ones with the highest requirement are serviced first until the system runs out of resources or the buffer is full;
4. one random session is selected from the buffer for service [5];
5. sessions are selected at random from the buffer until the system runs out of resources or the buffer is full;
6. first-come, first-served (FCFS).

During the system analysis, we store the size of the number of occupied resources for each session in the system in a vector of size N. However, only the aggregate value for the entire system is used in the analysis. Moreover, the time a session spends in service is recorded when it is accepted into the system.

The behavior of the system can be described as a random process $X(t) = (\xi(t), \delta(t), \theta(t))$, where $\xi(t)$ represents the number of devices in the system,$\delta(t)$ represents the number of occupied resources and $\theta(t)$ represents the number of sessions in the waiting buffer. The state space is given by the following form:

$$S = \bigcup_{0 \le n \le N} S_n, S_n = \left\{ (n, m, r) : 0 \le m \le M\,, 0 \le n \le N, 0 \le r \le R, p_m^{(n)} > 0 \right\}. \quad (1)$$

Let's sort the set of states by the number of occupied resources and denote $I(n, m, r)$ as the ordinal number of the state (n, m). The steady-state probabilities of $X(t)$ are given by the expression (2).

$$q_n(m) = \lim_{t\to\infty} P\{\xi(t) = n, \delta(t) = m, \theta(t) = r, (n, m) \in S_n\}. \quad (2)$$

.

3. Numerical Results

For numerical analysis we consider a system with $N = 100$ servers, and $R = 100$ resource units, $V = 50$ buffer size. The service time for each request is exponentially distributed with $\mu = 1$. The system is modeled using two different resource requirement distributions: the geometric distribution with parameter values of $p = 0.7$ and the Poisson distribution with parameter 4.

The results of the simulation are presented in tabular form for ease of comparison between the different selection methods. Overall, the performance of the system is heavily dependent on the resource requirement distribution and the method used to select sessions for service.

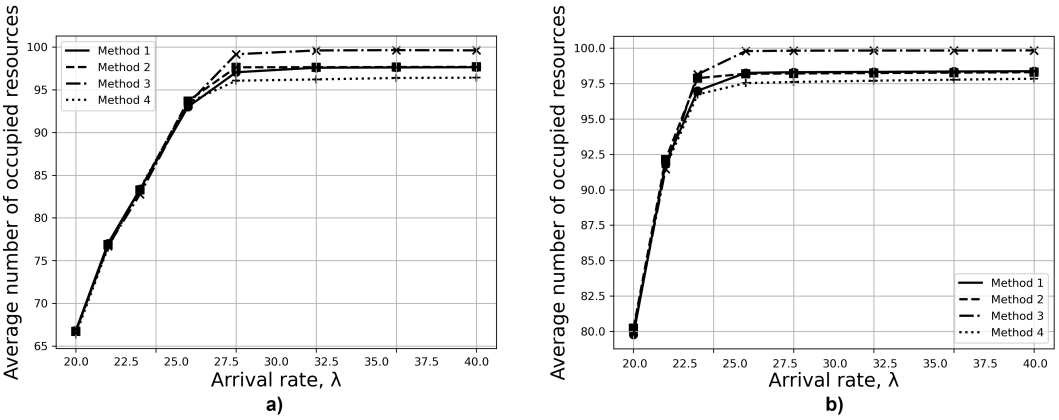


Fig. 2. Average number of occupied resources. a) Geometric distribution b) Poisson distribution

We start with Fig.2a and 2b illustrating the effect of the average number of occupied resourses on the arrival rate for the geometric and Poisson distributions, respectively. It is evident that under low load conditions, the methods exhibit approximately similar values. However, as the arrival rate increases, it becomes apparent that the method that prioritizes the most resource-demanding sessions from the waiting buffer demonstrates the highest average number of occupied resources in

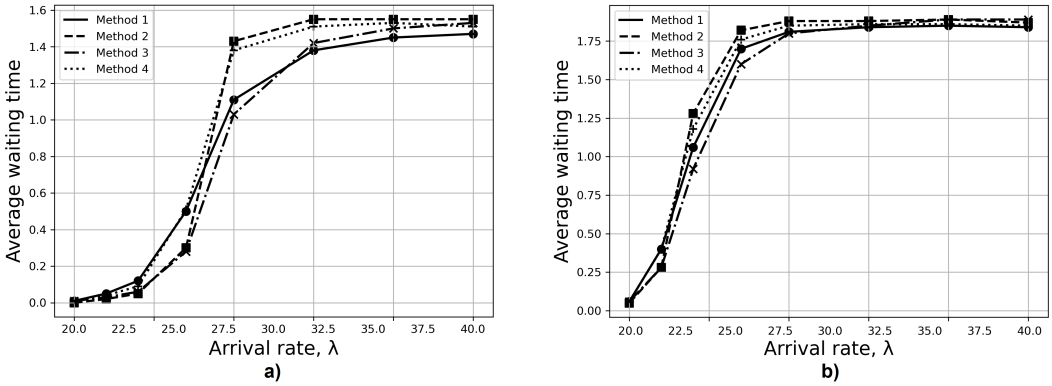


Fig. 3. Average waiting time. a) Geometric distribution b) Poisson distribution

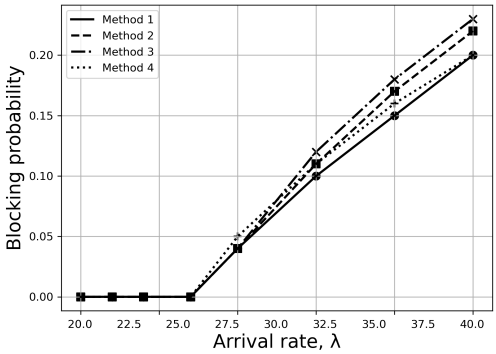


Fig. 4. Blocking probability. Geometric distribution

the system. This dependency is observed for both the geometric and Poisson distributions, with the only distinction being that the graphs for the Poisson distribution start to noticeably increase earlier.

Fig. 3a and 3b present the dependence of the average waiting time for sessions on the arrival rate. At low loads, a significant gain in time is observed for the method 3. In the case of the geometric distribution, at $\lambda = 26$, the waiting time is 67% less compared to other waiting buffer scheduling methods. At the incoming flow rate of $\lambda = 27$, the time difference is slightly less and amounts to about 20%. With further increase in the parameter, the average waiting time of method 3 relative to method 1 that provided to sessions in the waiting buffer until the system runs out of resources increases. For the Poisson distribution, the system behaves similarly, however with a smaller time advantage of method 3.

Fig. 4 shows the blocking probability for the geometric distribution. At low loads, the probability is zero because the number of occupied servers in the system has not

yet reached its maximum, as can be seen in Fig. 3a and 3b. With an increase in λ, the blocking probability also increases, which is due to the fact that a larger number of sessions enter the system, occupying resources and space in the waiting buffer.

4. Conclusion

In our paper we consider a model of a multi-server loss system with the multi-type of resources and waiting buffer and investigated the system performance indicators for various waiting buffer scheduling methods. The research revealed that, under specific parameters, the method that prioritized sessions with the highest resource requirements and served them until the system reached its capacity, demonstrated the most optimal system performance. The results of this study can be used in the design of 5G networks.

REFERENCES

1. 3GPP, TR 38.901, v.16.1.0. Evolved Universal Terrestrial Radio Access (E-UTRA); Carrier Aggregation; Base Station (BS) radio transmission and reception, 2020.
2. Naumov, V., et al.: 5G new radio system performance analysis using limited resource queuing systems with varying requirements. DCCN 2019, LNCS, vol. 11965, pp. 3–14. Springer, Cham (2019).
3. X. Lu et al.: Integrated Use of Licensed- and Unlicensed-Band mmWave Radio Technology in 5G and Beyond. IEEE Access, vol. 7, pp. 24376-24391. (2019).
4. Daraseliya, A.V. et al.: Analysis of 5G NR base stations offloading by means of NR-U technology. Inf. Appl. 15(3), 98–111 (2021).
5. Daraseliya, A., Sopin, E.S., Shorgin, S.Y., On Approximation of the Time-Probabilistic Measures of a Resource Loss System with the Waiting Buffer. DCCN 2022, CCIS, vol 1748. pp. 282–295. Springer, Cham (2023).
6. Buslenko N.P.,Simulation of complex systems [Modelirovaniye slojnih sistem], Ed. 2, revised. 1978.

УДК: 621.394/396

Решение для аутентификации устройств пользователей в корпоративной беспроводной сети университета с помощью механизма IEEE 802.1x

Д.К. Ярчук[1], М.А. Орлова[1], Л.И. Абросимов[1]

[1]Национальный Исследовательский Университет "Московский Энергетический Институт", Красноказарменная 14, с. 1, Москва, 111250, Россия

{YarchukDK, OrlovaMA, AbrosimovLI}@mpei.ru

Аннотация

В данной работе предложено решение для аутентификации пользователей с помощью механизма IEEE 802.1x с использованием разработанного веб-сервера для распределения цифровых сертификатов в корпоративной беспроводной сети университета.

Ключевые слова: Wi-Fi, IEEE 802.11, аутентификация, безопасность, идентификация пользователей

1. Введение

Идентификация пользователей в беспроводной сети согласно нормативным документам на 2023 г. [1, 2, 3] обязательна при предоставлении сетевых ресурсов пользователям. Идентификация проводится при подключении к сети и аутентификации. Бурный рост программного и аппаратного обеспечения приводит к появлению новых взломостойких методов аутентификации с новыми способами идентификации и аутентификации пользователя устройства имеющие различные преимущества и недостатки [4, 5, 6, 7, 8, 9]. В работе [4] авторы преложили поведенческий способ аутентификации на основании данных о сигнале и информации о сети на мобильном устройстве, собранных в течение 3 секунд, для определения легитимности устройства в сети. Но для применения данного метода необходимо выполнение продолжительных экспериментов для получения данных для составления прогноза и данный метод не применим для сети технического университета, в котором множество устройств, вносящих помеху и их включение/выключение не регламентировано по времени. В работе [5] авторы предложили архитектуру безопасного беспроводного города (SWC) с поддержкой Wi-Fi и WiMAX. На основе архитектуры SWC представлена аутентификация пользователя в сочетании

с расширяемым протоколом аутентификации, но в случае применения данного метода остаётся проблема быстрого, безопасного и удобного способа распределения ключей между клиентами и сервером аутентификации. В данной работе предложен способ модификации беспроводной сети университета для выполнения аутентификации широкого набора устройств пользователей не требующей длительной настройки и предлагающий способ распределения пользовательских сертификатов на мобильные пользовательские устройства.

2. Постановка задачи

Объектом данной работы является корпоративная беспроводная сеть (БКС) университета, обслуживающая различные устройства пользователей: портативные компьютеры и смартфоны. Устройства пользователей (УП) имеют различные операционные системы и различные версии программ, в связи с этим, можно считать каждое устройство уникальным. На рис. 1 представлена структура беспроводного сегмента корпоративной сети, включающая N точек доступа (ТД$_i$ ($i = 1 \dots N$)) различных производителей (Cisco, Eltex, D-Link, Ubiquiti) с разными версиями операционных систем, к которым подключаются M УП$_j$ ($j = 1 \dots M$) для обращения к ресурсам ЛВС и сети Internet. Требуется обеспечить доступ к ресурсам ЛВС и сети Internet для беспроводных УП, при условии ограничений нормативных документов для идентификации УП в публичной беспроводной сети на 2023 год.

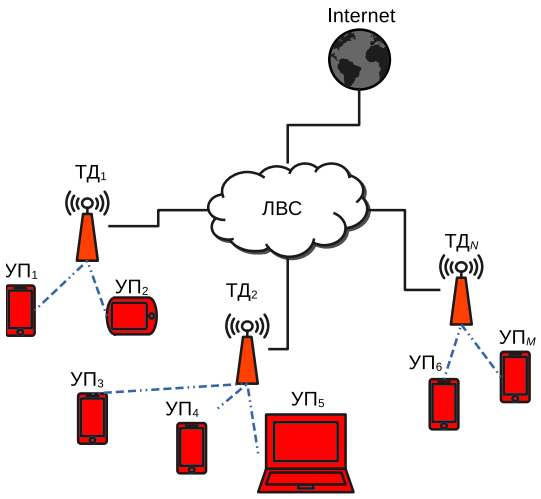


Рис. 1. Структура беспроводного сегмента корпоративной сети

3. Механизмы аутентификации в корпоративной беспроводной сети университета

Для обеспечения доступа УП к ресурсам ЛВС и сети Internet в беспроводном сегменте корпоративной сети согласно нормативным документам №126-ФЗ от 7.07.2003 г. в редакции от 21.04.2014 г., №801 от 12.08.2014 г., Постановления Правительства РФ №758 от 31.07.2014 г. необходимо провести идентификацию личности и устройства пользователя, подключившемуся к беспроводной сети. Процесс идентификации — это получение данных о пользователе из документов, удостоверяющих личность, номер сотового телефона или через аккаунт на сайте «Госуслуги». В корпоративной сети Университета расположены специфические защищённые информационные системы (ИС) для обработки персональных данных сотрудников и студентов. Анализ ИС и существующих в корпоративной сети механизмов аутентификации и авторизации, позволил установить связь между учётными записями на сервере аутентификации и авторизации пользователей (САА́П) и идентификационными данными необходимыми для подключения к беспроводной сети. Поэтому необходимо использовать существующие учётные записи для аутентификации УП в БКС Университета.

В табл. 1 перечислены существующие механизмы аутентификации в БКС стандарта IEEE 802.11 [6, 7, 8, 9]. Механизм аутентификации позволяет идентифицировать УП и ассоциировать (создать логический беспроводной канал (БК)) для передачи данных между УП и ТД. В связи с ограничениями задачи, подходящим методом аутентификации выбран механизм IEEE 802.1X.

Название механизма	Информация для идентификации
Open/Открытая	MAC адрес
PSK/Общий ключ	Общий ключ
IEEE 802.1X	Расширяемый, несколько схем аутентификации

Таблица 1. Обзор существующих механизмов аутентификации в БКС

Стандарт IEEE 802.1X определяет несколько схем для аутентификации для БКС. Схемы представлены в табл. 2.

В настоящее время в БКС университета настроен метод аутентификации PEAP-MSCHAPv2 при котором пользователю необходимо выполнить настройки подключения и ввести логин и пароль для САА́П. Аутентификация пользователя выполняется за 22 шага рис. 2. EAP-TLS и EAP-AKA выполняют аутентификацию быстрее остальных рассмотренных схем табл. 2. Но схема EAP-AKA поддерживается меньшим количеством УП, поэтому для разработки механизма аутентификации выбрана схема EAP-TLS, при которой аутентификация пользо-

Название схемы	Информация для идентификации	Необходимые компоненты	RFC
EAP-TLS	Цифровой сертификат	RADIUS, цифровые сертификаты, настройки ТД, настройки УП	[RFC5216]
EAP-TTLS	Имя пользователя и пароль	RADIUS, цифровой сертификат сервера, настройки ТД, настройки УП	[RFC5281]
PEAP	Имя пользователя и пароль	RADIUS настройки ТД настройки УП	[RFC 4017]
EAP-FAST	Имя пользователя и пароль и пароль	RADIUS, цифровой сертификат сервера	[RFC4851]
EAP-SIM	SIM-карта	RADIUS, HLR, настройки ТД, настройки УП	[RFC4186]
EAP-AKA	USIM-карта	RADIUS, HLR, настройки ТД, настройки УП	[RFC4187]

Таблица 2. Обзор схем аутентификации для механизма IEEE 802.1X

вателя выполняется за 12 шагов рис. 3. Для внедрения данной схемы аутентификации необходимо установить специальный сервер RADIUS. Для внедрения схемы EAP-TLS необходимо выполнить ряд сложных настроек на стороне пользователя. Для устранения данной проблемы разработано решение для создания и распределения пользовательских настроек и сертификатов на УП.

4. Разработка решения для распределения пользовательских сертификатов в корпоративной беспроводной сети

Цифровой сертификат RADIUS-сервера может быть выпущен центром сертификации, которому доверяет пользовательское устройство. Сертификаты пользователей хранятся на RADIUS-сервере. Возникает проблема создания и доставки сертификата пользователя на УП.

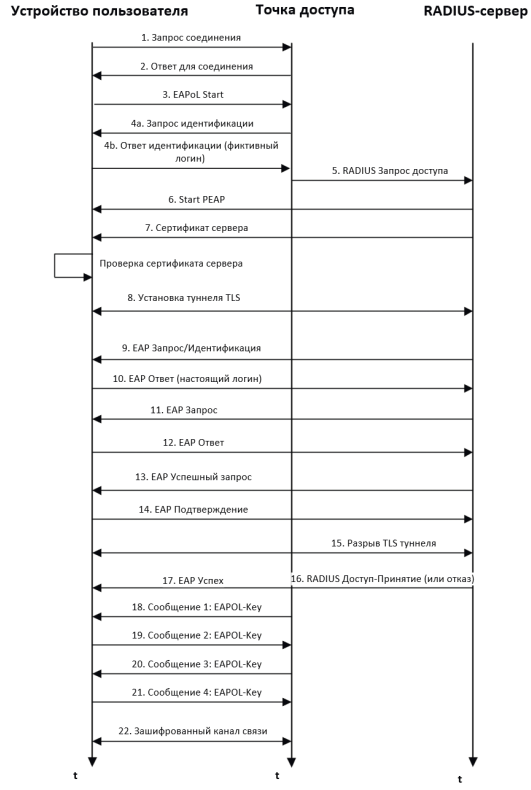


Рис. 2. Процесс аутентификации с помощью схемы PEAP-MSCHAPv2

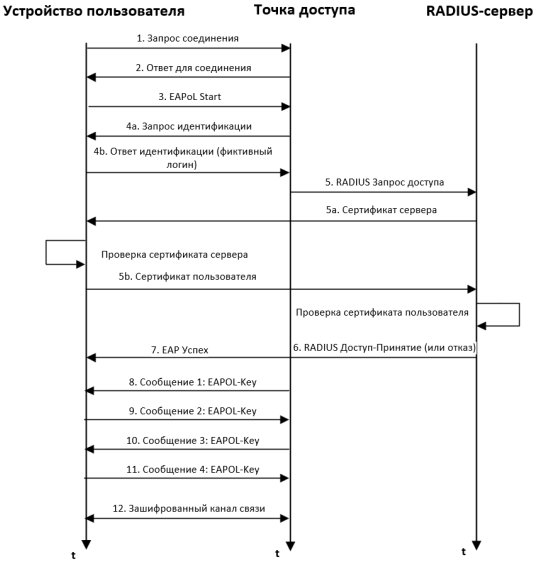


Рис. 3. Процесс аутентификации с помощью схемы EAP-TLS

Разработано следующее решение. Используются две беспроводных сети – 2 SSID, в одной из которых используется открытый метод аутентификации для получения сертификата пользователя через защищённый канал до веб-сервера, а вторая использует метод аутентификации EAP-TLS. Для получения сертификата пользователь заходит на сайт выдачи сертификатов, на котором он должен пройти аутентификацию, для идентификации своей личности. Затем выполнятся установка сертификата и настройка сети. Для автоматической установки сертификата используются разработанные программы. Для данной инфраструктуры используется САА́П, на котором хранятся учётные данные пользователей, RADIUS-сервер, на котором хранятся сертификаты пользователей и который проводит аутентификацию пользователей и веб-сервер, который обращается к САА́П для аутентификации пользователя на сайте и к RADIUS-серверу для получения сертификата пользователя. В случае использования пользователем

нестандартной операционной системой, имеется возможность скачать сертификат и вручную поставить его на свое устройство. Схема сети для получения сертификата пользователя представлена на рис. 4. Для синхронизации пользователей, для которых хранятся сертификаты на RADIUS-сервере и учетных записей на СААП, RADIUS-сервер обращается к СААП для запроса данных о пользователях 1 раз в день. Более подробная информация о компонентах решения представлена в материалах к докладу.

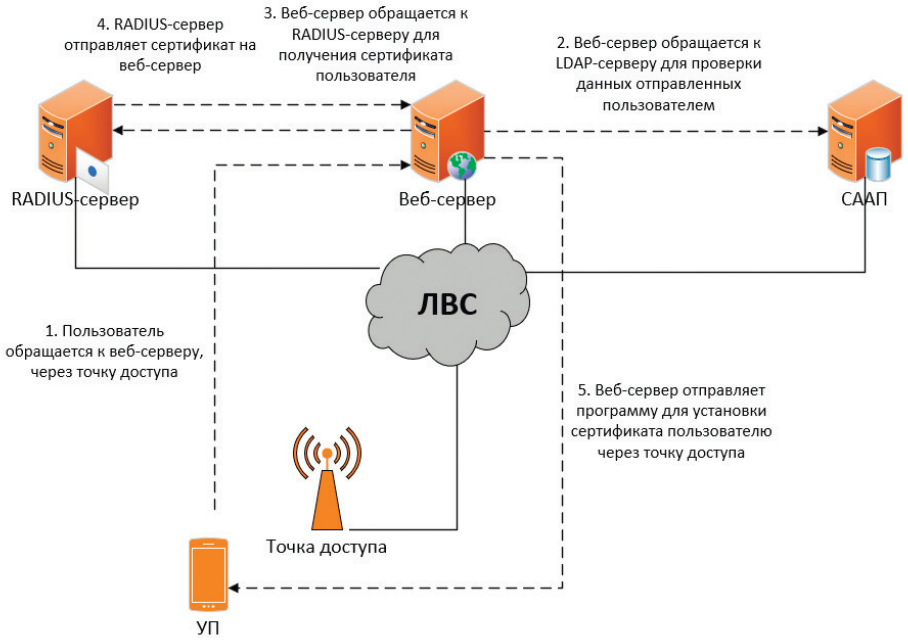


Рис. 4. Структура беспроводного сегмента для получения сертификата пользователя

5. Заключение

В результате работы разработано решение включающее: веб-приложение в открытой гостевой беспроводной корпоративной сети Университета для выдачи пользователю настроек для подключения к защищённой беспроводной корпоративной сети Университета, настройки компонентов для схемы аутентификации EAP-TLS и схему их взаимодействия. Разработанное решение позволяет выполнить настройку защищённого метода аутентификации на стороне пользователя за 3 шага. Настройка выполняется единожды при первом подключении к беспроводной сети.

Литература

1. Федеральный закон Российской Федерации от 7 июля 2003 года №126-ФЗ «О связи»
2. Постановление Правительства РФ от 12.08.2014 №801 (ред. от 31.12.2021) «О внесении изменений в некоторые акты Правительства Российской Федерации»
3. Постановление Правительства РФ от 31.07.2014 № 758 (ред. от 31.12.2021) "О внесении изменений в некоторые акты Правительства Российской Федерации в связи с принятием Федерального закона "О внесении изменений в Федеральный закон "Об информации, информационных технологиях и о защите информации"и отдельные законодательные акты Российской Федерации по вопросам упорядочения обмена информацией с использованием информационно-телекоммуникационных сетей"
4. Li, Guoqiang, and Patrick Bours. "Studying WiFi and accelerometer data based authentication method on mobile phones."In Proceedings of the 2018 2nd international conference on biometric engineering and applications, pp. 18-23. 2018.
5. Chen, Yu-Tso. "Achieve user authentication and seamless connectivity on wifi and wimax interworked wireless city."In 2007 IFIP International Conference on Wireless and Optical Communications Networks, pp. 1-5. IEEE, 2007.
6. Kumar, A., Paul, P. (2019). A Secure Three-Way Handshake Authentication Process in IEEE 802.11i. In: Nath, V., Mandal, J. (eds) Proceeding of the Second International Conference on Microelectronics, Computing & Communication Systems (MCCS 2017). Lecture Notes in Electrical Engineering, vol 476. Springer, Singapore. https://doi.org/10.1007/978-981-10-8234-4_58
7. Hidayat, T. N., & Riadi, I. (2021). Optimation wireless security IEEE 802.1 X using the extensible authentication protocol-protected extensible authentication protocol (EAP-PEAP). International Journal of Computer Applications, 174(11), 25-30.
8. Diwakar, M., Singh, P., Kumar, P., Tiwari, K., Bhushan, S. (2022). A Critical Review on Secure Authentication in Wireless Network. In: Tomar, A., Malik, H., Kumar, P., Iqbal, A. (eds) Machine Learning, Advances in Computing, Renewable Energy and Communication. Lecture Notes in Electrical Engineering, vol 768. Springer, Singapore. https://doi.org/10.1007/978-981-16-2354-7_55
9. Mubarak Baltabaevna Abdujapparova, and Saodat Uzakbayeva. "ANALYSIS OF WI-FI WIRELESS ACCESS METHODS"Science and innovation, vol. 2, no. Special Issue 3, 2023, pp. 406-410. doi:10.5281/zenodo.7856631

UDC: 654.021:654.029

Towards a new format of datasets in traffic analysis

O.P. Morozova[1], M.A. Orlova[1], N.A. Naumov[1], L.I. Abrosimov[1]

[1]National Research University "Moscow Power Engineering Institute", Moscow, Russia

MorozovaOP@mpei.ru, OrlovaMA@mpei.ru, NaumovNA@mpei.ru, AbrosimovLI@mpei.ru

Abstract

The purpose of this work is to develop a reliable method of obtaining datasets of network traffic with ground truth already defined. This method allows to get datasets with accurate ground truth while not violating data privacy since the critical data is stripped and replaced by traffic meta description, which makes its useful for a wide range of traffic analysis methods.

Keywords: Traffic Analysis, ground truth labeled datasets, a traffic analysis suite

1. Introduction

The need to classify network traffic arose with the proliferation of the internet itself. At first it was brought on by the necessity to identify malicious traffic and attack patterns to ensure the security of networks, and later on, with the development of Quality of Service (QoS) systems, the need for more granular traffic classification became apparent. First classifiers were port- and payload-based and used ports and payload pattern respectively to identify applications. But since traffic encryption and dynamic port assignment gained wide recognition, efficiency of those methods has been steadily declining. In their place, a variety of statistics-based methods have been developed, most of them use different machine learning techniques to identify traffic type. Nowadays, both network security monitoring and QoS systems require fast and precise classifiers that are able to meet the challenge of analyzing the ever-changing network environment in real time. However, the development of such classifiers is seriously stunted by unresolved issues in the field of traffic analysis. One of the most significant is collection of big and representative datasets with reliable ground truth that can be used for training and evaluating classifiers. [1]

reports lack of such datasets in multiple domains of traffic analysis: network analytics, intrusion detection and network functions in middleboxes. All those fields require correctly labeled and illustrative datasets, containing different types of network traces. [2] discusses the same issue in the analysis of Internet of Things (IoT) and [3] both points out the importance of datasets suitable for deep learning algorithms. Researchers in [4] use their database to analyze the current state of the field of traffic classification. They conclude that most used datasets are either outdated and do not represent the network environment accurately or private and therefore unavailable to be verified. They indicate that this issue stems from concern for data privacy and enterprise security. This systematic problem is also thoroughly discussed in [5] and [6]. Additionally, in research dedicated to development of new classifier [7] it is pointed out the lack of suitable datasets and the necessity to collect it from scratch.

There have been several studies dedicated to developing a reliable way to obtain big datasets with dependable ground truth. G'eza szab'o et. al [8] proposed an active measurement method that allowed to validate other classifiers based on their performance on emulated traffic. The ground truth obtained using this method was absolutely accurate since their algorithm recorded application for every packet and written that information into the packet's header. However, this came with a limitation on packet size - it had to be 4 bytes shorter than the maximum allowed packet size in order to write an application tag in its' header. Additionally, emulated traffic cannot replicate the variety of real network flows, therefore classifiers may show worse results in action than during the validation process since their ability to analyze traffic in all its complexity was not tested to the fullest. Similar systems based on emulated network traffic have been deployed in [9] and [10]. Their absolutely accurate ground truth allowed researchers to evaluate the most common ways to obtain labels for datasets to this day: port- and DPI-based. [9] proved that their accuracy even on emulated flows, without any additional complications that happen in real networks, has been lacking. Another approach to obtaining ground truth was introduced in [11] and [12]. Those systems are heuristics-based which increases the analysis speed, but decreases ground truth accuracy since agglomerating methods tend to be less exact in their classification. Therefore, those methods cannot be used as a tool for obtaining correct ground truth.

A different tool was introduced by authors of [13]. Tstat can capture traffic in real life or analyze previously recorded traces, creating a log file of flow-level measurements and statistic histograms which is quite useful for network monitoring. However, it does not save any packet-level information that could be used for traffic classification. Additionally, it obtains ground truth for unencrypted packets using DPI-based methods which has been proven ineffective. As for encrypted packets, Tstat implements a SPI-based classifier which also does not provide completely

accurate results. Therefore, ground truth obtained using Tstat cannot be called absolutely reliable. Another tool for the network monitoring was proposed in [14]. VBS is a traffic analysis system installed on end users' machines that listens to all connections and collects packets into flows. The application tag for the given flow is being determined via socket monitor which grants absolutely correct ground truth for every flow. However, for its usage on Windows, a root certificate installation is required and its compatibility with latest versions in unknown.

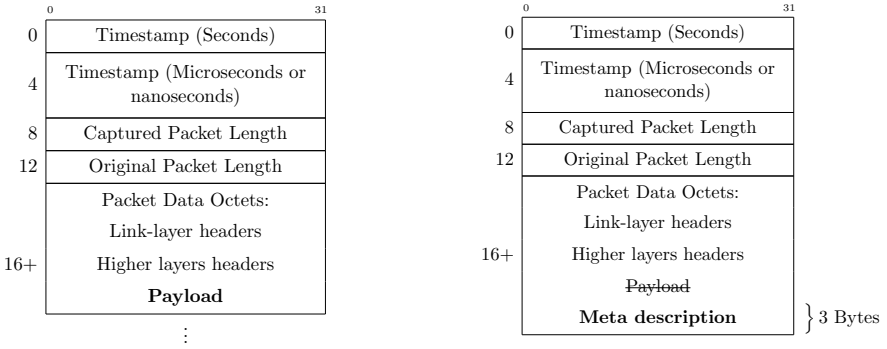
Another way to self-collect datasets is using widely known sniffer applications like Wireshark or TCPdump. However, those methods do not solve the problem of ground truth definition and often researches have to do it manually or resort to DPI-based classifiers which have multiple shortcomings discussed above. Additionally, the size of datasets gets disproportionally large as traces contain packets with full payload, that is nowadays rarely used in traffic analysis systems, therefore a great part of data in those traces is unimportant. Many researchers, faced with multiple shortcomings of self-collection of training data, turn to publicly available datasets. However, this method has its own issues. One of the most prominent is lack of necessary variety of traffic traces. Classifiers build for different purposes require different training datasets and public databases, even though they are growing in number, and are yet to satisfy this demand for diversity. Another point of concern is the definition of ground truth in those datasets. Many of them do not contain any correlation to applications and types of network traffic, and those that do tend to use port- and DPI-based methods to obtain it, and their lack of accuracy was discussed above.

As our survey of related works shows, the problem of dataset collection with reliable ground truth is yet to be fully resolved dispute the efforts of researchers.

2. Traffic meta description

Analysis of the current methods of obtaining training datasets showed that this direction needs further exploration in order to optimize this process. In light of this, we proposed a special formatting for captured packets to be stored in traffic meta description. Each captured packet is to be stored as shown in Fig.1b, as opposed to standard packet storing described in [15] and shown in Fig.1a.

In traffic meta description, each packet trace consists of 3 main parts: header, payload descriptor and an application tag. The header contains all the protocol layers the original packet had. Payload descriptor is what replaces the payload itself since it is removed from the trace. Descriptor contains payload's size and some statistical measurements that can be used for traffic classification. Finally, the application tag is what solves the problem of ground truth definition. This tag is added directly at the moment of packet collection using connection tracker, therefore there is no risk of misidentifying the application.



(a) standard .pcap file (b) modified .pcap file

Fig. 1. Standard and modified packet storing in .pcap files

This approach allows us to deal with multiple problems of dataset collection described above. Since most state-of-art classifiers rely on statistical features of packets and flows, extraction of payload does not affect the quality of classification, but it does decrease the size of one packet and the whole dataset significantly. For instance, it was observed that an hour of traffic capturing can produce a file up to a few gigabytes in size in .pcap format, if we are to use standard sniffing software. Using method depicted in fig.1b has shown promising file size reduction - the same amount of sniffing produces only a few megabytes of data in .pcap format. This approach makes it possible to collect longer traces using the same amount of memory storage. Additionally, the application tag eliminates the issue of defining ground truth - with traffic meta description obtained dataset already contains ground truth for each packet right after the collection finishes.

Since traffic meta description proved to be an advantageous method of storing packet traces, we used it to develop a traffic analysis suite that allows to inspect collected traces, sorting packets by generating application. Further description, alongside with performance tests' results, will be provided in presentation. However, here we detail an algorithm for creating trace file with meta description for each packet stored in .pcap format.

1) Incoming or outgoing packet gets captured using python library scapy;
2) Application that generated given packet is determined using a list of open connections on chosen interface accessed via tools provided by python library psutil. Application tag for meta description is formed. In case application is not found in the list of active processes, the packet get discarded;

3) Size of packet payload is determined and temporarily saved, the payload itself is deleted from the packet. Payload descriptor is formed. In the future we intend to provide an opportunity to customize this part of packet trace with additional payload features;
4) Already formed meta description is added to the packet data octets of the trace instead of payload, as shown in Fig.1b;
5) Modified packet is stored in the trace file in .pcap format.

3. Conclusion

In this paper we present a new formatting for datasets in traffic analysis - packets are stored without payload, but have a payload descriptor that ensures that no meaningful data is lost from payload itself and an application tag that serves as ground truth for the packet. We first demonstrate that there is an unresolved issue with the collection of datasets and obtaining ground truth for them. Then we describe our format of datasets where each packet is stored with traffic meta description instead of raw payload. We further describe an application that allows to collect datasets in said format and gather some basic statistics from them in real time - traffic analysis suite.

REFERENCES

1. Eva Papadogiannaki and Sotiris Ioannidis. 2021. A Survey on Encrypted Network Traffic Analysis Applications, Techniques, and Countermeasures. ACM Comput. Surv. 54, 6, Article 123 (July 2022), 35 pages. https://doi.org/10.1145/3457904
2. Hamid Tahaei, Firdaus Afifi, Adeleh Asemi, Faiz Zaki, Nor Badrul Anuar, The rise of traffic classification in IoT networks: A survey, Journal of Network and Computer Applications, Volume 154, 2020, 102538, ISSN 1084-8045, ttps://doi.org/10.1016/j.jnca.2020.102538.
3. S. Rezaei and X. Liu, "Deep Learning for Encrypted Traffic Classification: An Overview," in IEEE Communications Magazine, vol. 57, no. 5, pp. 76-81, May 2019, doi:10.1109/MCOM.2019.1800819.
4. Iglesias F, Ferreira DC, Vormayr G, Bachl M, Zseby T. NTARC: A Data Model for the Systematic Review of Network Traffic Analysis Research. Applied Sciences . 2020; 10(12):4307. https://doi.org/10.3390/app10124307
5. Getman, A.I., Ikonnikova, M.K. A Survey of Network Traffic Classification Methods Using Machine Learning. Program Comput Soft 48, 413–423 (2022). https://doi.org/10.1134/S0361768822070052

6. Deart V.Yu., Mankov V.A., Krasnova I.A. Analysis of promising approaches and research on traffic flow classification for maintaining QoS by ML methods in SDN networks. The Herald of the Siberian State University of Telecommunications and Information Science. 2021;(1):3-23. (In Russ.) https://doi.org/10.55648/1998-6920-2021-15-1-03-22
7. Klenilmar Lopes Dias, Mateus Almeida Pongelupe, Walmir Matos Caminhas, Luciano de Errico, An innovative approach for real-time network traffic classification, Computer Networks, Volume 158, 2019, Pages 143-157, ISSN 1389-1286, https://doi.org/10.1016/j.comnet.2019.04.004.
8. Szabó, G., Orincsay, D., Malomsoky, S., Szabó, I. (2008). On the Validation of Traffic Classification Algorithms. In: Claypool, M., Uhlig, S. (eds) Passive and Active Network Measurement. PAM 2008. Lecture Notes in Computer Science, vol 4979. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-79232-1_8
9. F. Gringoli, Luca Salgarelli, M. Dusi, N. Cascarano, F. Risso, and k. c. claffy. 2009. GT: picking up the truth from the ground for internet traffic. SIGCOMM Comput. Commun. Rev. 39, 5 (October 2009), 12–18. https://doi.org/10.1145/1629607.1629610
10. P. Lizhi, Z. Hongli, Y. Bo, C. Yuehui and W. Tong, "Traffic Labeller: Collecting Internet traffic samples with accurate application information," in China Communications, vol. 11, no. 1, pp. 69-78, Jan. 2014, doi: 10.1109/CC.2014.6821309.
11. Canini, M., Li, W., Moore, A.W., Bolla, R. (2009). GTVS: Boosting the Collection of Application Traffic Ground Truth. In: Papadopouli, M., Owezarski, P., Pras, A. (eds) Traffic Monitoring and Analysis. TMA 2009. Lecture Notes in Computer Science, vol 5537. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-01645-5_7
12. Arian Baer, Pedro Casas, Alessandro D'Alconzo, Pierdomenico Fiadino, Lukasz Golab, Marco Mellia, Erich Schikuta, DBStream: A holistic approach to large-scale network traffic monitoring and analysis, Computer Networks, Volume 107, Part 1, 2016, Pages 5-19, ISSN 1389-1286, https://doi.org/10.1016/j.comnet.2016.04.020.
13. A. Finamore, M. Mellia, M. Meo, M. M. Munafo, P. D. Torino and D. Rossi, "Experiences of Internet traffic monitoring with tstat," in IEEE Network, vol. 25, no. 3, pp. 8-14, May-June 2011, doi:10.1109/MNET.2011.5772055.
14. T. Bujlow, K. Balachandran, T. Riaz and J. M. Pedersen, "Volunteer-based system for classification of traffic in computer networks," 2011 19thTelecommunications Forum (TELFOR) Proceedings of Papers, Belgrade, Serbia, 2011, pp. 210-213, doi:10.1109/TELFOR.2011.6143528.

15. PCAP Capture File Format https://datatracker.ietf.org/doc/id/draft-gharris-opsawg-pcap-00.html

UDC: 004.94

Characterizing Blockage Statistics of Reflected Propagation Paths in sub-THz Indoor Communications

Anatoliy Prikhodko[1,2], Abdukodir Khakimov[3], Evgeny Mokrov[3], Vyacheslav Begishev[3], Alexander Shurakov[1,2], Gregory Gol'tsman[1,2]

[1]Moscow Pedagogical State University, Moscow, Russia
[2]National Research University Higher School of Economics, Moscow, Russia
[3]Peoples' Friendship University of Russia (RUDN University), Moscow, Russia

anprihodko@hse.ru, khakimov-aa@rudn.ru, mokrov-ev@rudn.ru, begishev-vo@rudn.ru, alexander@rplab.ru, goltsman@rplab.ru

Abstract

The future 6G cellular systems are expected to utilize the lower part of the terahertz frequency band, $100 - 300$ GHz. As a result of high path losses, the coverage of such systems will be limited to a few tens of meters making them suitable for indoor environments. As compared to outdoor deployments, indoor usage of THz systems is characterized by the need to operate over shorter distances using the reflected propagation paths. This paper aims to characterize the impact of blockage of reflected propagation paths in typical scenarios. Specifically, we carry out a detailed measurements campaign at 156 GHz and report reflection losses, blockage losses over the reflected path as well as blockage duration, signal fall and rise times. Our results show that signal polarization has a profound impact on the reflection losses with E-plane horizontally oriented signal losses being at least 8 dB higher as compared to H-plane signal horizontal orientation. Furthermore, the reflection material types do not affect the mean blockage attenuation over the reflected paths. Generally, the presence of a reflector neither quantitatively nor qualitative changes the mean attenuation induced by a blockage phenomenon.

Keywords: 6G, terahertz, reflections, blockage, attenuation, duration, signal fall and rise times

1. Introduction

Seeking for the capacity boost at the access interface in cellular systems, ITU-R and 3GPP utilize millimeter wave (mmWave) bands, $30 - 100$ GHz for 5G New Radio (NR) systems [1]. The next step in the evolution of such systems is the utilization

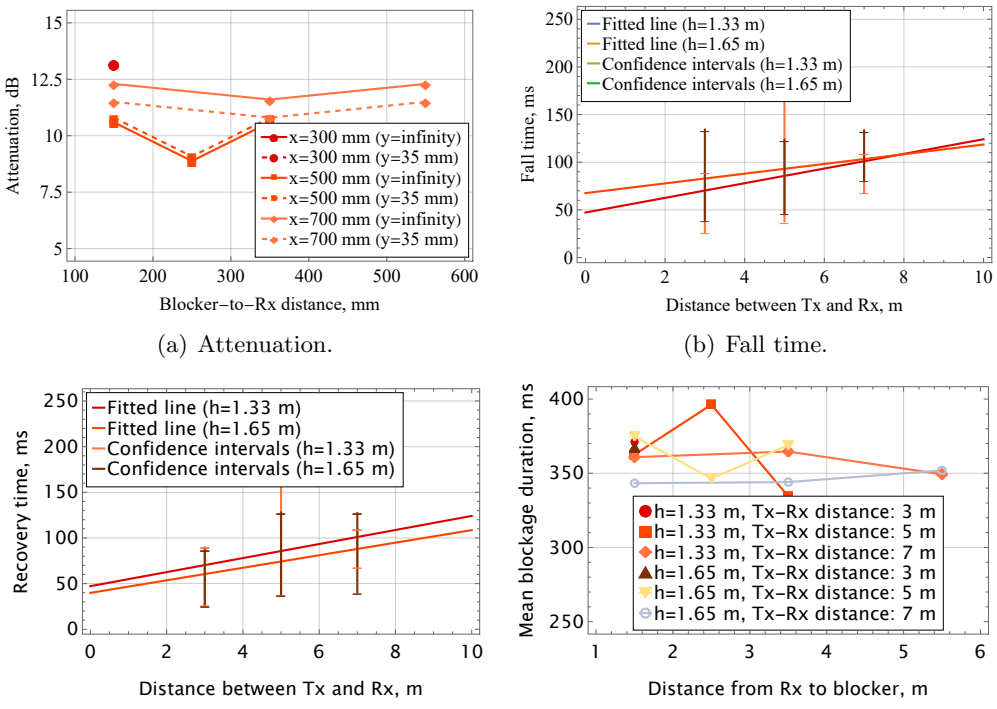
of the lower part of the terahertz (THz) frequency band, where large parts of the spectrum are still not regulated and tens of gigahertz of bandwidth can be allocated to 6G systems [2].

To partially compensate for the reduction in the effective antenna aperture that reduces with the increase of the carrier frequency, similarly to 5G NR mmWave systems, 6G sub-THz systems will heavily rely upon the use of antenna arrays at both base station (BS) and user equipment (UE) operating in beamforming mode. Nevertheless, the coverage of such systems will still be limited to tens or hundreds of meters making them a suitable choice for indoor areas, where most of the traffic demands originate. The landscape of applications in the indoor environment is rather large including conventional 4k/8k video watching, virtual/augmented reality (AR/VR) gaming, and forthcoming applications such as collective VR gaming, holographic communications [3], etc.

Indoor deployments of 6G sub-THz systems are characterized by several propagation specifics. First of all, the link distances are on average smaller as compared to those outdoors. Secondly, due to rather small heights of BS, human body blockage is more likely to occur. Finally, as a result of the complex geometry of indoor premises, communications over reflected paths are expected to be much more common. Specifically, short distances have been recently shown to lead to much smaller human body blockage attenuation at 156 GHz varying in the range of $8-13$ dB [4] as compared to $15-35$ dB losses over larger distances and at lower frequencies, e.g., as reported in [5, 6].

The aim of this paper is to characterize reflected propagation paths in indoor environments in the sub-THz frequency band. Specifically, by carrying out a large-scale measurement campaign at carrier frequency of 156 GHz, we characterize reflection losses of different materials and blockage attenuation of reflected propagation paths. In addition to attenuation, we also investigate time-related metrics such as blockage duration as well as signal rise and fall times. The main findings of our paper acquired empirically are:

- the orientation of the antenna polarization plane has a profound impact on the reflection losses with horizontally oriented polarization plane losses being at least 7 dB higher as compared to vertically orientated polarization plane;
- the reflection material types do not affect the mean blockage attenuation over the reflected paths;
- the presence of reflector neither quantitatively nor qualitative changes the mean attenuation induced by a blockage phenomenon;
- blockage, fall and rise times for drywall are characterized by slightly smaller mean values as compared to concrete for vertical polarization plane.



(a) Attenuation.

(b) Fall time.

(c) Rise time.

(d) Blockage time.

Fig. 1. Mean attenuation, mean signal blockage, fall and rise times.

The paper is organized as follows. First, in Section 2, we overview related studies reported in the literature. Next, in Section 3, we outline experimental setup. The main results of the conducted experiments are reported further in Section 4. And conclusions are provided in the last section.

2. Related Work

In this section, we outline related works. We start by briefly reminding the results for non-reflected path propagation and blockage. They are recapitulated for comparison purposes in our study. Then, we discuss results similar to those reported in our study which are related to reflection losses and blockage of reflected propagation paths in the mmWave and THz bands.

As one of the paper goals is to compare blockage statistics over reflected paths to that over primary line-of-sight (LoS) paths, we briefly introduce the latter as reported earlier in [4]. To this aim, Fig. 1 provides mean attenuation, mean signal blockage, fall and rise times for different Tx-to-Rx distances, x, LoS heights, h, and

Rx-to-blocker distances. By analyzing the presented results, one may observe that the mean attenuation varies between $8 - 13$ dB and is generally independent of blocker-to-Rx and Tx-to-Rx distances. Furthermore, both fall and rise times increase as a function of the distance potentially making it more feasible to detect blockage events timely. The absolute difference between the reported times is insignificant and lies within 2–4% of the nominal value (e.g., 60 ms for $x = 3$ m, 80 ms for $x = 5$ m and 100 ms for $x = 7$ m for fall times). It is worth noting that both fall and rise times have almost identical nominal values, and, in general, the rise time is 7–10% smaller than the fall time.

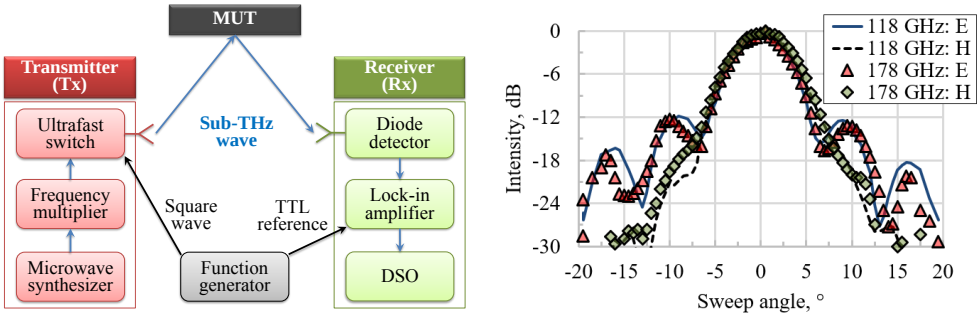
The attenuation caused by reflections from different materials was the subject of several studies. In [7], the authors reported the measurement results of the received signal reflected from aluminum, glass, plastic, hardboard and concrete using THz time-domain spectroscopy (THz-TDS) equipment for different angles of incidence. The time duration of utilized pulses was chosen such that the energy is mainly concentrated in the $0.1 - 4$ THz band. The observed losses were in the range of $10 - 60$ dB depending on the angle of incidence and type of the material. Specifically, in the $0.1 - 0.3$ THz band, aluminum demonstrated the least attenuation of around 25–35 dB for the angle of incidence of $\pi/4$. The rest of the materials provided higher attenuation.

The authors in [8, 9] reported the results of reflections from typical vehicle materials at 300 GHz for different configurations including front and rear reflections, side-lane and under-vehicle reflections. The front and rear reflections were reported to result in 24–42 dB and 15–30 dB attenuations, respectively, while side-lane reflections led to additional 16–20 dB losses. The authors also proposed to model the under-vehicle reflection losses of the asphalt by utilizing the $\alpha d^{-\beta}$ function, where d is the separation distances, while α and β are some coefficients tabulated in Table I in their manuscript.

Similar studies have been performed in the mmWave band. In [10], specifically, the authors investigated the impact of polarization properties of the indoor 38 GHz channel after single bounce reflection for different angles of incidence and observation and two types of materials: aluminum and concrete. They highlighted that, when the polarization of Tx and Rx antennas coincide, much smaller losses are experienced. The difference can reach 15–20 dB and is maximized for peculiar reflections. Concrete results in 8–12 dB higher losses as compared to aluminum.

3. Experimental Setup

In this section, we introduce our measurement setup and present details on the acquisition of experimental data.



(a) Measurement equipment.

(b) Tx/Rx radiation pattern.

Fig. 2. Experimental setup.

A schematic of the measurement setup employed for blockage studies is presented in Fig. 2(a). We use a 156 GHz constant waveform source with amplitude modulation at 25 kHz. It provides a 6° wide beam incident on the material under test (MUT) at a constant angle of 70°. The reflected beam is received by a low-barrier diode detector equipped with the same optics as the source. A lock-in amplifier is used to readout the detector response voltage. When a blocker walks across the reflected beam at the midpoint between MUT and the detector, the voltage-vs-time series is registered by a digital signal oscilloscope (DSO). The measurement covers a time frame of 4 s with a resolution of 100 μs. Referring to Fig. 2(a), list of the employed measurement equipment includes the following items:

- microwave synthesizer: Hittite HMC-T2220;
- frequency multiplier: RPG Tx-134-158-20;
- ultrafast switch: ELVA-1 VCVA-06;
- function generator: SRS DS345;
- diode detector: DOK WR-06;
- lock-in amplifier: SRS SR844;
- DSO: R&S RTO1012.

The source-to-detector (Tx-to-Rx) optical path of 3 m is chosen in all the measurements. The source provides 52 mW of power, and the setup ensures a signal-to-noise ratio of up to 3×10^4 at the detector output. Measured response voltages are further converted into power levels at the detector input via its responsivity, which is equal to 500 V/W at 156 GHz. MUT is successively presented by concrete, drywall and class samples with thicknesses of 50, 12.5 and 6 mm, respectively. The sample linear dimensions of 0.5 m × 0.5 m are chosen to overlap a 156 GHz beam upon reflection. The sample is installed in a wooden frame to set its center at 1.65 m

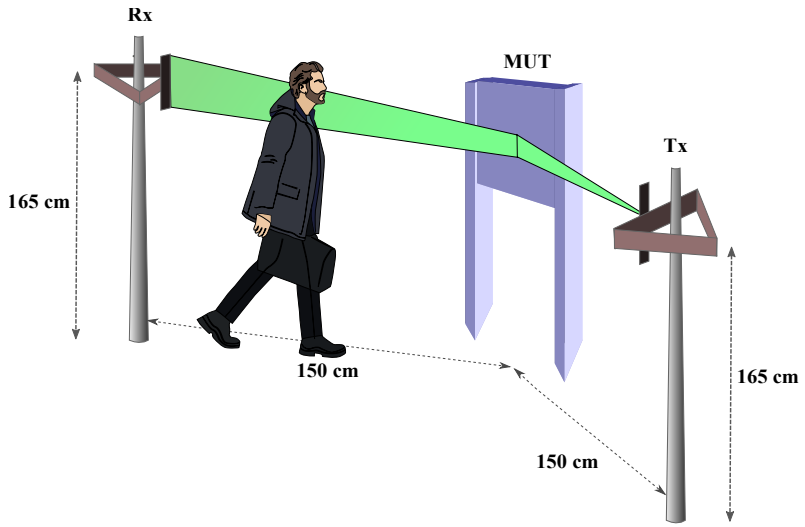


Fig. 3. Illustration of the considered scenario.

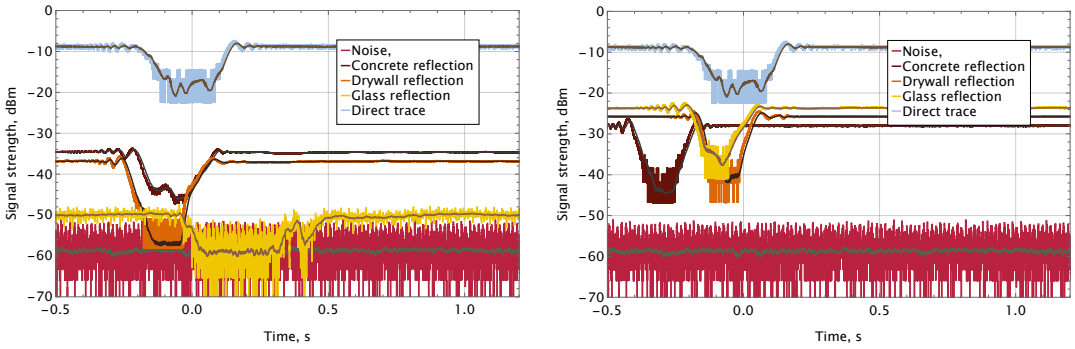
above the floor which corresponds to the LoS height between the source and the detector. Spurious reflections from the wall behind the sample are geometrically filtered out by the positioning of the measurement equipment. Input optics of the source and the detector rely on a pyramidal horn known as a wide-band antenna with different side lobe levels in H- and E-planes, see Fig. 2(b). We make use of this feature and conduct a series of measurements for two orientations of the horn antennas, when either their H- or E-planes are horizontally oriented, i.e., coinciding with the plane of incidence of the transmitted 156 GHz beam.

The schematic illustration of the scenario is shown in Fig. 3.

4. Measurements Results

In this section, we report our results. We start with visual illustrations of the blockage phenomenon over the reflected paths. Then, we characterize the mean attenuation caused by reflection for typical types of wall materials such as drywall, concrete and glass. Further, we investigate blockage attenuation over reflected paths. Finally, we report mean and cumulative distribution function (CDF) of blockage duration, signal fall and rise times for reflected paths.

4.1. Time Series. We start with a time series representation of the blockage over the reflected paths, as demonstrated in Fig. 4, for all the considered types of materials and orientations of the Tx/Rx antennas. Here, for comparison purposes, we also demonstrate blockage over the LoS path ("direct trace" in Fig. 4).



(a) Horizontal plane. (b) Vertical plane.

Fig. 4. Comparison of reflection and direct traces.

By visually inspecting the presented data, one may notice that the received signal level in no blockage condition is significantly lower as compared to direct LoS propagation. This is attributed to the reflection losses considered in details below. Furthermore, we observe significant difference between signal levels and blockage profiles for horizontally and vertically oriented E-planes of Tx/Rx antennas (the H-planes are respectively orthogonal). The former condition is referred to as horizontal plane (HP), and the latter – as vertical plane (VP). Specifically, we see that the signal strength for the HP reflection from glass is barely higher than the noise level. And for the VP reflection case, it is comparable to other materials. Finally, visual inspection does not allow clearly highlight any difference between attenuation- and time-related blockage profiles for different materials requiring detailed statistical analysis.

4.2. The Impact of Reflection. We start our analysis with the characterization of the reflection losses by comparing the reflected signal against the LoS signal studied earlier in [4]. In order to determine reflection losses, we utilize the propagation model, $L(x)$, with the coefficients y_1 and y_2 empirically derived from the signal difference between Tx and Rx at different distances.

$$L(x) = y_1 \log_{10} x + 20 \log_{10} f_c + y_2 + I_B L_B(x, d), \tag{1}$$

In the case of LoS propagation for the carrier frequency of 156 GHz, the coefficients are $y_1 = -22.04$ and $y_2 = -251.704$ [4].

By utilizing (1), we subtract the model's values from the path losses observed in non-LoS conditions obtaining the reflection losses. These losses are reported in Table 1 for different types of materials and polarizations. By analyzing the presented

data, one may deduce that the impact of the polarization plane orientation is critical. For the VP condition, all the materials behave similarly leading to 14–19 dB losses with glass having 4 dB gain on top of concrete. The HP condition, however, is characterized by at least 7–13 dB stronger attenuations. Specifically, there are 7 dB higher losses for concrete and 11 dB higher losses for drywall. In the HP condition, glass attenuates the signal on reflection by approximately 42 dB making the reflected path signal strength comparable to noise, see Fig. 4. We specifically note that attenuation of at least 30 dB may lead to the loss of connectivity depending on the propagation distance between Tx and Rx.

4.3. Blockage Attenuation over Reflected Paths. Now, we proceed to characterize the blockage attenuation over the reflected paths. To this aim, we subtract the value of the received signal with no blockage impairments from the average value in the blocked state. Table 2 presents the mean blockage attenuation and the mean values of the signal strength propagated directly to the Tx and reflected from the considered materials in blocked and non-blocked cases.

By analyzing the reported data, one may observe that, while the antenna polarization plane orientation greatly impacts on the received signal strength, the impact on the mean blockage value is rather limited. All the material types lead to almost constant blockage attenuation of approximately 9–10 dB for the VP orientation. By comparing the obtained results with those reported in Fig. 1 for non-reflected blockage, one may deduce that the presence of a reflector neither quantitatively nor qualitatively changes the mean attenuation induced by a human blockage phenomenon. The HP orientation is characterized by larger differences between mean blockage attenuations varying in the range of $7 - 10$ dB. However, these changes can be potentially attributed to smaller signal strengths accompanied by reduction in measurement accuracy for drywall and glass.

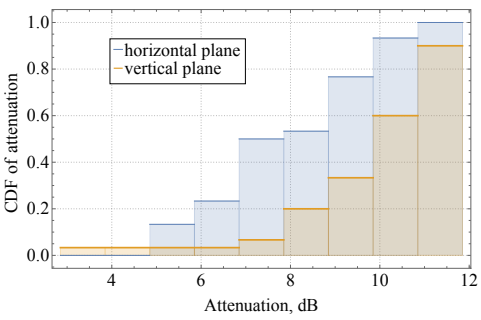
To provide additional information regarding the attenuation, Fig. 5 reports cumulative distribution functions (CDFs) of blockage attenuations for concrete

Table 1. Reflection and blockage losses.

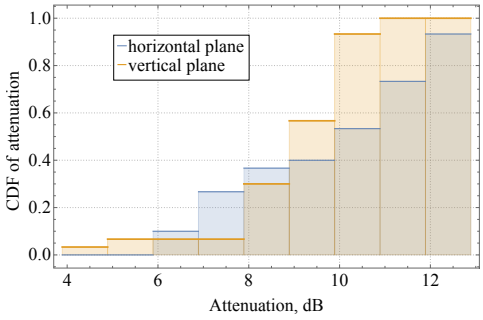
Type	Reflection loss, dB	Blockage loss, dB
Concrete, HP	25.85	33.84
Drywall, HP	27.97	37.96
Glass, HP	41.59	48.08
Concrete, VP	18.93	28.76
Drywall, VP	16.87	25.94
Glass, VP	14.79	23.60

Table 2. The mean blockage attenuation.

Type	Signal, dBm	Blocked signal, dBm	Blockage attenuation, dB
Direct, LoS	-8.88	-16.21	7.32
Concrete, HP	-34.74	-42.73	7.98
Drywall, HP	-36.86	-46.85	9.99
Glass, HP	-50.48	-56.97	6.48
Concrete, VP	-27.82	-37.65	9.83
Drywall, VP	-25.76	-34.83	9.07
Glass, VP	-23.68	-32.48	8.80



(a) Concrete.



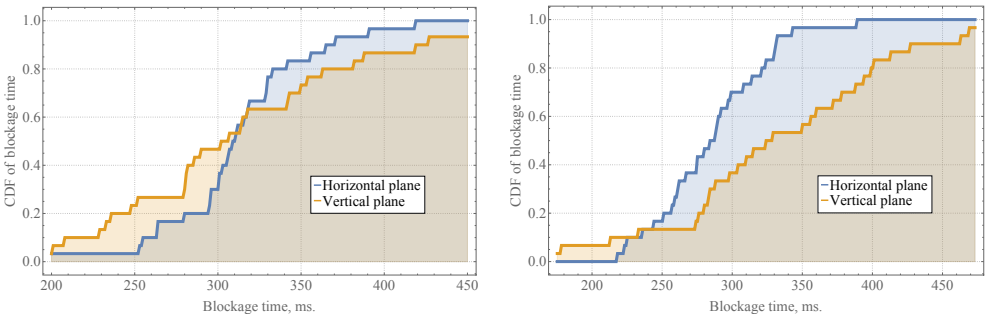
(b) Drywall.

Fig. 5. CDFs of blockage attenuation for concrete and drywall.

and drywall materials. The illustration highlights the difference between blockages for different orientations of the antenna polarization plane. Notably, the range of attenuation can be quite large varying between approximately 6 and 10 dB. These differences can be attributed to slight changes in the trajectory of a person crossing the LoS path.

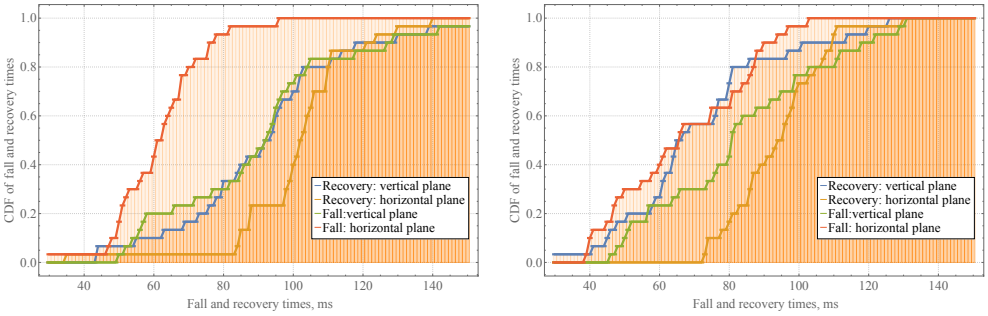
Table 3. Mean fall, rise and blockage times.

Type	Blockage time, ms	Rise time, ms	Fall time, ms
Concrete, HP	316	101	61
Drywall, HP	333	94	66
Concrete, VP	308	92	90
Drywall, VP	285	71	89



(a) Concrete. (b) Drywall.

Fig. 6. CDF of blockage times.



(a) Concrete. (b) Drywall.

Fig. 7. CDFs of fall and rise times.

4.4. Fall, Rise and Blockage times. In addition to attenuation statistics, time-related parameters of blockage are of importance for the design of sub-THz communications systems. In this section, we report blockage, fall and rise times including their mean values and CDFs. Note that the rise time is of special importance in the context of blockage detection [4, 11], while the blockage time is critical for designing algorithms to improve service reliability.

The mean values of the considered metrics are summarized in Table 3. By analyzing the reported values, one may observe that there is a 8% difference between blockage times for concrete and drywall for the VP antenna orientation. Mean rise time for drywall is 30% smaller than for concrete, and their fall times are almost identical. Note that for the HP orientation, all these values differ by 5–8% as evident from Table 3. Still, recalling the results for blockage remedies, these deviations are expected to not affect the design of blockage detection algorithms [4, 11].

Complementing the mean values, we present the CDFs of blockage, fall and rise times in Fig. 6 and 7. The blockage duration CDFs indicate that CDF for concrete is steeper as compared to drywall and also has a smaller range of approximately 200 ms. Also, the VP antenna orientation leads to a noticeably smoother increase in CDF behavior. In practice, it means that deviation in blockage duration for the HP antenna orientation is much more clustered around its mean-making.

By analyzing CDFs of fall and rise times demonstrated in Fig. 7, one may observe that drywall is characterized by a slightly wider range of values as compared to concrete. The difference in these times for different orientations of the antenna polarization plane is also noticeable. This generally means that the time budget for the detection of blockage events is higher for drywall.

5. Conclusions

As most of the traffic in cellular systems originates indoors, where the link distances are generally shorter while communications over reflected paths are more common as compared to outdoor deployments, in this paper, we performed a measurements campaign at 156 GHz characterizing the reflection and blockage losses of reflected paths. We considered different types of reflection materials typical for the indoor environment including concrete, drywall and glass.

Our main findings are: (i) the orientation of the antenna polarization plane has a profound impact on the reflection losses with horizontally oriented polarization plane losses being at least 7 dB higher as compared to vertically orientated polarization plane, (ii) the reflection material types do not affect the mean blockage attenuation over the reflected paths, (iii) the presence of reflector neither quantitatively nor qualitative changes the mean attenuation induced by a blockage phenomenon, (iv) blockage, fall and rise times for drywall are characterized by slightly smaller mean values as compared to concrete for vertical polarization planes. In general, blockage statistics over reflected paths are similar to that for LoS paths and do not require special communications algorithms design.

6. Acknowledgements

This study was conducted as a part of strategic project “Digital Transformation: Technologies, Effectiveness, Efficiency” of Higher School of Economics development programme granted by Ministry of science and higher education of Russia “Priority-2030” grant as a part of “Science and Universities” national project. Support from the Basic Research Program of the National Research University Higher School of Economics is gratefully acknowledged.

REFERENCES

1. A. Zaidi, F. Athley, J. Medbo, U. Gustavsson, G. Durisi, X. Chen, 5G Physical Layer: principles, models and technology components, Academic Press, 2018.
2. M. Matthaiou, O. Yurduseven, H. Q. Ngo, D. Morales-Jimenez, S. L. Cotton, V. F. Fusco, The road to 6G: Ten physical layer challenges for communications engineers, IEEE Communications Magazine 59 (1) (2021) 64–69.
3. D. Moltchanov, E. Sopin, V. Begishev, A. Samuylov, Y. Koucheryavy, K. Samouylov, A tutorial on mathematical modeling of 5G/6G millimeter wave and terahertz cellular systems, IEEE Communications Surveys & Tutorials (2022).
4. A. Shurakov, D. Moltchanov, A. Prikhodko, A. Khakimov, E. Mokrov, V. Begishev, I. Belikov, Y. Koucheryavy, G. Gol'tsman, Empirical blockage characterization and detection in indoor sub-thz communications, Computer Communications 201 (2023) 48–58.
5. S. Nie, G. R. MacCartney, S. Sun, T. S. Rappaport, 72 GHz millimeter wave indoor measurements for wireless and backhaul communications, in: 2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), IEEE, 2013, pp. 2429–2433.
6. G. R. MacCartney, T. S. Rappaport, S. Rangan, Rapid fading due to human blockage in pedestrian crowds at 5g millimeter-wave frequencies, in: GLOBECOM 2017-2017 IEEE Global Communications Conference, IEEE, 2017, pp. 1–7.
7. J. Kokkoniemi, V. Petrov, D. Moltchanov, J. Lehtomäki, Y. Koucheryavy, M. Juntti, Wideband terahertz band reflection and diffuse scattering measurements for beyond 5g indoor wireless networks, in: European Wireless 2016; 22th European Wireless Conference, VDE, 2016, pp. 1–6.
8. V. Petrov, J. M. Eckhardt, D. Moltchanov, Y. Koucheryavy, T. Kurner, Measurements of reflection and penetration losses in low terahertz band vehicular communications, in: 2020 14th European Conference on Antennas and Propagation (EuCAP), IEEE, 2020, pp. 1–5.
9. J. M. Eckhardt, V. Petrov, D. Moltchanov, Y. Koucheryavy, T. Kürner, Channel measurements and modeling for low-terahertz band vehicular communications, IEEE Journal on Selected Areas in Communications 39 (6) (2021) 1590–1603.
10. I. Gaspard, Co-and crosspolar scattering measurements at slightly rough walls for indoor propagation channels at mmwaves, in: 2019 IEEE-APS Topical Conference on Antennas and Propagation in Wireless Communications (APWC), 2019, pp. 038–041. doi:10.1109/APWC.2019.8870411.
11. S. Wu, M. Alrabeiah, A. Hredzak, C. Chakrabarti, A. Alkhateeb, Deep learning for moving blockage prediction using real mmwave measurements, in: ICC 2022-IEEE International Conference on Communications, IEEE, 2022, pp. 3753–3758.

UDC: 004.94

The simulation of finite-source retrial queues with two-way communication to the orbit using a backup server

Ádám Tóth[1] and János Sztrik[1]

[1]University of Debrecen, Debrecen 4032, Hungary

{toth.adam,sztrik.janos}@inf.unideb.hu

Abstract

This paper investigates a two-way communication retrial queuing system with a server that may experience random breakdowns. The system is a finite-source M/M/1//N type, and the idle server can make calls to the customers in the orbit, also known as secondary customers. The service time of the primary and secondary customers follow independent exponential distributions with rates of μ_1 and μ_2, respectively. The novelty of this study is to analyze the impact of various distributions of failure time on the key performance measures using a backup server, such as the mean response time of an arbitrary customer. One could think of a backup server as a primary server that operates at a reduced rate during periods of repair. To ensure a valid comparison, a fitting process is conducted so that the mean and variance of every distribution are equal. The self-developed simulation program provides graphical illustrations of the results.

Keywords: Simulation, Queueing system; Finite-source model, Sensitivity analysis, Backup server, Unreliable operation, Outgoing calls

1. Introduction

Nowadays, due to the growth of traffic and the increasing number of users, analyzing communication systems or designing optimal patterns for these schemes is a challenging task. Information exchange is essential in every aspect of life, and it is crucial to develop mathematical and simulation models of telecommunication systems or modify the existing ones to keep pace with these changes. Retrial queues are effective and appropriate tools for modeling real-life problems that arise in telecommunication systems, networks, mobile networks, call centers, and similar systems. Numerous papers and books have been dedicated to studying a variety of retrial queuing systems with repeated calls like in [1],[2].

We are investigating a retrial queuing system with two-way communication capabilities, which has become a popular research topic due to its resemblance to certain real-life systems. This is particularly relevant in call centers, where service units may engage in additional activities such as sales, promotion, and product advertising while attending to incoming calls. In our study, the primary server calls in customers from the orbit, known as secondary customers, when it becomes idle after a random period of time. The utilization of the service unit is monitored and has been extensively studied in previous works, for example in [3],[4].

In some scenarios, it is assumed by researchers that service units are available continuously, but failures or sudden events may happen during operation resulting in the rejection of incoming customers. Devices used in different industries are subject to breakdowns, and considering their reliable operation is quite an optimistic and unrealistic approach. Similarly, in wireless communication, various elements can affect the transmission rate, and interruptions may occur during packet transmission. The unreliable nature of retrial queuing systems greatly affects the system's operation and performance measures. At the same time, completely stopping production is not feasible as it can lead to delays in fulfilling the orders. Hence, during such failures, machines or operators with lower processing rates can continue to work to ensure a smoother operation. Additionally, the authors examined the possibility of having a backup server available to provide service at a reduced rate in cases where the main server is unavailable. Many recent papers have extensively studied retrial queuing systems with unreliable servers, [5],[6] are just a few examples.

The main objective of this study is to investigate the impact of the unreliable operation of a system by comparing various failure time distributions on performance measures such as the mean response time of a customer or the service unit utilization. This paper is a continuation of the previous work by the authors [7], where the system had an unreliable server, but now, if the server is unavailable a backup server takes its place to serve incoming requests. To obtain the desired performance measures, a simulation model was developed using SimPack [8], a set of C/C++ libraries and executable programs for computer simulation. Simulation is an excellent alternative to deriving exact formulas, particularly when it is problematic or almost impossible. The user can apply as many distributions as needed to approximate performance measures. In this paper, we present a sensitivity analysis of various failure time distributions on the main performance measures. We illustrate the results through graphical representations of interesting phenomena related to sensitivity problems.

2. System model

The system under consideration is a retrial queuing system with an unreliable server and a finite-source. The source contains N customers, each generating primary

customer requests with a rate of λ, such that inter-arrival times are exponentially distributed with a parameter of λ. Note that our model does not include waiting queues, so incoming customers occupy the server only when it is available and not busy. The service time of primary customers follows an exponential distribution with a parameter of μ_1. Following a successful service, the customer returns to the source. However, if an arriving customer (either from the source or orbit) encounters the server in a busy or failed state, the request is forwarded to the orbit. In the orbit, the customer may make an attempt to get its service requirement after an exponentially distributed random time with a parameter of σ. The system is assumed to have an unreliable server that can break down according to different distributions such as gamma, hypo-exponential, hyper-exponential, Pareto, and lognormal, each with different parameters but the same mean value. The repair process begins immediately after the server fails, and the repair time is exponentially distributed with parameter γ_2. If the server is busy and fails, the customer is immediately transferred to the orbit. All customers in the source can generate requests even if the service unit is unavailable, but these requests are directed to the backup server, which serves at a reduced rate (this is also an exponentially distributed random variable with parameter μ_3) when the main server is unavailable. The backup server is assumed to be reliable and works only if the main server is down. In the case of a busy backup server, the incoming requests are placed in the orbit. However, when the server is idle, it can initiate an outgoing call to the customers in the orbit after a random time, which is exponentially distributed with rate τ. The service time of these secondary customers follows an exponential distribution with parameters μ_2. The assumption made during model creation is that all random variables are completely independent of one another.

3. Simulation results

We utilized a statistical module class providing a statistical analysis tool that enables us to quantitatively estimate the mean and variance values of observed variables using the batch mean method. The method aggregates n successive observations of a steady-state simulation to generate a sequence of independent samples. The batch mean method is a common technique used to establish confidence intervals for the steady-state mean of a process. To ensure the sample averages are approximately independent, large batches are required. More information on the batch mean method can be found in [9]. We conducted simulations with a 99.9% confidence level, and the simulation run was halted when the relative half-width of the confidence interval reached 0.00001.

In this section, we aimed to set the parameters of failure time for each distribution in such a way that the mean value and variance would be equal. The fitting process

Table 1. Used numerical values of model parameters

$\mathbf{N}$	$\boldsymbol{\lambda}$	$\boldsymbol{\gamma_2}$	$\boldsymbol{\sigma}$	$\boldsymbol{\mu_1}$	$\boldsymbol{\mu_2}$	$\boldsymbol{\nu}$	$\boldsymbol{\mu_3}$
100	0.01	1	0.01	1	1.2	0.02	0.1;0.6

Table 2. Parameters of failure time

Distribution	**Gamma**	**Hyper-exponential**	**Pareto**	**Lognormal**
Parameters	$\alpha = 0.6$ $\beta = 0.5$	$p = 0.25$ $\lambda_1 = 0.41667$ $\lambda_2 = 1.25$	$\alpha = 2.2649$ $k = 0.67018$	$m = -0.3081$ $\sigma = 0.99037$
Mean	1.2			
Variance	2.4			
Squared coefficient of variation	1.6666666667			

used for this purpose can be found in the following paper [10]. Four different distributions were considered in order to investigate their impact on performance measures. The hyper-exponential distribution was chosen to ensure that the squared coefficient of variation is greater than one. Table 2 presents the input parameters of the various distributions, while Table 1 shows the values of other applied parameters.

The steady-state distribution for different failure time distributions is presented in Figure 1. On the X-axes i represents the number of customers located in the system, and on the Y-axes $P(i)$ denotes the probability that exactly i customer is found in the system. Upon closer examination of the curves, it can be observed that all of them resemble the normal distribution. Although the Pareto distribution appears to have more customers in the system, there are no significant differences among the various distributions tested. Including a backup server results in a lower mean number of customers in the system in comparison with the paper of [7].

Figure 2 illustrates the relationship between the mean response time of customers and the arrival intensity. Consistent with the observations from Figure 1, the highest mean response time is observed with the Pareto distribution. However, the differences among the other distributions are more noticeable. The gamma distribution yields the lowest mean response time. Interestingly, as the arrival intensity increases, the mean response time initially increases, but then starts to decrease after a certain point. This is a unique feature of retrial queuing systems with a finite source, and is a general characteristic when suitable parameter settings are used. Due to page limitations, other figures in connection with the effect of using a backup server and parameter setting can be watched in the extended version of the paper.

4. Conclusion

We present a retrial queuing system with finite source and two-way communication, where there is a primary server that is unreliable, and there is also a secondary service unit replacing it in the faulty periods. Moreover, we perform a sensitivity

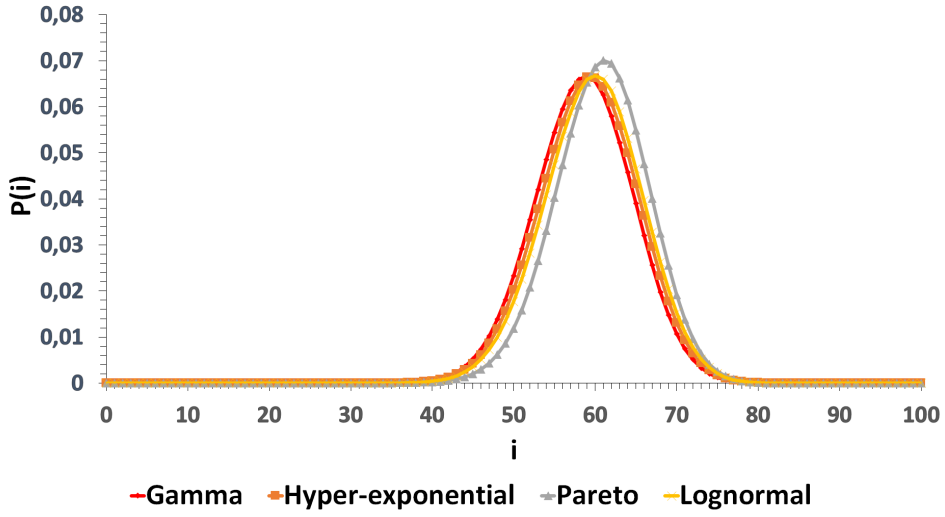


Fig. 1. Comparison of steady-state distributions

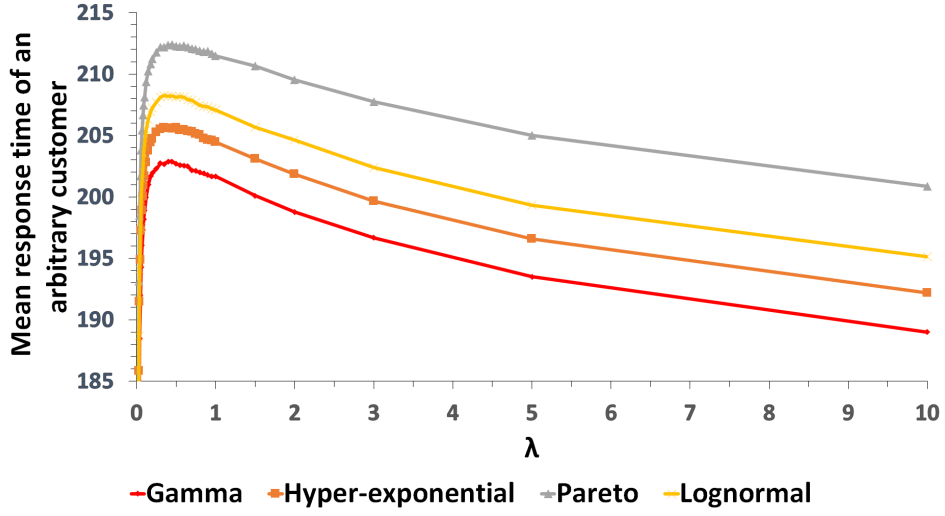


Fig. 2. Mean response time vs. arrival intensity

analysis using various random number generators to explore the effect of different distributions on the performance measures. We observe that when the squared coefficient of variation is greater than one, the mean response time of a customer

exhibits some disparity among the values, but the influence is negligible when it is less than one. Using a backup service unit may significantly decrease the time spent in the system of the customers, especially in those scenarios where Future work may include exploring additional distributions or incorporating new system features, such as vacation.

REFERENCES

1. V. I. Dragieva, Number of retrials in a finite source retrial queue with unreliable server., Asia-Pac. J. Oper. Res. 31 (2) (2014) 23. doi:10.1142/S0217595914400053.
2. D. Fiems, T. Phung-Duc, Light-traffic analysis of random access systems without collisions, Annals of Operations Research (2017) 1–17.
3. V. Dragieva, T. Phung-Duc, Two-way communication $M/M/1//N$ retrial queue, in: International Conference on Analytical and Stochastic Modeling Techniques and Applications, Springer, 2017, pp. 81–94.
4. A. Kuki, J. Sztrik, Á. Tóth, T. Bérczes, A Contribution to Modeling Two-Way Communication with Retrial Queueing Systems , in: Information Technologies and Mathematical Modelling. Queueing Theory and Applications, Springer, 2018, pp. 236–247.
5. N. Gharbi, B. Nemmouchi, L. Mokdad, J. Ben-Othman, The impact of breakdowns disciplines and repeated attempts on performances of small cell networks, Journal of Computational Science 5 (4) (2014) 633–644.
6. A. Krishnamoorthy, P. K. Pramod, S. R. Chakravarthy, Queues with interruptions: a survey, TOP 22 (1) (2014) 290–320. doi:10.1007/s11750-012-0256-6.
7. J. Sztrik, Á. Tóth, Á. Pintér, Z. Bács, The effect of operation time of the server on the performance of finite-source retrial queues with two-way communications to the orbit, Journal of Mathematical Sciences 267 (2022) 196–204. doi:10.1007/s10958-022-06124-z.
8. P. A. Fishwick, Simpack: Getting started with simulation programming in c and c++, in: In 1992 Winter Simulation Conference, 1992, pp. 154–162.
9. E. J. Chen, W. D. Kelton, A procedure for generating batch-means confidence intervals for simulation: Checking independence and normality, SIMULATION 83 (10) (2007) 683–694.
10. A. Toth, J. Sztrik, A. Kuki, T. Berczes, D. Effosinin, Reliability analysis of finite-source retrial queues with outgoing calls using simulation, in: 2019 International Conference on Information and Digital Technologies (IDT), 2019, pp. 504–511. doi:10.1109/DT.2019.8813419.

UDC: 621.391

Analysis of Procedures to Ensure the Required QoS Indicators in Multiservice Access Nodes

M.S. Stepanov[1], S.N. Stepanov[2], M.G. Kanischeva[3], F. S. Kroshin[4]

[1,2,3]Moscow Technical University of Communications and Informatics, Department of communication networks and commutation systems, 8A, Aviamotornaya str., Moscow, 111024, Russia

[2]Kotel'nikov Institute of Radio Engineering and Electronics of RAS, Mokhovaya 11-7, Moscow, 125009, Russia

mihstep@yandex.ru, stpnvsrg@gmail.com, m.g.kanishcheva@mtuci.ru, f.s.kroshin@mtuci.ru

Abstract

A generalized mathematical model of joint servicing of real-time traffic and data (files) traffic in a multiservice access node has been constructed and investigated. The model takes into account the presence of a priority for real-time traffic, as well as the group nature of the arrival, elastic properties, the possibility of waiting and aging of the transmitted information for data traffic. The coming of requests for the transmission of real-time traffic and groups of files obeys the Poisson laws, and the service times have an exponential distribution. The definitions of the indicators of the quality of the joint service of incoming requests are formulated and a method for their evaluation based on the solution of a system of equilibrium equations is considered. The use of the model for solving the problems of estimating the value of transmission resource required to serve the offered traffic for given QoS indicators and estimating the volume of the offered traffic offloaded in a situation of congestion to other access nodes in order to achieve the specified QoS indicators is considered.

Keywords: multiservice traffic, real-time traffic, elastic traffic, batch arrival of files, aging of transmitted information, transmission resource planning

1. Introduction

Achieving the specified indicators of the quality of service of incoming requests for the provision of various types of services is the main task of modeling networks and communication systems. It is especially important to find its solution for access nodes that perform the function of concentrating multiservice subscriber traffic. Depending on the formulation, the problem to be solved can be classified into the following two categories [1, 2].

- *Problem of strategic planning.* Development of software and analytical tools for estimating the volume of a resource that provides the required quality of service for given flows of information load.
- *Problem of operational planning.* Estimation of the maximum allowable amount of information load that can be served on a given resource with the required quality. This task is also called traffic offloading.

Both tasks are solved by constructing a mathematical model that takes into account the main features of the formation of incoming requests and the distribution of the access node information transmission resource during their service. The list of features includes: the multiservice nature of incoming requests, the priority of applications for the transmission of real-time traffic, as well as the group nature of arrival, elastic properties, the possibility of waiting and aging of the transmitted information for data traffic. We will assume that data traffic consists of files containing the results of observations and other information messages similar to them in terms of properties. Depending on the problem statement, all of the above features may be present in the analyzed model or considered in a smaller aggregate. Some of the listed characteristics of the multiservice access node were taken into account in publications [3–10]. The novelty of this work consist in construction of a generalized mathematical model of an access node having mentioned features and the analysis of the possibilities of its use to provide the required QoS indicators in the joint service of traffic of modern communication applications.

2. Model Description

Let us denote by C the throughput of the multiservice access node provided by the used communication standard. For the convenience of modeling, we introduce the concept of a virtual channel, which will be used to numerically characterize the resource for transmitting information provided to users. Let us denote by c the information transfer rate of one channel. The choice of the value of c depends on the problem statement, for example, it may be the minimum requirement for the transfer rate required to service the ordered services.

The access node processes a Poisson flow of requests for the provision of real-time services of intensity λ_r, divided into n service categories. With probability $p_{r,k}$, the request belongs to the k-th category, requires c_k bit/s, and occupies the resource random time that has an exponential distribution with the parameter α_k, $k = 1, \ldots, n$. To build the model, the C and c_k are converted to the virtual channel format. The v total number of resource units (r.u.) is calculated from expression $v = \left\lfloor \frac{C}{c} \right\rfloor$ r.u. The requests of the kth flow require $b_k = \left\lceil \frac{c_k}{c} \right\rceil$ resource units for their service, and the moments of their arrival form a Poisson flow of intensity $\lambda_{r,k} = \lambda p_{r,k}$, $k = 1, \ldots, n$

Together with the real-time traffic, the access node processes the Poisson group flow of requests for the transmission of elastic data traffic in the form of files. Denote by λ_d the intensity of this flow. The appearance of each request means that with probability f_j it is necessary to transfer a group of j files, $j = 1, 2, \ldots, m_g$ and $\sum_{j=1}^{m_g} f_j = 1$. Denote by b_m the average number of files in one group $b_m = \sum_{j=1}^{m_g} f_j\, j$. Files that are not accepted in service when they arrive may wait for a resource to be released. The number of waiting places will be denoted by w. The waiting time is limited by a random variable that has an exponential distribution with the parameter σ. If after this time the file has not been transferred, it is assumed that the transmitted information is outdated and the file is considered lost without renewal. The procedure for generating incoming requests can be considered in Fig. 1, if we put $p_r = p_d = 0$.

Let's assume that the size of each file has an exponential distribution with the mean value F expressed in bits, and the minimum amount of resource used to transfer one file is one channel. It is clear that the time until the end of the transmission of one file by one channel has an exponential distribution with the parameter α_d. Let's denote by ℓ the number of channels used to service real-time traffic, and by i_d we'll denote the total number of files being served and waiting. Files are served using $(v - \ell)$ channels. Let $0 < i_d \le v - \ell$ and $s = \left\lfloor \frac{v-\ell}{i_d} \right\rfloor$.

The procedure for sharing a resource between files under maintenance follows the provisions of the Processor Sharing discipline. As a result, free channels are divided between i_d files according to the following rule. To serve each of the $(v - \ell - si_d)$ files, $(s+1)$ channels are used, and to serve each of the $((s+1)i_d - (v-\ell))$ files, s channels. Since all $(v - \ell)$ channels are busy, it is easy to show that the time until the end of the transmission of one of the i_d files being served has an exponential distribution with the parameter $(v - \ell)\alpha_d$. It is also clear that when the inequality $i_d > v - \ell$ is fulfilled, in the considered access node model $(v - \ell)$ files are transmitted using the capabilities of one channel, and $(i_d - v + \ell)$ files will be waiting for service to start.

An incoming request for real-time traffic transmission has priority in occupancy of the resource, reducing, if necessary, the number of channels used for file transmission to the value of one channel. Let us denote by $i_{r,k}(t)$, $k = 1, \ldots, n$ the number of requests of the k-th flow for the transmission of traffic of real-time services that are in service at time t, and denote by $i_d(t)$ the number of files being served and waiting at time t. The dynamics of changes in the number of requests located in the access node at different stages of service is described by the Markov process $r(t) = (i_{r,1}(t), \ldots, i_{r,n}(t), i_d(t))$ defined on the finite state space S, which includes

the states $(i_{r,1}, \ldots, i_{r,n}, i_d)$, with components

$$i_{r,1} = 0, 1, \ldots, \left\lfloor \frac{v}{b_1} \right\rfloor; \quad i_{r,2} = 0, 1, \ldots, \left\lfloor \frac{v - i_{r,1}b_1}{b_2} \right\rfloor; \tag{1}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$i_{r,n} = 0, 1, \ldots, \left\lfloor \frac{v - i_{r,1}b_1 - \ldots - i_{r,n-1}b_{n-1}}{b_n} \right\rfloor;$$

$$i_d = 0, 1, \ldots, v + w - i_{r,1}b_1 - \ldots - i_{r,n}b_n.$$

3. Evaluation of Performance Measures

The quality of service for the requests of the k-th flow for the transmission of real-time traffic is determined by the portion of lost requests $\pi_{r,k}$ and the average number of busy virtual channels $m_{r,k}$. The value of the last characteristic makes it possible to calculate the average number of requests of the kth flow being serviced $y_{r,k} = m_{r,k}/b_k$ and the average used bandwidth of the access node occupied by them $z_{r,k} = m_{r,k}c$. The quality of service for the requests of elastic traffic is determined by the portion of files lost for all reasons analyzed in the model, π_d, the average number of busy virtual channels m_d, the average used bandwidth of the access node $z_d = m_d c$.

The introduced performance measures can be calculated if stationary probabilities $p(i_{r,1}, \ldots, i_{r,n}, i_d)$ of states $(i_{r,1}, \ldots, i_{r,n}, i_d) \in S$ are known. For the state $(i_{r,1}, \ldots, i_{r,n}, i_d)$ let ℓ denote the number of virtual channels used to service real-time traffic $\ell = i_{r,1}b_1 + \cdots + i_{r,n}b_n$ and assume that the maximum size of a group of incoming files is determined from the expression $m_g = v + w$. We have the following expressions

$$\pi_{r,k} = \sum_{\left\{(i_{r,1},\ldots,i_{r,n},i_d)\in S \mid \ell+i_d+b_k>v\right\}} p(i_{r,1}, \ldots, i_{r,n}, i_d); \tag{2}$$

$$m_{r,k} = \sum_{(i_{r,1},\ldots,i_{r,n},i_d)\in S} p(i_{r,1}, \ldots, i_{r,n}, i_d) i_{r,k} b_k; \quad k = 1, \ldots, n;$$

$$\pi_d = \left(\sum_{i=0}^{i_d} \sum_{\left\{(i_{r,1},\ldots,i_{r,n},i_d)\in S \mid \ell+i_d=v+w\right\}} p(i_{r,1}, \ldots, i_{r,n}, i_d - i) \lambda_d \sum_{j=i+1}^{v+w} f_j\,(j-i) + \right.$$

$$\left. + \sum_{\left\{(i_{r,1},\ldots,i_{r,n},i_d)\in S \mid \ell+i_d>v\right\}} p(i_{r,1}, \ldots, i_{r,n}, i_d)(\ell + i_d - v)\sigma \right) \frac{1}{\lambda_d b_m};$$

$$m_d = \sum_{\left\{(i_{r,1},\ldots,i_{r,n},i_d)\in S \mid i_d>0\right\}} p(i_{r,1},\ldots,i_{r,n},i_d)(v-\ell).$$

Similarly, other indicators of servicing of coming requests that are not included in the considered list are determined. These are: average number of files waiting, average queue time, average file transfer time, average bitrate used to transfer a file, etc. In order to evaluate the performance measures according to the definition introduced above, it is necessary to compose and solve a system of equilibrium equations (SEE). It looks as follows

$$P(i_{r,1},\ldots,i_{r,n},i_d)\Big\{\sum_{k=1}^{n}\Big(\lambda_{r,k}I(\ell+i_d+b_k\le v)+i_k\alpha_{r,k}I(i_{r,k}>0)\Big)+ \tag{3}$$

$$+\lambda_d I(\ell+i_d+1\le v+w)+(v-\ell)\alpha_d I(i_d>0)+(\ell+i_d-v)\sigma I(\ell+i_d>v)\Big\}=$$

$$=\sum_{k=1}^{n}P(i_{r,1},\ldots,i_{r,k}-1,\ldots,i_{r,n},i_d)\lambda_{r,k}I(i_{r,k}>0,\ell+i_d\le v)+$$

$$+\sum_{i=1}^{i_d}P(i_{r,1},\ldots,i_{r,n},i_d-i)\lambda_d\left(f_i+I(\ell+i_d=v+w)\sum_{j=i+1}^{v+w}f_j\right)+$$

$$+\sum_{k=1}^{n}P(i_{r,1},\ldots,i_{r,k}+1,\ldots,i_{r,n},i_d)(i_{r,k}+1)\alpha_{r,k}\times$$

$$\times\Big(I(\ell+b_k+i_d\le v)+I(\ell+b_k+i_d>v,\ell+b_k\le v,\ell+b_k+i_d\le v+w)\Big)+$$

$$+P(i_{r,1},\ldots,i_{r,n},i_d+1)\times\Big((v-\ell)\alpha_d I(\ell+i_d+1\le v+w)+$$

$$+(\ell+i_d+1-v)\sigma I(\ell+i_d+1\le v+w,\ell+i_d+1>v)\Big);$$

$$\sum_{(i_{r,1},\ldots,i_{r,n},i_d)\in S}p(i_{r,1},\ldots,i_{r,n},i_d)=1.$$

In the above expression, the indicator function $I(\cdot)$ is determined from the relation

$$I(\cdot)=\begin{cases}1, & \text{if the condition formulated in parentheses is fulfilled;}\\ 0, & \text{in opposite case.}\end{cases}$$

The matrix of the system of equilibrium equations does not have any properties that would allow the use of recursive or matrix methods for its solution. In this case,

the solution of the SEE can be obtained by the iterative Gauss-Seidel method for the number of unknowns in the SEE equal up to several millions. Details of the use of this approach can be found in [7].

4. Examples of Using the Model in Applications

Let us consider the use of the constructed model to solve the problems of estimating the value of transmission resource required to serve the offered traffic for given QoS indicators and estimating the volume of the offered traffic offloaded in a situation of congestion to other access nodes in order to achieve the specified QoS indicators. Assume that resource sufficiency is determined from the condition $\max(\pi_{r,1}, \ldots, \pi_{r,n}, \pi_d) < \pi$, where π — prescribed value of losses. Let's start with the solution of the first of the listed problems. The algorithm for selecting the required resource value depends on the structural parameters of the model. If the number of unknowns in the SEE does not exceed several millions, then the results of solving the SEE by the Gauss-Seidel algorithm are used to estimate the QoS indicators. If the number of unknowns in the SEE exceeds this value, then the decomposition method similar to one proposed in [10] is used to evaluate the characteristics of the model. It is based on a joint numerical analysis of two particular cases of the model under study: when the access node resource is occupied only by real-time service traffic or only by elastic data traffic. In each of the listed cases, efficient recursive algorithms are used to calculate the characteristics. It can be shown that the results of calculating the characteristics will be asymptotically exact in the region of low losses. This corresponds to the range of load parameters, where the problem of estimating the required throughput of the access node is solved.

Let's turn to the task of offloading traffic. Usually it is solved in a situation of local overloads of the access node [10]. Let's assume that the available traffic leads to exceeding the level of loss of requests. Excess traffic is directed to other access nodes. For wireless networks, these may be communication systems operating in unlicensed frequency bands. Let us denote by p_r the share of requests for servicing real-time services, by p_d we will denote the share of requests for the transfer of elastic data that will be redirected for servicing to other access nodes so that losses of the remaining part of the traffic on a given resource does not exceed normative values. The procedure for offloading traffic is shown in Fig. 1.

Let us give a numerical example illustrating the algorithm for estimating p_r and p_d. Let us consider the model of an overloaded access node with the following parameters: $C = 60$ Mbps; $n = 2$; $c_1 = 2$ Mbps; $c_2 = 5$ Mbps. Based on the assumptions made, we get the structural parameters of the model: $c = 1$ Mbps; $v = 60$ r.u.; $b_1 = 2$ r.u.; $b_2 = 5$ r.u. Let's assume that $F = 80$ Mbit. The average file transfer time by one channel is 80 s. When performing calculations, this time will be

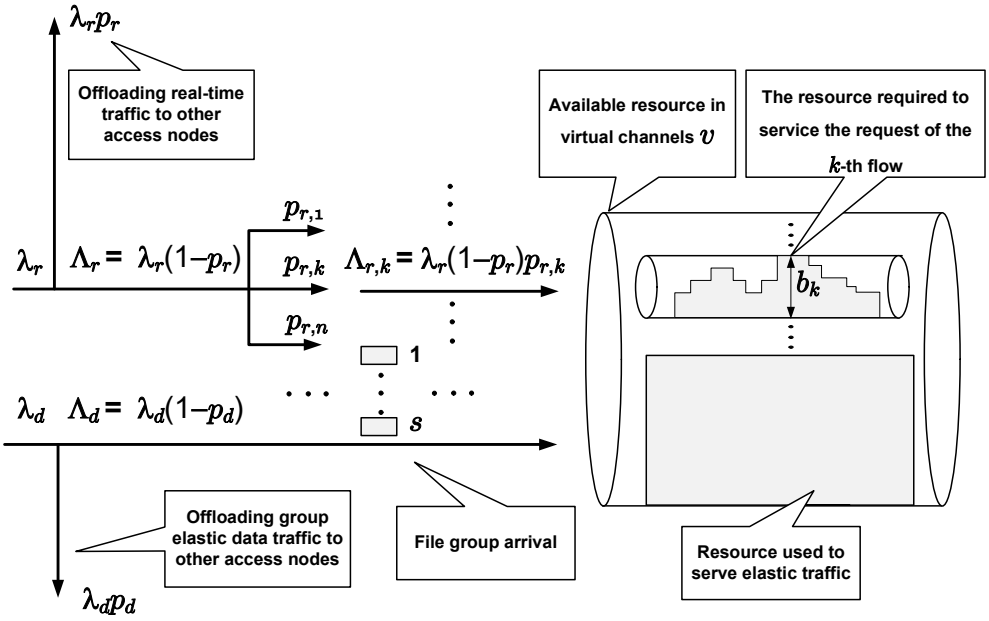


Fig. 1. The procedure for offloading traffic in a multiservice access node when servicing real-time service traffic and elastic data

taken as a unit. From this assumption $\alpha_d = 1$. For real-time services, the service time parameters are selected from the $\alpha_1 = 1$ and $\alpha_2 = 1$. Let us suppose that $w = 10$; $\lambda = 18$ req./s; $p_{r,1} = 5/6$; $p_{r,2} = 1/6$; $f_i = 1/30$, $i = 1, \ldots, 30$, $\sigma = 0{,}1$. Let's assume for simplicity that the offload is used only for real-time traffic, i.e. $p_d = 0$. After the implementation of the offloading procedure, the access node serves the flows of requests with the following rates: $\Lambda_{r,1} = \lambda_r(1 - p_r)p_{r,1}$; $\Lambda_{r,2} = \lambda_r(1 - p_r)p_{r,2}$; $\Lambda_d = \lambda_d$. Changing p_r, we find the value $p_r = 0{,}68$, at which the required level of requests losses $\max(\pi_{r,1}, \pi_{r,2}, \pi_d) < \pi$ is reached. In this case $\pi = 0{,}05$. The results of estimating the portion of offloaded traffic of real-time services are shown in Fig. 2.

5. Conclusion

A generalized mathematical model of joint servicing of real-time traffic and data (files) traffic in a multiservice access node has been constructed and investigated. The model takes into account the presence of a priority for real-time traffic, as well

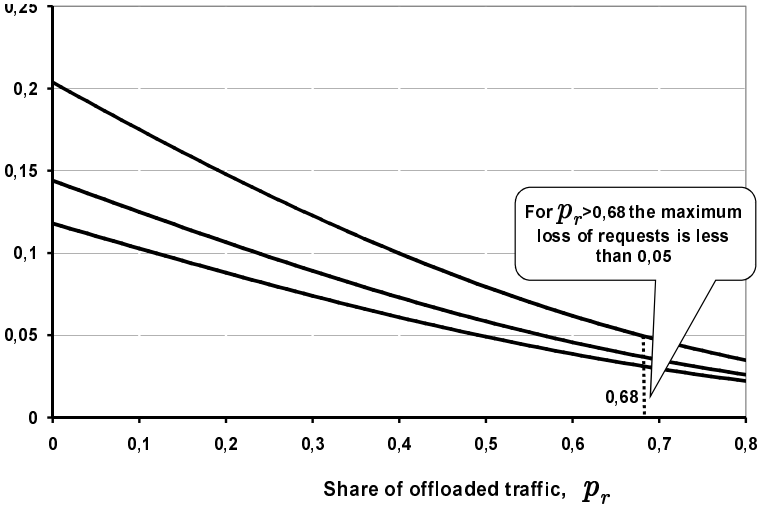


Fig. 2. Results of evaluating the share of offloaded traffic of real-time

as the group nature of the arrival, elastic properties, the possibility of waiting and aging of the transmitted information for data traffic. The arrivals of requests for the transmission of real-time traffic and groups of files obeys the Poisson laws, and the service times have an exponential distribution. The change of model states is described by a multidimensional Markov process with a finite state space. The definitions of the performance measures of the joint service of incoming requests are formulated and a method for their evaluation based on the solution of a system of equilibrium equations is considered. The constructed model and the results of its analysis can be used to solve the problems of estimating the value of the transmission resource required to serve offered traffic for given QoS indicators and estimating the volume of the offered traffic offloaded in a situation of congestion to other access nodes. Numerical examples illustrating this point are given. Additional results obtained for this model include approximate evaluation methods based on the implementation of model decomposition principles and the use of a system of simplified equilibrium equations. It has been established that the obtained estimates of the characteristics are asymptotically exact in the region of small and large losses [11]. The results obtained were also used to solve the problems of estimating the necessary information transmission resource of access nodes in satellite communication networks [12].

REFERENCES

1. Study on scenarios and requirements for next generation access technologies. 3GPP Technical Report (TR) 138.913 version 15.0.0 Release 15. (2018)
2. Network Slice Selection Services. 3GPP Technical Specification (TS) 129.531 version 15.5.0. Release 15. (2019)
3. Bonald, T., Virtamo, J.: A recursive formula for multirate systems with elastic traffic. IEEE Communicat. Lett. **9**. (8), 753–755 (2005)
4. Gudkova, I.A., Samouylov, K.E.: Modelling a radio admission control scheme for video telephony service in wireless networks. Lecture Notes Comput. Sci. **7469**, 208-215 (2012)
5. Begishev, V., Petrov, V., Samuylov, A., Moltchanov, D., Andreev, S., Koucheryavy, Y., Samouylov, K.: Resource Allocation and Sharing for Heterogeneous Data Collection over Conventional 3GPP LTE and Emerging NB-IoT Technologies. Comput. Communicat. **120** (2). 93–101 (2018)
6. Basharin, G. P., Gaidamaka, Yu. V., Samouylov, K. E.: Mathematical Theory of Teletraffic and Its Application to the Analysis of Multiservice Communication of Next Generation Networks. Automatic Control and Computer Sciences, **47** (2). 62-69 (2013)
7. Stepanov, S. N.: Teletraffic Theory: Concepts, Models, Applications (in Russian). Goriachay Linia-Telecom, Moscow, 2015
8. Stepanov, S.N., Stepanov, M.S.: Methods for Estimating the Required Volume of Resource for Multiservice Access Nodes. Autom, Remote Control. **81** (12). 2244–2261 (2020)
9. Stepanov, M.S., Stepanov, S.N., Andrabi, U.M., Petrov, D.S., Ndayikunda, J.: The Increasing of Resource Sharing Efficiency in Network Slicing Implementation. Lecture Notes Comput. Sci. (LNCS). **1552**. 18–35 (2022)
10. Chen, J., Chang, Z., Guo, X., Li, R., Han Z., Hamalainen, T.: Resource Allocation and Computation Offloading for Multi-Access Edge Computing With Fronthaul and Backhaul Constraints, IEEE Transactions on Vehicular Technology. **70** (8). (8037-8049) 2021
11. 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** (11). 2004–2015 (2017)
12. Stepanov, S.N., Andrabi, U.M., Stepanov, M.S., Ndayikunda, J.: Reservation Based Joint Servicing of Real Time and Batched Traffic in Inter Satellite Link. Proc. of 2020 Systems of Signals Generating and Processing in the Field of on Board Communications. Moscow. Russia. (1-5) (2020). doi: 10.1109/IEEECONF48371.2020.9078542

УДК: 519.2

Приоритетная система обслуживания с профилактиками прибора в общих предположениях на управляющие последовательности

Берговин Алексей Константинович[1], Ушаков Владимир Георгиевич[2]

[1]Россия, Москва, Московский государственный университет имени М.В. Ломоносова, факультет вычислительной математики и кибернетики, кафедра математической статистики, аспирант

[2]Россия, Москва, Московский государственный университет имени М.В. Ломоносова, факультет вычислительной математики и кибернетики, кафедра математической статистики, профессор, д.ф.-м.н

alexey.bergovin@gmail.com, vgushakov@mail.ru

Аннотация

Данная работа посвящена рассмотрению приоритетной системы массового обслуживания, в которой предусмотрены профилактики обслуживающего прибора. Интервалы между поступлениями требований являются независимыми в совокупности случайными величинами с произвольным абсолютно непрерывным распределением. Поступающие требования разделяются на несколько приоритетных классов независимо друг от друга и от состояния системы, между классами установлена дисциплина относительного приоритета. Распределение времен обслуживания каждого класса является произвольным абсолютно непрерывным. В те моменты времени когда система становится свободной, обслуживающий прибор отправляется на профилактику, которая длится случайное время с заданной функцией распределения, длительность профилактики не зависит ни от входящего потока, ни от времен обслуживания, ни от длительности других профилактик. В работе найдено нестационарное совместное распределение количества заявок каждого приоритета в системе.

Ключевые слова: относительный приоритет, профилактики прибора, одноканальная система, длина очереди

1. Введение

Интерес к таким системам вызван их обширным практическим применением, например, в компьютерных и телекоммуникационных сетях, колл-центрах.

Постановки задачи и существующие результаты могут быть найдены в обзоре [1], статье [2], а также в монографиях [3, 4]. Неприоритетная и приоритетная системы с гиперэкспоненциальным входящим потоком рассмотрены в работах [5] и [6] соответственно.

2. Описание системы и обозначения

Рассматривается однолинейная система обслуживания с неограниченным числом мест для ожидания, в которую поступает рекуррентный поток требований с функцией распределения интервалов между поступлениями $A(x)$. Требования разделяются на r приоритетных классов. Будем считать, что требования i-го класса обладают приоритетом над требованиями j-го класса, $1 \leqslant i < j \leqslant r$. Поступившее требование относится к i-му приоритетному классу с вероятностью p_i, $\sum\limits_{i=1}^{r} p_i = 1$, независимо от остальных требований и состояния системы. Длительности обслуживания — независимые в совокупности случайные величины с функцией распределения $B_i(x)$ для требований из i-го класса.

Если после завершения обслуживания заявки в системе отсутствуют требования всех классов, то обслуживающий прибор отправляется на профилактику, которая длится случайное время с функцией распределения $C(x)$. Если за время профилактики поступают требования, то после ее завершения прибор начинает их обслуживать. Если же требований не поступило, то прибор снова отправляется на профилактику и т.д.. Длительности различных профилактик являются независимыми случайными величинами и не зависят от входящего потока и времени обслуживания.

Предполагается, что $A(x) < 1$, $B_i(x) < 1$, $C(x) < 1$, $\forall x$, $1 \leqslant i \leqslant r$, и существуют плотности распределения $a(x)$, $b_i(x)$, $c(x)$.

Рассматриваются следующие случайные процессы:

$\mathbf{L}(t) = (L_1(t), \ldots, L_r(t))$, где $L_i(t)$ — число требований i-го класса в системе в момент времени t;

$i(t) = i, i \in \{0, 1, \ldots, r\}$ — либо номер класса ($1 \leqslant i \leqslant r$), требование которого обслуживается в момент времени t (если $\mathbf{L}(t) \neq \mathbf{0}$), либо $i(t) = 0$ (это означает, что в данный момент прибор находится на профилактике);

$x(t)$ — время, прошедшее с начала обслуживания до момента t, если $i(t) \neq 0$, или время, прошедшее с начала профилактики до момента t, если $i(t) = 0$;

$y(t)$ — время, прошедшее с момента последнего поступления требования до момента t.

Будем предполагать, что в момент начала наблюдения за системой обслуживания $t = 0$ в системе нет требований, и с начала профилактики прибора

прошло случайное время с плотностью $f(x)$, а с момента последнего поступления требования случайное время с плотностью $g(x) = \frac{1-A(x)}{\alpha_1}$, где $\alpha_1 = \int\limits_0^\infty u \cdot a(u)du$.

Обозначения, используемые далее:

$$\alpha(s) = \int\limits_0^\infty e^{-su}a(u)du,\ \beta_i = \beta_i(s) = \int\limits_0^\infty e^{-su}b_i(u)du,\ i = \overline{1,r},\ \gamma(s) = \int\limits_0^\infty e^{-su}c(u)du.$$

$$\nu(x) = \frac{a(x)}{1-A(x)}, \quad \eta_i(x) = \frac{b_i(x)}{1-B_i(x)}, \quad \mu(x) = \frac{c(x)}{1-C(x)}.$$

$$\mathbf{(x,y)}_k^n = \sum_{i=k}^n x_i y_i, \quad \mathbf{(x,y)} = \mathbf{(x,y)}_1^r, \quad \mathbf{z}_k = (z_k, z_{k+1}, \ldots, z_r), \mathbf{z} = (z_1, \ldots, z_r).$$

$$H_1(s,v) = e^{-sv}\int\limits_0^\infty \frac{f(u)c(u+v)}{1-C(u)}du,\ \delta_{i,j} = \begin{cases} 1, & i=j, \\ 0, & i \neq j, \end{cases} \quad \mathbb{I}(A(x,y)) = \begin{cases} 1, & (x,y) \in A, \\ 0, & (x,y) \notin A, \end{cases}$$

$$\chi(s,w) = \int\limits_0^\infty c(x)e^{-(s+w)x}\int\limits_0^x e^{(s+w)u}\frac{f(u)}{1-C(u)}dudx = \int\limits_0^\infty e^{-sw}H_1(s,v)dv.$$

Для функций $F(x)$ и $G(x)$, равных 0 на отрицательной полуоси:

$$F{*}G(x) = \int\limits_0^x F(u)G(x-u)du,\ F^{*1}(x) = F(x),\ F^{*n}(x) = \int\limits_0^x F(u)F^{*(n-1)}(x-u)du,\ n \geqslant 2,$$

$$P_i(\mathbf{n},x,y,t) = \frac{\partial^2}{\partial x \partial y}\mathbb{P}(\mathbf{L}(t) = \mathbf{n}, x(t) < x, y(t) < y, i(t) = i),\ i = \overline{0,r},$$

$$p_i(\mathbf{z},x,y,s) = \int\limits_0^\infty e^{-st}\sum_{n_1=0}^\infty \cdots \sum_{n_r=0}^\infty z_1^{n_1}\ldots z_r^{n_r}P_i(\mathbf{n},x,y,t)dt,\ i = \overline{0,r},$$

3. Математическая модель

Рассматривая возможные изменения состояний процесса $(\mathbf{L}(t), x(t), y(t), i(t))$ в интервале времени $(t, t+\Delta)$, и устремляя $\Delta \to 0$, получим следующую систему дифференциальных в частных производных:

$$\frac{\partial p_0(\mathbf{z},x,y,s)}{\partial x} + \frac{\partial p_0(\mathbf{z},x,y,s)}{\partial y} = -(s + \nu(y) + \mu(x))p_0(\mathbf{z},x,y,s) + f(x)g(y),$$

$$\frac{\partial p_i(\mathbf{z},x,y,s)}{\partial x}+\frac{\partial p_i(\mathbf{z},x,y,s)}{\partial y}=-(s+\nu(y)+\eta_i(x))p_i(\mathbf{z},x,y,s),\ i=\overline{1,r},$$

$$\sum_{i=1}^{r}p_i(\mathbf{z},0,y,s)=\int_0^\infty p_0(\mathbf{z},x,y,s)\mu(x)dx+\sum_{i=1}^{r}\frac{1}{z_i}\int_0^\infty p_i(\mathbf{z},x,y,s)\eta_i(x)dx-$$

$$-\sum_{i=1}^{r}\int_0^\infty\int_0^\infty e^{-st}P_i((0,\dots,1,\dots,0),x,y,t)\eta_i(x)dxdt-\int_0^\infty\int_0^\infty e^{-st}P_0(\mathbf{0},x,y,t)\mu(x)dxdt.$$

Для перехода от системы дифференциальных уравнений в частных производных к системе обыкновенных дифференциальных уравнений, предлагается использовать следующее интегральное преобразование

$$q_i(\mathbf{z},x,w,s)=\int_0^\infty e^{-wy}\frac{1-(\mathbf{p},\mathbf{z})\int_0^y e^{wu}a(u)du}{1-A(y)}\cdot p_i(\mathbf{z},x,y,s)dy,\ i=\overline{0,r}.$$

Выбор данного интегрального преобразования обусловлен тем, что применение классического преобразование Лапласа не помогает упростить исходную систему, поэтому выбрано преобразование с указанной весовой функцией. Применив его, получим:

$$\frac{\partial q_i(\mathbf{z},x,w,s)}{\partial x}=-(s+w+\eta_i(x))q_i(\mathbf{z},x,w,s),\quad i=\overline{1,r},$$

$$\frac{\partial q_0(\mathbf{z},x,y,s)}{\partial x}=-(s+w+\mu(x))q_0(\mathbf{z},x,y,s)+f(x)\cdot\frac{1-(\mathbf{p},\mathbf{z})}{\alpha_1 w},$$

$$\sum_{i=0}^{r}q_i(\mathbf{z},0,w,s)=\int_0^\infty q_0(\mathbf{z},x,w,s)\mu(x)dx+\sum_{i=1}^{r}\frac{1}{z_i}\int_0^\infty q_i(\mathbf{z},x,w,s)\eta_i(x)dx.$$

Решения данной системы можно явно выписать:

$$q_i(\mathbf{z},x,w,s)=q_i(\mathbf{z},0,w,s)\cdot(1-B_i(x))\cdot e^{-(s+w)x},\quad i=\overline{1,r},$$

$$q_0(\mathbf{z},x,w,s)=\left(q_0(\mathbf{z},0,w,s)+\frac{1-(\mathbf{p},\mathbf{z})}{\alpha_1\cdot w}\int_0^x f(b)\frac{e^{(s+w)b}}{1-C(b)}db\right)\cdot(1-C(x))\cdot e^{-(s+w)x},$$

$$\sum_{i=0}^{r}q_i(\mathbf{z},0,w,s)=\int_0^\infty q_0(\mathbf{z},x,w,s)\mu(x)dx+\sum_{i=1}^{r}\frac{1}{z_i}\int_0^\infty q_i(\mathbf{z},x,w,s)\eta_i(x)dx.$$

4. Основной результат

Теорема 1. *Функции* $p_i(\boldsymbol{z}_i, x, y, s),\ i = \overline{0,r}$, определяются следующими соотношениями:

$$p_i(\mathbf{z}_i, x, y, s) = (1 - A(y))f_i(\mathbf{z}_i, x, y, s),$$

где $f_i(\mathbf{z}, x, y, s),\ i = \overline{0,r}$, — единственные решения уравнений:

$$f_i(\mathbf{z}_i, x, y, s) = (\mathbf{p}, \mathbf{z}) \int\limits_y^\infty a(u - y)f_i(\mathbf{z}_i, x, u, s)du + d_i(\mathbf{z}_i, x, y, s),\ i = \overline{0,r}.$$

Для $i = \overline{1,r}$

$$d_i(\mathbf{z}_i, x, y, s) = \frac{1 - B_i(x)}{e^{sx}} \left(h_i(\mathbf{z}_i, y - x, s) - p_i z_i \int\limits_{y-x}^\infty a(v - y + x)h_i(\mathbf{z}_i, v, s)dv \right) \cdot \mathbb{I}(y \geqslant x),$$

$$d_0(\mathbf{z}, x, y, s) = \frac{1 - C(x)}{e^{sx}} \left(h_0(y - x, s) - (\mathbf{p}, \mathbf{z}) \int\limits_{y-x}^\infty a(v - y + x)h_0(v, s)dv \right) \cdot \mathbb{I}(y \geqslant x) +$$
$$+ \frac{1 - C(x)}{\alpha_1}(1 - (\mathbf{p}, \mathbf{z})) \int\limits_0^{\min(x,y)} e^{-sv} \frac{f(x - v)}{1 - C(x - v)}dv.$$

где

$$h_0(y, s) = \frac{p_0(y, s)}{1 - A(y)}, \qquad h_k(\mathbf{z}, y, s) = \frac{p_k(\mathbf{z}, 0, y, s)}{1 - A(y)}.$$

а функции $h_k(\mathbf{z}_k, y, s),\ k = \overline{1,r}$, являются единственными решениями уравнений:

$$h_k(\mathbf{z}_k, y, s) - \left((\mathbf{p}, \beta(s + w))_1^{k-1} + (\mathbf{p}, \mathbf{z})_k^r \right) \int\limits_0^\infty a(u) \cdot h_k(\mathbf{z}_k, y + u, s)du =$$
$$= p_k z_k \Bigg(\sum_{m=k+1}^{r} \frac{z_m - \beta_m(s + w)}{z_m} \int\limits_0^\infty e^{wu}a(u) \cdot h_m(\mathbf{z}_m, y + u, s)du + \int\limits_0^\infty a(u)h_0(y + u, s)du -$$
$$- \int\limits_0^y e^{-s(y-v)}c(y - v) \int\limits_0^\infty a(u)h_0(y + u, s)dudv - \frac{1}{\alpha_1} \int\limits_0^\infty H_1(s, v)dv \Bigg), \quad k = \overline{1,r}.$$

Функция $h_0(y,s)$ является единственным решением уравнения

$$h_0(y,s) - \sum_{i=1}^{r} p_i \int_0^{\infty} h_0(\tau,s)d\tau \int_0^{\min(y,\tau)} e^{-s(y-v)} b_i(y-v)a(\tau-v)dv =$$

$$= \frac{1}{\alpha_1}\left(H_1(s,y) + \sum_{k=1}^{\infty} \int_0^{y} (e^{-su}c(u))^{*k} H_1(s,y-u)du\right) \cdot \left(1 - \sum_{i=1}^{r} p_i \int_0^{y-\tau} e^{-su} b_i(u)du\right).$$

Единственность решения выписанных выше уравнений следует из того, что

$$\sup_y \left| \mathbf{(p,z)} \int_y^{\infty} a(y-u)du \right| = |\mathbf{(p,z)}| < 1 \quad \forall\ |z_k| < 1$$

$$\sup_y \left| \sum_{i=1}^{r} p_i \int_0^{\infty} d\tau \int_0^{\min(y,\tau)} e^{-s(y-v)} b_i(y-v)a(\tau-v)dv \right| = \left| (\mathbf{p}, \beta(s+w)) \right| < 1.$$

5. Заключение

В работе рассмотрена приоритетная система массового обслуживания с возможностью ухода обслуживающего прибора на профилактику. По результатам исследования были найдены интегральные уравнения, решения которых однозначно определяют нестационарное совместное распределение количества требований каждого приоритетного класса в системе.

ЛИТЕРАТУРА

1. Doshi B.T. Queueing systems with vacations – a survey // Queuing Systems. 1986 N1. P. 29–66.
2. Doshi B.T.: Single server queues with vacations. In: Takagi, H. Stochastic Analysis Computer and Communication Systems, North-Holland, Amsterdam (1990), 217-265.
3. N. Tian, Z. G. Zhang, Vacation Queueing Models: Theory and Applications, Springer-Verlag, New York, 2006
4. Takagi H. Queueing Analysis: A Foundation of Performance Analysis. Vol. 1: Vacation and Priority Systems. Part 1. Amsterdam: Elsevier Science Publishers B.V. 1991.
5. Кондранин Е.С., Ушаков В.Г. Система обслуживания с относительным приоритетом и профилактиками прибора // Информатика и ее применения 2018. 12. №4. С. 33–38.

6. Ушаков В. Г. Система обслуживания с гиперэкспоненциальным входящим потоком и профилактиками прибора //Информатика и ее применения. 2016. 10. №2. С. 92–97.

УДК: 519.872

Математическая модель двухэтапного производственного процесса в виде двухфазной СМО с двумя входящими пуассоновскими потоками и обратной связью

Л.А. Задиранова

Национальный исследовательский Томский государственный университет, Томск, Россия

zhidkovala@mail.ru

Аннотация

В работе представлена математическая модель двухэтапного производственного процесса в виде двухфазной системы массового обслуживания (СМО) с входящими пуассоновскими потоками на первую и вторую фазы. Каждая фаза состоит из бесконечного числа устройств. Время обслуживания заявок распределено согласно экспоненциального закона. После обслуживания на первой фазе, заявка с вероятностью r_1 переходит на вторую фазу, либо, с вероятностью $1 - r_1$ возвращается на первую для повторного обслуживания. По завершению обслуживания на второй фазе, заявка с вероятностью r_2 покидает систему, либо, с вероятностью $1 - r_2$ возвращается для повторного обслуживания на вторую фазу. Получен вид производящей функции для случайного процесса, характеризующего число событий в потоке повторных обращений.

Ключевые слова: двухфазная СМО с двумя входящими потоками и обратной связью, поток повторных обращений, метод производящих функций.

1. Введение

Во времена рыночной экономики, главной целью каждого производственного предприятия является получение максимальной прибыли. Для достижения указанной цели, организации стремятся выстроить оптимальный производственный процесс с учетом всех особенностей. Так, например, требуется сократить время простоя оборудования, обеспечить его эффективное использование, сократить процент брака, оптимизировать время изготовления продукции и т.д.

Одним из методов решения поставленных задач может служить математическое моделирование производственного процесса и решение соответствующих оптимизационных задач.

Существует множество работ демонстрирующих применение многолинейных систем массового обслуживания с целью описания математических моделей реальных систем. Так в публикации [1] авторы предлагают математическую модель процесса изменения демографической ситуации в виде СМО с двумя типами заявок и неограниченным числом приборов. Применение многолинейной СМО с неограниченным буфером с целью проектирования экономных систем энергопотребления показано в [2]. В работе [3] ресурсная СМО с неограниченным числом приборов и дискретным ресурсом конечного объема единиц ресурса используется для построения математической модели отдельно стоящей соты сети 5G NR. Отметим также использование бесконечнолинейных СМО с обратной связью для описания математических моделей торговых и страховых компаний [4, 5].

Из-за сложности функционирования многих реальных систем, задача построения адекватных математических моделей остается актуальной. В данной работе представлена математическая модель двухэтапного производства в виде двухфазной бесконечнолинейной СМО с двумя входящими потоками и мгновенной обратной связью.

2. Постановка задачи

Рассмотрим пример двухэтапного производства оборудования. Пусть в компанию поступают заявки на производство изделий. Моментом начала производства на первом этапе будем считать момент регистрации заказа на предприятии, т.о. количество поступающих заказов является неограниченным, в случае, если оборудование производства занято, то время ожидания начала производства будет считаться частью времени выполнения первого этапа производства. По окончанию первого этапа, производимое изделие проходит контроль, после чего с вероятностью r_1 начинается выполнение второго этапа производства, или, с вероятностью $1 - r_1$, изделие поступает повторно на первый этап для доработки. Положим, что кроме собственных заявок производства, на второй этап могут поступать внешние заявки. По окончанию второго этапа, изделие также проходит контроль и, с вероятностью r_2 отгружается покупателю, либо, с вероятностью $1 - r_2$ отправляется на доработку. Основной характеристикой представленного процесса будем считать число изделий, которые нуждаются в доработке. Вспомогательная характеристика – число изделий на каждом этапе производства.

3. Математическая модель

Математическую модель такого производственного процесса представим в виде двухфазной бесконечнолинейной СМО с двумя входящими потоками и обратной связью. Пусть на первую фазу поступает пуассоновский поток заявок

с параметром λ_1. Время обслуживания на первой и второй фазах распределено согласно экспоненциального закона с параметрами μ_1 и μ_2 соответственно. После обслуживания на первой фазе, заявка с вероятностью r_1 переходит на вторую фазу, либо, с вероятностью $1 - r_1$, возвращается на первую для повторного обслуживания. На второй фазе происходит обслуживание как заявок, поступающих с первой фазы, так и заявок второго входящего пуассоновского потока с параметром λ_2. Обслуженная на второй фазе заявка с вероятностью r_2 покидает систему, либо, с вероятностью $1 - r_2$, возвращается на вторую фазу для повторного обслуживания. Заявки, поступающие в систему с вероятностями $1 - r_1$ и $1 - r_2$ для повторного обслуживания, будем считать событиями потока повторных обращений. Ставится задача исследования потока повторных обращений в систему.

4. Производящая функция числа занятых приборов в системе

В качестве вспомогательного результата приведем лемму без доказательства о виде производящей функции числа занятых приборов в рассматриваемой системе.

Лемма 1. *Пусть λ_1, λ_2 параметры входящих пуассоновских потоков на первую и вторую фазы системы соответственно. Время обслуживания на каждой фазе распределено согласно экспоненциального закона с параметрами μ_1, μ_2. Параметры r_k, $k = 1, 2$ определяют вероятности покинуть k-ю фазу обслуживания, $1 - r_k$, $k = 1, 2$ – вероятность вернуться на k-ю фазу для повторного обслуживания.*

Тогда производящая функция двумерного случайного процесса $\{i_k(t)\}$, $k = 1, 2$, характеризующего число занятых приборов в системе в момент времени t имеет вид

$$G(x_1, x_2, t) = exp\left\{\frac{\lambda_1(x_1-1)}{r_1\mu_1}(1 - e^{-r_1\mu_1 t}) + \frac{(\lambda_1+\lambda_2)(x_2-1)}{r_2\mu_2}(1 - e^{-r_2\mu_2 t}) + \right.$$
$$\left. + \frac{\lambda_1(x_2-1)}{(r_2\mu_2 - r_1\mu_1)}(e^{-r_2\mu_2 t} - e^{-r_1\mu_1 t})\right\}. \quad (1)$$

Полагая в (1) $t \to \infty$ получим производящую функцию двумерного процесса $\{i_1, i_2\}$ при стационарном режиме функционирования системы

$$g(x_1, x_2) = exp\left\{\frac{\lambda_1(x_1 - 1)}{r_1\mu_1} + \frac{(x_2 - 1)}{r_2\mu_2}(\lambda_1 + \lambda_2)\right\}. \quad (2)$$

Следует отметить, что данный результат можно также получить применяя теорему ВСМР.

5. Производящая функция числа событий в потоке повторных обращений

Сформулируем и докажем теорему о виде производящей функции числа событий в потоке повторных обращений.

Теорема 1. *Пусть λ_1, λ_2 параметры входящих пуассоновских потоков на первую и вторую фазы системы соответственно. Время обслуживания на каждой фазе распределено согласно экспоненциального закона с параметрами μ_1, μ_2. Параметры r_k, $k = 1, 2$ определяют вероятности покинуть k-ю фазу обслуживания, $1 - r_k$, $k = 1, 2$ – вероятность вернуться на k-ю фазу для повторного обслуживания. Тогда производящая функция случайного процесса $\{n_k(t)\}$, $k = 1, 2$, характеризующего число событий потока повторных обращений, поступивших в систему за время t, определяется выражением*

$$
\begin{aligned}
G(y_1,y_2,t) = exp\Big\{ & \tfrac{(\lambda_1+\lambda_2)}{\mu_2 r_2}\left(\tfrac{\mu_2 r_2+\mu_2(1-y_2(1-r_2)-r_2)e^{-\mu_2(1-y_2(1-r_2))t}}{\mu_2(1-y_2(1-r_2))}-1\right)+\\
&+\tfrac{\lambda_1(1-e^{-\mu_1 t(1-y_1(1-r_1))})}{\mu_1(1-y_1(1-r_1))}\left(\tfrac{\mu_1 r_1((1-y_2(1-r_2))-r_2)}{(1-y_2(1-r_2))[\mu_2(1-y_2(1-r_2))-\mu_1 r_1(1-y_1(1-r_1))]}-\right.\\
&\left.-\tfrac{r_1 r_2}{(1-y_1(1-r_1))(1-y_2(1-r_2))}+1\right)+\tfrac{\lambda_1 r_2}{\mu_1(1-y_1(1-r_1))(1-y_2(1-r_2))}+\\
&+\tfrac{\lambda_1}{r_1\mu_1}\left(\tfrac{\mu_1 r_1((1-y_2(1-r_2))-r_2)e^{-\mu_1 t(1-y_1(1-r_1))}}{(1-y_2(1-r_2))[\mu_2(1-y_2(1-r_2))-\mu_1(1-y_1(1-r_1))]}-\right.\\
&\left.-\tfrac{r_1 r_2 e^{-\mu_1 t(1-y_1(1-r_1))}}{(1-y_1(1-r_1))(1-y_2(1-r_2))}+e^{-\mu_1 t(1-y_1(1-r_1))}-1\right)+\\
&+\tfrac{\lambda_1\mu_1 r_1(1-y_2(1-r_2)-r_2)\left(e^{-\mu_2(1-y_2(1-r_2))t}-1\right)}{\mu_2(1-y_2(1-r_2))^2[\mu_2(1-y_2(1-r_2))-\mu_1(1-y_1(1-r_1))]}-\\
&-\tfrac{\lambda_1(1-y_2(1-r_2)-r_2)e^{-\mu_2(1-y_2(1-r_2))t}}{(1-y_2(1-r_2))[\mu_2(1-y_2(1-r_2))-\mu_1(1-y_1(1-r_1))]}+\\
&+\lambda_1 t\left(\tfrac{r_1 r_2}{(1-y_1(1-r_1))(1-y_2(1-r_2))}-1\right)+\lambda_2 t\left(\tfrac{r_2}{(1-y_2(1-r_2))}-1\right)\\
&+\tfrac{\lambda_2\left(1-e^{-\mu_2(1-y_2(1-r_2))t}\right)(1-y_2(1-r_2)-r_2)}{\mu_2(1-y_2(1-r_2))^2}\Big\}.
\end{aligned}
\tag{3}
$$

Доказательство. Пусть в момент времени t состояние системы определяется вектором $\{i_k(t), n_k(t)\}$, $k = 1, 2$, где $i_k(t)$, $k = 1, 2$ - число занятых приборов на k-ой фазе в момент времени t, $n_k(t)$, $k = 1, 2$ - число событий потока повторных обращений, поступивших в систему за время t. Для распределения вероятностей $P(i_1, i_2, n_1, n_2, t) = P(i_1(t) = i_1, i_2(t) = i_2, n_1(t) = n_1, n_2(t) = n_2)$ запишем прямую систему дифференциальных уравнений Колмогорова

$$
\begin{aligned}
&\tfrac{\partial P(i_1,i_2,n_1,n_2,t)}{\partial t} = -(\lambda_1+\lambda_2+i_1\mu_1+i_2\mu_2)P(i_1,i_2,n_1,n_2,t)+\\
&+\lambda_1 P(i_1-1,i_2,n_1,n_2,t)+\lambda_2 P(i_1,i_2-1,n_1,n_2,t)+\\
&+r_1\mu_1(i_1+1)P(i_1+1,i_2-1,n_1,n_2,t)+r_2\mu_2(i_2+1)P(i_1,i_2+1,n_1,n_2,t)+\\
&+(1-r_1)\mu_1 i_1 P(i_1,i_2,n_1-1,n_2,t)+(1-r_2)\mu_2 i_2 P(i_1,i_2,n_1,n_2-1,t).
\end{aligned}
\tag{4}
$$

Определив производящую функцию $G(x_1, x_2, y_1, y_2, t) =$
$= \sum\limits_{i_1=0}^{\infty} \sum\limits_{i_2=0}^{\infty} \sum\limits_{n_1=0}^{\infty} \sum\limits_{n_2=0}^{\infty} x_1^{i_1} x_2^{i_2} y_1^{n_1} y_2^{n_2} P(i_1, i_2, n_1, n_2, t)$, из системы (4) получаем линейное дифференциальное уравнение в частных производных первого порядка для функции $G(x_1, x_2, y_1, y_2, t)$

$$\begin{array}{l} \frac{\partial G(x_1,x_2,y_1,y_2,t)}{\partial t} + \mu_1[x_1(1-y_1(1-r_1)) - r_1 x_2] \frac{\partial G(x_1,x_2,y_1,y_2,t)}{\partial x_1} + \\ + \mu_2[x_2(1-y_2(1-r_2)) - r_2] \frac{\partial G(x_1,x_2,y_1,y_2,t)}{\partial x_2} = \\ = [\lambda_1(x_1-1) + \lambda_2(x_2-1)]\, G(x_1, x_2, y_1, y_2, t). \end{array} \tag{5}$$

Решая уравнение (5), с начальным условием $G(x_1, x_2, y_1, y_2, 0) = g(x_1, x_2)$, где функция $g(x_1, x_2)$ имеет вид (2), получаем выражение для производящей функции четырехмерного процесса $\{i_1(t), i_2(t), n_1(t), n_2(t)\}$

$$\begin{array}{l} G(x_1, x_2, y_1, y_2, t) = exp\left\{ \frac{\lambda_1\mu_1 r_1 \left(e^{-\mu_2(1-y_2(1-r_2))t} - 1\right)}{\mu_2(1-y_2(1-r_2))^2[\mu_2(1-y_2(1-r_2)) - \mu_1(1-y_1(1-r_1))]} + \right. \\ + \lambda_1 t \left(\frac{r_1 r_2}{(1-y_1(1-r_1))(1-y_2(1-r_2))} - 1 \right) - \frac{\lambda_1(x_2(1-y_2(1-r_2)) - r_2) e^{-\mu_2(1-y_2(1-r_2))t}}{(1-y_2(1-r_2))[\mu_2(1-y_2(1-r_2)) - \mu_1(1-y_1(1-r_1))]} + \\ + \frac{\lambda_1 r_2}{\mu_1(1-y_1(1-r_1))(1-y_2(1-r_2))} + \frac{\lambda_1+\lambda_2}{\mu_2 r_2} \left(\frac{\mu_2(x_2(1-y_2(1-r_2)) - r_2) e^{-\mu_2(1-y_2(1-r_2))t} + \mu_2 r_2}{\mu_2(1-y_2(1-r_2))} - 1 \right) + \\ + \frac{\lambda_1\left(1-e^{-\mu_1(1-y_1(1-r_1))t}\right)}{\mu_1(1-y_1(1-r_1))} \left(\frac{\mu_1 r_1(x_2(1-y_2(1-r_2)) - r_2)}{(1-y_2(1-r_2))[\mu_2(1-y_2(1-r_2)) - \mu_1(1-y_1(1-r_1))]} - \right. \\ - \left. \frac{r_1 r_2}{(1-y_1(1-r_1))(1-y_2(1-r_2))} + x_1 \right) + \frac{\lambda_1}{\mu_1 r_1} \left(\frac{\mu_1 r_1(x_2(1-y_2(1-r_2)) - r_2) e^{-\mu_1(1-y_1(1-r_1))t}}{(1-y_2(1-r_2))[\mu_2(1-y_2(1-r_2)) - \mu_1(1-y_1(1-r_1))]} - \right. \\ - \left. \frac{r_1 r_2 e^{-\mu_1(1-y_1(1-r_1))t}}{(1-y_1(1-r_1))(1-y_2(1-r_2))} + x_1 e^{-\mu_1(1-y_1(1-r_1))t} - 1 \right) + \lambda_2 t \left(\frac{r_2}{(1-y_2(1-r_2))} - 1 \right) + \\ \left. + \frac{\lambda_2\left(1-e^{-\mu_2(1-y_2(1-r_2))t}\right)(x_2(1-y_2(1-r_2)) - r_2)}{\mu_2(1-y_2(1-r_2))^2} \right\}. \end{array} \tag{6}$$

Полагая в (6) $x_1 = x_2 = 1$, получим вид производящей функции числа событий в потоке повторных обращений

$$G(y_1,y_2,t)=exp\Big\{\frac{(\lambda_1+\lambda_2)}{\mu_2 r_2}\left(\frac{\mu_2 r_2+\mu_2(1-y_2(1-r_2)-r_2)e^{-\mu_2(1-y_2(1-r_2))t}}{\mu_2(1-y_2(1-r_2))}-1\right)+$$
$$+\frac{\lambda_1(1-e^{-\mu_1 t(1-y_1(1-r_1))})}{\mu_1(1-y_1(1-r_1))}\left(\frac{\mu_1 r_1((1-y_2(1-r_2))-r_2)}{(1-y_2(1-r_2))[\mu_2(1-y_2(1-r_2))-\mu_1 r_1(1-y_1(1-r_1))]}-\right.$$
$$\left.-\frac{r_1 r_2}{(1-y_1(1-r_1))(1-y_2(1-r_2))}+1\right)+\frac{\lambda_1 r_2}{\mu_1(1-y_1(1-r_1))(1-y_2(1-r_2))}+$$
$$+\frac{\lambda_1}{r_1\mu_1}\left(\frac{\mu_1 r_1((1-y_2(1-r_2))-r_2)e^{-\mu_1 t(1-y_1(1-r_1))}}{(1-y_2(1-r_2))[\mu_2(1-y_2(1-r_2))-\mu_1(1-y_1(1-r_1))]}-\right.$$
$$\left.-\frac{r_1 r_2 e^{-\mu_1 t(1-y_1(1-r_1))}}{(1-y_1(1-r_1))(1-y_2(1-r_2))}+e^{-\mu_1 t(1-y_1(1-r_1))}-1\right)+ \qquad (7)$$
$$+\frac{\lambda_1\mu_1 r_1(1-y_2(1-r_2)-r_2)\left(e^{-\mu_2(1-y_2(1-r_2))t}-1\right)}{\mu_2(1-y_2(1-r_2))^2[\mu_2(1-y_2(1-r_2))-\mu_1(1-y_1(1-r_1))]}-$$
$$-\frac{\lambda_1(1-y_2(1-r_2)-r_2)e^{-\mu_2(1-y_2(1-r_2))t}}{(1-y_2(1-r_2))[\mu_2(1-y_2(1-r_2))-\mu_1(1-y_1(1-r_1))]}+$$
$$+\lambda_1 t\left(\frac{r_1 r_2}{(1-y_1(1-r_1))(1-y_2(1-r_2))}-1\right)+\lambda_2 t\left(\frac{r_2}{(1-y_2(1-r_2))}-1\right)$$
$$\left.+\frac{\lambda_2\left(1-e^{-\mu_2(1-y_2(1-r_2))t}\right)(1-y_2(1-r_2)-r_2)}{\mu_2(1-y_2(1-r_2))^2}\right\}.$$

Что и требовалось доказать. ■

Этот результат, позволяет получить как распределение числа событий потока повторных обращений в систему, так и маргинальные распределения вероятностей числа событий потоков повторных обращений на каждую фазу, что дает возможность определить основные числовые характеристики исследуемого процесса.

6. Заключение

В работе представлена математическая модель двухэтапного производственного процесса в виде двухфазной СМО с двумя входящими потоками и обратной связью. Результат проведенного исследования потенциально может быть использован для решения оптимизационных задач по минимизации количества продукции, не прошедшей контроль качества. В дальнейшем планируется провести исследование при условии произвольного времени обслуживания заявок, а также учесть факт, что доработка некачественных изделий может потребовать время, отличное от времени производства изделия.

ЛИТЕРАТУРА

1. Назаров А. А., Носова М. Г. Многофазная автономная система массового обслуживания и ее применение к задачам демографии // Известия Томского политехнического университета. Серия Управление, вычислительная техника и информатика. 2009. Том 315. №5. С. 183-186

2. Клименок В. И. Многолинейная система массового обслуживания с резервными приборами // Журнал Белорусского государственного университета. Математика. Информатика. 2019. 3. С. 57–70.
3. Бесчастный В. А., Гайдамака Ю. В. Модель обслуживания трафика одноадресных и многоадресных соединений высокочастотной сети 5G // Современные информационные технологии и ИТ-образование. 2019. Т. 15, № 2. С. 263-273. DOI: 10.25559/SITITO.15.201902.263-273
4. Даммер Д. Д. Математическая модель страховой компании в виде системы массового обслуживания с неограниченным количеством приборов с учетом единовременных страховых выплат // Материалы XV Международной конференции имени А.Ф. Терпугова «Информационные технологии и математическое моделирование». Томск: Изд-во Том. ун-та, 2016. С. 18–23.
5. Жидкова Л. А., Моисеева С. П. Математическая модель потоков покупателей двухпродуктовой торговой компании в виде системы массового обслуживания с повторными обращениями к блокам // Известия Томского политехнического университета. 2013. Т. 322. № 6. С. 5–9.

UDC: 519.176

About heuristic algorithm for Correlation Clustering problem solving

A.A. Soldatenko[1], D.V. Semenova[1], E.I. Ibragimova[1]

[1]Siberian Federal University, 79 Svobodny pr., Krasnoyarsk, Russia

ASoldatenko@sfu-kras.ru, DVSemenova@sfu-kras.ru, IbragimovaEI@mail.ru

Abstract

The Correlation Clustering (CC) problem is traditionally defined as a problem of partitioning a signed graph without specifying the number of clusters in advance. In this paper, CC problem is considered for undirected and unweighted signed graphs without multiple edges and loops, where error functional is linear combination of intercluster and intracluster errors. In this formulation, the CC problem is NP-complete. Exact algorithms for this problem are time-consuming. Approximate algorithms for solving CC problem often lead to unsatisfactory results, and heuristic algorithms are often non-deterministic in the number of steps leading to a solution. We propose a new heuristic algorithm $SGClust_\alpha$ for the CC problem solving. The main idea of this algorithm is in intracluster error minimizing and optimization of error functional according to the greedy strategy. It was proved that this algorithm takes polynomial time. Numerical experiments were carried out on randomly generated signed graphs.

Keywords: Correlation Clustering, Signed graph, Heuristic Greedy algorithm, Structural balance, Graph partition.

1. Introduction

Correlation clustering (CC) problem is a well-known unsupervised learning problem aimed at finding a vertices partition in a signed graph with a disagreements minimum number or the maximum number of consistent edges [2]. A disagreement occurs when a positive edge connects vertices from different clusters or a negative edge connects vertices from the same cluster. In such cases, the edge is called inconsistent, and in all others it is called consistent.

Currently, various formulations of the CC problem can be found in the literature. They are classified by following parameters: the number of clusters (set initially,

This work is supported by the Krasnoyarsk Mathematical Center and financed by the Ministry of Science and Higher Education of the Russian Federation (Agreement No. 075-02-2023-936).

limited by any number or unlimited) [6], the size of clusters (fixed, limited or unlimited) and the type of error functional.

Our attention is focused on the CC problem for undirected and unweighted signed graphs without multiple edges and loops with an error functional in the form of a linear combination of the intercluster and intracluster errors. This formulation was proposed by Doreian et al. in the work [4]. Bansal et al. in the [2] have proved that this problem is NP-hard. Let's note the CC problem may have more than a single solution. Due to the difficulty of the problem under consideration, research to develop heuristic algorithms that find a solution in an acceptable time are being conducted [8]. To solve the CC problem, we propose a new heuristic algorithm $SGClust_\alpha$ based on a greedy strategy.

The CC problem arises in various scientific fields such as Community Identification in Social Network, Image Processing, Data Mining, Computational Biology, Telecommunication and Control Systems. So, Abdelnasser et al. [1] have proposed a clustering, sub-channel and power allocation framework to be implemented in a semi-distributed fashion in a two-tier OFDMA cellular network. They have proposed correlation clustering approach, which considers the trade-off between the bandwidth and interference, offers a performance that is very close to that of the optimal clustering; however, with much reduced complexity. Maatouk et al. [7] addressed the problem of clustering and user scheduling with massive MIMO technology, which is regarded as a key factor in the development of 5G networks.

2. Problem formulation

2.1. Signed Graph. In the paper the type $\Sigma = (G, \sigma)$ of signed graphs are considered, where $G = (V, E)$ is an undirected, unweighted graph without multiple edges and loops with set of vertices V, $|V| = n \geq 2$ and set of edges E, $|E| = m \geq 1$. In the graph G, each edge is uniquely represented by an unordered pair $e = (u, v)$, where $e \in E$, $u, v \in V$. In this case, it is said that the edge e is incident to the vertices u or v, and the vertices u and v are adjacent. Under the degree of a vertex v the number of edges incident to it is traditionally understood. Then $\Delta = \max\limits_{v \in V} \delta(v)$ is the graph degree. On the edges $(u, v) \in E$ of the graph G, the sign function $\sigma : E \to \{+, -\}$ is given, it generates a graph edges set partition $E = E^+ \cup E^-$, where E^+ is a set of positive edges, E^- is a set of negative edges.

A signed graph is called k-balanced if the set of its vertices can be divided into k pairwise disjoint nonempty clusters so that all positive edges are inside the clusters and negative edges are between the clusters [5].

2.2. Correlation Clustering problem. Let's denote the sets system forming a partition of the vertices set V into k subsets by

$$\mathcal{C} = \left\{ C_i \subseteq V : \bigcup_{i=1}^{k} C_i = V, C_i \cap C_j = \emptyset, i \neq j;\ i = \overline{1,k} \right\}. \tag{1}$$

It is known that for an arbitrary signed graph, the k-balance property may have no place. In this case, it is interesting to search for such a partition of the graph set vertices for which it is possible to obtain a k-balanced graph by changing the sign of the edges minimum number. This problem is considered as a graph clustering problem with a special kind of the error functional. The elements of the $C_i \in \mathcal{C}$ partition will be called clusters.

Under a positive error $P(\mathcal{C})$ of the partition (1) a number of positive edges between subsets of $C_1, \ldots, C_k$ will be understood. Let's note that $P(\mathcal{C})$ is the intercluster error calculated by the formula $P(\mathcal{C}) = \sum\limits_{i=1}^{k} \sum\limits_{u \in C_i} \sum\limits_{v \in V \setminus C_i} [(u,v) \in E^+]$, where hereinafter $[\cdot]$ is Iverson's Convention.

Under a negative error $N(\mathcal{C})$ a number of negative edges inside subsets for the partition (1) will be understood. The negative error is the intracluster error calculated by the formula $N(\mathcal{C}) = \sum\limits_{i=1}^{k} \sum\limits_{\{u,v\} \subseteq C_i} [(u,v) \in E^-]$.

In the [4] was proposed to represent the total error as a convex combination of positive and negative errors, this total error depends on the parameter $\alpha \in [0,1]$:

$$Q_\alpha(\mathcal{C}) = \alpha N(\mathcal{C}) + (1-\alpha) P(\mathcal{C}). \tag{2}$$

An error of the form (2) will be called an α-error of the partition $\mathcal{C}$.

In this paper, the clustering problem of a signed graph is considered in the following formulation [4].

Correlation Clustering (CC) problem

Condition:	a signed graph $\Sigma = (G, \sigma)$ is given, where $G = (V, E)$ is an undirected graph, $n = \lvert V \rvert \geq 2, m = \lvert E \rvert \geq 1$.
Problem:	for a given $\alpha \in [0,1]$, it is required to find the partition $\mathcal{C}$ of the vertices set V of the signed graph Σ with the minimum total error $Q_\alpha(\mathcal{C})$.

In the work [2] it was shown that Correlation Clustering with an error functional in the form (2) at $\alpha = 0.5$ is NP-complete.

Let Φ_k be the set of partitions into k subsets, and $\Phi = \bigcup\limits_{k=1}^{n} \Phi_k$ be the set of all possible partitions of V [4]. Then the cardinality of the solution space Φ equals to

the Bell number B_n. The CC problem solution is a set of clusters $\mathcal{C}^*$, which gives the minimum value of the error function (2):

$$\mathcal{C}^* = \arg\min_{\mathcal{C}\in\Phi}\big[\alpha N(\mathcal{C}) + (1-\alpha)P(\mathcal{C})\big] \tag{3}$$

and $k = |\mathcal{C}^*|$ is a number of clusters. It should be noted that the solution (3) may not be the only one.

3. Main results

The paper proposes the heuristic algorithm $SGClust_\alpha$ for solving the Correlation Clustering problem. As input, the algorithm takes a signed graph $\Sigma = (G, \sigma)$ and some initial partition $\mathcal{C}_0$. The algorithm result is a partition $\mathcal{C}$ with an error not exceeding the error of the original partition: $Q_\alpha(\mathcal{C}) \le Q_\alpha(\mathcal{C}_0)$.

The strategy of the proposed algorithm is based on the intracluster error $N(\mathcal{C})$ minimizing without increasing the total error $Q_\alpha(\mathcal{C})$. Each the algorithm iteration consists of the following four stages.

Search:	to find the vertex that makes the greatest contribution to the intracluster error.
Removing:	to remove this vertex from its cluster in order to reduce the intracluster error $N(\mathcal{C})$.
Optimization:	to find such cluster for this vertex (it can be the initial cluster), joining to which will minimize the total error (2), if such cluster does not exist, then this vertex can be allocated to a new cluster
Fixation:	to fix the given vertex from further movements between clusters.

Four stages of the algorithm form a conditional cycle. This cycle guarantees that the algorithm will move each vertex not more than once and then be finished its work by finding some partition $\mathcal{C}$. During the search and optimization stages, the algorithm follows a greedy strategy. And the following theorem is true.

Theorem 1. The algorithm $SGClust_\alpha$ is correct, i. e. if as input a signed graph $\Sigma = (G, \sigma)$, some initial partition $\mathcal{C}_0$ and a parameter α are given, then this algorithm will find a new partition $\mathcal{C}$ with α-error $Q_\alpha(\mathcal{C}) \le Q_\alpha(\mathcal{C}_0)$ in time $\mathcal{O}(n^2 \cdot \Delta)$.

Since the strategy proposed above moves each vertex between clusters only once, a successful initial clustering will play a large role in its efficiency. In this paper, the initial partition is chosen as the connected components of the graph $G = (V, E^+)$, i.e. each C_i cluster will contain only vertices from one connected component. It is easy to prove that vertices belonging to different connected components of a graph built only on positive edges, but lying in the same cluster, will contribute equal to or

greater than zero to the total error E_α. Therefore, it is expedient to take the initial partition such that the clusters contain only vertices from one connected component of the graph $G = (V, E^+)$.

4. Computational experiments

Computational experiments were carried out to evaluate the $SGClust_\alpha$ algorithm effectiveness. All experiments were carried out on a computer with 16GB RAM, an AMD Ryzen 5 3600 6-Core 3.60 GHz processor running the Windows 10 operating system using single-threaded mode. Comparison was performed with the algorithm *relocation heuristic* (RH) [3] on randomly generated graphs. The difference between $SGClust_\alpha$ and RH algorithms consist in the number of testable partitions. The RH algorithm checks all possible transfers of vertices between clusters without considering their contribution to the total error. For the current partition the $SGClust_\alpha$ algorithm finds the vertex with largest contribution to the total error and then checks whether moving this vertex will improve the error value.

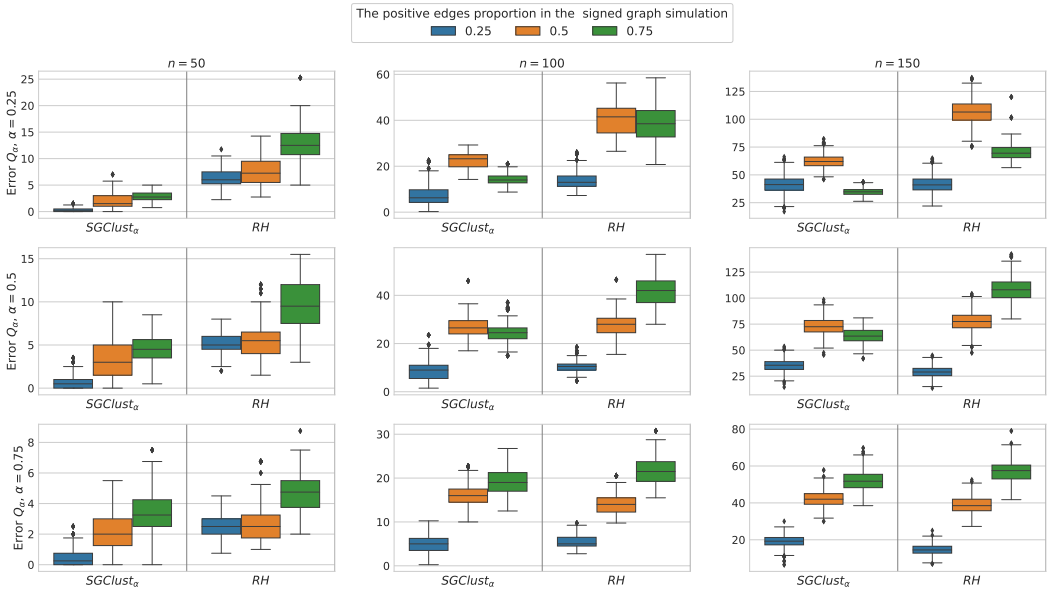


Fig. 1. Comparison of algorithms $SGClust_\alpha$ and RH for different parameters

The Fig. 1 shows results of computational experiments. They were carried out for randomly generated connected graphs using the Waxman method with parameters 0.15 and 0.4. Experiments were carried on 1000 of graphs with a vertices fixed number $n \in \{100, 150\}$ and a approximate proportion of positive edges $p \in \{0.25, 0.5, 0.75\}$.

Algorithms $SGClust_\alpha$ and RH for α various parameters were compared. The algorithm RH was given as input the number of clusters obtained by the algorithm $SGClust_\alpha$. From the Fig. 1 it can be seen that the errors of the algorithms are close, while the algorithm $SGClust_\alpha$ works better with a larger proportion of positive edges and with $\alpha = 0.25$, also in most cases the algorithm $SGClust_\alpha$ reduces the positive error better, and the RH reduces the negative error better. At the same time, the operating time of the $SGClust_\alpha$ is significantly less.

5. Conclusion

The $SGClust\alpha$ algorithm developed by us for solving CC problem is correct and finds a solution in an acceptable time. Experiments have shown that the optimization quality of positive and negative errors depends on the structure of the original graph. The algorithm $SGClust_\alpha$ wins in the ratio of clustering quality and running time in comparison to the RH algorithm. A peculiarity of this algorithms class is the solution quality dependence on an input graph vertices renumbering, that is one of the promising directions for further research.

REFERENCES

1. Abdelnasser A., Hossain E., Kim D. I. Clustering and resource allocation for dense femtocells in a two-tier cellular OFDMA network // IEEE Trans Wireless Communications. 2014. V. 13(3), 1628–1641.
2. Bansal N., Blum A., Chawla S. Correlation Clustering // Machine Learning. 2002. V. 56. P. 89–113.
3. Brusco M. J., Doreian P. Partitioning signed networks using relocation heuristics, tabu search, and variable neighborhood search // Social Networks. 2019. V. 56. P. 70–80.
4. Doreian P., Mrvar A. A partitioning approach to structural balance // Social Networks. 1996. V. 18. P. 149–168.
5. Harary F. Structural Balance: A Generalization of Heider's Theory // Psychological Review. 1956. V. 63(5). P. 227–293.
6. Il'ev V., Il'eva S., Kononov A. Short Survey on Graph Correlation Clustering with Minimization Criteria // DOOR-2016, Lecture Notes in Computer Science. 2016. V. 9869. P. 25–36.
7. Maatouk A., Hajri S. E., Assaad M., Sari H.: On optimal scheduling for joint spatial division and multiplexing approach in fdd massive mimo // IEEE Trans Signal Process. 2018. V. 67(4). P. 1006–1021.
8. Wahid D. F., Hassini E.A Literature Review on Correlation Clustering: Cross-disciplinary Taxonomy with Bibliometric Analysis // SN Operations Research Forum. 2022. V. 3(47).

UDC: 004.94

Application Identification in mmWave/THz Systems via Machine Learning Algorithms

Svetlana Dugaeva[1], Vyacheslav Begishev[1], Nikita Stepanov[2]

[1]Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Str., Moscow, 117198, Russia

[2]Saint-Petersburg State University of Aerospace Instrumentation (GUAP University), St-Petersburg, Russia, Bolshaya Morskaya str., Saint Petersburg, 190000, Russia

dugaeva-sa@rudn.ru, begishev-vo@rudn.ru, stepanov.nikita@guap.ru

Abstract

Beamtracking is a critical functionality in modern millimeter wave (mmWave) 5G New Radio (NR) systems and is expected to become even more critical in future 6G systems operating in terahertz (THz) frequency band. To enable uninterrupted connectivity base stations (BS) need to invoke this procedure periodically. Due to the use of massive antenna arrays in 6G THz systems, the amount of resources consumed by beamtracking will be extremely large making the time interval between sweeping beam configurations a very critical parameter. One of the phenomena affecting the choice of this interval is a user equipment (UE) micromobility – quick displacements and rotations of UE in the hands of a user happening even when the latter is in a stationary position. In this paper, by utilizing machine learning (ML) algorithms, we propose a procedure for the detection of the beam center at the BS side for applications characterized by different types of micromobility. We demonstrate that one can safely differentiate between applications characterized by low as well as distinctively different micromobility speeds. All the considered classifieds including the tree, random forest, and neural network perform qualitatively similarly. For applications having fast and similar micromobility speeds such as VR and gaming, the classification accuracy stays at around $85 - 90\%$. However, this loss in accuracy does not affect the ultimate goal of the remote application detection algorithm – understanding how often the beam alignment procedure must be invoked at UE and BS.

Keywords: 6G, terahertz, cellular systems, micromobility, application detection

Sections 2 and 4 were written by Vyacheslav Begishev under the support of the Russian Science Foundation, project No. 21-79-10139. This publication has been supported by the RUDN University Scientific Projects Grant System, project No. 021928-2-074 (recipients Svetlana Dugaeva, Sections 3 and 5).

1. Introduction

Enabling the capacity boost at the access interface of future cellular systems is only feasible by providing more bandwidth. To this aim, 5G New Radio systems specified operation in the so-called millimeter frequency band at carriers around 28, 38, and 72 GHz, where multiple channels each 400 MHz wide are available [1]. Future 6G terahertz (THz) systems are expected to push the frequency bands even higher to the $100 - 300$ GHz range with bandwidth expected to the on the order of 2 GHz per single channel [2].

To compensate for small antenna aperture at mmWave/THz frequencies massive antenna arrays operating in beamforming mode need to be used at base stations (BS) in 5G/6G systems [3]. Such an antenna creates extremely directional radiation patterns towards user equipment (UE) allowing for high gains in transmit and receive direction and compensating for the limited emitted power of a single array element [4]. As a side effect, these arrays posse very directional main lobe whose half-power beamwidth (HPBW) may reach a fraction of a degree [5]. As a result, as opposed to 4G and the previous generation of cellular systems beamtracking procedure keeping BS and UE synchronized at all times needs to be utilized.

There are two critical phenomena affecting continuous connectivity between BS and UE. The first one is blockage by small moving objects and large stationary objects that have been deeply studied in the past [6]. To efficiently avoid this problem, 3GPP has recently proposed a multiconnectivity solution [7] that allows for fairly well avoid the blockage events [8, 9, 10]. In addition to blockage, the micromobility phenomenon referring to quick displacements and rotations in the hand of a user has been recently identified [11]. Although, similarly to the blockage, this effect may lead to the temporal loss of connectivity and its characteristics are somewhat documented in the literature, the measures against this effect has not been developed so far. Specifically, the authors in [12] characterized the channel capacity in the presence of micromobility showing that for sub-seconds HPBWs on-demand beamtracking outperforms the current regular beamtracking approach utilized currently in 5G NR systems. In [13], it has been shown that the array switching time needs to be drastically decreased (by three orders of magnitude) if one wants to completely remove the connectivity losses occurring as a result of multiconnecitity.

Having much larger antenna arrays 6G THz systems would require extremely large resources for periodic beamtracking procedure in the presence of micromobility. To decrease the resource usage of may attempt to optimize the interval between beamtracking time instants. However, as demonstrated in [14, 15] different applications utilized by modern smartphones are characterized by principally different statistical characteristics including time to outage. Thus, based on the knowledge of the application type utilized at the UE may provide BS with appropriate information

on how often the beamtracking procedure needs to be invoked. Unfortunately, 3GPP stack does not currently provide the signaling information for notifying the BS about the currently utilized application.

By utilizing the measurements reported in [14] in this paper we propose a method for remote detection of the type of application utilized at UE. To this aim, we utilize the information about the trajectory of the center of the beam that is indirectly available at the BS when the antenna arrays at UE and BS are both known. We then apply ML approaches including trees, random forests, and neural networks to differentiate between 4 types of applications including video watching, virtual reality (VR) gaming, racing gaming, and phone calling.

The main contributions of our paper are:

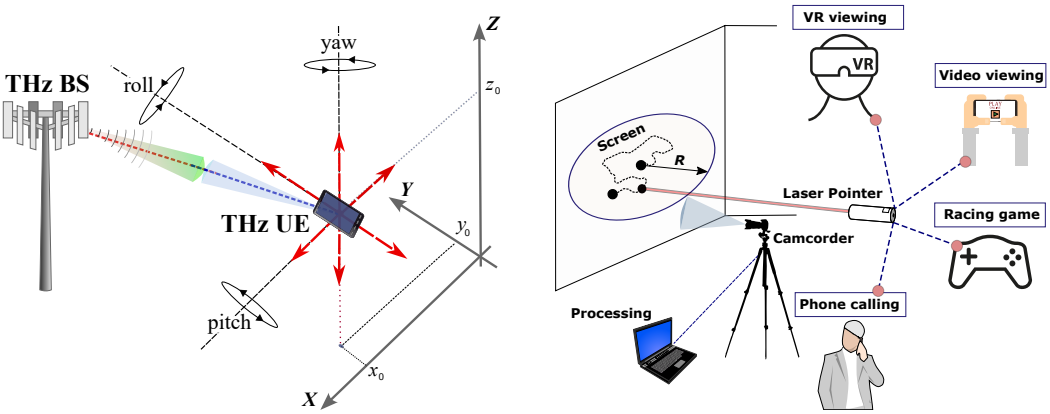
- the non-intrusive procedure that does not rely upon signaling information to determine the type of applications utilized at the UE for further adaptation of beamtracking interval at the BS;
- numerical results showing that: (i) the accuracy of differentiating between applications having low speeds or distinctively different speeds stays at 100% (ii) all the considered algorithms including the tree, random forest, and neural network perform qualitatively similary.

The rest of the paper is organized as follows. We start in Section 2 reporting the summary of the results of the micromobility measurements and modeling campaign. Then, in Section 3, we introduce the proposed approach. In Section 4, we report our numerical results. The conclusions are drawn in the last section.

2. Micromobility Measurements, Properties, and Modeling

In this paper, we utilize the micromobility measurement results and model reported in [14]. In this section, we will briefly recall the measurement setup, statistical characteristics, and the utilized modeling framework.

2.1. Measurements Setup. We consider a point-to-point communication between a base station (BS) and a UE, as shown in Fig. 1(a). Both UE and BS are assumed to operate utilizing arrays creating directional antenna radiation patterns. In the proposed THz system, the BS appears to be firmly fixed while the user associated with UE is in a stationary position. However, even in stationary conditions, UE in hands of a user is expected to perform small displacements and rotations that are modulated by the perceived content. Specifically, these are small displacements along Ox and Oy axes and rotations over the vertical and transverse axes. These types of movements may cause the loss of connectivity between UE and BS. Note that the micro changes along Oz-axis, as well as rotations along the longitudinal axis do not affect the state of connection.



(a) The considered scenario

(b) Setup and applications

Fig. 1. The considered scenario, setup, and applications.

To capture the micromobility of the UE, it is sufficient to characterize the motion pattern of the UE's beam center. The measurements performed in [14] were conducted by utilizing a laser pointer rigidly connected to a smartphone. A laser pointer with a beam diameter of 3 mm, an output power of 5 mW and a long wavelength of light of 650 ± 10 nm was used. The laser spot detection was facilitated by utilizing 1280×720 30 frames-per-second video camera, as shown in Fig. 1(b). Four different applications were considered: (i) watching video, (ii) phone calling, (iii) watching virtual reality (VR) video, and (iv) playing a racing game. We carried out our experiments at a distance of 2 m from the UE to the screen. The observed area is 1.5×2.5 m.

To obtain statistical data 10 independent experiments were carried out for each considered application. The duration of each experiment is set to 10 s. The original camera resolution was 30 frames per second, resulting in a source trace for each application containing approximately 3300 sample points.

2.2. Statistical results. A summary of the statistics and simulation results is in Table 1. As one may observe, none of the applications are characterized by radial symmetry. The speed and drift of random mobility patterns are inherently distance dependent. Thus, in order to accurately capture the stochastic properties of UE micromobility models, these properties require complex modeling approaches involving 2D random walks with independent increments.

2.3. Modeling. The authors in [14] also proceeded offering 2D Markov models capturing micromobility patterns of applications. They utilized a direct approach by segmenting the screen into N by N grid for different values of N ranging from 50 to 200. Then they proceeded to parameterize the Markov model, determining

Table 1. Summary of micromobility characteristics and modeling [14]

Characteristics	Video	Phone	VR	Racing
Radial symmetry	Yes	No	No	No
Velocity	$3-10$ m/s	7 m/s	$9-13$ m/s	$9-5$ m/s
Drift to the origin	$0.17-0.11$	$0.17-0.3$	0.17	0.17
Axes dependence	Negligible	Moderate	Negligible	Strong
Correlation coefficient	0.0	-0.2	0.0	-0.4
X-axis velocity	Increases, $1-6$ m/s	$3-6$ m/s	$6-9$ m/s	$7-4$ m/s
Y-axis velocity	Increases, $2-8$ m/s	$3-5$ m/s	$5-8$ m/s	$3-2$ m/s
X-axis drift	$0.17-0.05$	$0.17-0.13$	0.17	$0.13-0.21$
Y-axis drift	$0.17-0.21$	$0.17-0.25$	0.17	$0.19-0.14$
Markov modeling	Yes	Yes	Limited	No

the transition probabilities p_{ij}, $i, j = 1, 2, \ldots, N^2$. They further demonstrated that these models are suitable for applications with low and purely random dynamics, such as video viewing, VR viewing, and phone calling and may not fully resemble the properties of the more complex patterns when the application directly controls the user's behavior, as in the case of racing game.

3. The Proposed Approach

3.1. Proposal. The core of the proposed idea is to differentiate between different types of applications having distinctively different micromobility speeds, and thus requiring different frequencies of beam alignments, remotely, without modifying the current signaling interface between BS and UE. To this aim, we propose BS to constantly monitor the current received signal strength. By knowing this information and antenna radiation patterns at both UE and BS sides one may deduce the location of the beam center. BS may track the current trajectory of the beam center and utilize the pre-trained ML algorithms to classify the type of applications.

To evaluate the proposed approach, we utilize the two-dimensional Markov models reported in [14]. We utilize these models to generate random traces of beam center movement of different applications. These traces are then utilized to train the ML models. Finally, to assess the accuracy of the proposal, we provide the classification of applications based on the remaining sets of generated traces.

3.2. Description of the Utilized Algorithms. In our study we consider three types of ML algorithms for remote application detection. These are decision trees, random forests, and neural networks. Below, we provide their short description and discuss parameterization.

Decision Tree. The decision tree algorithm works by recursively splitting the dataset into subsets based on the values of the features, with the aim of maximizing the information gain at each split. Information gain is a measure of the reduction in entropy or disorder in the dataset, and it is calculated as

$$I(D, F) = H(D) - H(D|F), \tag{1}$$

where $I(D, F)$ is the information gain, D is the dataset, F is the feature being considered, $H(D)$ is the entropy of the dataset, and $H(D|F)$ is the conditional entropy of the dataset given the feature F.

The entropy of a dataset is calculated as

$$H(D) = -\sum(p_i \log_2(p_i)), \tag{2}$$

where p_i is the proportion of instances in the dataset that belong to class i.

The decision tree algorithm split the dataset based on the features with the highest information gain until a stopping criterion is met, such as a maximum tree depth or a minimum number of instances in each leaf node.

Random Forest. Random forest is an ensemble learning algorithm that constructs a multitude of decision trees at training time and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. The basic idea behind the random forest is to combine multiple decision trees in determining the final output, with each decision tree being constructed on a different random subset of the training data.

Assuming a classification problem with a dataset of n examples with m features, the algorithm works as follows: (i) randomly select k features from m total features, where $k << m$, creating the so-called "feature subspace" for the given decision tree, (ii) use the selected features to build a decision tree, where each node in the tree splits the data based on the value of one of the k randomly selected features, (iii) repeat steps (i) and (ii) t times to generate a total of t decision trees, (iv) to make a prediction for a new example, classify the example using each of the t decision trees and take the majority vote as the final prediction.

In random forests, the predicted output of each decision tree is aggregated to determine the final prediction. Let Y denote the class label of an input example, and let T be the number of decision trees in the random forest. Then, the output of the random forest, denoted as $\hat{Y}$, is computed as

$$\hat{Y} = M(Y_1, Y_2, \ldots, Y_t), \tag{3}$$

where M is the statistical mode.

Neural network. The most basic form of a neural network is a feedforward neural network, which consists of an input layer, one or more hidden layers, and an output layer. Each layer contains a set of neurons, and each neuron is connected to all neurons in the adjacent layers by a set of weights.

The output of a neuron in a neural network is determined by the weighted sum of its inputs, passed through an activation function. The activation function is a nonlinear function that introduces non-linearity into the model, allowing it to learn complex patterns in the data. The most commonly used activation functions are the sigmoid function and the rectified linear unit (ReLU) function. Specifically, the output of a neuron in layer j can be computed as

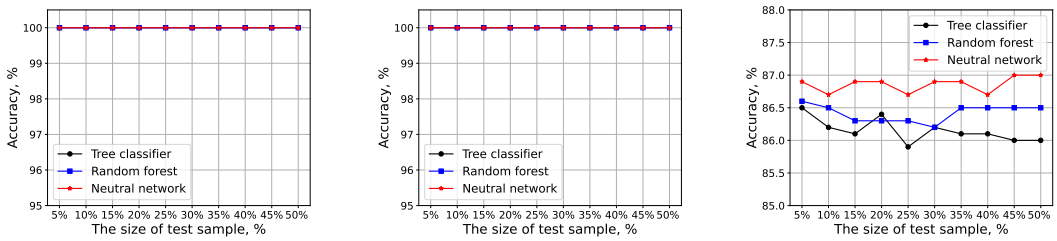
$$z_j = \sum (w_{i,j} * x_i) + b_j,\ a_j = g(z_j), \tag{4}$$

where $w_{i,j}$ is the weight of the connection between neuron i in layer $j-1$ and neuron j in layer j, x_i is the input value to neuron i, b_j is the bias term for neuron j, $g(\cdot)$ is the activation function, z_j is the weighted sum of inputs to neuron j, and a_j is the output value of neuron j.

The output of the neural network is obtained by applying the feedforward process to the input layer and propagating the computed values through each layer until the output layer is reached. During training, the weights and biases of the neural network are adjusted to minimize the difference between the predicted output and the actual output, using a loss function. The most commonly used loss function for regression problems is the mean squared error (MSE) function, which measures the average squared difference between the predicted and actual outputs. The weights and biases are updated using an optimization algorithm, such as gradient descent, which iteratively adjusts the weights in the direction of the steepest descent of the loss function. This process is called backpropagation, and it allows the neural network to learn the optimal weights and biases that minimize the loss function.

3.3. Implementation. To generate traces, the initial data must be presented in the following format: (i) the starting point (x and y-coordinates), (ii) followed by the transition point (x, y), and (iii) the probability of transitioning from the initial position to the next. Next, we calculate the trajectory of the UE, generating 10^4 states for each trace by applying the conventional procedure for generating trajectories of the Markov process. We generated 10^5 traces for each application.

The following attributes have been selected for classification: (i) mean distance from the center, (ii) mean speed of movement, (iii) total distance traveled, (iv) mean distance from the center on the Ox axis, (v) mean distance from the center on the Oy axis. Once the set of attributes for each trace was calculated, the next step was to use them for classification. For this purpose, Matlab was used, specifically its



(a) Call and video (b) Video and gaming (c) VR and gaming

Fig. 2. The classification accuracy for pairwise applications classification.

built-in *Classification Learner (CL)* function. To classify the data, the attributes have been provided to the CL function, where each feature was marked with the scenario of use to which it belonged.

The classification process involved selecting the size of the test sample to train the models, followed by the selection of the necessary classification algorithms. The following decision tree, random forests, and neural network algorithms have been considered. The models were then trained and the classification was carried out.

To determine the weight of the contribution of each feature to the accuracy of classification, the *Feature Selection (FS)* function was used. Note that the model is trained on the selected sample size, and the classification accuracy is calculated on the entire data, including test data. Finally, as we also want to understand how useful each feature was for classification, the weight of the contribution of each of the attributes to the accuracy of classification was calculated.

4. Numerical Results

In this section, we report our numerical results for the classification of applications at the UE. We first consider three types of applications' classification: (i) video and gaming (ii) voice and video (iii) VR and gaming. Then, we proceed to report the classification results for all four considered applications.

Overall, 10^4 traces were generated for each application. A fraction of those was utilized to train the considered models, while the rest – for assessing the performance. In all the cases, we utilized the prediction (classification) accuracy metric defined as the number of correctly predicted applications to the overall number of test samples. In addition, we also report the attributes' contribution to the decision making.

4.1. Pairwise Classification. Recalling that the ultimate goal of identifying the type of applications utilized at UE is to understand how often beam alignment procedure needs to be invoked, we start with the pairwise classification of the types of applications including: (i) video and gaming (ii) voice and video (iii) VR and

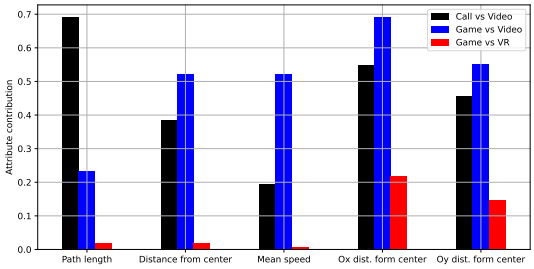


Fig. 3. The attributes contribution to the pairwise applications classification.

gaming. Note that, based on Table 1, voice and video applications are characterized by similar statistical properties including the speed of motion of the beam center. VR and gaming applications also have qualitatively comparable statistics with the exception that gaming applications are characterized by much higher speed over the Ox axis. Thus, we may expect that it might be difficult to distinguish between voice and video applications, while the classification should work much better for video and gaming applications.

The classification accuracy metric is reported in Fig. 2 for all the considered pairs of applications and three utilized algorithms as a function of the percentage of generated traces utilized for testing. By analyzing the presented data we see that for the classification of call and video as well as video and gaming applications the accuracy remains perfect at 100% even when the size of the training sample decreases. However, for VR and gaming applications the results are not perfect leading to approximately 13-15% of erroneously predicted cases. Out of all the considered models, the neural network shows the most promising results slightly outperforming tree and forest classifiers by approximately 0.5-1%. However, this difference is not relevant in practical applications implying that the choice of the classifier can be solely based on implementation complexity. Finally, we also notice that the tree classifier is the only one showing slight degradation in performance as compared to the other two algorithms when the training sample size decreases.

The abovementioned results of pairwise applications classification can be explained by analyzing attributes' contribution to the decision making shown in Fig. 3 for all the pairwisely classified applications. Here, we see that for call and video as well as video and gaming applications, all the attributes play significant roles in the final decision. However, gaming and VR applications appear to be quite similar in terms of three attributes – length traversed by a beam center, distance from the center, and mean speed. The main contributions in this case come from different

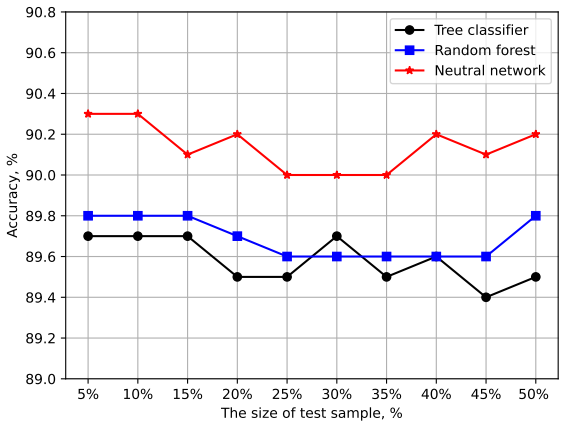


Fig. 4. The classification accuracy for all the applications.

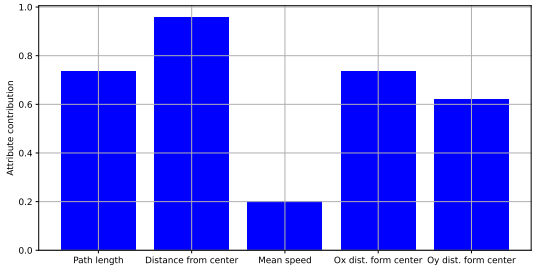


Fig. 5. The attributes contribution for all applications classification.

distances over Cartesian axes. However, by noticing that all the considered algorithms perfectly distinguish between different applications having low micromobility speeds and applications having distinctively different speeds, we observe that failure to perfectly distinguishing between applications characterized by high micromobility speeds does not affect the ultimate goal consisting in understanding how often the beam alignment procedure needs to be invoked.

4.2. Joint Classification. Having studied the pairwise applications classification, we now proceed to report the results of classifying all the applications altogether. To this aim, Fig. 4 shows the classification accuracy for all the considered algorithms as a function of the fraction of traces utilized for testing. Here, we again see that the neural network is characterized by negligibly better accuracy while the tree

classifiers are more sensitive to the size of the training set. By comparing these results to those reported for pairwise comparison we may deduce that the loss in classification accuracy is mainly attributed to the presence of 50% of traces with high and comparable micromobility speed – VR and gaming applications. This is implicitly confirmed by the contribution of the micromobility speed to the overall decision making shown in Fig. 5. Here, as the overall collection of traces becomes quite heterogeneous, logically, the distance to the center plays the most significant role.

5. Conclusions

Motivated by the need to optimize regular beamtracking interval at mmWave/THz BSs in 5G/6G systems utilizing massive antenna arrays, in this paper we proposed a non-intrusive procedure for remote detection of the application running at UE. Specifically, by utilizing the trajectory of the center of HPBW available at BS, we proposed to use ML techniques to discriminate the type of application. The proposed approach can be used to optimize the time interval between beamtracking time instants such that continuous connectivity between UE and BS is ensured consuming the minimum amount of resources.

Our numerical results demonstrate that all the considered classifiers, including tree, random forest and neural network allow us to perfectly distinguish between applications having low micromobility speed and applications having distinctively different speeds. However, the accuracy of classification of those applications having fast and similar speeds such as VR and gaming is not perfect staying at approximately 85-90%. Nevertheless, as the main goal of the proposed classification is to understand how often the beam alignment procedures must be invoked, the abovementioned observation does not lead to significant challenges. Finally, we note that all the considered classifiers including the tree, random forest, and neural network perform qualitatively similary with the tree being more sensitive to the training sample size.

REFERENCES

1. A. Zaidi, F. Athley, J. Medbo, U. Gustavsson, G. Durisi, X. Chen, 5G Physical Layer: principles, models and technology components, Academic Press, 2018.
2. M. Matthaiou, O. Yurduseven, H. Q. Ngo, D. Morales-Jimenez, S. L. Cotton, V. F. Fusco, The road to 6G: Ten physical layer challenges for communications engineers, IEEE Communications Magazine 59 (1) (2021) 64–69.
3. V. Petrov, A. Pyattaev, D. Moltchanov, Y. Koucheryavy, Terahertz band communications: Applications, research challenges, and standardization activities, in: 2016 8th international congress on ultra modern telecommunications and control systems and workshops (ICUMT), IEEE, 2016, pp. 183–190.

4. C. A. Balanis, Antenna theory: analysis and design, John wiley & sons, 2015.
5. I. F. Akyildiz, J. M. Jornet, Realizing ultra-massive mimo (1024× 1024) communication in the (0.06–10) terahertz band, Nano Communication Networks 8 (2016) 46–54.
6. M. Gapeyenko, A. Samuylov, M. Gerasimenko, D. Moltchanov, S. Singh, E. Aryafar, S.-p. Yeh, N. Himayat, S. Andreev, Y. Koucheryavy, Analysis of human-body blockage in urban millimeter-wave cellular communications, in: 2016 IEEE International Conference on Communications (ICC), IEEE, 2016, pp. 1–7.
7. 3GPP, NR; Multi-connectivity; stage 2 (Release 16), 3GPP TS 37.340 V16.0.0, 3GPP (December 2019).
8. E. Sopin, D. Moltchanov, A. Daraseliya, Y. Koucheryavy, Y. Gaidamaka, User association and multi-connectivity strategies in joint terahertz and millimeter wave 6g systems, IEEE Transactions on Vehicular Technology 71 (12) (2022) 12765–12781.
9. V. Begishev, E. Sopin, D. Moltchanov, R. Pirmagomedov, A. Samuylov, S. Andreev, Y. Koucheryavy, K. Samouylov, Performance analysis of multi-band microwave and millimeter-wave operation in 5G NR systems, IEEE Transactions on Wireless Communications 20 (6) (2021) 3475–3490.
10. V. Begishev, E. Sopin, D. Moltchanov, R. Kovalchukov, A. Samuylov, S. Andreev, Y. Koucheryavy, K. Samouylov, Joint use of guard capacity and multiconnectivity for improved session continuity in millimeter-wave 5G NR systems, IEEE Transactions on Vehicular Technology 70 (3) (2021) 2657–2672.
11. V. Petrov, D. Moltchanov, Y. Koucheryavy, J. M. Jornet, The effect of small-scale mobility on terahertz band communications, in: Proceedings of the 5th ACM International Conference on Nanoscale Computing and Communication, 2018, pp. 1–2.
12. V. Petrov, D. Moltchanov, Y. Koucheryavy, J. M. Jornet, Capacity and outage of terahertz communications with user micro-mobility and beam misalignment, IEEE Transactions on Vehicular Technology 69 (6) (2020) 6822–6827.
13. D. Moltchanov, Y. Gaidamaka, D. Ostrikova, V. Beschastnyi, Y. Koucheryavy, K. Samouylov, Ergodic outage and capacity of terahertz systems under micromobility and blockage impairments, IEEE Transactions on Wireless Communications 21 (5) (2021) 3024–3039.
14. N. Stepanov, D. Moltchanov, V. Begishev, A. Turlikov, Y. Koucheryavy, Statistical analysis and modeling of user micromobility for THz cellular communications, IEEE Transactions on Vehicular Technology 71 (1) (2022) 725–738. doi:10.1109/TVT.2021.3124870.
15. N. Stepanov, A. Turlikov, V. Begishev, Y. Koucheryavy, D. Moltchanov, Accuracy assessment of user micromobility models for THz cellular systems, in: Proceedings

of the 5th ACM Workshop on Millimeter-Wave and Terahertz Networks and Sensing Systems, 2021, pp. 37–42.

УДК: 004.056.53

Применение аппаратных счетчиков производительности для выявления угроз информационной безопасности приложений

В.А. Галатенко[1], К.А. Костюхин[1]

[1]ФГУ ФНЦ НИИСИ РАН, Нахимовский пр-т, д. 36, корп. 1, Москва, Россия

galat@niisi.ras.ru, kost@niisi.ras.ru

Аннотация

Работа посвящена исследованию возможностей применения аппаратных счетчиков производительности для выявления потенциальных угроз безопасности критически важных систем и комплексов. Приведено описание прикладного программного интерфейса измерения производительности, доработанного авторами для отечественной аппаратно-программной платформы.

Ключевые слова: Счетчики производительности, информационная безопасность, атаки по сторонним каналам, PAPI

1. Введение

Многие современные процессоры поддерживают профилирование программ посредством использования аппаратных счетчиков – специальных регистров, фиксирующих аппаратные события определенного типа. В качестве примера аппаратных событий можно привести общее число процессорных тактов, общее число выполненных команд, число выполненных операций с плавающей точкой, число промахов при обращении к виртуальной и кэш-памяти и др.

Изначально аппаратные счетчики предназначались именно для построения профилей выполнения и последующей оптимизации, однако они могут выполнять и другую важную функцию – помогать разработчикам и системным архитекторам своевременно выявлять так называемые атаки по сторонним каналам (side-channels attacks [1]). В рамках данной работы авторы исследовали возможность использования аппаратных счетчиков для обнаружения такого рода атак, а также портировали программный интерфейс измерения производительности (Perfomance Application Programming Interface, PAPI [2]) на отечественные

Работа выполнена в рамках государственного задания по проведению фундаментальных исследований по теме «Исследование и реализация программной платформы для перспективных многоядерных процессоров» (FNEF-2022-002)

процессоры архитектуры MIPS, работающие под управлением отечественной операционной системы реального времени.

2. Проект PAPI

Целью проекта PAPI является разработка, стандартизация и реализация переносимого и эффективного прикладного программного интерфейса для доступа к аппаратным счетчикам. Проект PAPI поддерживается консорциумом Parallel Tools Consortium (`http://www.ptools.org`). На сегодняшний день PAPI, по сути, является стандартом de facto, для разработчиков ПО, осуществляющего доступ к аппаратным счетчикам.

На рисунке 1 представлена архитектура PAPI.

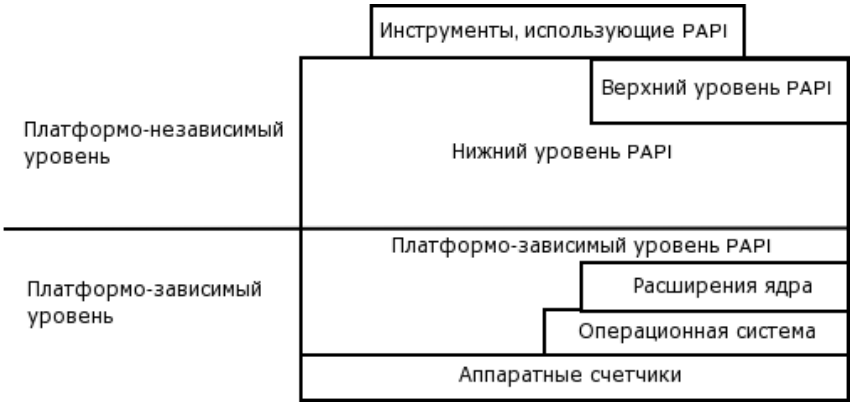


Рис. 1. Архитектура PAPI

В PAPI выделяется два основных уровня: платформо-независимый и платформо-зависимый, скрывающий от пользователя детали реализации доступа к аппаратным счетчикам конкретного процессора. Для этого, в частности, функции PAPI платформо-зависимого уровня могут использовать расширения ядра, функции целевой операционной системы или непосредственно язык ассемблера.

3. Интерфейсы PAPI

Пользователям PAPI предоставляются интерфейсы верхнего и нижнего уровней, различающиеся сложностью настройки и использования. В настоящее время существуют, например, реализации интерфейсов для таких языков высокого уровня, как Си и Фортран.

3.1. Высокоуровневый интерфейс. В состав высокоуровневого интерфейса входит всего 7 функций, предоставляющих базовые операции над аппаратными

счетчиками: запуск, останов, чтение со сбросом и без. При этом можно использовать только события, предопределенные стандартом PAPI.

Функции интерфейса верхнего уровня используют низкоуровневый интерфейс PAPI, освобождая пользователя от явных вызовов, например, функции инициализации библиотеки PAPI.

int PAPI_num_counters(void)

Инициализирует PAPI (если требуется). Возвращает число аппаратных счетчиков.

int PAPI_start_counters(int *events, int len)

Инициализирует PAPI (если требуется). Связывает множество событий с аппаратными счетчиками. Запускает счетчики.

int PAPI_stop_counters(long_long *vals, int alen)

Останавливает счетчики и сохраняет их значения в массиве vals.

int PAPI_accum_counters(long_long *vals, int alen)

Прибавляет значения счетчиков к значениям в массиве vals и обнуляет счетчики.

int PAPI_read_counters(long_long *vals, int alen) Считывает значения счетчиков в массив vals и обнуляет счетчики.

int PAPI_flips (float *real_time, float *proc_time, long long *flpins, float *mflpins)

int PAPI_flops (float *real_time, float *proc_time, long long *flpins, float *mflpins)

int PAPI_ipc (float *real_time, float *proc_time, long long *ins, float *ipc)

Упрощенные вызовы для измерения числа команд и операций с плавающей точкой, а также частоты выполнения команд процессора. Кроме того, эти функции возвращают реальное время работы процессора, а также виртуальное время, то есть время выполнения пользовательского процесса.

3.2. Низкоуровневый интерфейс. Низкоуровневый интерфейс обладает по сравнению с высокоуровневым интерфейсом расширенной функциональностью и большей эффективностью. В его состав входит более 50 различных функций, которые можно условно разделить на следующие группы:

- инициализация библиотеки PAPI
- функции измерения времени
- функции получения информации
- служебные функции
- функции управления множествами событий
- функции управления аппаратными счетчиками

В примере 1 показано использование низкоуровневого интерфейса PAPI для подсчета общего числа тактов процессора, а также числа команд сопроцессора плавающей арифметики во время вызова функции do_work.

Пример 1.

```
#include <papi.h>

#define NUM_EVENTS 2

int Events[NUM_EVENTS] =
{PAPI_FP_INS,PAPI_TOT_CYC};
int EventSet;
long long values[NUM_EVENTS];

/* Инициализация библиотеки PAPI */
ret = PAPI_library_init(PAPI_VER_CURRENT);

/* Создать новое множество событий */
ret = PAPI_create_eventset(&EventSet);

/* Добавить новые события в множество */
ret = PAPI_add_events(&EventSet,Events,NUM_EVENTS);

/* Стартовать счетчики */
ret = PAPI_start(EventSet);

do_work();  /* Искомая функция */

/* Остановить счетчики и сохранить результат в
массиве values */
ret = PAPI_stop(EventSet,values);
```

Для используемой целевой системы авторами была доработана и портирована версия papi-c 3.9.0. Такой выбор объясняется слабой зависимостью этой версии от системных вызовов современных ОС семейства Windows или Linux.

4. Анализ потенциальных угроз, которые можно выявлять с помощью аппаратных счетчиков

4.1. Исполнение внедренного вредоносного кода. Такого рода атаки можно назвать классическими с точки зрения информационной безопасности. Как

будет показано далее, анализ профилей выполнения, построенных на аппаратных счетчиках производительности, может своевременно выявлять эти атаки.

4.2. Атаки через кэш-память. Атаки этого типа используют измеренное время доступа к общей кэш-памяти, а затем извлекают конфиденциальную информацию у жертвы. Существует множество хорошо известных атак, таких как Flush+Reload или Prime+Probe [1].

Используя атаки этого типа, злоумышленник определяет использование кэш-памяти жертвой, измеряя время доступа к строке кэш-памяти, которая является общей для злоумышленника и жертвы. Практически все эти атаки являются скрытыми и незаметными для жертвы.

5. Роль аппаратных счетчиков в обеспечении информационной безопасности

Авторы проводили эксперименты на процессоре Intel Core I5, под управлением ОС Fedora Core. Для имтитации атаки использовался инструмент Mastik ([3]). Поскольку атаки по сторонним каналам с использованием кэш-памяти предполагают увеличение числа промахов по кэш-памяти 3-го уровня (L3), то вполне логично использовать это событие (L3_MISS) в качестве индикатора потенциальной угрозы. Однако одного его недостаточно. Сама логика приложения может предполагать большое число событий L3_MISS, например, при работе с большим числом данных, нехранящихся локально. Поэтому было предложено еще одно событие (L1_REPL), показывающее, как часто замещаются строки в кэш-памяти 1-го уровня. Теперь если взять их отношение $\frac{L3\_MISS}{L1\_REPL}$, то полученный индикатор будет означать, что приложение значительно использует память, но при этом часто очищает кэш. Что может свидетельствовать о проводимой атаке.

Последующие эксперименты показали, что за время измерения атакуемые приложения имели в несколько раз более высокое значение предложенного индикатора (в среднем, в 5 раз), по сравнению с его же значением в обычном режиме работы.

Также были построены профили типичного выполнения задач на отечественной аппаратной платформе с архитектурой MIPS под управлением операционной системы реального времени собственной разработки. Поскольку задачи, в основном, вычислительные, то в качестве критерия было предложено использовать отношение числа выполненных команд сопроцессора плавающей арифметики к числу всех выполненных команд (профиль строился для каждого критического потока управления). Замеры проводились в определенных заранее контрольных точках. Эксперименты показали, что разброс в значениях критерия при разных запусках не превысил 1%. Такой подход позволяет устранить проблему недетерминизма аппаратных счетчиков [4]. В профиль были включены и

события по выполнению перехода и выполнению инструкции ветвления. Учитывая недерминизм аппаратных счетчиков, сравнение профилей выполнения задачи в эксплуатационном режиме с построенными в ходе настройки системы мониторинга эталонными профилями проводилось по контрольным событиям перехода и ветвления (были выделены контрольные последовательности событий, нарушение которых является признаком потенциального сбоя или атаки). Стресс-тестирование показало, что система мониторинга успешно выявляла некорректные последовательности событий.

6. Заключение

Авторами был проведен анализ потенциальных угроз, реализован прикладной программный интерфейс доступа к аппаратным счетчикам производительности, проведены исследования, подтверждающие изначальное предположение о том, что профиль выполнения атакуемой системы, построенный на определенных аппаратных событиях, меняется, что позволяет построить эффективную систему мониторинга и защиты.

В проекте PAPI предложено стандартизованное, переносимое решение для профилирования кода посредством управления аппаратными счетчиками событий. На сегодняшний день существуют реализации PAPI в виде библиотек для многих современных платформ. Следует отметить также хорошую документированность проекта и простоту использования предлагаемых интерфейсов, что позволило осуществить его реализацию для отечественной аппаратно-программной критически важной системы.

ЛИТЕРАТУРА

1. F. Liu, Y. Yarom, Q. Ge, G. Heiser, R.B. Lee. Last-Level Cache Side-Channel Attacks are Practical, Security Privacy. In Proceedings of the 2015 IEEE Symposium on Security and Privacy, San Jose, CA, USA, 17–21 May 2015
2. PAPI User's Guide, `http://icl.cs.utk.edu/papi/`
3. Mastik: A Micro-Architectural Side-Channel Toolkit, `https://github.com/0xADE1A1DE/Mastik`
4. S. Das, J. Werner, M. Antonakakis, M. Polychronakis and F. Monrose. SoK: The Challenges, Pitfalls, and Perils of Using Hardware Performance Counters for Security. 2019 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 2019, pp. 20-38, doi: 10.1109/SP.2019.00021.

UDC: 65.011.56

On Reports in Open Transport Data Analysis Platform

M.V. Bulygin[1] and D.E. Namiot[1]

[1]Moscow State University, Leninskiye Gory 1, Moscow, Russian Federation

messimm@yandex.ru, dnamiot@gmail.com

Abstract

According to UN studies, more than 60% of the population will live in cities by 2030. To change existing and build new transport systems, it is important to receive reports on changes in the transport behavior of city residents. In this article, we talk about the automated report generation module for an open platform for transport data analysis. The classification of transport changes in the city is given. The work of the module is demonstrated in the example of the analysis of changes in traffic flows in Moscow during the celebrations in early May.

Keywords: Data Analysis, Transport Data, Data Processing Platforms, Big Data, Digital Urbanism

1. Introduction

According to UN studies, the share of the urbanized population will reach 60% by 2030. The urban population is increasing, and new areas are being built up. At the same time, the transport needs of city residents are also changing. City services need to receive reports on changes in the traffic situation in the city in order to take appropriate measures: changes in metro and bus schedules, adjustments to existing routes, construction of new metro stations, and others. As part of our research, we are developing an open platform for transport data analysis. Using this platform, it is possible to develop a solution for the automated generation of reports on the state of the city's transport system.

2. On the modern data sources

Previously, censuses, surveys, questionnaires, and manual data collection (with the help of enumerators) were the data sources for solving problems in the transport planning field. Thanks to the development of Internet of Things technologies, cellular communications, and big data storage, many new data sources for research have appeared.

2.1. Validators data. In modern public transport systems in metropolitan areas, smart cards are the primary means of payment. RFID technology is the basis for the production of these cards. Users bring their cards to the validators to pay for the fare. Validators are usually a PoS (Point of Sale) payment terminal. Payment data is captured and stored. Such data most often contains the place identifier and the exact validation time, card identifier, and may also contain some additional data, for example, the type of card or the number of remaining trips.

2.2. Cellular Operators Data. Currently, cellular operators are installing base stations in the metro to improve the level of service. During their operation, cell phones communicate with these stations about signal strength and delay. From this data, cellular operators can determine at which station the cell phone owner began his trip and the endpoint station of this trip.

2.3. Data types. Post-collection data contains records of each trip. Such individual data is suitable for analyzing the trajectories of individual passengers and allows researchers to make more accurate conclusions. However, such data are too voluminous and inconvenient for storage and analysis. There are approaches that personalize the data of cellular operators. For avoiding these shortcomings, data is aggregated by time (for example, by half-hour intervals) and by space, for example, by metro stations or hoods.

3. Literature review

F. Calabrese was one of the first to suggest using the data of cellular operators to measure traffic flows in the articles [1, 2]. He and his co-authors presented a system for solving several problems of digital urbanism according to the platform developed by Telecom Italia. Currently, many authors support the idea of transforming urban infrastructure systems into smart city systems using data analysis platforms of telecom operators, public transport system validators, and IoT system sensors. [3, 4, 5, 6]. Currently, there are several such platform solutions [7, 8, 9] that exist and are being developed at once.

4. On the open platform architecture

Currently, there are many platforms for solving digital urbanism problems based on analyzing individual and aggregated data. The architectures of these platforms, according to our review, are fixed. They do not provide the ability to dynamically extend functionality.

As part of our research, we are developing a modular platform, the architecture of which is shown in the Fig.1 below [10].

In this architecture, the general tasks of storing and processing data, as well as visualizing the results, are solved by dedicated modules. New models within a given

platform can be defined using an API module. This makes the system architecture open and allows platform users to create new solutions without re-writing data modules.

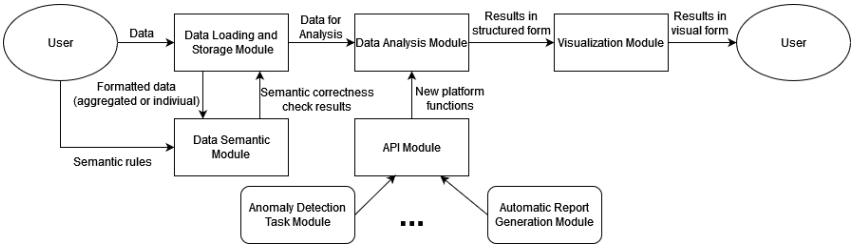


Fig. 1. Platform architecture

5. Reports on transport usage

For timely response to events taking place in the city, transport services need to receive reports about them. Events according to their impact on the city transport network are divided by us into three types: short-term, medium-term, and long-term.

5.1. Short-term changes. Short-term changes in the city traffic flow may be associated with incidents in the transport network. Examples of such incidents can be a person falling on the rails in the subway, smoke in the subway tunnels, or major accidents on the route of city buses. Such changes cannot be predicted in advance. City services need to understand the number of passengers affected by such an incident to take action, for example, allocating compensation buses. Such changes do not affect the overall volume of the traffic flow, but may only cause delays. To provide reports on such events, a digital twin of the existing transport system can be created. During an incident, the operator can view the workload of the affected areas and assess the scope of necessary measures.

Another reason for short-term changes in the city's transport network may be important social events (concerts, fireworks, football matches). Information about such events is usually known in advance, so there is no need to predict them either. To provide additional transport, it is important to understand the amount of traffic flow generated by such an event. To generate reports on similar events, it is necessary to extract information about the change in traffic flows during similar events in the past. Anomaly detection algorithms can be used to identify points of change in traffic flows. It is important to note that they must be applied taking into account the type of day of the week (weekend/weekday) and the time of observation.

5.2. Mid-term changes. Medium-term changes include periodic changes in traffic flow. Such changes may be associated with long holidays (New Year and May

holidays in Russia, Chinese New Year and Mid-Autumn Festival in China), as well as vacations for schoolchildren and students. The start and end dates of such changes are usually known in advance, but it is necessary to adjust the transport system to changes in the transport requests of city residents. Changes in the traffic flow are continuous, but then there is a return to normal.

5.3. Long-term changes. Long-term changes are associated with major changes in the city. These changes are permanent and then become the new normal. Major changes may take place over a long period. An example of a continuous change in traffic flow would be a new residential complex opening. The traffic flow will increase as the complex is occupied and then after the end of the settlement, it will stabilize.

Changes can occur quite quickly in the case of the opening of new transport facilities, for example, the launch of a new metro station, the introduction of new bus routes, or the construction of new interchange hubs

All described types of changes are presented in Fig.2

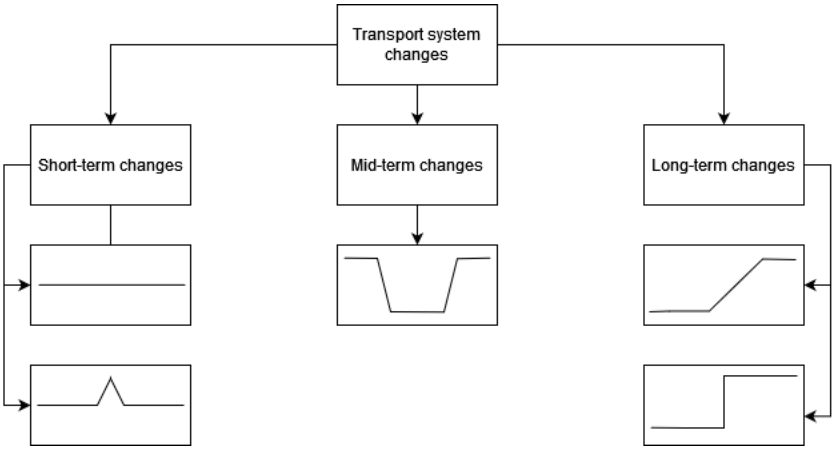


Fig. 2. Type of changes in transport systems

6. Example of platform usage

In Russia, the Spring and Labor Day (May 1) and Victory Day (May 9) are celebrated annually at the beginning of May. Each of these holidays corresponds to several days off in the production calendar. Working people often take vacations on intermediate working days to get a long holiday. The traffic flow of people moving to their workplaces in the morning and to their place of residence from work in the evening is significantly reduced these days. With the end of the holidays, traffic flows return to normal. According to our classification, this event belongs to the

medium-term ones. We applied the developed module to the individual transport data of metro users for May 2021. An example of a text report compiled based on the analysis of the traffic flow of the Elektrozavodskaya metro station for May 4 is shown in the Fig.3.

Station: **Elektrozavodskaya**
Date: 04.05.2021
Event: May Celebrations
Day type: **Weekday**

Report: The total outgoing traffic flow per day was reduced by 8177 people.
Typical traffic flow volume value: 18532. Observed value: 10355

Fig. 3. Example of text report on mid-term change

A similar change in the traffic flow was observed throughout all working days of the period under review. Report generation is also available for individual time intervals, which allows transport services to identify time intervals with the greatest changes in the traffic flow.

7. Future development

In the future, we plan to expand the set of tasks solved using the platform. In modern public transport systems, there are concession tickets for various social groups such as schoolchildren, students, pensioners, and the disabled. The fare for these concessionary tickets is significantly lower than the fare for a regular ticket. Social cards can be used by fraudsters to save money on travel. Each of the social groups has its pattern of transport behavior. For example, for schoolchildren in the Moscow region, it is typical to make one trip to the place of study in the interval from 07:30 to 08:00 in the morning, as well as a return trip before 15:00 to the place of residence. The constant use of a social card for trips between 09:00 and 10:00 and return trips between 18:00 and 19:00 may be a signal of fraudulent activities.

We also plan to expand the set of supported data. Sets of both individual and aggregated data can be collected from car-sharing or scooter/bike rental systems. Data on the place and time of the beginning and end of trips can be used to identify areas with low and high demand for sharing facilities. This information can be used to maximize profits.

8. Conclusion

We have developed a module for the automated generation of reports on the city's transport system state for an open transport data analysis platform. This module

allows transport services to generate different types of reports following the proposed classification of changes in the city's transport network. The reports generated using this module were used to study the changes in traffic in the Moscow metro during the working days between holidays in May in Russia.

REFERENCES

1. Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., and Ratti, C. (2010). Real-time urban monitoring using cell phones: A case study in Rome. IEEE transactions on intelligent transportation systems, 12(1), 141-151.
2. Calabrese, F., Di Lorenzo, G., Liu, L., and Ratti, C. (2011). Estimating Origin-Destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area.
3. Young, M., and Farber, S. (2019). The who, why, and when of Uber and other ride-hailing trips: An examination of a large sample household travel survey. Transportation Research Part A: Policy and Practice, 119, 383-392
4. Becker, H., Balac, M., Ciari, F., and Axhausen, K. W. (2020). Assessing the welfare impacts of Shared Mobility and Mobility as a Service (MaaS). Transportation Research Part A: Policy and Practice, 131, 228-243.
5. Holguín-Veras, J., Leal, J. A., Sánchez-Diaz, I., Browne, M., and Wojtowicz, J. (2020). State of the art and practice of urban freight management: Part I: Infrastructure, vehicle-related, and traffic operations. Transportation Research Part A: Policy and Practice, 137, 360-382.
6. Xu, Z. (2022). UAV surveying and mapping information collection method based on Internet of Things. Internet of Things and Cyber-Physical Systems, 2, 138-144.
7. Golubev, A., Chechetkin, I., Solnushkin, K. S., Sadovnikova, N., Parygin, D., and Shcherbakov, M. (2015, December). Strategway: web solutions for building public transportation routes using big geodata analysis. In Proceedings of the 17th international conference on information integration and web-based applications and services (pp. 1-4).
8. Hailin Feng, Haibin Lv, Zhihan Lv (2023). Resilience towarded Digital Twins to improve the adaptability of transportation systems. Transportation Research Part A: Policy and Practice, Volume 173
9. Bogomolov, Y., and Sobolevsky, S. (2023). A Scalable Spatio-Temporal Analytics Framework for Urban Networks. In Networks in the Global World VI: Proceedings of NetGloW 2022 (pp. 68-78). Cham: Springer International Publishing.
10. Bulygin, M. V., Namiot, D. E., and Pokusaev, O. N. (2023). On the analysis of individual data on transport usage. Proceedings ISA RAN, 73(1):24–33.

UDC: 621.391

Estimation of MAP component decoding of product codes in two-state channels

Anna Fominykh[1] and Andrei Ovchinnikov[2]

[1]Skolkovo Institute of Science and Technology Bolshoy Boulevard 30, bld. 1. Moscow, Russia

[2]HSE University, Kantemirovskaya Street 3, k. 1, lit. a., Saint-Petersburg, Russia

anna.fominykh@skoltech.ru, a.ovchinnikov@hse.ru

Abstract

Product code construction is a powerful error-correcting tool for both channels with and without memory. The common approach to decoding product (iterative) code is to apply consequent decoders in a sequential manner. The paper examines the influence of memory in the channel on iterative decoding for hard decision, soft decision, and trellis-based decoding algorithms. Also, the attainable performance of iterative schemes with and without knowledge of channel state information is presented.

Keywords: Iterative codes, channels with memory, MAP decoding

1. Introduction

Research in coding theory began with the work of C. Shannon in 1948. Coding theory develops methods for information processing in order to protect it against errors appearing during data transmission, storage, and processing. Transmission, storage, and processing procedures are often described by mathematical channel models for the convenience of the analysis. The mathematical models of the channel are divided into channels with memory and channels without memory (memoryless channels). In memoryless channels, the errors appearing are independent, while when transmitted over a channel with memory, the errors are not independent and appear in groups, composing error bursts. With the development of information and coding theory, most of the research efforts were aimed at studying discrete channels without memory, while the direction of channels with memory turned out to be a less investigated area, despite the fact that the presence of the memory effect in the channel leads to an "increase" in channel capacity.

Several coding techniques could be applied to channels with memory. One of the common techniques is an iterative design that allows to build longer codes from

shorter codes, most often Bose–Chaudhuri–Hocquenghem (BCH) codes or Hamming codes. Some research has been done on the application of low-density parity-check and polar codes in channels with memory [1]. Iterative codes, as opposed to other types of codes, have an appropriate structure for burst error correction without the need for additional interleaving. A widely used method for iterative code decoding is the Chase-Pyndiah algorithm. At the same time, the use of short component codes in an iterative design allows for optimal trellis decoding methods. A Berlekamp-Massey (BM) algorithm is another well-known approach to decoding up to half the minimum code distance [2].

The purpose of this paper is to compare the error probability provided by the BM algorithm, the Chase-Pyndiah algorithm, and the optimal trellis-based maximum a posteriori probability (MAP) algorithm, as well as to evaluate the influence of the presence of memory in the channel on the degradation of channel parameters.

2. Model overview

2.1. Channels with memory. One of the common ways to represent channels with memory is by Markov chains, which describe the channel that transits between states [3]. Most frequently, it is assumed that the number of states is finite. Important examples of channels described by Markov chains with two states are the Gilbert and Gilbert-Elliott models [4]. The first state is called a good state (denoted as G) and the second state is called a bad state (denoted as B). Each state is described by its own error probability. Gilbert model supposes that the error probability in the good state is zero. In 1963, the Gilbert model was generalized by E. Elliott, which suggested the error probability of a good state be non-zero. Being one of the first memory channel models, the Gilbert-Elliott model is still relevant and frequently used to describe actual systems. These models states that a previous state of the channel dictates its current state. Gilbert and Gilbert-Elliot channels are described by a set of probabilities $(P_B, P_G, P_{BG}, P_{GB})$. The probabilities P_B and P_G represent the bit error probabilities in a bad and good state, respectively. The pair of probabilities P_{BG} and P_{GB} is called transition probabilities and represent a probability of transitioning from bad to good and from good to bad state, respectively.

2.2. Iterative codes. Iterative or product codes were introduced by P. Elias in [5]. The iterative coding technique is an effective approach to composing long and powerful codes by using short component codes. The simplest case of iterative code with two component codes can be visualized as a matrix where each row corresponds to a codeword of one component code and each column corresponds to a codeword of another code. Generally, more than two component codes may be used to construct iterative code, but each extra code lowers the overall code rate.

2.3. Decoding algorithms. The common approach to decoding iterative codes is to perform subsequent decodings of the component codes. Lower error probabilities may be achieved if subsequent decoders exchange information about the reliability of decisions made (the so-called soft decisions) in addition to hard decoding results.

The Chase-Pyndiah algorithm is a well-known technique for such soft decision decoding and is suitable for discrete channels as well. At the same time, the Bahl-Cocke-Jelinek-Raviv (BCJR) algorithm, which guarantees optimal symbol-by-symbol MAP decoding, is one of the well-known optimal trellis-based decoding algorithms for sufficiently short block component codes.

3. Experiments and results

This section presents the simulation results for several scenarios in Gilbert and Gilbert-Elliott channels with parameters $(P_B, P_G, P_{BG}, P_{GB})$. The considered scenarios are

- Error probability comparison for 1 and 3 decoding iterations for Chase-Pyndiah, BCJR, and BM algorithms in binary-symmetric channel (BSC).
- Error probability comparison for 3 decoding iterations for Chase-Pyndiah, BCJR, and BM algorithms in Gilbert and Gilbert-Elliott channels without knowledge about channel states.
- Error probability comparison for 3 decoding iterations for Chase-Pyndiah, BCJR, and BM algorithms in Gilbert and Gilbert-Elliott channels with perfect knowledge about channel states.

In the simulation figures, the horizontal axis is plotted in terms of channel bit error probability (CBEP), which is calculated as $\text{CBEP} = (P_{GB}P_B + P_{BG}P_G)/(P_{GB} + P_{BG})$. The simulation parameters are: $P_B = 0.5$, $P_{GB} = 0.01$, P_{BG} varies, and $P_G = 0$ in the Gilbert channel and $P_G = 0.01$ in the Gilbert-Elliott channel. BCH codes are considered as component codes with a code length 31 and 21 information bits.

The error probability comparison for 1 and 3 outer decoding iterations for decoding algorithms in the BSC is given in Figure 1. The simulation results show that the performance of the BM algorithm with iterations improves not as much as the performance of BCJR and Chase-Pyndiah.

The error probability comparison for 3 outer decoding iterations for decoding algorithms in Gilbert and Gilbert-Elliott channels without knowledge about channel states is presented in Figures 2 and 3, respectively. The performance of the decoders is slightly worse in the Gilbert channel compared to the Gilbert-Elliott channel. The error probability comparison for 3 outer decoding iterations for decoding algorithms in Gilbert and Gilbert-Elliott channels with perfect knowledge about channel states is presented in Figures 4 and 5, respectively. Knowledge about channel state information is more beneficial for the BCJR algorithm in the Gilbert channel compared to the

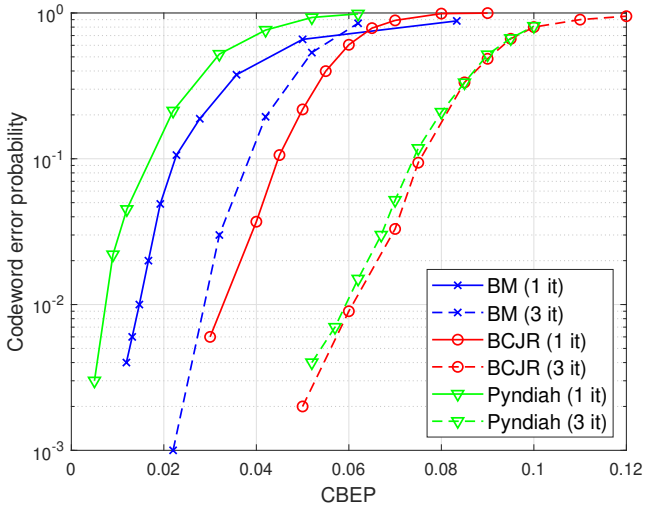


Fig. 1. Error probability comparison for 1 and 3 iterations in BSC

Gilbert-Elliott channel. The opposite behavior is observed for the Chase-Pyndiah algorithm.

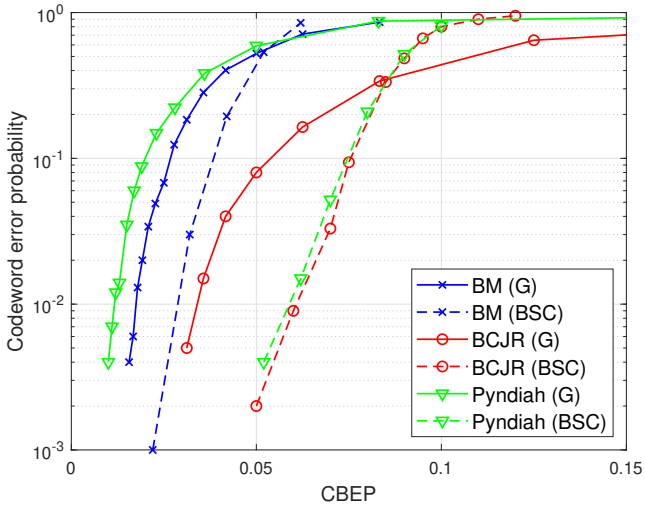


Fig. 2. Error probability comparison for 3 iterations in Gilbert and BSC channels

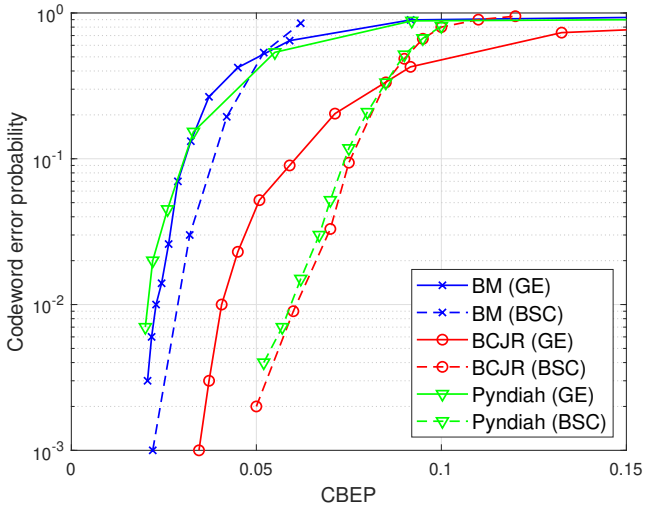


Fig. 3. Error probability comparison for 3 iterations in Gilbert-Elliott and BSC channels

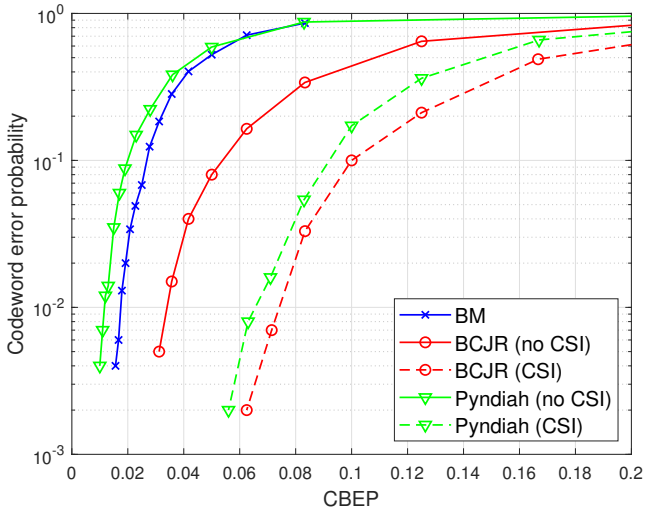


Fig. 4. Error probability comparison for unknown CSI and known CSI cases for 3 iterations in Gilbert channel

4. Conclusion

In this work, the error probabilities of the BM, Chase-Pyndiah, and BCJR algorithms were compared in channels with and without memory. An investigation of the influence of channel parameters on error probability degradation is provided.

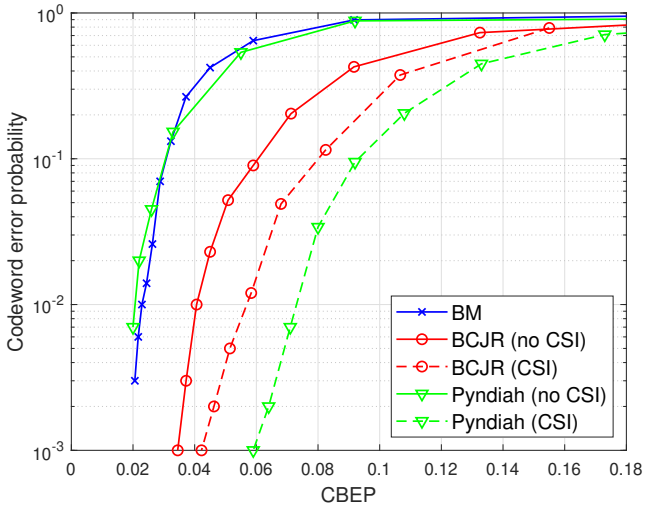


Fig. 5. Error probability comparison for unknown CSI and known CSI cases for 3 iterations in Gilbert-Elliott channel

Also, the attainable performance of iterative schemes with and without knowledge of channel state information is analyzed.

The article was prepared within the framework of the Basic Research Program at HSE University

REFERENCES

1. Ovchinnikov, A. A., Fominykh, A. A. Evaluation of error probability of iterative schemes for channels with memory // 2023 Wave Electronics and its Application in Information and Telecommunication Systems (WECONF). 2023. St. Petersburg, Russia.: IEEE. P. 1-5.
2. Lin, S. Li, J. Fundamentals of Classical and Modern Error-Correcting Codes. Cambridge: Cambridge University Press, 2022.
3. Bremaud, P. Markov Chains. Springer Cham, 2020.
4. Elliott, E. O. Estimates of error rates for codes on burst-noise channels // The Bell System Technical Journal. 1963. V. 42. P. 1977-1997.
5. Elias, P. Error-free coding // IRE Transactions on Information Theory. 1954. V 4, P. 29-37.

UDC: 004.94

The Peak Age of Information of URLLC service in 5G NR Systems

E. Zhbankova[1], V. Manaeva[1], E. Markova[1], Yu. Gaidamaka[1,2]

[1]RUDN University, 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation
[2]Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (FRC CSC RAS), 44-2 Vavilov St, Moscow, 119333, Russian Federation
zhbankova-ea@rudn.ru, 1032201197@rudn.ru,
markova-ev@rudn.ru, gaydamaka-yuv@rudn.ru

Abstract

Ultra-reliable low latency (URLLC) service is one of the cornerstone services that needs to be supported in fifth-generation (5G) cellular systems. For a class of applications utilizing URLLC service – periodic state updates, a vital metric of interest is Age of Information (AoI) characterizing the timeliness of updates received. The analysis of latency-related metrics in 5G cellular systems is however complicated by the orthogonal-frequency division multiple access (OFDMA) specifics resulting in batch service of packets. In this paper, we formalize the solve a queueing model capturing the specifics of URLLC service over OFDMA-enabled systems by explicitly accounting for batch service. For this model, we derive the mean peak AoI (PAoI).

Keywords: AoI, Age of Information, Peak Age of Information, 5G, URLLC, OFDMA

1. Introduction

The fifth generation (5G) systems needs to support a plethora of services with different quality of service requirements (QoS) including enhanced mobile broadband (eMBB), massive machine type communications (mMTC), ultra-reliable low latency service (URLLC) as well as different intermediate services. The recently standardized New Radio (NR) interface incorporates various advanced link level capabilities such as flexible frame structure and numerologies, different operational bands, network slicing for traffic isolation, etc. However, aside from eMBB service, the performance of mMTC and URLLC applications running over 5G NR radio interface is still loosely addressed in the literature.

The research was funded by the Russian Science Foundation, project No.22-79-10053, (`https://rscf.ru/en/project/22-79-10053/`).

A critical use-case for URLLC service in 5G systems is the exchange of state update information between end devices (ED) and the control center [1, 2]. Such use-case finds its application in many mission-critical fields such as telemedicine, control of production lines in industrial environments, automotive networks, video surveillance systems, energy grids. The main metric of interest for such services is the timeliness of the remote system updates available at the receiver.

Conventionally, the performance of time-constrained applications has been evaluated by utilizing latency as the main metric. However, this metric depends on the traffic load in the network that in turn is a function of the update interarrival time at the sources. Recently, a new measure of timeliness for state update services, the so-called Age of Information (AoI) has been proposed [3]. AoI quantifies how fresh the information available at the receiver with respect to the last update generated at the source. The metric is an explicit function of the update interarrival times and presumes that only timely received updates can reflect the current state of the system. AoI allows to describe the detailed behavior of the URLLC service operating over the 5G NR systems and can be considered as a new measure of QoS.

In spite of the significant interest AoI and peak AoI (PAoI) metrics attracted over the last few years, the models developed so far does not account for specifics of URLLC service over 5G systems. First of all, most of the models proposed so far assumed a single ED as an input. Furthermore, orthogonal-frequency division multiple access scheme specifics of 5G interface organization has not been addressed in the existing literature. The rationale is that this access scheme naturally leads to queueing formulations with batch service process. Such queueing systems have been loosely studied in the literature. We aim to fill these gap.

The aim of this paper is to characterize URLLC service performance operating over the 5G cellular network interface with orthogonal-frequency division multiple access (OFDMA) channel organization. To this aim, we first develop a system model that accounts for the specifics of URLLC service organization in 5G systems. Then, we proceed formalizing this model as a queueing system in discrete time and further provide its continuous approximation. We solve the latter for the mean PAoI.

The main contributions of our study are

- system model for service process of URLLC service in 5G NR systems with OFDMA channel access;
- mathematical formalization of the service process of URLLC EDs as a queueing model with multiple input flows and batch service time in both discrete and continuous time.

The rest of the paper is organized as follows. In section 2, we offer a brief overview of the related work s, formalize our system and specify two models in discrete and

continuous time. We then proceed solving the latter in Section 3 for AoI and PAoI as the metrics of interest. The conclusions are drawn in the last section.

2. System Model

Consider the system shown in Figure 1. The 5G NR base station (BS) serves N URLLC service end-devises (ED). We assume that URLLC traffic is completely isolated from other types of traffic at the BS. This can be achieved by using the Network Slicing (NS) function based on resource reservation [4, 5]. The length of one subframe, which is the transmission time interval (TTI), is $T = 10^{-6}$ s, the bandwidth available to URLLC traffic is W MHz. We assume the OFDMA access scheme with the size of one resource block (RB) of R_W MHz. Overall, there are K RBs available at each subframe.

We assume that each URLLC ED generates a packet of length L bits in each TTI with probability a, independently of other EDs. In URLLC applications the value L is often significantly smaller than the size of the RB. Therefore, we assume that one packet is transmitted in exactly one RB, regardless of the type of modulation and coding scheme (MSC) utilized. Finally, we assume that all N URLLC EDs are characterized by the same service priority. By considering these two facts one may observe that the utilized service discipline at the BS is first come first served (FCFS). The size of the buffer at BS is limited to R packets.

Due to the use of OFDMA channel organization the service process of packets is batch in nature. Specifically, we assume that at most K packets can be transmitted per TTI, where $K = \lfloor W/R_W \rfloor$. Here, K can be calculated using the assumptions about the blockage and micro-mobility of user equipment according to the methodology proposed in [6]. We also take into account the requirements for the reliability of

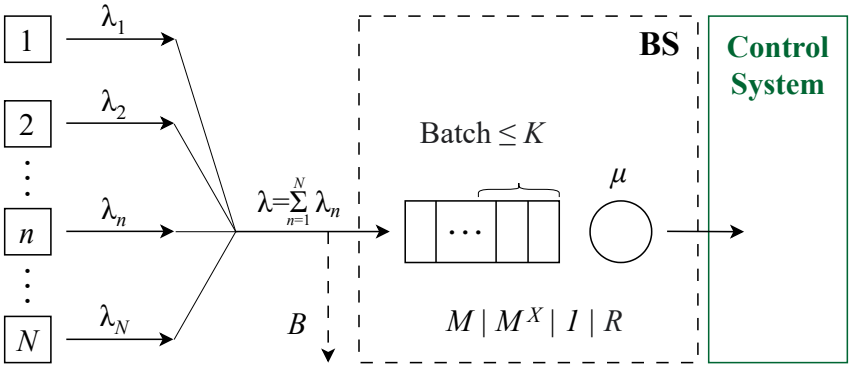


Fig. 1. The considered 5G BS serving URLLC end devices

URLLC traffic delivery. Specifically, we assume that the probability of erroneous reception of a packet is small which is achieved via repetition coding [7].

That considered technical system can be represented as a queueing system of the form $D^X/D^X/1/R$, where X indicates the batch arrivals and service [8] with binomial distribution of an arrival batch length $B(N, a)$ and at most K customers in a batch under service. Such a model is a special case of more general systems considered in [9, 10].

The problem with discrete queueing analysis is that (i) there might be no closed-form solution for stationary probabilities of the number of packets in the system, (ii) the delay distribution is often expressed in terms of infinite sums. Thus, in our paper, we approximate such a discrete system by a continuous system of the form $M/M^X/1/R$ in a way described in [11]. In our case the arrival rate $\lambda = E[A]/T$, where $E[A] = Na$ is the mean arrival batch length, T is the frame duration, the batch service rate is $\mu = 1/T$ and each batch has no more than K packets.

3. Performance Evaluation

We consider an $M/M^X/1/R$ queueing system of capacity R with a single server. The arrival process is a homogeneous Poisson with intensity λ. Packets are served in batches of no more than K packets depending on how many packets are available in the queue at the instant when service starts. The batch service time distribution is exponential with intensity μ. The service starts at the instant the server is released if there are requests in the queue or at the instant the first request arrives if the server is not busy. We consider the FCFS service discipline.

Let $X(t)$ be the Markov process of the packets number in the queue over the state space

$$\mathcal{X} = \{0, 1, \ldots, R\}, \tag{1}$$

with stationary state probabilities

$$q_i = \lim_{t\to\infty} P\{X(t) = i\},\ i \in \mathcal{X}. \tag{2}$$

The transition intensity diagram of the considered systems is shown in Fig.2.

Since the queueing system $M/M^X/1/R$ is stable for finite state space $\mathcal{X}$, one may utilize this diagram to define the system of equilibrium equations:

$$\begin{cases} \lambda q_0 = \mu \sum_{i=1}^{K} q_i, \\ (\lambda+\mu) q_i = \lambda q_{i-1} + \mu q_{i+K}, i = \overline{1, R-K}, \\ (\lambda+\mu) q_i = \lambda q_{i-1}, i = \overline{R-K+1, R-1}, \\ \mu q_R = \lambda q_{R-1}, \end{cases} \tag{3}$$

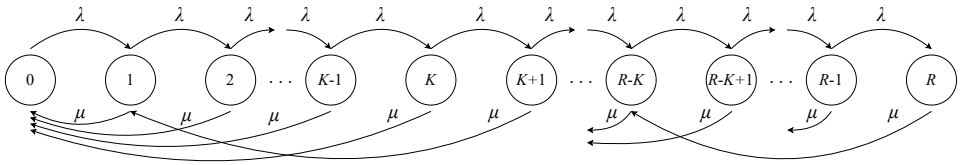


Fig. 2. Transition intensity diagram for $M/M^X/1/R$ queueing system

with the normalization condition $\sum_{i=0}^{R} q_i = 1$.

The defined system can be solved numerically using conventional approaches, e.g. Gauss-Zeidel method. Once the stationary distribution is found, the performance measures immediately follow:

- the mean number of packets in the queue

$$Q = \sum_{i=1}^{R} i q_i, \tag{4}$$

- the packet loss probability

$$B = q_R, \tag{5}$$

- the mean waiting time

$$\overline{\omega} = Q/\lambda(1-B), \tag{6}$$

- the mean PAoI

$$PAoI = \lambda^{-1} + \overline{\omega} + \mu^{-1}. \tag{7}$$

4. Conclusion

In this paper, we investigated the PAoI performance of URLLC services in 5G NR systems. To this aim, we first proposed a system model of a 5G NR BS serving multiple URLLC UEs. Then, we proceeded formalizing the queueing models in both discrete and continuous times by taking into account the specifics of the OFDMA-based access. For the latter model, we derived the mean PAoI time.

In the extended version of the paper, we will report the numerical results. To this aim, we will assume that 5G system utilizes the network slicing to provide a certain fixed amount of resources to URLLC service [12, 13]. Then, we will first determine the operational regime of the system by estimating the required amount of resources that need to be allocated for an URLLC slice, such that the packet loss probability satisfies the requirements of URLLC service, i.e. less than 10^{-5}. Then, we proceed assessing the delay and PAoI performance of the system.

REFERENCES

1. M. T. Okano, IoT and industry 4.0: the industrial new revolution, in: International Conference on Management and Information Systems, Vol. 25, 2017, pp. 75–82.
2. P. Schulz, M. Matthe, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S. A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, et al., Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture, IEEE Communications Magazine 55 (2) (2017) 70–78.
3. S. Kaul, M. Gruteser, V. Rai, J. Kenney, Minimizing age of information in vehicular networks, in: 2011 8th Annual IEEE communications society conference on sensor, mesh and ad hoc communications and networks, IEEE, 2011, pp. 350–358.
4. N. Yarkina, A. Gaydamaka, D. Moltchanov, Y. Koucheryavy, Performance Assessment of an ITU-T Compliant Machine Learning Enhancements for 5 G RAN Network Slicing, IEEE Transactions on Mobile Computing (2022) 1–17.
5. I. Kochetkova, A. Vlaskina, S. Burtseva, V. Savich, J. Hosek, Analyzing the effectiveness of dynamic network slicing procedure in 5g network by queuing and simulation models, Lecture Notes in Computer Science 12525 (2020) 71–85.
6. E. Khayrov, V. Prosvirov, A. Platonova, Traffic arrival model for millimeter wave 5g nr systems, Lecture Notes in Computer Science 13766 (2022) 161–175.
7. D. Ivanova, E. Markova, D. Moltchanov, R. Pirmagomedov, Y. Koucheryavy, K. Samouylov, Performance of priority-based traffic coexistence strategies in 5g mmwave industrial deployments, IEEE Access 10 (2022) 9241–9256.
8. M. Chaudhry, J. G. Templeton, First course in bulk queues., John Wiley & Sons, 1983.
9. D. Moltchanov, Y. Koucheryavy, J. Harju, Loss performance model for wireless channels with autocorrelated arrivals and losses, Computer Communications 29 (13-14) (2006) 2646–2660.
10. A. S. Alfa, Queueing theory for telecommunications: discrete time modelling of a single node system, Springer Science & Business Media, 2010.
11. G. Basharin, V. Efimushkin, Issledovanie odnolineynoy sistemy s zayavkami neskol'kikh tipov v diskretnom vremeni [study of a single-line system with requests of several types in discrete time], Problemy Peredachi Informatsii [Problems of Information Transmission] 20 (1) (1984) 95–104.
12. N. Yarkina, L. M. Correia, D. Moltchanov, Y. Gaidamaka, K. Samouylov, Multi-tenant resource sharing with equitable-priority-based performance isolation of slices for 5g cellular systems, Computer Communications 188 (2022) 39–51.

13. Y. Koucheryavy, E. Lisovskaya, D. Moltchanov, R. Kovalchukov, A. Samuylov, Quantifying the millimeter wave new radio base stations density for network slicing with prescribed slas, Computer Communications 174 (2021) 13–27.

UDC: 519.8

On queuing systems with N policy and various server activation strategies

Greeshma Joseph[1], Varghese Jacob[2], Achyutha Krishnamoorthy[3]

[1,3]CMS College, Kottayam, India
[2]Government College, Kottayam, India

Abstract

In this paper, we consider two single server queues under N policy with different activation strategies of the server. Customers arrive according to a Markovian Arrival Process and the service time follows a phase type distribution. The activation time of the server is exponentially distributed. We obtain stationary distribution of the queuing process using the Matrix Geometric Method. Using these distribution we calculate performance measures of the system. We also analyse this model numerically in order to have a comparison of the performance measures associated with it. An illustrative numerical example is discussed.

Keywords: queuing systems under N policy, activation time, forced activation, Matrix Geometric Method

1. Introduction

Many real-world situations involve queuing systems in which the server may be unavailable occasionally when the system becomes empty. In such systems, the server's idle times or service facilities can be utilised for other important purposes in the system to enhance its efficiency. These queuing systems have been investigated and applied extensively in various engineering systems, like production units, inventory systems, computers networks, flexible manufacturing domains, and telecommunication systems.

In literature there are a large number of policies regulating vacations in queuing systems like single vacation policy, working vacation policy, multiple vacation policy, N policy, T policy, D policy Krishnamoorthy, Ushakumari [1], and some of their combinations Chakravarthy et al. [2]. The concept of N policy was first introduced by Yadin and Naor [3] in an $M/G/1$ queueing system. Baker [4] examined operating policies in the $M/M/1$ queue with exponential start-up. Tian et al. [5] worked on an $M/M/1$ queue with single working vacation where the server gives service at a lower rate during it's vacation. Recently Sreenivasan et al. [6] extended it to $MAP/PH/1$

queue with working vacations, vacation interruptions and N policy. Greeshma et al. [7] have conducted a comparative study of queuing systems under N policy with variant of activation times. In this model, the authors have introduced the notion of forced activation into the queuing systems, where the server requires positive amount of time to activate before service. The work of [7] is extended in the following manner in this paper. Firstly we use a more general point process for modelling the arrivals, known as Markovian Arrival Process (MAP). Then for services we use the phase type distribution, which is the generalised version of some of the prominent distributions like exponential, hyperexponential and Erlang.

2. Model description

We consider two single server queuing models to which customers arrive according to a Markovian arrival process with parameter matrices D_0 and D_1 of dimension m. The service times follows a phase type distribution with representation (α, T) of order n. Both the queues are considered under N policy. The customers are served in the order of their arrival.

2.1. Model 1. In this model server requires a positive amount of time to start its service (that is, not instantaneous service). As soon as the total number of arrivals in the queue reaches the pre-determined threshold N $(1 \leq N < \infty)$, the server initiates the activation process. But it takes a random duration of time to get activated. When the sever is in activation mode there will be no service. Soon after completing the activation process, server begins its service. And the server is deactivated when all the customers present in the system are served. We assume that activation time of the server follows an exponential distribution with parameter θ.

Let $X(t)$ denotes the number of customers in the system at time t, $J_1(t)$ denotes the status of the server at time t,

$$J_1(t) = \begin{cases} 0 & \text{if the server is idle at time } t, \\ 1 & \text{if the server is in activation mode at time } t, \\ 2 & \text{if server is busy at time t.} \end{cases}$$

$S(t)$ denotes the phase of the service process when the server is busy, and $M(t)$ denotes the phase of the arrival process at time t.

Then $\{(X(t), J_1(t), S(t), M(t)) : t \geq 0\}$ is a four-dimensional Continuous Time Markov Chain with state space:

$$\Omega_1 = \bigcup_{i=0}^{\infty} l(i)$$

where $l(0) = \{(0,0,*,1),(0,0,*,2),\ldots,(0,0,*,m)\}$
and for $1 \le h \le n$ and $1 \le l \le m$,
$l(i) = \{(i,0,h,l) : 1 \le i \le N-1\} \cup \{(i,1,h,l) : i \ge N\} \cup \{(i,2,h,l) : i \ge 1\}$

2.2. Model 2. In our previous model, the server requires activation time which is exponentially distributed with parameter θ. But the server may take long duration of time (even though finite) to get activated. This long waiting for service can cause many constraints to the system like time shortage, energy consumption, holding cost of customers etc. Taking into account the impact of these constraints on the system, it would be better to activate the server by giving an extra force. So we consider forced activation of the server if the server does not get activated until the realization of a certain stage. In this model the server is forcefully activated, if the server does not get naturally activated up to the accumulation of a predetermined number $N+k$ ($k \ge 0$) of customers. As a result, the server will be in busy state when there are more than $N+k$ number of customers in the system.

Then $\{(X(t), J_2(t), S(t), M(t)) : t \ge 0\}$ is a four-dimensional Continuous Time Markov Chain with state space:

$$\Omega_2 = \bigcup_{i=0}^{\infty} l(i)$$

where $l(0) = \{(0,0,*,1),(0,0,*,2),\ldots,(0,0,*,m)\}$
and for $1 \le h \le n$ and $1 \le l \le m$,
$l(i) = \{(i,0,h,l) : 1 \le i \le N-1\} \cup \{(i,1,h,l) : N \le i \le N+k\} \cup \{(i,2,h,l) : i \ge 1\}$

3. Steady-state analysis

We can study the models described above as a quasi-birth-and-death(QBD) process. Using the lexicographical sequence for the states, the infinitesimal generator matrix Q of the QBD process has the form,

$$Q = \begin{pmatrix} B_{00} & B_{01} & & & \\ B_{10} & Q_1 & Q_0 & & \\ & Q_2 & Q_1 & Q_0 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}$$

3.1. Stability condition. Let π denote the steady-state probability vector of the generator $Q_0 + Q_1 + Q_2$. That is, $\boldsymbol{\pi}(Q_0 + Q_1 + Q_2) = \mathbf{0}$, $\boldsymbol{\pi}\mathbf{e} = \mathbf{1}$. The LIQBD description of the model indicates that the queuing system is stable (see, Neuts [12]) if and only if

$$\boldsymbol{\pi} Q_0 \mathbf{e} < \boldsymbol{\pi} Q_2 \mathbf{e}$$

3.2. Stationary distribution. The stationary distribution of the Markov chain under consideration is obtained by solving the set of equations

$$\mathbf{x}Q = \mathbf{0};\ \mathbf{x}\mathbf{e} = \mathbf{1} \tag{1}$$

where Q is the infinitesimal generator of the LIQBD describing the above process. Then $\mathbf{x}$ can be partitioned as,

$$\mathbf{x} = (\mathbf{x_0}, \mathbf{x_1}, \mathbf{x_2}, \ldots) \tag{2}$$

From $\mathbf{x}Q = \mathbf{0}$, we obtain the following equations;

$\mathbf{x_0}B_{00} + \mathbf{x_1}B_{10} = 0$
$\mathbf{x_0}B_{01} + \mathbf{x_1}Q_1 + \mathbf{x_2}Q_2 = 0$
$\mathbf{x_{i-1}}Q_0 + \mathbf{x_i}Q_1 + \mathbf{x_{i+1}}Q_2 = 0,\ i \geq 2$

By Matrix Geometric Method (Neuts [12]), the sub vectors $\mathbf{x_i}$'s has the form

$$\mathbf{x_i} = \mathbf{x_1}R^{i-1},\ i \geq 2$$

We obtained the rate matrix R explicitly by solving the matrix quadratic equation,

$$R^2Q_2 + RQ_1 + Q_0 = O \tag{3}$$

and $\mathbf{x_0}$ and $\mathbf{x_1}$ satisfies

$$(\mathbf{x_0}, \mathbf{x_1}) \begin{pmatrix} B_{00} & B_{01} \\ B_{10} & Q_1 + RQ_2 \end{pmatrix} = (0, 0)$$

The normalizing condition of (1) results in

$$\mathbf{x_0}\mathbf{e} + \mathbf{x_1}(I - R)^{-1}\mathbf{e} = 1.$$

After obtaining the rate matrix R, the vector $\mathbf{x}$ can be determined by exploiting the special structure of the coefficient matrices.

4. System performance measures of the models

We evaluate important performance characteristics of the models as follows.

- The probability that the server is idle, $p_{idle} = \mathbf{x_0}^*\mathbf{e} + \sum_{i=1}^{N-1} \mathbf{u_i}^*\mathbf{e}$
- The probability that the server is in activation mode,

- ⋆ for Model 2, $p_{act} = \sum_{i=1}^{\infty} \mathbf{w_i e}$
- ⋆ for Model 3, $p_{act} = \sum_{i=N}^{N+k} \mathbf{w_i}^* \mathbf{e}$

- The probability that the server is busy,
  - ⋆ for Model 2, $p_{busy} = \sum_{i=1}^{N-1} \mathbf{v_i}^* \mathbf{e} + \sum_{i=1}^{\infty} \mathbf{v_i e}$
  - ⋆ for Model 3, $p_{busy} = \sum_{i=1}^{N+k} \mathbf{v_i}^* \mathbf{e} + \sum_{i=1}^{\infty} \mathbf{x_i e}$
- The expected number of customers in the system when server is idle, $E_{idle} = \sum_{i=1}^{N-1} i\mathbf{u_i}^* \mathbf{e}$
- The expected number of customers in the system when server is in activation mode,
  - ⋆ for Model 2, $E_{act} = N + \sum_{i=1}^{\infty} (N-1+i)\mathbf{w_i e}$
  - ⋆ for Model 3, $E_{act} = N + \sum_{i=N}^{N+k} i\mathbf{w_i}^* \mathbf{e}$
- The expected number of customers in the system when server is busy,
  - ⋆ for Model 2, $E_{busy} = \sum_{i=1}^{N-1} i\mathbf{v_i}^* \mathbf{e} + \sum_{i=1}^{\infty} (N-1+i)\mathbf{v_i e}$
  - ⋆ for Model 3, $E_{busy} = \sum_{i=1}^{N+k} i\mathbf{v_i}^* \mathbf{e} + \sum_{i=1}^{\infty} (N+k+i)\mathbf{x_i e}$

5. Conclusion

In this paper, we studied two single server queues under N policy with MAP arrivals and phase type services. We have considered different activation strategies of the server in these two models. All the models are exhaustively analysed. Matrix Geometric Method is used to find the stationery probability vector, which makes it easy to obtain some key performance measures. The influence of various parameters on the system measures are also investigated through numerical example.

REFERENCES

1. Krishnamoorthy, A., Ushakumari, P.V.: k-out-of-n: G system with repair : the D-policy. Computers and Operations Research 28, 973-981 (2001)
2. Chakravarthy, S.R., Krishnamoorthy, A., Ushakumari, P.V.: A k-Out-of-n reliability system with an unreliable server and phase type repairs and services : The (N, T) policy. Journal of Applied Mathematics and Stochastic Analysis 14, 361-380 (2001)
3. Yadin, M., Naor, P.: Queueing system with a removable service station. Operational Research Quarterly 14, 393-405 (1963)
4. Baker, K.R.: A note on operating policies for the queue M/M/1 with exponential start up. Information Systems and Operational Research 11, 71-72 (1973)
5. Tian, N., Zhao, X., Wang, K.: The M/M/1 Queue with Single Working Vacation. International Journal of Information and Management Sciences 19 (4), 621-634 (2008)
6. Sreenivasan, C., Chakravarthy, S.R., Krishnamoorthy, A.: MAP/PH/1 queue with working vacations, vacation interruptions and N policy. Applied Mathematical Modelling 37, 3879–3893 (2013)
7. Joseph, G., P N, J., Jacob, V.: A comparative study of queuing systems with variant of activation times under N policy. (communicated) (2023)
8. Neuts, M.F.: A versatile Markovian point process. J. Appl. Probab. 16(4), 764–779 (1979)
9. Chakravarthy, S.R.: Markovian Arrival Processes. Wiley Encyclopedia of Operations Research and Management Science (2011)
10. Neuts, M.F.: Probability distributions of phase type. Liber Amicorum Prof. Emeritus H. Florin (1975)
11. He, Q.: Fundamentals of Matrix-Analytic Methods. Springer, New York (2014)
12. Neuts, M.F.: Matrix-Geometric Solutions in Stochastic Models- An Algorithmic Approach. 2nd ed. Dover Publications Inc., New York (1994)
13. Dudin, A.N., Klimenok, V.I., Vishnevsky, V.M.: The Theory of Queuing Systems with Correlated Flows. Springer. (2020)
14. Jacob, V.: Analysis of Customer Induced Interruption in a Retrial Queuing System with Classical Retrial Policy. International Journal of Applied Engineering Research 15(5), 445-451 (2020)
15. Neuts, M.F., Rao, B.M.: Numerical Investigation of a multiserver retrial model. Queueing Systems 7, 169-190 (1990)
16. Lucantoni, D.M.: New results on the single server queue with a batch Markovian arrival process. Commun. Stat. Stoch. Model. 7(1), 1–46 (1991)

УДК: 004.724.4

Поиск корректного отображения параллельных вычислений на систему с коммутационной средой RapidIO методами ЦЛП

Н.И. Вьюкова[1], В.А. Галатенко[1], А.Н. Павлов[1], С.В. Самборский[1]

[1]ФГУ ФНЦ НИИСИ РАН, Нахимовский пр-т, д. 36, корп. 1, Москва, Россия

qniva@yandex.ru, galat@niisi.ras.ru, antony@niisi.msk.ru, sambor@niisi.ras.ru

Аннотация

Работа посвящена вопросам отображения параллельных вычислений на распределенные системы с коммутационной средой RapidIO. Под параллельным вычислением понимается множество процессов, взаимодействующих посредством передачи потоков данных. Требуется назначить вычислительные узлы для процессов и определить маршруты для потоков данных, а также построить таблицы маршрутизации для коммутаторов. В работе предложен подход, основанный на методах целочисленного линейного программирования (ЦЛП), позволяющий корректно отобразить параллельное вычисление на распределенную систему и оптимизировать отображение по ряду характеристик.

Ключевые слова: Распределенная система, параллельные вычисления, RapidIO, целочисленное линейное программирование

1. Введение

В ФГУ ФНЦ НИИСИ РАН разработаны процессорные и мезонинные модули на базе микропроцессоров архитектуры MIPS, а также коммутаторы и другие аппаратные средства для построения многопроцессорных комплексов, связанных посредством коммуникационной сети RapidIO [1].

При использовании многопроцессорных комплексов возникает вопрос оптимального отображения параллельного вычисления на распределенную вычислительную систему. Процессы взаимодействуют между собой путем передачи сообщений. Для эффективного выполнения параллельного вычисления в целом необходимо, чтобы сетевые маршруты обеспечивали определенный темп передачи данных между взаимодействующими процессами. Этого можно добиваться как

Работа выполнена в рамках государственного задания по проведению фундаментальных исследований по теме «Исследование и реализация программной платформы для перспективных многоядерных процессоров» (FNEF-2022-002)

путем варьирования назначения вычислительных узлов для процессов, так и путем изменения схемы маршрутизации в сети RapidIO. При этом вычислительный узел, назначаемый для каждого процесса, должен обладать определенными свойствами, такими как уровень производительности или наличие некоторого сопроцессора. Если число процессов исчисляется десятками, подбор вручную отображения и схемы маршрутизации оказывается весьма трудоемким и не гарантирует оптимального результата.

Работа посвящена вопросам точного решения задачи об отображении заданного параллельного вычисления на распределенную вычислительную систему с сетью произвольной топологии и статической маршрутизацией сетевой среды. Данная задача является NP-полной, и в настоящей работе для ее решения предложен подход, использующий методы целочисленного линейного программирования (ЦЛП). В качестве инструмента для описания и решения ЦЛП-задач применялся пакет GLPK [2].

2. Формальные модели

2.1. Модель распределенной вычислительной системы. Коммуникационная среда RapidIO представляет собой сеть с коммутацией пакетов, состоящую из узлов, соединенных физическими каналами связи [1]. Структура распределенных систем на базе аппаратуры линейки Комдив под управлением ОСРВ Багет, подробно описана в [3]. Формальная модель такой распределенной системы представлена в [4]. В данной работе, как и в [5], мы будем рассматривать упрощенную модель, включающую узлы двух типов:

- вычислительный узел - оконечное устройство, имеющее числовой идентификатор и, как правило, только один порт;

- коммутатор - устройство, предназначенное для маршрутизации, которое всегда имеет несколько портов.

Коммутатор работает под управлением таблицы маршрутизации, которая записывается во время инициализации коммуникационной среды RapidIO. Есть два типа коммутаторов. Коммутатор первого типа имеет одну общую таблицу маршрутизации для всех портов, в коммутаторе второго типа имеется индивидуальная таблица для каждого порта.

В коммутаторах первого типа выходной порт, с которого будет отправлен полученный пакет данных, однозначно определяется указанным в этом пакете идентификатором вычислительного узла - получателя. Это означает, что если маршруты, по которым передаются данные одному получателю, пересекаются на некотором коммутаторе, то далее они совпадают.

В коммутаторах второго типа выходной порт, с которого будет отправлен полученный пакет данных, определяется идентификатором получателя и номером

порта, на который этот пакет пришел. Таким образом, пакеты к одному получателю могут уйти с коммутатора по разным соединениям, если они поступили в него по разным соединениям.

Последовательность соединений, через которые проходит пакет от отправителя до получателя, называется маршрутом. Множество маршрутов для всех пар отправителей и получателей, между которыми происходит передача данных, называется *схемой маршрутизации*.

Представим теперь формальную модель распределенной системы, которая будет использована в формулировках ЦЛП-задач в разделе 3.

Распределенная система представляется в виде ориентированного размеченного (нагруженного) графа

$G = (H, C_1, C_2, L, b, perf)$

где H - множество вычислительных узлов, C_1 - множество коммутаторов первого типа, C_2 - множество коммутаторов второго типа. Обозначим через $Node$ множество вычислительных узлов и коммутаторов: $Node = H \cup C_1 \cup C_2$. Тогда $L \subset Node \times Node$ - множество дуг, соответствующих соединениям между узлами распределенной системы. Отображения $b : L \to \mathbb{N}$ и $perf : H \to \mathbb{N}$ задают, соответственно, пропускную способность соединений и производительность вычислительных узлов в некоторых условных единицах.

На практике в среде RapidIO соединения двунаправленные, с одинаковой пропускной способностью в обе стороны. Поэтому граф G можно было бы определить как неориентированный. Тем не менее мы рассматриваем его как ориентированный, поскольку в формулировках ЦЛП-задач соединения (n, n') и (n', n) удобно определять независимо друг от друга.

2.2. Модель параллельного вычисления. Параллельное вычисление мы будем представлять в виде ориентированного размеченного графа

$PG = (P, S, v, req)$

где P - множество процессов, реализующих параллельное вычисление, $S \subset P \times P$ множество дуг, связывающих взаимодействующие процессы: $s = (p, p') \in S$, если процесс p отправляет сообщения процессу p'. Допускаются как ациклические графы, так и графы с циклами. Дуги $s \in S$ мы далее будем называть потоками данных. Отображения $v : S \to \mathbb{N}$ и $req : P \to \mathbb{N}$ определяют, соответственно, требуемую для каждого потока данных пропускную способность сетевого маршрута и требуемую для каждого процесса производительность вычислительного узла.

3. Формулировки ЦЛП-задач

Были построены две формулировки ЦЛП-задач, связанных с отображением параллельного вычисления на распределенную вычислительную систему.

В основной задаче требовалось найти отображение множество процессов P параллельного вычисления $PG = (P, S, v, req)$ на множество вычислительных узлов H распределенной вычислительной системы $G = (H, C_1, C_2, L, b, perf)$ с учетом производительности узлов. При этом также требуется вычислить схему маршрутизации, обеспечивающую заданную пропускную способность для потоков данных между процессами, а также вычислить таблицы маршрутизации для коммутаторов.

Допускается отображение нескольких процессов на один узел, имеющий достаточную производительность. Но отображение одного процесса на несколько узлов недопустимо. Если возникает такая необходимость, например, когда требуемая для процесса производительность превышает производительность имеющихся вычислительных узлов, такой процесс должен быть заранее распараллелен на несколько процессов.

Маршруты и таблицы коммутации должны удовлетворять ряду ограничений. Это ограничения, связанные со спецификой коммутационной среды RapidIO, ограничения, обеспечивающие заданную пропускную способность для потоков данных между процессами, ограничения, исключающие маршруты с петлями.

В более простой задаче предполагается, что множество процессов P параллельного вычисления уже отображено на множество вычислительных узлов H распределенной вычислительной системы G. Требуется только вычислить корректную схему маршрутизации и таблицы маршрутизации для коммутаторов.

3.1. Константы, переменные, отображения и подмножества, используемые в формулировках ЦЛП-задач. Константный массив $LN : L \times Node \to \{-1, 0, 1\}$ определяет, как связаны соединения $l \in L$ и вершины $n \in Node$ графа G. Пусть $l = (n_1, n_2)$ - сетевое соединение, направленное от узла n_1 к узлу n_2, тогда:

$$LN[l, n] = \begin{cases} -1, & \text{if } n = n_1 \\ 1, & \text{if } n = n_2 \\ 0, & \text{if } n \notin \{n_1, n_2\} \end{cases}$$

Аналогично введем константный массив SP, определяющий связь между потоками данных и и процессами параллельного вычисления: $SP : S \times P \to \{-1, 0, 1\}$, описывающий связь между потоками данных и процессами параллельного вычисления. Пусть $s = (p_1, p_2)$ - поток данных от процесса p_1 к процессу p_2, тогда:

$$SP[s, p] = \begin{cases} -1, & \text{if } p = p_1 \\ 1, & \text{if } p = p_2 \\ 0, & \text{if } p \notin \{p_1, p_2\} \end{cases}$$

Введем также обозначения для следующих подмножеств соединений: $Lout_1$ - множество соединений, исходящих из коммутаторов первого типа; $Lout_2$ - мно-

жество соединений, исходящих из коммутаторов второго типа; Lin_2 - множество соединений, входящих в коммутаторы второго типа.

Искомое отображение множества процессов P на множество вычислительных узлов H представим в виде переменной M, имеющей тип массива бинарных значений: $M : H \times P \to \{0, 1\}$. Для $h \in H$, $p \in P$ значение $M[h, p] = 1$, если процессу p назначен вычислительный узел h, иначе $M[h, p] = 0$.

Основная переменная, описывающая маршруты для потоков данных из S, это переменная R, которая имеет тип массива бинарных значений: $R : L \times S \to \{0, 1\}$. Для $l \in L$, $s \in S$ значение $R[l, s] = 1$, если соединение l используется в маршруте для потока данных s, в противном случае $R[l, s] = 0$.

Массив переменных TC_1 определяет таблицы маршрутизации для коммутаторов первого типа, $TC_1 : Lout_1 \times H \to \{0, 1\}$. $TC_1[l, h] = 1$, если соединение $l \in Lout_1$ используется для передачи сообщений, адресованных вычислительному узлу $h \in H$.

Массив переменных TC_2, который определяет таблицы маршрутизации для коммутаторов второго типа определяется аналогично с учетом зависимости от входного соединения.

3.2. Примеры ЦЛП-ограничений. Показать полную формулировку ЦЛП-задачи здесь невозможно. Приведем только пару простых ограничений. С отображением M связаны очевидные требования, первое - каждому процессу должен быть назначен в точности один вычислительный узел, это гарантируется следующим набором равенств:

$\forall p \in P : \quad \sum_{h \in H} M[h, p] = 1$

Второе - сумма потребностей в производительности процессов, назначенных на один вычислительный узел, не должна превышать производительности этого узла. Для этого задаем неравенства:

$\forall h \in H : \quad \sum_{p \in P} req(p) \cdot M[h, p] \leq perf(h)$

Подобным образом можно сформулировать ограничения описывающие все остальные условия, требующиеся для корректности отображения параллельного вычисления на распределенную систему и соответствующего заполнения таблиц маршрутизации.

3.3. Целевая функция. С помощью целевой функции можно обеспечить оптимизацию решения по различным параметрам. Для данной задачи предлагается минимизировать линейную комбинацию следующих выражений: максимальная длина маршрута, суммарная длина маршрутов, суммарное число записей во всех таблицах маршрутизации (минимизация числа записей позволяет избавиться от "мусора"в таблицах маршрутизации).

Отметим, что используемые ограничения не запрещают несвязные маршруты с циклами, отдельными от основного маршрута (запрет сложно сформулировать

без порядка или нумерации узлов на маршруте). Но такие маршруты легко исключаются оптимизацией по суммарной длине маршрутов в целевой функции.

4. Заключение

Предложенный в работе подход был опробован на небольших примерах (до нескольких десятков процессов и узлов), и оказался работоспособным. Поскольку ручной подбор требуемого отображения весьма трудоемок и не всегда обеспечивает оптимальный результат, то можно значительно упростить процедуру отображения сложных параллельных вычислений на распределенные системы, использующие коммуникационную среду RapidIO. Применение методов ЦЛП или других подобных методов гарантирует нахождение отображения, если оно существует, и позволяет оптимизировать его по различным параметрам.

Тем не менее, в этой области остается еще множество нерешенных вопросов, требующих дальнейших исследований. К их числу относятся вопросы масштабируемости предложенных формулировок ЦЛП-задач, дополнительные параметры оптимизации искомого отображения, более реалистичные модели параллельного вычисления и распределенной системы, наконец, задача построения распределенной системы по заданному параллельному вычислению из набора вычислительных узлов и коммутаторов, как в [6].

ЛИТЕРАТУРА

1. RapidIO Interconnect Specification (Revision 1.3). http://www.rapidio.org/rapidio-specifications.
2. GNU Linear Programming Kit. https://www.gnu.org/software/glpk.
3. Годунов А.Н., Солдатов В.А. Конфигурирование многопроцессорных систем в операционной системе реального времени Багет // Программная инженерия. 2016. Т. 7. № 6. с. 243-251.
4. Бакулин А. А. Проверка допустимости схемы маршрутизации в системе RapidIO // Программные продукты и системы. 2011. № 4. с. 20-23.
5. Годунов А.Н., Солдатов В.А., Хоменков И.И. Передача сообщений в коммуникационной среде RapidIO для семейства операционных систем реального времени Багет // Программная инженерия. 2020. Т. 11. № 1. с. 26-33.
6. Bobda C., Ishebabi H., Mahr P., Mbongue J.M., Saha S.K. MeXT: A Flow for Multiprocessor Exploration // IEEE High Performance Extreme Computing Conference (HPEC). 2019. p. 1-7.

УДК: 519.85

Метод марковского суммирования для исследования суммарного потока повторных обращений в многофазной системе массового обслуживания с обратной связью

А.В. Подгайнов[1], М.А. Шкленник[1]

[1]Национальный исследовательский Томский государственный университет, проспект Ленина 36, г. Томск, Россия

artem.podgaynov1414@gmail.com, shklennikm@yandex.ru

Аннотация

В статье рассматривается многофазная система массового обслуживания с неограниченным числом приборов. Каждый переход заявки с фазы на фазу будем считать повторным обращением заявки к системе. В связи с чем, исследуемую многофазную систему будем считать системой с мгновенной обратной связью. В работе исследуется суммарный поток повторных обращений заявок к системе, то есть число всех повторных обращений всех заявок, поступивших в систему за определенный интервал времени. Получен вид характеристической функции суммарного числа повторных обращений за определенный интервал времени при нестационарном режиме работы системы.

Ключевые слова: *многофазная система массового обслуживания, метод марковского суммирования, характеристическая функция, повторное обслуживание*

1. Введение

Системы массового обслуживания (СМО) с обратной связью широко используются при исследовании процессов, протекающих в различных сферах деятельности, таких как социально-экономическая, демографическая и, конечно же, в последнее время очень актуальная и активно развивающаяся область телекоммуникаций. Поэтому моделирование процессов, учитывающих возможность повторного обращения заявки для обслуживания [1, 2], является очень востребованным. Возникающий при этом поток повторных обращений заявок в систему можно считать дополнительным потоком и для его исследования при нестационарном режиме работы системы будем использовать метод марковского

суммирования [3, 4]. Исследуемую многофазную систему можно рассматривать как СМО с неограниченным числом приборов и обратной связью, учитывающую различия во времени обслуживания при каждом повторном обращении заявки. Число всех повторных обращений всех заявок, поступивших в систему за определенный период времени, будем называть r-потоком. Число повторных обращений, реализованных одной заявкой, поступившей в какой-либо момент времени периода наблюдения за системой, до окончания этого периода, будем называть локальным r-потоком. Ставится задача нахождения вида характеристической функции для r-потока, имея вид характеристической функции для локального r-потока [5], методом марковского суммирования.

2. Математическая модель

Рассмотрим многофазную систему массового обслуживания с неограниченным числом приборов и мгновенной обратной связью (Рисунок 1). На вход поступает простейший поток заявок с параметром λ.

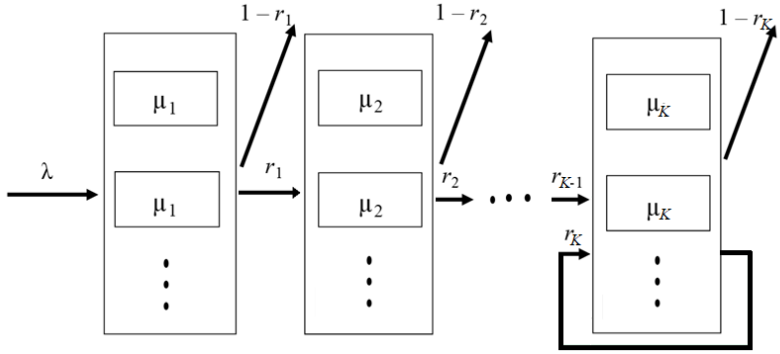


Рис. 1. Многофазная СМО

Время обработки заявки на каждой фазе является экспоненциально распределенной случайной величиной с параметром μ_i, i=1, 2, 3,..., K. После обработки на первой фазе заявка с вероятностью $(1 - r_1)$ покидает систему - завершает обслуживание, либо с вероятностью r_1 переходит на следующую фазу - повторно обращается к системе. После обслуживания на последней фазе, заявка с вероятностью r_K снова обращается к K-ой фазе, а с вероятностью $(1 - r_K)$ выходит из системы.

Рассмотрим процесс $\{n(t)\}$ - суммарное число переходов всех заявок с фазы на фазу за время t. Ставится задача исследования процесса $\{n(t)\}$ в нестационарном режиме, при условии, что в начальный момент времени система пуста.

3. Метод марковского суммирования

Для случайного процесса $\{n(t)\}$, определяющего число событий в r-потоке за интервал времени $[0,\ t]$, запишем дифференциальное уравнение Колмогорова для распределения вероятностей $P(n,t)$

$$\frac{\partial P(n,t)}{\partial t} = -\lambda P(n,t) + \lambda \sum_{i=0}^{n} P(n-i,t)g(i,t), \tag{1}$$

Переходя к характеристическим функциям вида

$$H(u,t) = \sum_{i=0}^{\infty} e^{jun} P(n,t), \tag{2}$$

$$G(u,t) = \sum_{i=0}^{\infty} e^{jui} g(i,t), \tag{3}$$

запишем дифференциальное уравнение для характеристической функции исследуемого процесса $\{n(t)\}$

$$\frac{\partial H(u,t)}{\partial t} = -\lambda H(u,t)(G(u,t)-1), \tag{4}$$

с начальными условиями

$$H(u,0) \equiv 1. \tag{5}$$

Вид характеристической функции для $G(u,t)$ процесса $\xi(t)$ определяется выражением (6), полученным в работе [5]

$$\begin{aligned} G(u,t) = \frac{\phi_1(u)}{\mu_1}(1-e^{-\mu_1(T-t)}) + e^{-\mu_1(T-t)}\psi_1(u) - \frac{\phi_2(u)}{\mu_2}(1-e^{-\mu_2(T-t)}) - \\ -\frac{e^{-\mu_2(T-t)}}{d_{21}}\psi_2(u) - ... - \frac{\phi_{K-1}(u)}{\mu_{K-1}}(1-e^{-\mu_{K-1}(T-t)}) - \frac{e^{-\mu_{K-1}(T-t)}}{d_{21}}\psi_{K-1}(u) - \\ -\frac{\phi_K(u)}{\mu_K(1-r_K eju)}(1-e^{-\mu_K(1-r_K e^{ju})(T-t)}) - \frac{e^{-\mu_K(1-r_K e^{ju})(T-t)}}{d_{21}}\psi_K(u), \end{aligned} \tag{6}$$

где

$$\gamma_k = \prod_{l=1}^{k} \mu_l r_l,$$

$$\phi_1(u) = \mu_1(1-r_1) + \sum_{m=1}^{K-1} \mu_{m+1}(1-r_{m+1})\gamma_m e^{mju}(\prod_{k=2}^{m+1} d_{k1})^{-1},$$

$$\phi_2(u) = \mu_2(1-r_2)\frac{\gamma_1 e^{ju}}{d_{21}} + \frac{1}{d_{21}}\sum_{m=2}^{K-1}\mu_{m+1}(1-r_{m+1})\gamma_m e^{mju}(\prod_{k=3}^{m+1} d_{k2})^{-1},$$

$$\phi_3(u) = \mu_3(1-r_3)\frac{\gamma_2 e^{2ju}A_3}{d_{21}} + \frac{A_3}{d_{21}}\sum_{m=3}^{K-1}\mu_{m+1}(1-r_{m+1})\gamma_m e^{mju}(\prod_{k=4}^{m+1} d_{k3})^{-1},$$

$$,...,$$

$$\phi_{K-1}(u) = \mu_{K-1}(1-r_{K-1})\frac{\gamma_{K-2}e^{(K-2)ju}A_{K-1}}{d_{21}} + \mu_K(1-r_K)\frac{\gamma_{K-1}e^{(K-1)ju}A_{K-1}}{d_{21}d_{KK-1}},$$

$$\phi_K(u) = \mu_K(1-r_K)\frac{\gamma_{K-1}e^{(K-1)ju}A_{K-1}}{d_{21}},$$

$$\psi_1(u) = 1 + \sum_{m=1}^{K-1}\gamma_m e^{mju}(\prod_{k=2}^{m+1} d_{k1})^{-1}, \psi_2(u) = \gamma_1 e^{ju} + \sum_{m=2}^{K-1}\gamma_m e^{mju}(\prod_{k=3}^{m+1} d_{k2})^{-1},$$

$$\psi_3(u) = \gamma_2 e^{2ju}A_3 + A_3\sum_{m=3}^{K-1}\gamma_m e^{mju}(\prod_{k=4}^{m+1} d_{k3})^{-1},$$

$$,...,$$

$$\psi_{K-1}(u) = \gamma_{K-2}e^{(K-2)ju}A_{K-1} + A_{K-1}\frac{\gamma_{K-1}e^{(K-1)ju}}{d_{KK-1}}, \psi_K(u) = \gamma_{K-1}e^{(K-1)ju}A_K.$$

Параметры $A_3, A_4, ..., A_K$ вычисляются по рекурентным формулам

$$A_3 = \frac{1}{d_{31}} - \frac{1}{d_{32}}, A_4 = \frac{1}{d_{31}d_{41}} - \frac{1}{d_{32}d_{42}} - \frac{1}{d_{43}}A_3,$$

$$A_5 = (\prod_{k=3}^{5} d_{k1})^{-1} - (\prod_{k=3}^{5} d_{k2})^{-1} - A_3(\prod_{k=4}^{5} d_{k3})^{-1} - A_4(\prod_{k=5}^{5} d_{k1})^{-1}, ...,$$

$$A_{K-1} = (\prod_{k=3}^{K-1} d_{k1})^{-1} - (\prod_{k=3}^{K-1} d_{k2})^{-1} - A_3(\prod_{k=4}^{K-1} d_{k3})^{-1} - ... - A_{K-2}(\prod_{k=K-1}^{K-1} d_{kK-2})^{-1},$$

$$A_K = (\prod_{k=3}^{K} d_{k1})^{-1} - (\prod_{k=3}^{K} d_{k2})^{-1} - A_3(\prod_{k=4}^{K} d_{k3})^{-1} -$$

$$-... - A_{K-2}(\prod_{k=K-1}^{K} d_{kK-2})^{-1} - \frac{A_{K-1}}{d_{KK-1}}.$$

Тогда подставим в уравнение (4) выражение для характеристической функции локального r-потока (6) и получим дифференциальное уравнение для характе-

ристической функции исследуемого процесса $\{n(t)\}$

$$\frac{\partial H(u,t)}{H(u,t)} = \lambda\{\frac{\phi_1(u)}{\mu_1}(1-e^{-\mu_1(T-t)}) + e^{-\mu_1(T-t)}\psi_1(u) - \frac{\phi_2(u)}{\mu_2}(1-e^{-\mu_2(T-t)}) - \\ -\frac{e^{-\mu_2(T-t)}}{d_{21}}\psi_2(u) - ... - \frac{\phi_{K-1}(u)}{\mu_{K-1}}(1-e^{-\mu_{K-1}(T-t)}) - \frac{e^{-\mu_{K-1}(T-t)}}{d_{21}}\psi_{K-1}(u) - \\ -\frac{\phi_K(u)}{\mu_K}(1-e^{-\mu_K(1-r_K e^{ju})(T-t)}) - \frac{e^{-\mu_K(1-r_K e^{ju})(T-t)}}{d_{21}}\psi_K(u)\}dt. \quad (7)$$

Проинтегрируем уравнение (7), подставим начальное условие (5) и получим вид характеристической функции для суммарного числа всех повторных обращений заявок за интервал времени $[0,T]$

$$H(u,T) = exp[\lambda\{-T - (\frac{\psi_1(u)}{\mu_1} - \frac{\phi_1(u)}{\mu_1^2})(1-e^{-\mu_1 T}) + \\ +(\frac{\psi_2(u)}{\mu_2 d_{21}} - \frac{\phi_2(u)}{\mu_2^2})(1-e^{-\mu_2 T}) + ... + (\frac{\psi_{K-1}(u)}{\mu_{K-1}d_{21}} - \frac{\phi_{K-1}(u)}{\mu_{K-1}^2})(1-e^{-\mu_{K-1}T}) + \\ +(\frac{\psi_K(u)}{\mu_K(1-r_K e^{ju})d_{21}} - \frac{\phi_{K-1}(u)}{\mu_K^2(1-r_K e^{ju})^2})(1-e^{-\mu_K(1-r_K eju)T}) + \\ +T(-\frac{\phi_1(u)}{\mu_1} + \frac{\phi_2(u)}{\mu_2} + ... + \frac{\phi_{K-1}(u)}{\mu_{K-1}} + \frac{\phi_K(u)}{\mu_K(1-r_K e^{ju})})\}]. \quad (8)$$

На рисунке 2 приведен результат численного алгоритма, реализованного в среде MathCAD, для построения распределения вероятностей числа событий в исследуемом r-потоке со следующими параметрами системы:λ=10, T=1, μ_1=1, r_1=0.8, μ_2=2, r_2=0.65, μ_3=3, r_3=0.4, μ_4=4, r_4=0.3, μ_5=5, r_5=0.1.

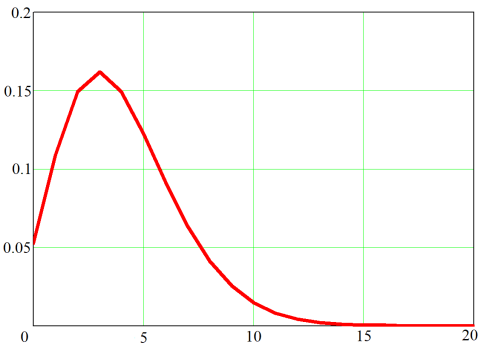


Рис. 2. Распределение вероятностей числа событий в потоке повторных обращений

4. Заключение

В данной работе был получен вид характеристической функции для суммарного потока повторных обращений в многофазной системе массового обслуживания методом марковского суммирования. С использованием обратного преобразования Фурье построено распределение вероятностей суммарного числа событий в потоке повторных обращений с помощью среды MathCAD. Полученное выражение может быть использовано для анализа вероятностных характеристик дополнительной нагрузки, возникающей в системах при повторных обращениях заявок для обслуживания.

ЛИТЕРАТУРА

1. Моисеева С. П. Исследование потока повторных обращений в бесконечнолинейной СМО с повторным обслуживанием / А. С. Морозова, С. П. Моисеева // Вестник Том-ского государственного университета. – 2005. – № 287. – С. 46–51.
2. Назаров А. А., Моисеева С. П., Морозова А. С. Исследование СМО с повторным об-ращением и неограниченным числом обслуживающих приборов методом предельной декомпозиции // Вычислительные технологии. 2008. Т. 13, спец. вып. 5. С. 88–92.
3. Nazarov A., Dammer D. Methods of limiting decomposition and Markovian summation in queueing system with infinite number of servers. In: A. Dudin, A. Nazarov, A. Moiseev, eds. Information Technologies and Mathematical Modelling. Queueing Theory and Applic-tions. ITMM 2018, WRQ 2018. Communications in Computer and Information Science, vol. 912. Springer, Cham, 2018. pp. 71–82.
4. Шкленник М. А. Метод марковского суммирования для исследования потока повтор-ных обращений в двухфазных системах M/GI/ → GI/ / М. А. Шкленник, А. Н. Мо-исеев, // Известия Саратовского университета. Новая серия. Серия: Математика. Механика. Информатика. 2021. – Т. 21, № 1. С. 125–137.
5. Подгайнов А. В. Исследование локального потока повторных обращений в многофазной системе массового обслуживания с неограниченным числом приборов / А. В. Подгайнов, М. А. Шкленник // Системы управления, информационные технологии и математическое моделирование : материалы V Всерос. науч.-практ. конф. с междунар. участием (Россия, Омск, 25–26 апреля 2023 г.) / Минобрнауки России, Ом. гос. техн. ун-т, каф. ММиИТЭ ; отв. ред. В. А. Бадрызлов. – Омск : Изд-во ОмГТУ, 2023. - С. 316–323.

УДК: 681.3.06 (075.32)

Вероятностное моделирование сопровождаемого цифрового двойника фрагментов магистральной трубопроводной сети для упреждающего противодействия природным угрозам

А.А. Нистратов[1]

[1]Федеральное государственное учреждение Федеральный исследовательский центр "Информатика и управление" Российской академии наук, ул. Вавилова, д. 44, корп. 2., Москва, Российская Федерация

andrey.nistratov@gmail.com

Аннотация

Цифровые двойники все шире охватывают характеристики различных объектов, систем и процессов. Сопровождение цифровых двойников во времени позволяет использовать накапливаемые исходные данные и существующие математические модели для упреждающего противодействия разнородным угрозам. Цифровые двойники фрагментов магистральных трубопроводных сетей сами по сути представляют собой распределенные компьютерные системы, подлежащие прагматичному использованию в интересах бизнеса. Исходя из этого в работе рассмотрены прикладные вопросы вероятностного моделирования сопровождаемого цифрового двойника фрагментов магистральной трубопроводной сети для упреждающего противодействия природным угрозам. Предложены вероятностные модели, методы их применения и интерпретации получаемых результатов моделирования.

Ключевые слова: модель, риск, сеть, система, цифровой двойник

1. Введение

Под цифровым двойником фрагмента магистральной трубопроводной сети, предназначенной для транспортировки на большие расстояния газа, воды, нефти, нефте- или иной продукции, понимается виртуальная модель этой сети, воспроизводящая в цифровом виде форму этого фрагмента в среде эксплуатации и хранимая в компьютерной системе. Так, цифровой двойник описывает: характеристики фрагмента трубы (диаметр, толщину, проектное давление, покрытие, внутритрубное устройство и др.), проектную и рабочую документацию на строительство трубопроводной сети с привязкой ко времени, характеристики среды

эксплуатации (месторасположение, характеристики местности, например – болото, переходы через водные преграды, автомобильные и железнодорожные пути и др.). Под сопровождением цифрового двойника понимается актуализация реального состояния эксплуатируемых фрагментов магистральной трубопроводной сети для использования обновляемых цифровых данных с целью моделирования различных процессов и вероятностного прогнозирования рисков. В свою очередь используемые вероятностные модели позволяют в упреждающем режиме по единой вероятностной шкале количественно спрогнозировать и затем сравнить эффективность упреждающего противодействия природным угрозам. Например, с использованием специальных технических требований по проведению внутритрубного диагностирования, по анализу эффективности катодной защиты и иных мер противодействия природным угрозам, в т.ч. коррозионной агрессивности грунтов (воздействие которых описывается с помощью временных характеристик). Таким образом цифровые двойники фрагментов магистральных трубопроводных сетей по сути сами представляют собой распределенные компьютерные системы, подлежащие прагматичному использованию в интересах бизнеса.

2. Пример вероятностного моделирования цифрового двойника

Предлагаемые вероятностные модели отражены в авторских работах [1, 2, 3], а также см. ГОСТ Р 59991–2022 «Системная инженерия. Системный анализ процесса управления рисками для системы», в котором эти модели реализованы.

Необходимыми исходными данными для моделирования являются:

- логическая структура для анализа (выделяются критичные фрагменты);

- по каждому составному фрагменту (в общем случае): частота возникновения угроз; среднее время развития угроз; период между диагностиками; длительность диагностики; среднее время восстановления целостности.

Положим, по результатам внутритрубного диагностирования выделены критичные фрагменты: на 1-м и 4-м фрагментах обнаружена зона продольных трещин, определен ремонт путем замены трубы; на 2-м и 3-м фрагментах обнаружена язвенная коррозия, определен ремонт заменой катушки; на 5-м и 6-м фрагментах, располагаемых в болотистой местности, выявлены продольные канавки и обширная коррозия с эквивалентом потери металла до 30%; на 7-м фрагменте выявлен корррозионный износ глубиной более 10%. Эти данные учтены при определении частота возникновения и среднего времени развития угроз.

Тем самым для моделирования сформирована логическая структура сопровождаемого цифрового двойника 7 фрагментов магистральной трубопроводной сети в виде последовательно объединяемых 7 элементов исследуемой системы. Интерпретация такова – вся трубопроводная сеть из 7 перечисленных фрагмен-

тов считается находящейся в состоянии целостности в течение заданного периода прогноза, если каждый из составных фрагментов в течение этого периода прогноза находится в состоянии целостности.

Исходя из производственных возможностей для всех фрагментов период между диагностиками равен 4 годам, длительность диагностики – 1 неделя, среднее время восстановления целостности – 1 месяц. Различающиеся для прогнозирования исходные данные по каждому из 7 элементов, определенные с учетом природных особенностей месторасположения фрагментов, сведены в Таблицу 1.

Параметр	По фрагментам 1, 7	По фрагментам 2, 3	По фрагментам 4, 6	По фрагменту 5
Частота возникновения угроз	1 раз в 15 лет	1 раз в 8 лет	1 раз в 5 лет	1 раз в 5 лет
Среднее время развития угроз	5 лет	4 года	4 года	3 года

Таблица 1. Исходные данные для моделирования

Этих исходных данных достаточно для моделирования.

Главный прогноз делается на 5 лет, полагая, что после каждой диагностики должны приниматься принципиальные решения по восстановлению требуемого уровня безопасности трубопроводной сети в условиях природных угроз. При этом оценивается интегральный риск нарушения целостности в диапазоне -50%+100% от задаваемых исходных данных. Вспомогательный прогноз для сравнения – на 2 года.

Допустимый уровень риска согласно требованиям ГОСТ Р 55999-2014, ГОСТ Р 59991-2022 полагается не выше 0.1, что соответствует вероятности успешного функционирования трубопроводной сети не ниже 0.9.

Результаты моделирования на уровне зависимости функции распределения времени нарушения целостности сопровождаемого цифрового двойника фрагментов магистральной трубопроводной сети показали следующее.

1. Риск нарушения целостности для всей сети из 7 критичных фрагментов в течение 5 лет составит 0.77. Это означает, что вероятность успешного функционирования сети в течение 5 лет (0.23) более, чем в 3.3 раза ниже, чем риск реального нарушения на каком-либо из фрагментов.

2. Зависимость интегрального риска от периода прогноза приведен на рис. 1. Анализ зависимости показывает, что лишь для прогнозного периода 1 год интегральный риск составит около 0.1.

3. Анализ обобщенных результатов моделирования при прогнозе на 5 лет

позволил сделать следующие выводы: к зоне допустимого риска относятся фрагменты 1, 7; к зоне недопустимого риска относятся вся сеть в целом и фрагменты 2-6; наивысший риск, равный 0.3, относится к фрагменту 5, этот риск на 20% выше риска для фрагментов 4, 6. Это объясняется меньшим временем активизации угроз из-за корррозионного износа и корррозионно-агрессивных условий ее расположения (3 года вместо 4-х лет для соседних труб).

4. Анализ обобщенных результатов моделирования при прогнозе на 2 года позволил установить: к зоне допустимого риска (не выше 0.1) относятся все фрагменты; к зоне недопустимого риска относится вся сеть в целом (риск=0.33); наивысший риск, равный 0.09, по-прежнему относится к фрагменту 5. При этом приблизительное среднее время наработки на нарушение целостности для фрагмента 5 составит 13.08 года.

при прогнозе на 5 лет

при прогнозе на 2 года

Рис. 1. Зависимость интегрального риска нарушения целостности от периода прогноза, изменяемого в диапазоне от 1 до 10 лет

Детальный анализ чувствительности интегрального риска к изменению исходных характеристик фрагмента 5, использованных при моделировании, можно проследить по зависимостям, отраженным на рис. 2-3.

Анализ детальных результатов прогнозных расчетов показывает обоснованность следующих рекомендаций в области противодействия угрозам, в т.ч. в условиях коррозионной агрессивности грунтов.

Чтобы не превышать риск 0.1 (т.е. обеспечивать успешность эксплуатации фрагмента трубопровода с вероятностью выше 0.9), необходимо:

- после 20 лет эксплуатации, при выявлении аномалий и эксплуатации в

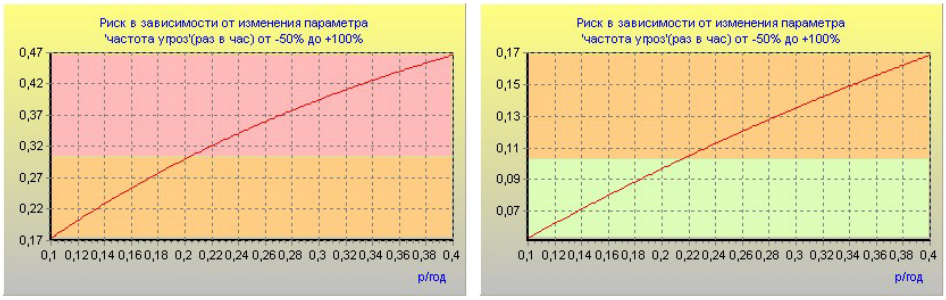


Рис. 2. Зависимость риска нарушения целостности от характеристики 'частота возникновения угроз' при прогнозе на 5 лет (слева) и 2 года (справа)

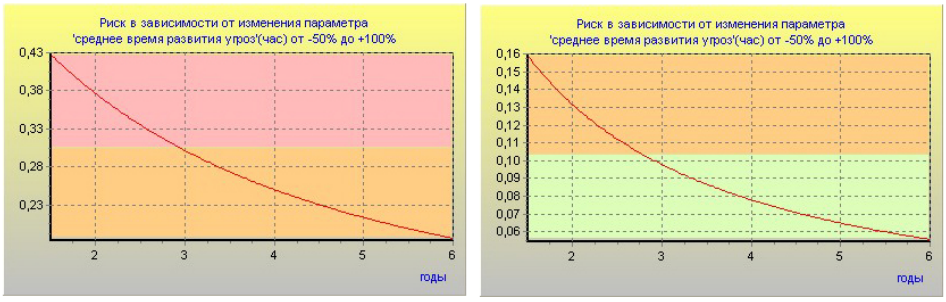


Рис. 3. Зависимость риска нарушения целостности от характеристики 'среднее время развития угроз', при прогнозе на 5 лет (слева) и 2 года (справа)

каррозионно-агрессивных условиях осуществлять внутритрубное диагностирование необходимо не через 4 года, а каждые 2 года;

- для ликвидации аномалий необходимо применять такие меры, которые гарантированно обеспечивают противодействие негативным природным воздействиям на срок не менее 3-х лет;

- для поддержки принятия управленческих решений вероятностные прогнозы осуществлять на срок, соизмеримый не только с долгосрочными планами (5-10 лет), но и со среднесрочными планами (2-4 года), а при выявлении для этих прогнозных сроков рисков, количественно превышающих допустимый уровень, осуществлять вероятностное прогнозирование рисков на период до 1 года для текущего планирования и упреждающего противодействия угрозам.

Эти рекомендации, полученные в результате вероятностного моделирования сопровождаемого цифрового двойника фрагментов магистральной трубопроводной сети, служат дополнением к техническим мерам, востребуемым по итогам регулярного внутритрубного диагностирования реальных сетей.

3. Заключение

В результате вероятностного моделирования сопровождаемого цифрового двойника фрагментов магистральной трубопроводной сети в настоящей работе предложены:

- вероятностные модели и методы прогнозирования и управления рисками, применимые для сложных логических структур и характеризуемые общими исходными данными по каждому из составных фрагментов: частотой возникновения угроз, средним временем развития угроз, периодом между диагностиками, длительностью диагностики, средним временем восстановления целостности;

- аналитические результаты вероятностного моделирования на уровне зависимостей функции распределения времени нарушения целостности сопровождаемого цифрового двойника фрагментов магистральной трубопроводной сети от учитываемых факторов, вплоть до характеристик конкретного фрагмента сети;

- интерпретации получаемых результатов вероятностного моделирования, системно дополняющие технические меры, востребуемые по итогам регулярного внутритрубного диагностирования реальных сетей.

Предложенные в работе вероятностные модели, методы их применения и интерпретации получаемых результатов моделирования в приложении к сопровождаемым цифровым двойникам фрагментов магистральной трубопроводной сети обладают аналитической новизной. Их применение обеспечивает прослеживаемость прогнозных рисков от влияющих факторов. Это предоставляет возможности для системного дополнения технических мер, востребуемых по итогам регулярного внутритрубного диагностирования, и способствует повышению безопасности эксплуатации реальных сетей.

ЛИТЕРАТУРА

1. Kostogryzov A., Nistratov G., Nistratov A. Some Applicable Methods to Analyze and Optimize System Processes in Quality Management. Total Quality Management and Six Sigma, InTech, 2012, pp. 127-196, https://www.intechopen.com/chapters/38088
2. Vsevolod Kershenbaum, Leonid Grigoriev, Petr Kanygin and Andrey Nistratov / Probabilistic modeling in system engineering. Probabilistic modeling processes for oil and gas systems. IntechOpen, 2018, P. 55-79. http://dx.doi.org/10.5772/intechopen.74963
3. Нистратов А.А. Аналитическое прогнозирование интегрального риска нарушения приемлемого выполнения совокупности стандартных процессов в жизненном цикле систем высокой доступности. Часть 1. Математические модели и методы // Системы высокой доступности. 2021. Т.17 №3, с. 16-31,

Часть 2. Программно-технологические решения. Примеры применения // Системы высокой доступности. 2022. Т.18 №2, с. 42-57

УДК: 681.3.06 (075.32)

Об архитектурных решениях, ориентированных на прогнозирование и рациональное управление рисками в системной инженерии

А.А. Нистратов[1]

[1]Федеральное государственное учреждение Федеральный исследовательский центр "Информатика и управление" Российской академии наук, ул. Вавилова, д. 44, корп. 2., Москва, Российская Федерация

andrey.nistratov@gmail.com

Аннотация

Перспективная системная инженерия охватывает широкий спектр областей функционального применения систем, ориентирована на компьютерные системы и сети, становящиеся более разумными, самоорганизующимися, ресурсоэффективными и безопасными, устойчивыми в эксплуатации и реагирующими на постоянно растущий и разнообразный спектр общественных потребностей. Учитывая это, в работе рассмотрены концептуальные вопросы архитектурного представления прикладных решений, ориентированных на прогнозирование и рациональное управление рисками в жизненном цикле систем. В качестве рассматриваемых систем могут выступать: государство, регионы, органы управления, инфраструктура (инженерная, транспортная, энергетическая, коммуникационная), предприятия, машины и механизмы, производимая продукция, процессы и пр. Примерами типовых задач, подлежащих решению, являются: прогнозирование рисков; обоснование допустимых рисков; выявление существенных угроз; обоснованию мер, направленных на достижение целей и противодействие угрозам; определение сбалансированных решений при средне- и долгосрочном планировании; обоснование предложений по совершенствованию и развитию систем. Приведены варианты некоторых архитектурных решений.

Ключевые слова: архитектура, риск, система, системная инженерия, управление

1. Введение

Под системной инженерией согласно ISO/IEC/IEEE 15288 (в России ГОСТ Р 57193) понимается сосредоточение междисциплинарных научно-технических и организационных усилий, требуемых для преобразования ряда потребностей

заинтересованных сторон, ожиданий и ограничений в прикладные решения и для поддержки этих решений в жизненном цикле системы. Система определена как комбинация взаимодействующих элементов, упорядоченная для достижения одной или нескольких поставленных целей. Подразумеваются сложные системы, создаваемых человеком для любой области приложений: в интересах органов государственной власти и корпораций, энергетических, финансово-экономических, страховых и промышленных структур (включая отдельные предприятия, строительные, нефтегазовые и транспортные комплексы, объекты опасного производства), предприятия авиационно-космической отрасли, служб по чрезвычайным ситуациям, жилищно-коммунального хозяйства и др. В настоящей работе рассматриваются архитектурные решения, ориентированные на прогнозирование и рациональное управление рисками. Согласно ISO/IEC/IEEE 42010 (в России ГОСТ Р 57100) под архитектурой понимаются основные понятия или свойства системы в ее окружающей среде, воплощенные в элементах, отношениях и конкретных принципах ее проекта и развития. В свою очередь под риском понимается сочетание вероятности нанесения ущерба и тяжести этого ущерба.

Проведенный анализ (в т.ч. анализ взглядов Международного совета по системной инженерии - INCOSE) показал, что перспективная системная инженерия охватывает широкий спектр областей функционального применения систем, ориентирована на компьютерные системы и сети, становящиеся более разумными, самоорганизующимися, ресурсоэффективными и безопасными, устойчивыми в эксплуатации и реагирующими на постоянно растущий и разнообразный спектр общественных потребностей в жизненном цикле систем. Системная инженерия должна поддерживаться всеобъемлющей теоретической основой, методами и инструментариями исследований, основанными на моделях, позволяющих лучше понимать все более сложные системы и решения, принимаемые в условиях неопределенности с учетом долгосрочных рисков. Системы должны создаваться с использованием эффективных инструментариев, реализующих инновации для поддержания необходимой конкурентоспособности.

Учитывая перспективы развития системной инженерии в условиях разнородных неопределенностей для систем различного функционального назначения в настоящей работе рассмотрены концептуальные вопросы архитектурного представления прикладных решений, ориентированных на прогнозирование и рациональное управление рисками [1–4].

2. Предлагаемые архитектурные решения

Предлагаемый вариант архитектурного представления определенных системных интересов высшего мета-уровня для решения задач системной инженерии представлен на рис. 1. В общем случае для достижения конкретных целей си-

стемы и выполнения задаваемых требований заинтересованных сторон с учетом разнородных угроз, различных ограничений и условий решению подлежат задачи, связанные с анализом рисков. Примерами типовых задач, подлежащих решению, являются: прогнозирование рисков; обоснование допустимых рисков; выявление существенных угроз; обоснованию управляющих мер, направленных на достижение целей и противодействие угрозам; определение сбалансированных решений при средне- и долгосрочном планировании; обоснование предложений по совершенствованию и развитию системы. Исходя из приведенного на рис. 1 архитектурного представления определенных системных интересов высшего мета-уровня ниже предлагается архитектурное представление системы для построения математических и программно-технологических инструментариев, ориентированных на достижение ожидаемых прагматических эффектов – см. рис. 2.

Рис. 1. Предлагаемый вариант архитектурного представления определенных системных интересов высшего мета-уровня

На уровне целевых аналитических потребностей пользователя определяются цели и назначение системы, выбираются задачи анализа, обоснования требований или оптимизации, подлежащие решению в жизненном цикле системы. При необходимости на уровне отдельного процесса дается описание выбранного процесса, реализуемого в системе и подлежащего анализу, выбирается модель для прогнозирования рисков, формируются необходимые исходные данные для моделирования. В качестве исходных данных может выступать не только статистика, но и данные, формируемые в режиме реального времени, например, в системе дистанционного контроля промышленной безопасности (для чего используются дополнительные специальные средства обработки телеметрической информации) – по ГОСТ Р 58494. На уровне интеграции процессов, которые должны быть учтены при прогнозировании интегральных рисков, дается описание выбранной совокупности реализуемых процессов и осуществляется задуманное моделирование. В частном случае на этом уровне может быть рассмотрен один единственный процесс. На уровне расчетных показателей отображаются частные и интегральные показатели рисков в их зависимости от исходных данных, дается интерпретация вероятностных расчетов применительно к целям системы и решаемым задачам.

При построении программных инструментариев в качестве математических моделей выбраны авторские модели [1–4], доведенные до реализации и рекомендуемые к использованию в стандартах системной инженерии ГОСТ Р 59347, ГОСТ Р 59989-ГОСТ Р 59991, ГОСТ Р ГОСТ Р 58494 и др. В итоге моделирования получаемые рекомендации предназначены для использования лицами, принимающими решение, в целях выработки рациональных упреждающих мер противодействия угрозам и достижения прагматических эффектов. Пример (вариант) архитектурного решения по формированию структуры сложной системы для моделирования с привязкой к технологической схеме обогатительной фабрики представлен на рис. 3. Нумерация в архитектурном представлении означает номер подсистемы в формальной структуре моделируемой системе для последовательно-параллельного объединения. При прогнозировании рисков в терминах функций распределения используются элементы логики «И», «ИЛИ». Значения о состоянии анализируемых объектов могут быть взяты из статистики, а также могут поступать от систем дистанционного контроля, автоматизированных систем управления, датчиков, сенсоров.

Предложенные архитектурные решения охватывают системы любой области приложения. Их применение обеспечивает аналитическую прослеживаемость интегрального и частных рисков от влияющих факторов. Это предоставляет возможности для поиска эффективных решений в системной инженерии с исполь-

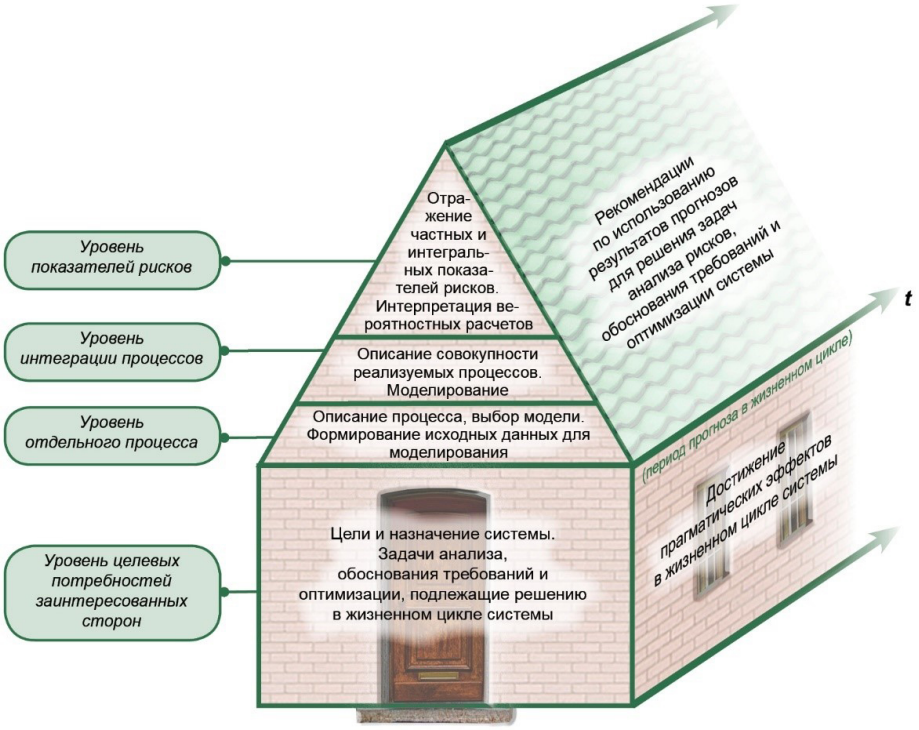


Рис. 2. Архитектурное представление для построения математических и программно-технологических инструментариев, ориентированных на достижение эффектов

зованием непрерывного аналитического прогнозирования рисков и обоснования способов их снижения или удержания в допустимых пределах [3, 4].

3. Заключение

Учитывая перспективы развития системной инженерии в условиях разнородных неопределенностей для систем различного функционального назначения в настоящей работе предложены:

- вариант архитектурного представления определенных системных интересов высшего мета-уровня для решения задач системной инженерии с применением методов прогнозирования и управления рисками;

- архитектурное представление системы для построения математических и программно-технологических инструментариев, ориентированных на достижение прагматических эффектов;

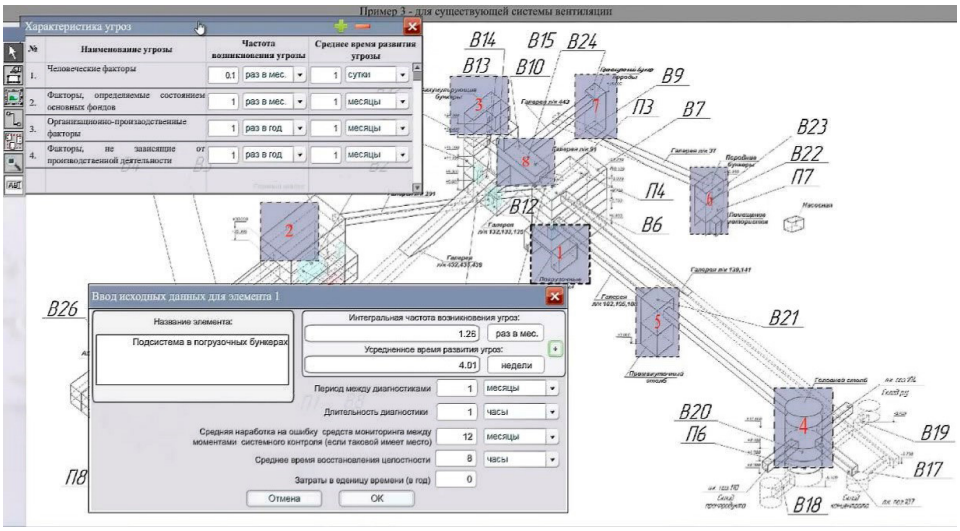


Рис. 3. Пример архитектурного решения по формированию структуры сложной системы для моделирования с привязкой к фотографии, скану или иному изображению бумажного документа (исходные данные приведены для элемента 1 с привязкой к технологической схеме обогатительной фабрики)

- вариант архитектурного решения по формированию структуры сложной системы для моделирования с привязкой к фотографии, скану или иному изображению бумажного документа.

Охватывая системы любой области приложения, предложенные архитектурные решения обладают аналитической новизной, за счет чего их применение обеспечивает прослеживаемость прогнозных рисков от влияющих факторов. Это предоставляет возможности для поиска эффективных решений в системной инженерии.

ЛИТЕРАТУРА

1. Kostogryzov A., Nistratov G., Nistratov A. Some Applicable Methods to Analyze and Optimize System Processes in Quality Management. Total Quality Management and Six Sigma, InTech, 2012, pp. 127-196.
2. Vsevolod Kershenbaum, Leonid Grigoriev, Petr Kanygin and Andrey Nistratov / Probabilistic modeling in system engineering. Probabilistic modeling processes for oil and gas systems. IntechOpen, 2018, P. 55-79. http://dx.doi.org/10.5772/intechopen.74963

3. Нистратов А.А. Аналитическое прогнозирование интегрального риска нарушения приемлемого выполнения совокупности стандартных процессов в жизненном цикле систем высокой доступности. Часть 1. Математические модели и методы // Системы высокой доступности. 2021. Т.17 №3, с. 16-31, Часть 2. Программно-технологические решения. Примеры применения // Системы высокой доступности. 2022. Т.18 №2, с. 42-57
4. Нистратов А.А. О математических, программно-технологических и методических решениях, ориентированных на рациональное управление рисками в системной инженерии. Сборник материалов Всероссийской научно-практической конференции «Россия в XXI веке в условиях глобальных вызовов: проблемы управления рисками и обеспечения безопасности социально-экономических и социально-политических систем и природно-техногенных комплексов», 26-27.04.2022, Президиум РАН. Под общ. ред. Проф. Я.Д. Вишнякова. – М.: Государственный университет управления. 2022. С. 251-255

UDC: 519.218

The regenerative stability analysis of some vacation models

S.S. Rogozin[1,2]

[1]Institute of Applied Mathematical Research Karelian Research Centre RAS, Petrozavodsk, Russia

[2]Petrozavodsk State University, Petrozavodsk, Russia

Abstract

In this research we consider two different vacation models which have been proposed and studied in previous works. First, we consider the stability of a multi-server vacation model with the *self-sustained servers* and some general distribution of the service time. Also we consider a single-server vacation model, in which there are consecutive periods of the attachment (glue), service and vacation. We consider the stability conditions of these systems obtained in previous works by various methods. The purpose of this note is to show how the regenerative methodology allows to deduce these stability conditions of these systems by a unified way.

Keywords: Regenerative method, stability analysis, vacation model, self-sustained servers, glue period systems

1. Introduction

Queueing systems with server vacations have a lot of applications in the real-world systems and telecommunication systems. The main goal of the vacation is to minimize the cost of the resources spend to maintain servers in the working state. Some literature of vacation queueing models can be found in the work [3].

In the present paper, we study the stability of a multi-server vacation model with Poisson input and a general distribution of service time. We assume that servers can go on vacation after serving a customer, and it happens independently on the queue size. Also if the queue has a medium size then the server interrupts vacation after exponential time. Moreover if the queue is large then the server manager starts to search of a volunteer server who agrees to interrupt vacation with some probability. Such a vacation model in Markovian setting has been studied in the paper [4] (A.Dudin et al, 2021). A distinctive feature of our research is to consider general service time and prove that in this case the stability condition remains the

same. This system, for example, can be motivated by a delivering problem, when company hires some pool of independent individual vehicle carriers (freelancers) having their own truck for the delivering the products (see [4]).

Also we consider another single-server vacation model in which the customers arrive according with Poisson input and the service time has a general distribution. From the beginning, arriving customers enter the queue while the server remains to be 'frozen' (not working). After a deterministic time G, the server switches on. In other words, all accumulated customers are served and then the server switches off for a random time I having a general distribution, and so on. Such a vacation model has been studied in the work [1]. To reprove the stability condition of this model, we again propose a proof based on the regenerative method. Such a system with glue periods is motivated by the performance modelling and analysis of optical networks.

Now we briefly describe the basic idea of the regenerative method we have used to analyse the stability of the systems. First we note that regenerative method allows us to establish stability condition of various non-Markovian models. In a regenerative model, the regeneration cycle is the length of random interval between arrivals in an empty system. When the system becomes idle, then its future behavior turns out to be independent of the past. This fact implies that the regeneration cycles are independent identically distributed (iid) random variables, and it allows to analyse the *positive recurrence* (stability) of the system by a classic technique. The positive recurrence means that the mean regeneration cycle is finite. For more detail on the theory of regenerative processes see [8, 7, 2, 5, 6].

2. A vacation multi-server model with self-sustained servers

Now we give a detailed description of the studying model. First we suppose that customers are arrived in an infinite capacity queue according to Poisson input process with rate λ. We assume that there are N identical servers, and let $\{S_n\}$ be the service time of the n-th arriving customer. Then we denote by S the generic service time and denote by $\mu = 1/\mathsf{E}S$. Now we suppose that after each service (at the departure instant) a server goes on vacation with the probability $1 - p$ and with probability p it tries to serve another customer in the queue. If there are no more customers in the queue then the server go in vacation. When the server goes to the vacation state, it tries to return in the working state with exponential time with rate γ, however it occurs, only if the queue size exceeds some threshold $J_1 > 0$. Otherwise it remains be in vacation and continues trying to start service. Then we assume that if the queue size exceeds some threshold $J_2 > J_1$ then the server manager starts to search of a volunteer server who agrees to interrupt vacation. The search time has exponential distribution with rate β. When one volunteer is found, then the search is stopped. A volunteer server agrees to interrupt vacation with the probability $1 - q$

(and declines the offer with probability q). In regeneration analysis the stability of the system is equivalently to the positive recurrence of the system, means that the mean generic regeneration cycle (denoted by T) is finite, that is $\mathsf{E}T < \infty$. Now we denote by P_i the steady-state probabilities that i servers are not in vacation. Then the following statement holds.

Theorem 1. If condition

$$\lambda < \sum_{i=1}^{N} P_i[ip\mu + (N-i)\gamma + \beta(1-q^{N-i})], \tag{1}$$

holds, then the initially idle system is positive recurrent, $\mathsf{E}T < \infty$.

In this work we establish this stability condition by means of the regenerative approach which also allow to obtain a nice intuitive interpretation. We recall that λ is the rate of customers arriving to the system. Actually the right-hand side of (1) is the rate of customers' departure from the buffer provided the system is overloaded. It is clear that, under the condition that i servers are busy, the rate of customers departure from the queue consists from the following three terms. First is the rate $ip\mu$ of service completion of customers, after which server does not take a vacation. Then we add the rate $(N-i)\gamma$ of the end of vacation and starting service by one of the $N-i$ vacated servers. And finally we add the rate $\beta(1-q^{N-i})$ of agreement of one of $N-i$ vacated servers to interrupt vacation and start service. Therefore (1) reflects the evident condition of stability of the system: when the system is overloaded, customers arriving rate must be less than the average departure rate from the buffer.

3. A system of type saturation-service-vacation

In this section we consider another single-server vacation model for which we obtain stability condition by the regeneration method. We assume that customers are arrived according to Poisson input process with rate λ. We also denote by $\{S_n\}$ the service time of the n-th arriving customer. Then we denote by S the generic service time and denote by $\mu = 1/\mathsf{E}S$. We consider a system with one server and one class of customers. The system has one of the following states. At the beginning, all arriving customers enter the queue (of infinite capacity), but the server is not working (attachment or glue period). After a deterministic value G, the server switches on and starts to service customers from the queue (service period). When all customers from the queue have been served, the server switches off for a random time I with some general distribution (vacation period). While the server is vacated, all arriving customers go into orbit, and each customer tries to capture the server after an exponentially distributed time ν. In other words all orbital

customers make attempts independently. We note that is the so-called *classic retrial policy*. When the vacation period ends, the server does not immediately start to service arriving customers, but again waits during the time G (glue period), and so on. Now we denote by $\rho = \lambda/\mu$. Then we outline the proof of the following statement.

Theorem 2. If condition

$$\rho < 1 \tag{2}$$

holds, then the initially idle system is positive recurrent.

The main idea of the proof of this statement is as follows. First, in interval $[0, t]$ we denote $V(t)$ the received work, that customers bring in the system, $I(t)$ the total vacation time and $G(t)$ the total glue period time. Also we denote by $W(t)$ the remaining workload at instant t. Then we construct the following balance equation connecting the arrived work $V(t)$ the remaining work $W(t)$, and also $I(t)$ and $G(t)$:

$$W(t) = V(t) - t + I(t) + G(t), \quad t \geq 0. \tag{3}$$

We denote by $Q(t)$ the orbit size. Then we assume that under condition (2) the system is overloaded, that is

$$Q(t) \Rightarrow \infty, \ \ t \to \infty. \tag{4}$$

We note that by SLLN w.p.1

$$\lim_{t\to\infty} \frac{V(t)}{t} = \rho. \tag{5}$$

Then we show that, w.p.1,

$$\lim_{t\to\infty} \frac{I(t)}{t} = \lim_{t\to\infty} \frac{G(t)}{t} = 0. \tag{6}$$

This is the most challenging step but one can show it using a regenerative approach and renewal theory. By intuition it is clear, because, when the system is saturated, the infinity amount of customers will be attached to the queue during the glue period and then the service period will be infinitely large. Our analysis implies that the assumption (4) is false. In turn this allows to show, using a standard unloading procedure (see [5]), that a regeneration instant can be reached in a finite interval of time with a positive probability. Then it follows that under assumption $\rho < 1$, by a characterization of the limiting behaviour of the remaining regeneration time (time up to the next regeneration instant), indeed $\mathsf{E}T < \infty$ that is the system is positive recurrent.

4. Conclusion

We consider several queuing models and analyse them by the unique regenerative method. The first model we consider is a multi-server vacation model with self-sustained servers. The second model is a single-server vacation model in which there are consecutive periods of the attachment (glue), service and vacation. We apply the unique regenerative method to reprove the known stability conditions of these systems, when service time becomes general.

REFERENCES

1. Abidini M., Boxma O., Resing J., Analysis and optimization of vacation and polling models with retrials // Performance Evaluation. 98. 10.1016/j.peva.2016.02.001.
2. Asmussen S., Applied Probability and Queues, Wiley, N.Y., 1987.
3. Chakravarthy, S.R., A Comparative Study of Vacation Models Under Various Vacation Policies: A Simulation Approach // In Mathematical Modeling and Computation of Real-Time Problems; CRC Press: Boca Raton, FL, USA, 2021; pp. 3–20.
4. Dudin A., Dudina O., Dudin S., Samouylov K., Analysis of Multi-Server Queue with Self-Sustained Servers // Mathematics 2021, 9, 2134. https://doi.org/10.3390/math9172134
5. Morozov E., Steyaert B., Stability Analysis of Regenerative Queueing Models. Springer, 2021. https://link.springer.com/book/10.1007/978-3-030-82438-9
6. Morozov E. and Delgado R., Stability analysis of regenerative queues // Automation and Remote control, vol. 70, pp. 1977-1991, 2009.
7. Sigman K. and Wolff R. W., A review of regenerative processes // SIAM Review, vol. 35, pp. 269-288, 1993.
8. Smith W. L., Regenerative stochastic processes // Proceedings of Royal Society, Ser. A, vol. 232, pp. 6-31, 1955.

UDC: 004.94

Adaptive redistribution of heterogeneous traffic with acceptable delays with transmission replication during route reconfiguration in nodes connecting segments of multipath networks

V.A. Bogatyrev [1,3], A.V. Bogatyrev [2], S.V. Bogatyrev [2,3]

[1]Department Information Systems Security, Saint-Petersburg State University of Aerospace Instrumentation, Saint Petersburg, Russia
[2]Yadro Cloud Storage Development Center, Saint Petersburg, Russia
[3]ITMO University, Saint Petersburg, Russia

Abstract

The article explores the possibilities of improving the reliability and fault tolerance of multipath networks based on a simplified version of reconfiguration implemented by means of switching nodes connecting segments of several communication paths of request source nodes available in the network with addressed nodes.

Servers with their possible association into clusters act as addressees. In the considered simplified version, the reconfiguration is carried out only according to the information available in the group of switching nodes of the segment switches, each of which is connected to all or part of the available communication paths with the addressees. As information for reconfiguration, the length of the queues of packets for transmission over various segments or the waiting time in the specified queues, as well as the availability of segments for transmission, is used. If there is a group of switches, traffic redistribution is implemented when the relevant information is exchanged between them.

Keywords: adaptive redistribution, multipath, route reconfiguration, criticality of requests to delays.

1. Introduction

Distributed computer systems and networks, as the basis of the digital economy, are currently being intensively introduced into various areas of economics and management, which is accompanied by their constant structural and functional complication with increasing requirements for reliability, fault tolerance performance while reducing transmission delays. These requirements are especially acute for real-time industrial networks involved in the management of industrial facilities,

technological and other processes associated with environmental safety risks and significant economic losses [1-3]. Reliability and fault tolerance of computer systems and networks with low transmission delays is achieved by introducing redundancy with the possibility of reconfiguration in case of accumulation of failures and traffic changes [3-6] .

The construction of reconfigurable networks and systems requires the implementation of system state monitoring and is based on the optimization of the structure and redistribution of traffic in the conditions of its change and the accumulation of network failures. Reconfigurable networks involve the introduction of special nodes of reconfiguration controllers associated with network nodes through reconfiguration agents.

The article explores the possibilities of improving the reliability and fault tolerance of multipath networks based on a simplified version of reconfiguration implemented by means of switching nodes connecting segments of several communication paths of request source nodes available in the network with addressed nodes. Servers with their possible association into clusters act as addressees. In the considered simplified version, the reconfiguration is carried out only according to the information available in the group of switching nodes of the segment switches, each of which is connected to all or part of the available communication paths with the addressees.

Reconfiguration is implemented as a result of traffic redistribution, taking into account its changes and loss of connectivity along some path segments, as a result of the accumulation of failures of switching nodes. A path is a sequence of nodes used to transfer data from sources to destinations. A segment is understood as part of the path from the source to the switch, from the switch to the destination. Segments can link daisy-chained path switches as needed.

2. Variants of multipath networks with reconfiguration nodes

In the proposed simplified version, the reconfiguration is implemented based on the information available in the segment switch nodes, such information is the length of the packet queues for transmission over various segments or the waiting time in the specified queues, as well as the availability of segments for transmission. If there is a group of switches, traffic redistribution is implemented when the relevant information is exchanged between them.

Consider some configurations of multipath networks with nodes connecting segments of different paths (path reconfiguration nodes). Figure 1 shows variants of multipath networks in the presence of one group of request sources (I), one reconfiguration node (R) in the presence of one (a) and two clusters (F) (b) of the same functionality. Figure 2 shows options for two groups of sources, one cluster, one (a) and two reconfiguration nodes, unconnected (b) and interconnected (c). Figure 3

shows options for two groups of sources, two clusters, two reconfiguration nodes, unconnected (a) and interconnected (b, c). Figure 3c shows the case of complete connectivity of reconfiguration nodes and clusters. For the case according to Fig.1. 2, 3 a, b and fig. 3a, reconfiguration allows you to adapt to the accumulation of failure segments of communication with clusters and their constituent servers. The rest of the network structures allow, during reconfiguration as a result of traffic redistribution, to adapt to its changes from different sources of requests.

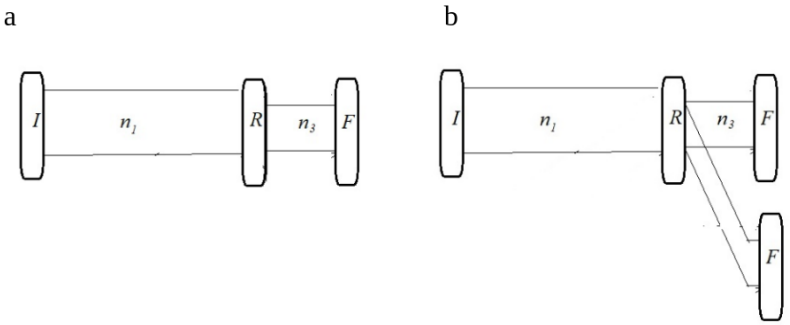


Fig. 1. Variants of multipath networks in the presence of one group of request sources, one reconfiguration node in the presence of one (a) and two clusters (b) of the same functionality

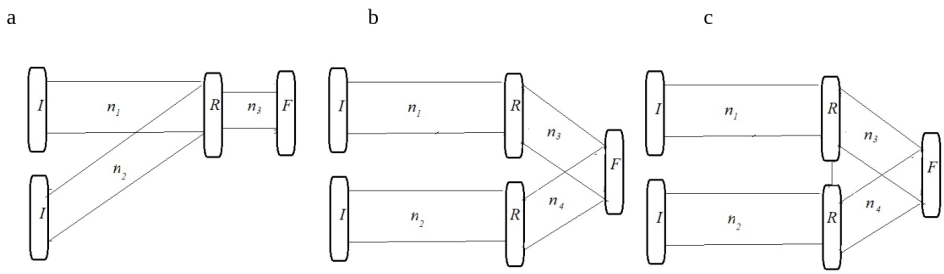


Fig. 2. Network options for two groups of sources, one single cluster, one (a) and two reconfiguration nodes, unconnected (b) and interconnected (c)

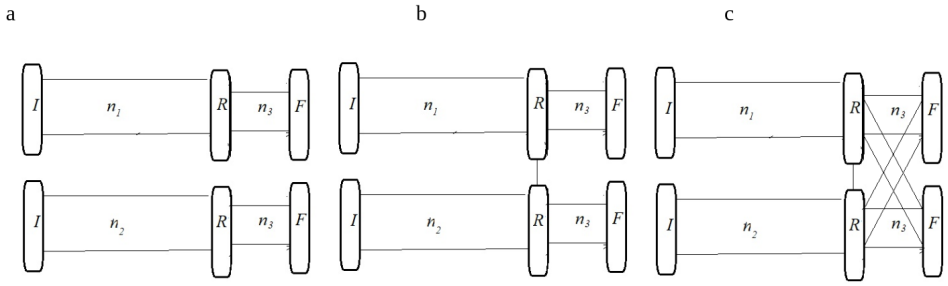


Fig. 3. Network options with two groups of sources, two clusters, two reconfiguration nodes, unconnected (a) and interconnected (b, c), option c corresponds to the complete connectivity of reconfiguration nodes and clusters

3. Multipath network model with traffic redistribution in reconfiguration nodes

When creating models, we will assume that the sources generate the simplest flow with two gradations of requests, one of which is more critical to network delays. For the first thread, the maximum allowable waiting time is t1, and for the second, $t2$. Let us represent the network nodes with the simplest models of the M/M/1 type.

We will carry out the calculation on the example of the structure according to fig. 2 in. when transmitting packets through the segments connecting the reconfiguration nodes with the cluster. We assume? that the source connected to the upper node of the reconfiguration is more critical to the delays of the transmitted packets. Let us determine the dependence of the timely transmission of the first and second streams on the share of the first stream redirected through the lower reconfiguration node.

The probability of not exceeding the allowable time for the first source of requests transmitted through the connection of the upper and lower segment switch with the cluster is defined as

$$P_{13} = 1 - \frac{\Lambda\alpha\beta v}{n3} e^{(\frac{\Lambda\alpha\beta}{n3} - v^{-1})(t_1 - \frac{v}{1-\frac{v\alpha}{n1}})}.$$

Where v is the average packet transmission time, α is the share of the first stream, and β is the share of packets of the first stream redistributed for transmission to the cluster through the second group of channels

$$P_{14} = 1 - \frac{\Lambda[\alpha(1-\beta)+(1-\alpha)]v}{n_4} e^{(\frac{\Lambda[\alpha(1-\beta)+(1-\alpha)]}{n_4} - v^{-1})(t_1 - \frac{v}{1-\frac{v\alpha}{n1}})}.$$

The average probability of timely transmission of packets from the first source through two groups of communication channels with the cluster is defined as

$$P_{134} = \beta P_{13} + (1 - \beta) P_{14}$$

The probability of timely transmission of packets from the second source through the second group of communication channels with the cluster is defined as

$$P_{24} = 1 - \frac{\Lambda[\alpha(1-\beta)+(1-\alpha)]v}{n_4} e^{\frac{\Lambda[\alpha(1-\beta)+(1-\alpha)]}{n_4} - v^{-1})(t_2 - \frac{v}{1-\frac{v\alpha}{n_2}})}$$

The probability of timely delivery of the first and second packets to the cluster is defined as

$$R = P_{134} P_{24}$$

The dependence of the probability of timely transmission from the first and second source on the proportion of packets from the first source redistributed during reconfiguration is shown in Fig. 4, in which curves 1-6 correspond to $\alpha = 0.2, 0.3, 0.4, 0.5, 0.6, 0.7$. The calculation was performed at $v = 0.01$ s, $t_1 = 0.01$ s, $t_2 = 0.5$ s, and $\Lambda = 80$ 1/s.

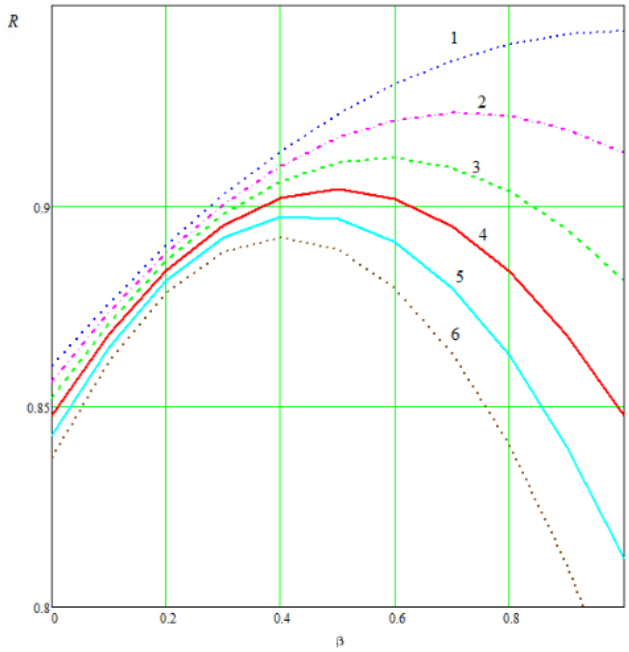


Fig. 4. Dependence of the probability of timely transmission from the first and second source on the proportion of first source packets redistributed during reconfiguration

The presented dependencies show the effectiveness of the implementation of reconfiguration in communication nodes that connect different segments of communication paths with recipients.

4. Conclusion

The article shows the possibilities of increasing the probability of timely delivery of traffic that is heterogeneous in terms of permissible delays as a result of reconfiguration and redistribution of traffic in communication nodes connecting all or part of the communication paths with the addressee.

REFERENCES

1. Malik V., Barde C.R. Live migration of virtual machines in cloud environment using prediction of CPU usage // International Journal of Computer Applications. 2015. V. 117 N 23. P. 1–5. doi: 10.5120/20691-3604

2. Tatarnikova T.M., Poymanova E.D.: Differentiated capacity extension method for system of data storage with multilevel structure. //Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 2020, 20(1), . 66–73 DOI: 10.17586/2226-1494-2020-20-1-66-73
3. Houankpo H.G., Kozyrev D.V., Nibasumb E. M. Mouale N.B. Mathematical Model for Reliability Analysis of a Heterogeneous Redundant Data Transmission System International Congress on Ultra Modern Telecommunications and Control Systems and Workshops 2020-October,9222431, . 189-194
4. Bogatyrev V.A. , Bogatyrev A.V. , Bogatyrev, S.V.: Reliability and probability of timely servicing in a cluster of heterogeneous flow of query functionality //Wave Electronics and its Application in Information and Telecommunication Systems (WECONF 2020) - 2020, pp. 9131165
5. Bogatyrev V.A. , Bogatyrev A.V. , Bogatyrev, S.V. : Redundant Servicing of a Flow of Heterogeneous Requests Critical to the Total Waiting Time During the Multi-path Passage of a Sequence of Info-Communication Nodes. Lecture Notes in Computer Science// Including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2020. Vol. 12563. pp. 100-112. DOI 10.1007/978-3-030-66471-8 9

УДК: 519.2

Об экстремальном индексе стационарного времени ожидания в системах M/G/1 с неоднородными входными потоками

И.В. Пешкова[1,2]

[1]Петрозаводский государсьтвенный университет, пр. Ленина, 33, Петрозаводск, Россия

[2]Институт прикладных математических исследований Карельского научного центра РАН, ул. Пушкинская, 11, Петрозаводск, РОссия

iaminova@petrsu.ru

Аннотация

Для исходной системы $M/G/1$ с неоднородным входным потоком (с вероятностью p заявка поступает из первого потока, с веротяностью $1-p$ заявка поступает из второго потока) предлагается построить мажорантную и миноратную системы, в которых времена обслуживания идентичны, а входные потоки соответствуют одному из входных потоков исходной системы. Сформулировано свойство монотонности экстремального индекса стационарного времени ожидания в трех системах в случае, когда входные потоки связаны отношением порядка (упорядочены по интенсивности отказов). В статье приведены результаты численного моделирования для случая гиперэкспоненциального распределения входного потока.

Ключевые слова: неоднородный входной поток, стационарное время ожидания, экстремальный индекс

1. Введение

Изучение экстремльных значений характеристик производительности коммуникационных систем иммет важное значение в связи с увеличением их сложности. Превышение высоких пороговых значений может означать отказы технических устройств из-за сбоев оборудования, приводить к потерям данных при передаче. Причем большие значения, как правило, появляюится не по одному, а группами, образуя так называемые кластеры [1]. Одним из показателей, характеризующих такую кластеризацию, является *экстремальный индекс* [2]. Он определяет предельное распределение экстремальных значений стационарных случайных последовательностей [3] и отражает кластерную структуру базовой последовательности или ее локальную зависимость. Одна из наиболее практически важных

интерпретаций экстремального индекса состоит в том, что его значение обратно пропорциаонально среднему размеру кластера. Для оценки экстремального индекса существуют различные способы, например блочный метод или метод регенеративного моделирования.

2. Монотонность экстремального индекса

В качестве исходной системы Σ рассмотрим систему обслуживания типа $GI/G/1/\infty$ с одним сервером и неограниченным буфером. Дисциплина обслуживания – первым пришел – первым обслужен. Предположим, что входной поток в систему состоит из двух потоков $\{\tau_n^{(1)},\, n \geq 1\}$ и $\{\tau_n^{(2)},\, n \geq 1\}$ с интенсивностями λ_1 и λ_2, с функциями распределения (ф.р.) $F_{\tau^{(1)}}(x)$ и $F_{\tau^{(2)}}(x)$, соответственно. Времена обслуживания $\{S_n,\, n \geq 1\}$ – независимые и одинаково распределенные (н.о.р.) случайные величины (с.в.) и не зависят от $\{\tau_n^{(1)},\, n \geq 1\}$ и $\{\tau_n^{(2)},\, n \geq 1\}$. Для простоты опустим индекс n в обозначениях.

Пусть $\lambda = \lambda_1 + \lambda_2$ – интенсивность обобщенного входного потока τ. Тогда $F_\tau(x) = pF_{\tau^{(1)}}(x)+(1-p)F_{\tau^{(2)}}(x)$ является конечной смесью двух входных потоков, где смешивающий коэффициент $p = \lambda_1/\lambda$. Пусть t_n – моменты поступления заявок в систему из обобщенного входного потока.

Для исходной системы построим две новые системы $\Sigma^{(1)}$ и $\Sigma^{(2)}$, в которых времена обслуживания совпадают, $S^{(1)} = S^{(2)} = S$, а входные потоки заданы $\tau^{(1)}$ и $\tau^{(2)}$, соответственно. Пусть $t_n^{(i)}$ – моменты прихода в i-ю систему, $i = 1, 2$.

Обозначим $W_n^{(i)}$ – время ожидания заявки с номером n в i-й системе в момент времени $t_n^{(i)-}$ прихода n-й заявки, W_n – незавершенная работа (время ожидания) в исходной системе Σ в момент времени t_n^-.

Предположим, что существуют следующие пределы (по распределению):

$$W_n \Rightarrow W, \quad W_n^{(i)} \Rightarrow W^{(i)},\ n \to \infty, \quad i = 1, 2.$$

Известно, что такие пределы существуют, если выполнено условие стационарности систем [6]

$$\mathbf{E}S < \mathbf{E}\tau^{(i)}, \quad i = 1, 2, \tag{1}$$

а интервалы между приходами заявок $\tau^{(i)}$ являются нерешетчатыми с. в, $i = 1, 2$.

Будем говорить, что с.в. упорядочены по интенсивности отказов, $X \underset{r}{\leq} Y$, если $r_X(x) \geq r_Y(x),\ \forall x \geq 0$, где $r_X(x) = f_X(x)/\overline{F}_X(x)$ – интенсивность отказов.

Предположим, что выполнены условия, гарантирующие существование предельных распределений максимумов стационарных времен ожидания [4, см. теорему 2]. Следующая теорема вытекает из теоремы 5 Витта [5], в которой

вместо стохастической упорядоченности входных потоков рассматривается упорядоченность по интенсивности отказов, а также теоремы 1 [4].

Теорема 1. *Пусть в системах* $\Sigma^{(1)}$ *и* $\Sigma^{(2)}$ *выполнены условия стационарности* (1)

Если выполнены соотношения

$$W_1^{(1)} \underset{st}{=} W_1^{(2)} = 0,\ \tau^{(1)} \underset{r}{\geq} \tau^{(2)},\ S^{(1)} \underset{st}{=} S^{(2)} \underset{st}{=} S \tag{2}$$

то экстремальные индексы стационарной незавершенной нагрузки в системах $\Sigma^{(1)}$, $\Sigma^{(2)}$ *и* Σ *удовлетворяют неравенствам:*

$$\theta_{W^{(1)}} \geq \theta_W \geq \theta_{W^{(2)}}, \tag{3}$$

где $\theta_W, \theta_{W^{(1)}}, \theta_{W^{(1)}}$ *– экстремальные индексы* $W_n, W_n^{(1)}, W_n^{(2)}$, *соответственно.*

Упорядоченность по интенсивности отказов является более строгим условием, но часто оказывается более удобной для практического применения чем стохастическая упорядоченность, поскольку, позволяет для многих распределений найти соотношения между параметрами (см., например, [7]).

3. Численное моделирование системы с гиперэкспоненциальным входным потоком

Пусть в системе Σ входной поток τ задан гиперэкспоненциальным распределением с хвостом ф.р. $\overline{F}_\tau(x) = pe^{-\lambda_1 x} + (1-p)e^{-\lambda_2 x},\ 0 < p < 1$.

Очевидно, что функции интенсивности отказов равны $r_{\tau_1(x)} = \lambda_1,\ r_{\tau_2(x)} = \lambda_2$. Если выполнено соотношение между параметрами входных потоков $\lambda_1 \leq \lambda_2$, то условие (2) теоремы выполнено и, следовательно, экстремальные индексы стационарной незавершенной работы в системах $\Sigma^{(1)}$, $\Sigma^{(2)}$ и Σ удовлетворяют соотношению (3).

Рисунок 1 демонстрирует результаты численного эксперимента для системы Σ с гиперэкпоненциальным распределением входного потока с параметрами $\lambda_1 = 5, \lambda_2 = 9,\ p = 0.5$. В минорантной системе входной поток - Пуассоновский с параметром $\lambda_1 = 5$, в мажорантной системе входной поток - Пуассоновский с параметром $\lambda_2 = 9$. Времена обслуживания во всех системах имеют показательное распределение с параметром $\mu = 10$. При таких параметрах условия стационарности (1) выполнены. Численное моделирование экстремального индекса стационарного времени ожидания проводилось для разных квантилей пороговых значений регенеративным методом.

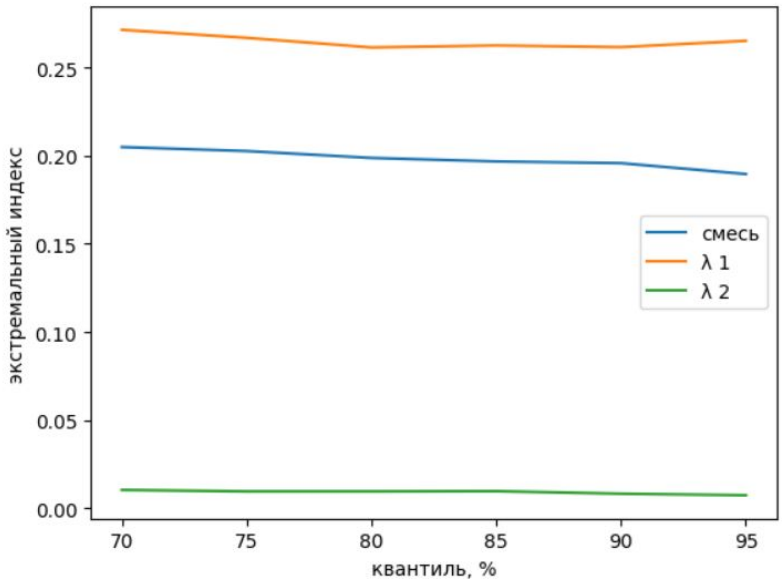


Рис. 1. Зависимость экстремального индекса стационарного времени ожидания от квантиля порогового значения в системах $\Sigma, \Sigma^{(1)}$ и $\Sigma^{(2)}$.

4. Заключение

Для системы $M/G/1$ с неоднородным входным потоком (с вероятностью p заявка поступает из первого потока, с вероятностью $1-p$ заявка поступает из второго потока) предложено построить мажорантную и миноратную системы, в которых времена обслуживания идентичны, а входные потоки соответствуют одному из входных потоков исходной системы. В Теореме 1 утверждается, что если входные потоки упорядочены по интенсивности отказов, то экстремальные индексы стационарных времен ожидания также упорядочены. Результаты приведенного численного эксперимента подтверждают теоретические выводы.

Работа выполнена при поддердке Российского научного фонда (проект 21-71-10135).

ЛИТЕРАТУРА

1. *Bertail P., Clemencon S., Tressou J.* Extreme values statistics for Markov chains via the (pseudo-) regenerative method // Extremes, 2009. Vol. 12. Iss. 4. P. 327–360. doi: 10.1007/s10687-009-0081-y.

2. *Resnick S.* Extreme Values, Regular Variation and Point Processes. – New York, NY, USA: Springer, 1987. 320 p.
3. *Embrechts P., Kluppelberg C., Mikosch T.* Modelling Extremal Events for Insurance and Finance. Applications of Mathematics. – Berlin, Heidelberg: Springer, 1997. 660 p.
4. Пешкова И. Границы экстремального индекса времени ожидания в системе M/G/1 с распределением времени обслуживания в виде конечной смеси // Информатика и её применения, 2022. Том. 16. Вып. 4. doi: 10.14357/19922264220405 (Scopus Q3, BAK)
5. *Whitt W.* Comparing counting processes and queues // Adv. Appl. Probab., 1981. Vol. 13. P. 207–220. doi: 10.2307/1426475.
6. *Asmussen S.* Applied Probability and Queues. Stochastic Modelling and Applied Probability. – New York, NY, USA: Springer-Verlag, 2003. 438 p.
7. Morozov, E., Peshkova, I., Rumyantsev, A. (2019). On Failure Rate Comparison of Finite Multiserver Systems. In: Vishnevskiy, V., Samouylov, K., Kozyrev, D. (eds) Distributed Computer and Communication Networks. DCCN 2019. Lecture Notes in Computer Science(), vol 11965. Springer, Cham. https://doi.org/10.1007/978-3-030-36614-8_32

UDC: 004.94

Readiness for timely execution of requests in fault-tolerant clusters with information recovery based on replication and backup

V.A. Bogatyrev [1,3], A.V. Bogatyrev [2], S.V. Bogatyrev [2,3]

[1]Department Information Systems Security, Saint-Petersburg State University of Aerospace Instrumentation, Saint Petersburg, Russia
[2]Yadro Cloud Storage Development Center, Saint Petersburg, Russia
[3]ITMO University, Saint Petersburg, Russia

Abstract

The possibilities of increasing the readiness of a cluster system for the timely execution of functional requests with a phased restoration of information based on replication and backup have been studied. Preliminary entry into the memory node intended to replace the failed one is provided with the results of the last backup. The influence of resource sharing of computing nodes on the execution of functional requests and the restoration of relevant information in memory is studied. The risks of using information based on the results of the last backup in the process of servicing functional requests are assessed. The effectiveness of the proposed solutions for restoring the system after failures is evaluated by the intensity of profit from servicing information requests, taking into account the risks of using outdated information.

Keywords: reliability, replication, Markov model, cluster, fault tolerance, backup.

1. Introduction

For real-time infocommunication systems, research aimed at ensuring their high fault tolerance and availability, with low computational delays with support for the continuity of the computational process, is relevant. In this regard, it is important to study the capabilities of cluster architecture computer systems to ensure fault tolerance and the probability of readiness at an arbitrary moment to service the flow of required functions, taking into account restrictions on allowable service delays and the inadmissibility of losing unique data accumulated during the operation of the system [1-4]. For failover clusters, it is effective to build duplicated cluster nodes, which are composed of two computers and two dual-input memory nodes. Markov

models for the reliability of duplicated cluster nodes while maintaining functioning based on the migration of computers are proposed in [5,6]. For computer systems containing memory nodes, a feature is the need for a phased restoration of memory: first, physical, and then informational. To restore information, replicas created in healthy cluster memory nodes can be used, while migrating replicas within a duplicated node requires the use of resources of computing nodes. Since the resources of computing nodes must also be used to service the flow of functional requests, the problem of distributing computing resources arises. Studies of the distribution of resources of serviceable computing nodes for restoring information in memory and for solving functional problems were carried out in [7]. Research [7] is aimed at reaching a compromise between the desire to increase the availability factor and the probability of timely execution of the incoming stream of functional requests. At present, the possibility of recovering memory by first entering the results of backup into a node designed to replace failed memory is not studied. As a result of this decision, the replaced memory node requires updating the information accumulated after backup. Up-to-date information can be entered into the restored memory node based on replication of only new information accumulated after backup. The memory node restored on the basis of backup can immediately be included in the work on servicing functional requests. This solution is intended to increase the likelihood of timely service of functional requests, although it is associated with the risks of servicing requests that require post-replication data.

Thus, the purpose of the study is to increase the readiness of the cluster system for the timely execution of functional requests based on the preliminary entry into the memory node intended to replace the failed one, the results of the last backup, with further phased restoration of relevant information.

2. Markov model of a cluster node taking into account phased information recovery based on replication and backup

As cluster nodes, we consider duplicated computer systems containing two calculators, each of which is connected to two two-input memory nodes. Two memory nodes are used to replicate the information accumulated during the operation of the system. Periodically, the system is backed up. To replace a failed memory node, a node is prepared in advance, into which the results of the last backup are applied. After connecting a backup memory block based on replication from a healthy node, the information accumulated since the last backup is updated. Data replication requires the use of resources from one of the healthy computers. If information is lost in two memory nodes, then a more complex and lengthy procedure for recovering information is provided.

The state and transition diagrams for the proposed Markov model of the system under study are shown in Fig. 1.1. The state of the cluster node is set by the matrix 2×2.

The upper line of the matrix displays the state of the calculators, and the lower line of the memory. The operability of the nodes corresponds to "1", and the failure "0". The s indicates the state of the memory with the last backup data entered.

The diagram in Fig. 1 shows the failure rates of the computing node and memory Λ_1, Λ_2, the recovery of the computer - μ_1 and memory with the preliminary entry of the results of the last backup μ_2. The intensity of restoring relevant information in memory based on replication from a working memory node μ_3, and in case of loss of relevant information in two memory nodes μ_4.

The set of states of the cluster node according to the possibilities of servicing the flow of functional tasks is divided into states in which two-channel or single-channel servicing of the flow of requests is possible, states with a risk associated with outdated data in memory, and also states in which the servicing of functional requests is not possible.

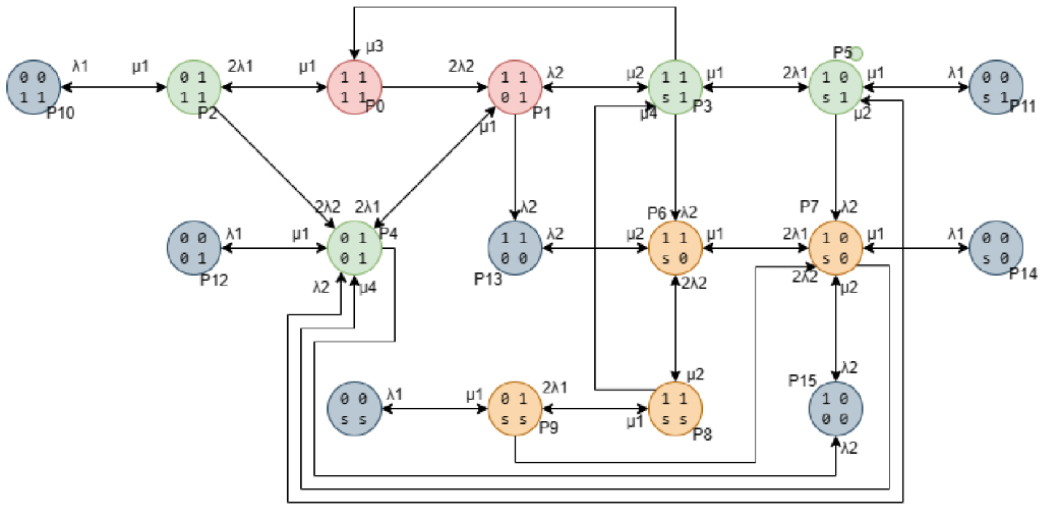


Fig. 1. Markov model of a duplicated computer node of a cluster

According to the diagram shown in Fig. 1, according to known rules, a system of algebraic equations is compiled [8], solving which by known methods using computer mathematics means, one can find the probabilities of all states. Summing up the probabilities of the system being in operable states, one can find the stationary

availability factor [8]. However, this indicator does not reflect the difference in the system's capabilities to provide one- and two-channel request servicing. This difference can be taken into account by the efficiency retention coefficient [8], but it does not allow estimating the probabilities of timely servicing of requests, which is important for real-time systems. This estimate can be obtained on the basis of the coefficient of readiness for timely fulfillment of functional requests proposed in [7]. However, the coefficient proposed in the index does not take into account the risks associated with serving functional requests that require data generated since the last backup. To overcome this shortcoming, it is proposed to evaluate the efficiency of the system by the intensity of profit received from servicing functional requests. A comprehensive performance indicator based on the intensity of profit, having a physical meaning, allows you to resolve the contradictions on the use of the above indicators.

We estimate the intensity of profit from servicing functional requests as:

$$\begin{aligned} S = \Lambda\{[R_1B_2 + R_2B_1 + aR_3B_2 + aR_4B_1]c_1 + \\ +[R_1(1-B_2) + R_2(1-B_1) + aR_3(1-B_2) + aR_4(1-B_1)]c_2 + \\ +R_0c_3 + (1-a)R_5c_5\}, \end{aligned}$$

where

$$R_1 = P_0 + P_1,$$

$$R_2 = P_2 + P_3 + P_4 + P_5,$$

$$R_3 = P_6 + P_8,$$

$$R_4 = P_7 + P_9,$$

$$R_0 = P_{10} + P_{11} + P_{12} + P_{13} + P_{14} + P_{15} + P_{16},$$

$$R_6 = R_3 + R_4.$$

Λ is the intensity of the flow of functional tasks,
c_1 - profit from timely servicing of requests,
c_2 - penalty for not timely servicing requests,
c_3 - penalty for refusing to service requests,

c_4 - penalty for servicing using outdated data,

B_1 and B_2 are the probabilities of timely execution of requests for a time less than the maximum allowable

t_0, respectively, with two and one calculator,

$$B_1 = 1 - \Lambda v \exp(t_0(\Lambda - v^{-1})),$$
$$B_2 = 1 - \frac{\Lambda}{2} v \exp(t_0(\frac{\Lambda}{2} - v^{-1}),$$

where v is the average execution time of a functional request.

3. Calculation example

The results of calculating the profit from servicing functional requests are shown in Fig.2. Curves 1-3 correspond to the profit received from servicing functional requests, the maximum allowable waiting time of which should not exceed $t_0 = v, 2v, 4v$.

The calculation was made at $v = 0.1$ s, $\Lambda_1 = \Lambda_2 = 10^{-4}$ 1/h. $\mu_1 = 1$ 1/h, $\mu_2 = 1$ 1/h, $\mu_3 = 0.2$ 1/h, $\mu_4 = 0.1$ 1/h, $c_1 = 30$ c.u., $c_2 = -10$ c.u., $c_3 = -20$ c.u., $c_4 = -50$ c.u. It can be seen from the figure that with a certain increase in the intensity of the flow of functional requests, it is rational to stop the provision of services for servicing the flow (if possible).

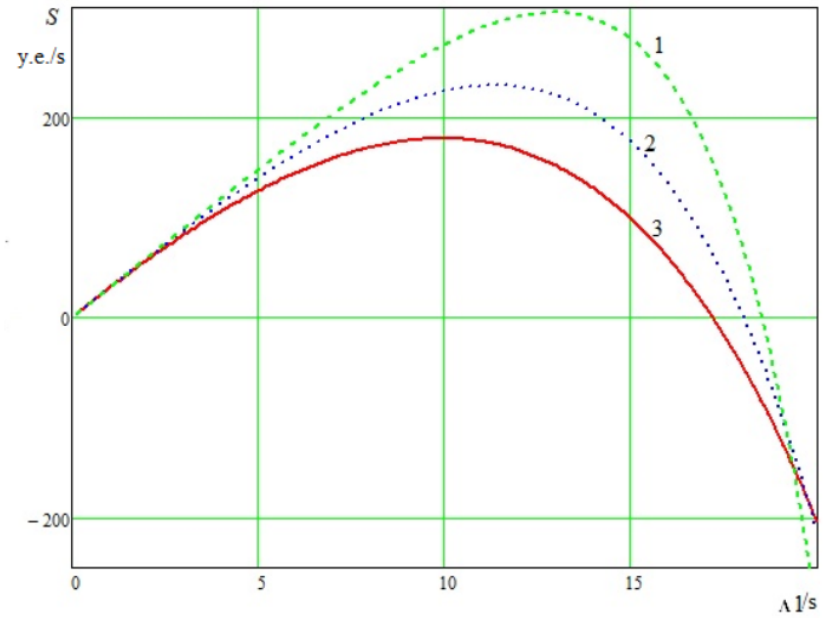


Fig. 2. Profit from serving functional requests

4. Conclusion

For computing systems of a cluster architecture built on the basis of duplicated nodes containing two computers and two nodes of dual-input memory, the possibilities of increasing the readiness of the cluster for the timely execution of requests that are critical to service delays are analyzed, taking into account the multi-stage information recovery of memory after failures.

Markov models of duplicated cluster nodes are proposed, on the basis of which the dependences of the system readiness for timely execution of requests on resource allocation options that have retained the operability of computing nodes to perform functional tasks and restore information in memory, including the stages of entering the results of the last backup and replicating relevant information from memory of healthy nodes.

The effectiveness of the proposed solutions for restoring the system after failures is estimated by the intensity of profit from servicing information requests, taking into account the risks of accessing irrelevant information by information.

REFERENCES

1. Afyouni I., Khan A., Aghbari Z.A. Deep-eware: spatio-temporal social event detection using a hybrid learning model //Journal of Big Data. 2022. V. 9. No. 1. S. 1-21
2. Afzal Sh., Kavitha G. Load balancing in cloud computing – a hierarchical taxonomical classification//Journal of Cloud Computing. 2019. V. 8. No. 1. S. 1-24
3. Tatarnikova T.M., Sverlikov A.V. Methodology for detecting anamalies in the traffic of the internet of things//Wave Electronics and Its Application in Information and Telecommunication Systems. 2022. V. 5. No. 1. S. 476-479.
4. Verzun N.A., Kolbanev M.O., Romanova A.A. Indicators of the effectiveness of the process of information interaction in the Internet of things // Izvestiya SPbGETU LETI. 2022. No. 3. S. 5-14.
5. Tatarnikova T.M., Ivanova A. Access control system for the premises using the "smart home" technology. // Wave Electronics and its Application in Information and Telecommunication Systems, WECONF 2020. 2020. P. 9131442
6. Bogatyrev V.A., Bogatyrev A.V., Bogatyrev S.V. Redistribution of requests between computing clusters during their degradation//Izvestia of higher educational institutions. Instrumentation. 2014. V. 57. No. 9. S. 54-58.
7. Bogatyrev V. A. Information systems and technologies. Reliability theory: textbook for undergraduate and graduate students. - M.: Yurayt, 2016
8. Bogatyrev V.A., Derkach A.N. Evaluation of a Cyber-Physical Computing System with Migration of Virtual Machines during Continuous Computing // Computers - 2020. V. 9. N 2. P. 42.
9. Bogatyrev V.A., Bogatyrev S.V., Bogatyrev A.V. Assessment of the readiness of a computer system for timely servicing of requests when combined with information recovery of memory after failures. Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 2023, vol. 23, no. 3, pp. (in Russian). doi:10.17586/2226-1494-2023-23-3-

UDC: 519.23

Model of operation of a cell of a mobile communication network with adaptive modulation schemes and batch arrivals

A.N. Dudin[1] and O.S. Dudina[1]

[1]Department of Applied Mathematics and Computer Science, Belarusian State University, 4, Nezavisimosti Ave., Minsk, 220030, Belarus

dudin@bsu.by, dudina@bsu.by

Abstract

A model of a cell of a mobile communication network divided into a finite number of zones with dependence of the users' service rate on a zone where the user is currently residing is considered. The arrival flow of users is defined by a batch marked Markovian process. The number of users that can receive service in a cell simultaneously is restricted. The operation of the cell is described by a multi-server queueing system whose dynamics is defined by a multi-dimensional Markov chain. The generator of the chain having the upper-Hessenbergian structure is derived. The stationary distribution of the chain and the main characteristics of the quality of service are found.

Keywords: adaptive modulation, correlated arrival process, stationary distribution

1. Introduction

The motivation for this research and a brief survey of the related literature can be found in [1]. The model considered in the present paper is a practically important generalization of the model from [1] to the case of possibility of batch generation of user requests. This generalization causes that the Markov chain under study does not belong to the case of quasi-birth-and-death processes, which essentially complicates the analysis.

2. Mathematical model

We consider the user's processing in the cell of a mobile communication network as service in an N-server queueing system without a buffer. This means that no more than N users can receive service simultaneously.

The cell is divided into R, $1 < R < \infty$, zones. The strength of a signal from a base station, which mainly defines the user's service rate, depends on the zone in the cell in which the user is currently residing. To take this into account, we formally divide all arriving users into R types corresponding to the zone where the user starts service. The users' arrival process is assumed to be defined by the $BMMAP$ (batch marked Markov arrival process).

This process is the generalisation of the more well-known marked Markov arrival process ($MMAP$) to the case of batch arrivals of heterogeneous users. The possible moments of user's arrivals in the $BMMAP$ are defined as the moments of the jumps of an irreducible continuous-time Markov chain ν_t, $t \geq 0$, with a finite state space $\{1, 2, ..., W\}$. The $BMMAP$ is defined by the set of the square matrices D_0, $D_r^{(l)}$, $r = \overline{1, R}$, $l = \overline{0, L}$, where L is the maximum batch size. The entries of the matrix $D_r^{(l)}$ define the intensities of the transitions of the chain ν_t which are accompanied by the arrival of l type-r users. Let us denote $D = \sum\limits_{r=1}^{R} \sum\limits_{l=1}^{L} D_r^{(l)}$. The non-diagonal entries of the matrix D_0 define the intensities of the transitions of the chain ν_t that are not accompanied by a user's arrival. The diagonal entries of the matrix D_0 define the rate of the exit of the chain ν_t from the corresponding states. The matrix $D(1) = D_0 + D$ is the generator of the Markov chain ν_t.

The average intensity λ_r of type-r users arrival is defined as $\lambda_r = \boldsymbol{\theta} \sum\limits_{l=1}^{L} l D_r^{(l)} \mathbf{e}, r = \overline{1, R}$, where $\boldsymbol{\theta}$ is the invariant probability vector of the chain ν_t. The average total intensity λ of users arrival is defined as $\lambda = \sum\limits_{r=1}^{R} \lambda_r$. For more information about the $BMMAP$, see, e.g., [2].

Due to the possibility of batch arrivals, we have to fix the user's acceptance discipline. We assume the partial admission discipline. If an arbitrary group that consists of l users of any type arrives when there are k, $k \leq l$, free servers, then $l - k$ users start service and the rest of users leave the system without service (are lost), $l = \overline{1, L}$.

Let us consider a user who generates a connection in the rth zone or transits into this zone from another one during an established connection. We assume that the user's sojourn time in the rth zone has an exponential distribution with the rate $\mu_r, r = \overline{1, R}$.

There are various reasons for a user to leave zone r listed below:

- the service of the user is successfully completed, and he/she leaves the cell. The probability of this event is assumed to be $p_{r,out}$;
- the user transits to the kth zone (and becomes a type-k user). The probability of this event is assumed to be $p_{r,k}$;

- the service of the user is terminated due to a loss of connection, and the user leaves the cell without complete service. The probability of this event is assumed to be $p_{r,loss}$;
- the service of the user is terminated due to handover to another cell. The probability of this event is assumed to be $p_{r,hand}$.

Note, that $p_{r,out} + p_{r,loss} + p_{r,hand} + \sum\limits_{k=1,k\neq r}^{R} p_{r,k} = 1$ for each r. By multiplying μ_r with the probabilities of these events, we obtain intensities $\mu_{r,serv}$, $\mu_{r,k}$, $\mu_{r,loss}$ and $\mu_{r,hand}$ of successful completion of the service, transition of the user to the kth zone, service termination without complete service and handover, correspondingly. It is clear that $\mu_r = \mu_{r,serv} + \mu_{r,loss} + \mu_{r,hand} \sum\limits_{k=1,k\neq r}^{R} \mu_{r,k}$.

We define the total service time of a user in the cell as the time during which an irreducible Markov chain η_t, $t \geq 0$, with the transient states $\{1, 2, \ldots, R\}$ and three absorbing states reaches one of the absorbing states. This Markov chain is described as follows. The initial state of the chain η_t is chosen among the transient states depending on the type of a user accepted for service. If this is a type-r user, the initial state of the chain η_t is the state r, $r = \overline{1, R}$.

The intensities of the transitions of the process η_t between the transient states are defined by the sub-generator

$$S = \begin{pmatrix} -\mu_1 & \mu_{1,2} & \mu_{1,3} & \cdots & \mu_{1,R} \\ \mu_{2,1} & -\mu_2 & \mu_{2,3} & \cdots & \mu_{2,R} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_{R,1} & \mu_{R,2} & \mu_{R,3} & \cdots & -\mu_R \end{pmatrix}.$$

When the process η_t stays in the rth transient state, the user receives type-r service. The transition of the process η_t from one transient state to another corresponds to the corresponding change of zones by the user.

The transition to the first absorbing state corresponds to successful service completion. The intensities of the transition to this absorbing state are defined by the components of the column vector $\mathbf{S}_{serv} = (\mu_{1,serv}, \mu_{2,serv}, \ldots, \mu_{R,serv})^T$.

The transition to the second absorbing state corresponds to leaving the cell without finishing the service (owing to impatience or dissatisfaction by the quality). The intensities of the transition to this absorbing state are defined by the column vector $\mathbf{S}_{loss} = (\mu_{1,loss}, \mu_{2,loss}, \ldots, \mu_{R,loss})^T$.

The transition to the third absorbing state corresponds to leaving the cell due to handover. The intensities of the transition to this absorbing state are defined by the column vector $\mathbf{S}_{hand} = (\mu_{1,hand}, \mu_{2,hand}, \ldots, \mu_{R,hand})^T$.

Having fully defined the user arrival and service processes, we can start the stationary analysis of the behaviour of the system.

3. Process of system states

The behaviour of the system under study can be described by the regular irreducible continuous-time Markov chain $\xi_t = \{n_t, \nu_t, \mathbf{m}_t\}$, $t \geq 0$, where during the epoch t the component n_t defines the number of users in the system, ν_t is the state of $BMMAP$ arrival process, and $\mathbf{m}_t = (m_t^{(1)}, \ldots, m_t^{(R)})$, where $m_t^{(r)}$ is the number of servers at phase r of the service (the number of users in the rth zone), $m_t^{(r)} = \overline{0, n_t}$, $\sum_{r=1}^{R} m_t^{(r)} = n_t$, $r = \overline{1, R}$.

The Markov chain $\xi_t,\ t \geq 0$, is irreducible and has a finite state space. Therefore, the stationary probabilities of the system states

$$\pi(n, \nu, m^{(1)}, \ldots, m^{(R)}) = \lim_{t \to \infty} P\{n_t = n, \nu_t = \nu, m_t^{(1)} = m^{(1)}, \ldots, m_t^{(R)} = m^{(R)}\}$$

always exist.

Let us form the row vectors $\boldsymbol{\pi}_n$, $n = \overline{0, N}$, of these probabilities which are enumerated in the reverse lexicographic order of the components $m_t^{(1)}, \ldots, m_t^{(R)}$ and the direct lexicographic order of the component ν_t.

It is well known that the probability vectors $\boldsymbol{\pi}_n$, $n = \overline{0, N}$, satisfy the following system of linear algebraic equations:

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_N) Q = \mathbf{0},\ (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_N)\mathbf{e} = 1 \tag{1}$$

where Q is the infinitesimal generator of the Markov chain $\xi_t,\ t \geq 0$. By analysing all possible transitions of the Markov chain $\xi_t, t \geq 0$, during an interval of infinitesimal length and rewriting the intensities of these transitions in block matrix form, we obtain the following result.

Theorem 1. The infinitesimal generator Q of the Markov chain $\xi_t,\ t \geq 0$, has the following block structure

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & Q_{0,2} & Q_{0,3} & \ldots & Q_{0,N-1} & Q_{0,N} \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & Q_{1,3} & \ldots & Q_{1,N-1} & Q_{1,N} \\ O & Q_{2,1} & Q_{2,2} & Q_{2,3} & \ldots & Q_{2,N-1} & Q_{2,N} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & O & \ldots & Q_{N,N-1} & Q_{N,N} \end{pmatrix}. \tag{2}$$

The blocks $Q_{i,l}$, $i = \overline{0, N}$, $l = \overline{\max\{0, i-1\}, N}$, are defined as follows:

$$Q_{0,0} = D_0,\ Q_{n,n} = D_0 \oplus [A_n(S) + \Delta^{(n)}],\ n = \overline{1, N-1},$$

$$Q_{N,N} = D_0 \oplus [A_N(S) + \Delta^{(N)}] + D \otimes I_{T_N},$$

$$Q_{n,n+l} = \sum_{r=1}^{R} D_r^{(l)} \otimes \prod_{i=n}^{n+l-1} P_i(\boldsymbol{\beta}_r),\ n = \overline{0, N-2},\ 1 \le l \le \min\{N-1-n, L\},$$

$$Q_{n,n+l} = O_{WT_n \times WT_{n+l}},\ n = \overline{0, N-2},\ L+1 \le l \le N-1-n,$$

$$Q_{n,N} = \sum_{r=1}^{R} \sum_{l=N-n}^{L} D_r^{(l)} \otimes \prod_{i=n}^{N-1} P_i(\boldsymbol{\beta}_r), n = \overline{\max\{0, N-L\}, N-1},$$

$$Q_{n,N} = O_{WT_n \times WT_N},\ n = \overline{0, N-L-1},$$

$$Q_{n,n-1} = I_W \otimes [L_n(\mathbf{S}_{serv}) + L_n(\mathbf{S}_{loss}) + L_n(\mathbf{S}_{hand})],\ n = \overline{1, N},$$

where I is the identity matrix and O is a zero matrix of an appropriate dimension; $\otimes$ and $\oplus$ indicate the symbols of the Kronecker product and sum of matrices, respectively; $T_n = \frac{(n+R-1)!}{n!(R-1)!}$; $\boldsymbol{\beta}_r = (\underbrace{0, \ldots, 0}_{r-1}, 1, \underbrace{0, \ldots, 0}_{R-r})$, $r = \overline{1, R}$. The detailed description of the matrices $P_n(\boldsymbol{\beta}_r)$, $L_n(\mathbf{S}_{serv})$, $L_n(\mathbf{S}_{loss})$, $L_n(\mathbf{S}_{hand})$, $A_n(S)$, $\Delta^{(n)}$ and the algorithms used for their calculation are presented in [1].

The number of equations of the finite system (1) with the matrix having the upper-Hessenberg form (2) may be high. To solve it, we recommend the numerically stable algorithm from [3] that effectively uses the structure of the generator Q.

4. PERFORMANCE MEASURES

The average number of users in the system at an arbitrary moment is computed by $N^{sys} = \sum_{n=1}^{N} n\boldsymbol{\pi}_n \mathbf{e}$.

The loss probability of an arbitrary user upon arrival to the r-th zone due the system overflow is $P_r^{ent-loss} = \frac{1}{\lambda} \sum_{l=1}^{L} \sum_{n=N-l+1}^{N} \boldsymbol{\pi}_n [(l-(N-n)) D_r^{(l)} \otimes I_{T_n}] \mathbf{e}$, $r = \overline{1, R}$.

The loss probability of an arbitrary user upon arrival due to the system overflow is computed by $P^{ent-loss} = \frac{1}{\lambda} \sum_{r=1}^{R} \sum_{l=1}^{L} \sum_{n=N-l+1}^{N} \boldsymbol{\pi}_n [(l-(N-n)) D_r^{(l)} \otimes I_{T_n}] \mathbf{e}$.

The intensity of the output flow from the r-th zone is computed by $\lambda_r^{out} = \sum_{n=1}^{N} \boldsymbol{\pi}_n (I_W \otimes L_n(\mathbf{S}_r^{serv})) \mathbf{e}$, $r = \overline{1, R}$.

The intensity of the output flow of successfully served users is computed by $\lambda^{out} = \sum_{n=1}^{N} \boldsymbol{\pi}_n (I_W \otimes L_n(\mathbf{S}_{serv})) \mathbf{e}$.

The loss probability of an arbitrary user due to a loss of connection in the r-th zone is $P_r^{loss-connection} = \frac{1}{\lambda} \sum_{n=1}^{N} \boldsymbol{\pi}_n (I_W \otimes L_n(\mathbf{S}_r^{loss}))\mathbf{e}$, $r = \overline{1, R}$.

The loss probability of an arbitrary user due to a loss of connection is $P^{loss-connection} = \frac{1}{\lambda} \sum_{n=1}^{N} \boldsymbol{\pi}_n (I_W \otimes L_n(\mathbf{S}_{loss}))\mathbf{e}$.

The loss probability of an arbitrary user due to handover to another cell is $P^{loss-handover} = \frac{1}{\lambda} \sum_{n=1}^{N} \boldsymbol{\pi}_n (I_W \otimes L_n(\mathbf{S}_{hand}))\mathbf{e}$.

The loss probability of an arbitrary user due to handover to another cell from the r-th zone is computed by $P_r^{loss-handover} = \frac{1}{\lambda} \sum_{n=1}^{N} \boldsymbol{\pi}_n (I_W \otimes L_n(\mathbf{S}_r^{hand}))\mathbf{e}$.

Here, the column vectors $\mathbf{S}_r^{serv}$, $\mathbf{S}_r^{loss}$ and $\mathbf{S}_r^{hand}$ have all zero entries except r-th entry which is equal to $\mu_{r,serv}$, $\mu_{r,loss}$, $\mu_{r,hand}$, $r = \overline{1, R}$, respectively.

The probability of an arbitrary user loss is computed by

$$P^{loss} = 1 - \frac{\lambda^{out}}{\lambda} = P^{ent-loss} + P^{loss-connection} + P^{loss-handover}.$$

The probability of an arbitrary user loss from the r-th zone is computed by

$$P_r^{loss} = P_r^{ent-loss} + P_r^{loss-connection} + P_r^{loss-handover},\ r = \overline{1, R}.$$

The probability that the system is empty is computed by $P_{idle} = \boldsymbol{\pi}_0 \mathbf{e}$.

5. Conclusion

Obtained results supplemented with the developed by authors software can be used for analysis of the performance of the cell and user's quality of service under the fixed configuration of the equipment and for optimization of the limit N of the number of simultaneously serviced users under various criteria of the quality of the system's operation.

REFERENCES

1. Kim C. et al. Mathematical model of operation of a cell of a mobile communication network with adaptive modulation schemes and handover of mobile users // IEEE Access. – 2021. – V. 9. – P. 106933-106946.
2. Dudin A. N., Klimenok V. I., Vishnevsky V. M. The theory of queuing systems with correlated flows. – Cham : Springer, 2020.
3. Dudin A. et al. Analysis of single-server multi-class queue with unreliable service, batch correlated arrivals, customers impatience, and dynamical change of priorities // Mathematics. – 2021. – V. 9(11). – P. 1257.

УДК: 519

Управление рисками при проектировании топологии компьютерных сетей

А.А. Широкий [1]

[1] ИПУ РАН, Профсоюзная, 65, Москва, Россия

shiroky@ipu.ru

Аннотация

Решение задач обеспечения безопасности компьютерных сетей часто сводится к выбору оптимального набора контрмер. При этом возможность модификации топологии сети обычно не рассматривается. В то же время для некоторых классов сетей (в частности, беспроводных mesh-сетей) перестройка топологии является естественным процессом. В связи с этим представляется целесообразным рассмотреть возможность применения результатов, полученных в рамках исследования влияния структуры сложных систем на их интегральный риск для решения задачи проектирования минимизирующей риск топологии компьютерной сети. В настоящей работе рассмотрен частный случай точного решения этой задачи для одноранговой распределённой сети топологии «шина», а также предложен алгоритм проектирования минимизирующей риск топологии при рассмотрении сценариев атак, структурно являющихся простыми цепями.

Ключевые слова: компьютерные сети, проектирование топологии, управление рисками

1. Введение

Решая задачи выбора оптимального набора контрмер при планировании безопасности компьютерных сетей, специалисты и исследователи часто используют риск-ориентированный подход. В его рамках обычно проводят идентификацию угроз, вероятных сценариев атаки, формируют множество возможных контрмер, а затем получают искомый набор контрмер как решение задачи оптимизации (см., например, [1, 2, 3]), либо как условие равновесия в игре [4, 5, 6].

При управлении безопасностью компьютерных сетей важную роль играет их топология. Во многих случаях она определяется требованиями к функциональности сети и не может быть изменена, в связи с чем возможностям снижения рисков за счёт перекоммутации узлов или их перестановки уделяется минимум

внимания. В то же время эта задача актуальна для самоорганизующихся сетей, в том числе, например, одноранговых распределённых сетей стандарта IEEE 802.11s [7].

В настоящей работе обсуждается возможность применения принципов управления рисками сложных сетей с учётом их структуры к задачам проектирования топологии компьютерных сетей. В частности, будет рассмотрено влияние на интегральный риск положения узлов относительно периметра в сети с топологией «шина». Полученный результат может быть использован как напрямую (например, при проектировании беспроводной mesh-сети), так и при сценарном анализе возможных атак. Также будет предложен алгоритм проектирования топологии компьютерной сети, минимизирующей риск относительно сценариев атаки, структурно являющимися простыми цепями.

2. Общая постановка задачи

Рассмотрим сложную систему, состоящую из конечного множества элементов (объектов, пока произвольной природы): $S = \{s_1, ..., s_i, ..., s_n\}, i \in N = \{1, ..., n\}$. Будем предполагать, что элементы $s_i \in S, i \in N$, системы S являются автономными и, в частности, не могут оказывать влияние на состояния друг друга.

Предположим, что существуют два субъекта (также пока произвольной природы), которых мы будем называть игрок A (иначе, Атакующий, *Attacker*) и игрок D (иначе, Защитник, *Defender*), имеющие несовпадающие интересы относительно состояния системы S. Определим для каждого элемента $s_i \in S$ функцию локального риска $\rho_i(x_i, y_i) : \mathbb{R}^0_+ \times \mathbb{R}^0_+ \to \mathbb{R}^0_+$, где $\mathbb{R}^0_+$ — множество действительных неотрицательных чисел.

Пусть на множестве элементов системы S задана структура $W = \langle G(S, E), T\rangle$, где $G(S, E)$ — граф на множестве вершин-элементов S со множеством рёбер E, а $T \subseteq S$ — некоторое подмножество вершин, которое будем называть периметром системы S.

Будем считать, что игрок A атакует элементы рассматриваемой системы по выбранной им цепи $c = \langle u, v\rangle, u \in T, v \in S$, причём переход из некоторой вершины $s_i \in c$ по инцидентному ей ребру в смежную вершину $s_j \in c$ осуществляется только в случае успешной атаки элемента s_i.

Пусть $x = (x_1, ..., x_n), y = (y_1, ..., y_n)$ — некоторые допустимые распределения ресурсов игроками D и A соответственно между вершинами — элементами системы S . Будем рассматривать ограниченные, неотрицательные и монотонные функции локального риска вида

$$\rho_i(x, y) = u_i(x, y) \cdot p_i(x, y) \tag{1}$$

для каждой вершины $s_i \in S$. Здесь $u_i(x, y) : \mathbb{R}^n_+ \times \mathbb{R}^n_+ \to \mathbb{R}^0_+$ — функция, описывающая зависимость ожидаемого ущерба в случае успешной атаки элемента s_i в зависимости от распределений ресурсов x и y, а $p_i(x, y) : \mathbb{R}^n_+ \times \mathbb{R}^n_+ \to (0, 1]$ — вероятность успешной атаки элемента s_i в зависимости от распределений ресурсов x и y.

Базовая модель управления рисками сложной системы со структурой и периметром задаётся следующим кортежем:

$$\langle S = \{s_i\}_{i \in N}, T, E, D, A, X, Y, \{\rho_i(\cdot, \cdot)\}_{i \in N}, \rho(\cdot, \cdot)\rangle, \tag{2}$$

где X и Y — объёмы ресурсов, располагаемых игроками D и A соответственно.

Если структура $W = \langle G(S, E), T\rangle$ фиксирована, то

- *целью Защитника* является распределение доступного ему ресурса X между элементами системы S с тем, чтобы добиться максимально возможного снижения значения функции интегрального риска $\rho(x, y)$;
- *целью Атакующего* — наоборот: распределить доступный ему ресурс Y между элементами системы S таким образом, чтобы добиться максимально возможного увеличения значения функции интегрального риска $\rho(x, y)$.

Обозначим $\mathscr{X}(X)$ множество допустимых распределений ресурса X между элементами системы S игроком D, а $\mathscr{Y}(Y)$ — множество допустимых распределений ресурса Y между элементами системы S игроком A:

$$\mathscr{X}(X) = \left\{ (x_1, ..., x_n) \in \mathbb{R}^n_+ : x_i \geq 0, i \in N, \sum_{i=1}^{n} x_i \geq X \right\}, \tag{3}$$

$$\mathscr{Y}(Y) = \left\{ (y_1, ..., y_n) \in \mathbb{R}^n_+ : y_i \geq 0, i \in N, \sum_{i=1}^{n} y_i \geq Y \right\}. \tag{4}$$

Тогда задача игрока D («задача Защитника») заключается в нахождении распределения ресурса $x^* \in \mathscr{X}$, минимизирующего интегральный риск, и формально может быть записана в виде:

$$x^* = \operatorname*{Arg\,min}_{x \in \mathscr{X}} \rho(x, y) = \operatorname*{arg\,min}_{x \in \mathscr{X}} \sum_{i=1}^{n} \rho_i(x, y). \tag{5}$$

Аналогично, задача игрока A («задача Атакующего») заключается в нахождении распределения ресурса $y^* \in \mathscr{Y}$, максимизирующего интегральный риск, и может быть записана в виде:

$$y^* = \operatorname*{Arg\,max}_{y \in \mathscr{Y}} \rho(x, y) = \operatorname*{arg\,max}_{y \in \mathscr{Y}} \sum_{i=1}^{n} \rho_i(x, y). \tag{6}$$

Если же структуру $W = \langle G(S, E), T\rangle$ можно изменять (например, модифицируя множества E и/или T), то прежде, чем решать задачу (5), Защитник может дополнительно снизить риски, решив задачу построения структуры, оптимальной в смысле минимизации рисков.

3. Задача минимизации риска при проектировании одноранговой распределённой сети топологии «шина»

Вначале рассмотрим частную задачу, заключающуюся в выборе размещения узлов сети, минимизирующего интегральный риск в сценарии последовательной атаки узлов начиная с некоторого узла-периметра.

Предположим, что мы располагаем множеством $S = \{s_1, ..., s_n\}$ узлов, каждый из которых может выполнять функции маршрутизатора. Для каждого элемента множества s_i нам известна вероятность p_i, что он будет успешно атакован (в некотором смысле) злоумышленником. Также будем считать, что нам известны величины ущерба u_i, причиняемого защищаемой системе в случае успешной атаки соответствующего узла. В рассматриваемом случае игрок A последовательно атакует узлы защищаемой сети, начиная с некоторого выбираемого игроком D узла-периметра и продвигаясь далее по соседям. Каждый узел игрок A посещает ровно один раз.

Для записи формальной постановки задачи воспользуемся модифицированными обозначениями и определениями, введёнными в работе [8].

Определение 1. *Пусть задан граф* $G\left(V = \{v_1, ..., v_n\}, E = \{(v_i, v_{i+1})\}_{i=1}^{n-1}\right), n \in \mathbb{N}$ *и периметр* $T = \{v_1\}$. *Тогда будем говорить, что кортеж* $W_n = \langle G(V, E), T\rangle$ *задаёт сценарий атаки длины* n.

Отметим, что граф G является простой цепью.

Определение 2. *Взаимно-однозначное отображение* $M^{-1} : S \to V, S = \{s_1, ..., s_n\}, n \in \mathbb{N} : \forall i \le n\ \exists! j \le n : v_j = M^{-1}(s_i)$ *будем называть размещением узлов* S *в сценарии* W_n. *Соответствующее обратное отображение* $M : V \to S$ *будем называть проекцией сценария* W_n *на множество узлов* S.

В дальнейшем для удобства мы будем опускать нижний индекс у обозначения сценария.

Для произвольного заданного размещения $M^{-1} : S \to V$ можно рассчитать значение интегрального риска

$$\rho\left(S, W, M^{-1}\right) = \sum_{i=1}^{n} \rho_{M(v_i)}, \tag{7}$$

где $\rho_{M(v_i)}$ — значение локального риска для узла $M(v_i)$, и записать задачу минимизации интегрального риска, заключающуюся в поиске множества $\mathbf{M}^{-1}_{min}$ таких размещений, для каждого из которых достигается минимальное значение интегрального риска риска ρ_{min}:

$$\mathbf{M}^{-1}_{min} = \underset{M^{-1}}{\operatorname{Arg\,min}} \rho\left(S, W, M^{-1}\right) : \rho_{min} = \sum_{i=1}^{n} \rho_{M(v_i)} \forall M^{-1} \in \mathbf{M}^{-1}_{min}. \quad (8)$$

Определение 3. *Будем говорить, что узлы $s_i, s_j \in S, i, j \in N, i \neq j$ нестрого упорядочены по возрастанию (убыванию) локального риска и записывать $s_i \preceq s_j$ ($s_i \succeq s_j$) если при заданном сценарии атаки W для любых размещений M^{-1}, K^{-1} и любых таких индексов p, q, k, l, $p < q$, $k > l$, что $s_i = M(v_p) = K(v_k)$, $s_j = M(v_q) = K(v_l)$ выполняется неравенство $\rho\left(S, W, M^{-1}\right) \leq \rho\left(S, W, K^{-1}\right)$ $\left(\rho\left(S, W, M^{-1}\right) \geq \rho\left(S, W, K^{-1}\right)\right)$.*

В работе [8] были доказаны следующие утверждения.

Утверждение 1. *Пусть $N = \{1, ..., n\}, S = \{s_1, ..., s_n\}$. Тогда $\forall i \in N \setminus \{n\}$ $s_i \preceq s_{i+1} \iff \frac{u_i}{u_{i+1}} \leq \frac{p_{i+1}(1-p_i)}{p_i(1-p_{i+1})}$; $s_i \succeq s_{i+1} \iff \frac{u_i}{u_{i+1}} \geq \frac{p_{i+1}(1-p_i)}{p_i(1-p_{i+1})}$.*

Утверждение 2. *Пусть $N = \{1, ..., n\}, S = \{s_1, ..., s_n\}$. Тогда $\forall i, j, k \in N$: $i < j < k$ $s_i \preceq s_j \preceq s_k \implies s_i \preceq s_k$.*

Эти утверждения задают транзитивный критерий упорядочивания узлов в сценарии атаки и позволяют решить задачу (8) для любого рассматриваемого сценария в общем виде.

4. Алгоритм проектирования минимизирующей риск топологии компьютерной сети

Рассмотрим компьютерную сеть с n узлами, задающими множество $S = \{s_1, ..., s_n\}$. Вначале будем предполагать, что все элементы доступны для атакующего, то есть модель сети на начальном этапе представляет собой полный граф $G(V, E), V = S, E = \cup_{i \neq j}(s_i, s_j), 1 \leq i, j \leq n$, а возможные сценарии атаки — маршруты в нём. В настоящей работе рассматриваются только сценарии, являющиеся простыми путями.

Предположим, что Атакующий хочет нанести рассматриваемой сети максимальный ущерб. Тогда он должен успешно атаковать все узлы без исключения, решив при этом задачу, обратную задаче (8):

$$\mathbf{M}^{-1}_{max} = \underset{M^{-1}}{\operatorname{Arg\,max}} \rho\left(S, W, M^{-1}\right) : \rho_{max} = \sum_{i=1}^{n} \rho_{M(v_i)} \forall M^{-1} \in \mathbf{M}^{-1}_{max}, \quad (9)$$

где $\mathbf{M}_{max}^{-1}$ — множество размещений, для каждого из элементов которого достигается максимальное значение интегрального риска риска ρ_{max}.

С учётом изложенного в предыдущем параграфе результата решение строится тривиальным образом и заключается в выборе простого пути $\left(v_1^A, v_2^A, ..., v_n^A\right)$, включающего все вершины модельного графа G, причём $v_i \succeq v_{i+1}\ \forall i < n$.

Задача Защитника, в свою очередь, заключается в том, чтобы направить Атакующего по наименее «выгодной» для последнего траектории. Эта траектория также легко вычисляется и, как и в предыдущем случае, представляет собой простой путь $\left(v_1^D, v_2^D, ..., v_n^D\right)$, включающий в себя все вершины графа G, причём $v_i \preceq v_{i+1}\ \forall i < n$. Отметим, что в случае, когда выполнено условие

$$\frac{1-p_i}{u_i p_i} = \frac{1-p_j}{u_j p_j} \iff i = j, i, j \in \{1, ..., n\}, \tag{10}$$

обе задачи имеют единственное решение, причём $v_i^D = v_{n-i+1}^A$. Иными словами, Атакующий и Защитник стремятся к реализации противоположных траекторий.

Тогда алгоритм решения задачи Защитника при выполнении условия (10) выглядит следующим образом:

1) Обеспечить единственный шлюз во внешнюю сеть (задать периметр) в узле v_1^D (он же v_n^A).
2) Назначить узел v_1^D текущим (положить i равным 1).
3) Последовательно удалять рёбра, соединяющие текущий узел v_i^D с узлами $v_n^D, v_{n-1}^D, ..., v_{i+1}^D$.
4) Если $i < n$, то назначить текущим узел v_{i+1}^D (положить i равным $i+1$).
5) Перейти к пункту 3.

Отметим, что при выполнении условия

$$\frac{1-p_i}{u_i p_i} = \frac{1-p_j}{u_j p_j}\ \forall i, j \in \{1, ..., , n\}, \tag{11}$$

узлы становятся нейтральными к перестановкам в смысле утверждения 1. В то же время, если Защитник знает о том, что в дальнейшем будет располагать неким ограниченным ресурсом, с помощью которого он сможет снижать удельные вероятности успешной атаки узлов, то у него появляется дополнительный критерий упорядоченности. А именно, Защитник будет заинтересован в том, чтобы вынести ближе к периметру узлы, наиболее отзывчивые к выделению ресурса в смысле повышения их стойкости к действиям Атакующего. В рамках настоящей работы эта задача пока не рассматривается, но представляется интересной для будущих исследований.

5. Заключение

В настоящей работе рассматривается задача поиска минимизирующей риска топологии при проектировании компьютерной сети. Содержательно полученный результат представляет собой правило размещения узлов в зависимости от их удалённости от периметра сети. Также предложен алгоритм проектирования топологии, минимизирующей риск относительно сценариев атаки, структурно являющимися простыми цепями.

Дальнейшим развитием работы видится изучение возможности формулировки правил построения топологий, минимизирующих риски при более сложных сценариях вероятных атак. Также представляет интерес решение задач динамической оптимизации топологии при атаках различного профиля.

ЛИТЕРАТУРА

1. Kavallieratos G., Spathoulas G., Katsikas S. Cyber risk propagation and optimal selection of cybersecurity controls for complex cyberphysical systems // Sensors. 2021. V. 21, No. 5, https://www.mdpi.com/1424-8220/21/5/1691.
2. Khouzani M.H., Liu Z., Malacaria P. Scalable min-max multi-objective cybersecurity optimisation over probabilistic attack graphs // European Journal of Operational Research. 2019. V. 278, No. 3. P. 894–903.
3. Sawik T. Selection of optimal countermeasure portfolio in IT security planning // Decision Support Systems. 2013. Vol. 55, No. 1. P. 156–164.
4. Zhang Y., Malacaria P. Bayesian Stackelberg games for cyber-security decision support // Decision Support Systems. 2021. V. 148. e113599.
5. Ait Temghart A., Outanoute M. H., Marwan M. Game Theoretic Approaches to Mitigate Cloud Security Risks: An Initial Insight // in Business Intelligence: 6th International Conference, CBI 2021, Beni Mellal, Morocco, May 27–29, 2021. Proceedings 2021 May 16. 2021. P. 335–347.
6. Graf J., Batchelor W., Harper S., Marlow R., Carlisle E., Athanas P. A practical application of game theory to optimize selection of hardware trojan detection strategies // Journal of Hardware and Systems Security. 2020. V. 4. P. 98–119.
7. Hiertz G. R, Denteneer D., Max S., Taori R., Cardona J., Berlemann L., Walke B. IEEE 802.11s: the WLAN mesh standard // IEEE Wireless Communications. 2010 V. 17, No. 1. P. 104–111.
8. Shiroky A. A., Kalashnikov A. O. Mathematical Problems of Managing the Risks of Complex Systems under Targeted Attacks with Known Structures // Mathematics. 2021. V. 9, No. 19, https://www.mdpi.com/2227-7390/9/19/2468.

UDC: 004.85

On the Evasion Attack Detector

Li Huayui, Vasily Kostyumov, Oleg Pilipenko, Dmitry Namiot[1]

[1]Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, 119991, Russian Federation

leesir1996@hotmail.com, kostyumov@yandex.ru, piligol1995@gmail.com, dnamiot@gmail.com

Abstract

The paper deals with the issue of detecting adversarial attacks on machine learning models. Such attacks are understood as deliberate (special) data changes at one of the stages of the machine learning pipeline, which is designed to either prevent the operation of the machine learning system, or vice versa, to achieve the desired result for the attacker. Contention attacks pose a great threat to machine learning systems because they do not guarantee the results and quality of the system. And such guarantees are, for example, mandatory for the use of a machine learning (artificial intelligence) system in critical areas such as avionics, automatic driving, special applications, etc. The article considers one of the possible detectors for the so-called evasion attacks.

Keywords: machine learning, adversarial attacks, explainable artificial intelligence, evasion attacks

1. Introduction

Machine learning systems (and, at least now, it is a synonym for artificial intelligence systems) depend on data. This tautological statement leads, in fact, to quite serious consequences. Changing the data then, generally speaking, changes the performance of the model. But, machine learning models are always trained on some subset of data (training set). And only then, at the operational stage, the model meets with real data from the general population. And these data may, in their characteristics, generally speaking, differ from those on which the model was trained and tested. That is, in the most usual (natural) way, a machine learning model, when working on real data, may not show the metrics that were achieved when the model was trained and confirmed during testing. But what if there is an opposing party for the model being used, which will purposefully look for such input data that affects the operation of the model? This issue can be especially acute for

critical applications (avionics, automatic driving, cybersecurity, etc.). Such targeted actions are called attacks on machine learning systems [1].

Adversarial attacks, which are possible for any discriminant machine learning models, pose a great threat to machine learning systems, since they do not guarantee the results and quality of the system. And such guarantees are, for example, mandatory for the use of a machine learning (artificial intelligence) system in critical areas such as avionics, automatic driving, special applications, etc. [2, 3].

Detection (determination of the fact of implementation) of such attacks is the subject of this work. The remainder of the article is structured as follows. In Section 2, we dwell on the taxonomy of attacks. Section 3 deals with attack detection. Section 4 presents the developed algorithm for detecting evasion attacks.

2. On the taxonomy of attacks

The following classification is quite general:

- Place and time of attack
- Knowledge about the attacked system
- Goals and objectives of attacks
- Application subject: digital or real objects
- Subject area (domain)

Attacks can be carried out at different stages: training the system (attacking the algorithm) or using it (attacking the model). Methods for carrying out attacks can be different. The goals of the attacker can be different - espionage (theft of information), sabotage (obstruction of work). All this can lead to different attacks.

Attacks can use evasions, poisoning, trojans (backdoors), reprogramming, and extracting hidden information. Dodge (commonly referred to as adversarial attacks) and poison are the most common attacks these days. Physical and digital attacks differ in the place of data change: real or digital objects. Attacks (harmful distortions) can differ in their goals: targeted (targeted) attacks, non-targeted or universal. For example, if we want to change the result of a particular classification, this is a targeted attack, if we want to worsen the performance of the classifier altogether, it is a non-targeted attack. Targeted attacks are more complex than non-targeted attacks, but a complete list might look like this:

1) the attacker is trying to influence the machine learning system (to prevent it from working correctly)

2) the attacker wants to achieve a special result in the work of the model

Or in another wording:

1) the attacker manipulates the data in order to influence the machine learning system (prevent it from working correctly)

2) the attacker manipulates the logic of the work in order to achieve a special result in the work.

If we talk about subject areas, then the following should be noted. All discriminant machine learning models are attackable. The biggest problem is obviously critical applications (avionics, automatic driving, etc.). Most applications in such areas are classifiers. According to the study of projects in the field of sustainable models [4], subject areas are described by the following list: Data (for critical applications):

- images (video) - classification,
- sound - distortion (change) of meaning, classification,
- text - classification
- time series - search for anomalies

Among the models, attacks on graph models of machine learning can be singled out separately.

Speaking about the significance of attacks, it can be noted that attacks by evasion (modification of input data) are potentially the most frequent. If the model requires input data to work, then you can try to modify it as needed. Attack schemes are considered in more detail in our work [4]. Evasion attacks on image classification systems are considered in detail in [5, 6]

3. On the detection of attacks

Speaking of adversarial attacks, most of the work focuses on the success of such attacks for given maximum changes (usually in one of the L-norms) [7], certified stability [8], and adversarial training [9].

In our work, we proceed from a situation where any changes to the model or its overtraining are not an option. For example, we proceed from the need to protect some ready-made MLaaS solutions. Evasion attacks are modifications of the input data that, for example, fool the classifier. Classifiers are mentioned because it is precisely this kind of system that is typical for critical applications. Such modifications are usually built as modifications (with a given change budget) of legitimate data. And the task of the detector is precisely to evaluate the input data in binary form - is it an adversarial instance of the data or not?

4. On the proposed algorithm

The evasion attack detection algorithm proposed in this paper is based on the difference in explaining how the model works for adversarial and "normal" examples. Approaches to explaining how machine (deep) learning models actually work are becoming increasingly important [20]. In many applications, the classic black box is not acceptable. In particular, we cannot, for example, talk about any evidence of the stability (reliability) of the model if we operate with a black box.

Currently, depending on the domain of interpretability, we can divide the methods of "model interpretability" into two different categories: "global interpretability" and "local interpretability" [10]. Global interpretability methods aim to reveal the behavioral characteristics of the model as a whole in order to better understand how the model makes decisions in various input situations. This type of interpretability focuses on the overall performance and characteristics of the model, which can help us analyze the generalizability, potential defects, and possible areas for improving the model. Local interpretation methods aim to focus on the predictive behavior of a model given specific inputs in order to better understand how the model makes a decision in a particular situation. Such methods can help analyze the performance of the model in individual cases, identify possible anomalies in behavior, and identify potential flaws in the model. One frequently used approach to local interpretation is the CAM family of methods (CAM, Grad-CAM, Grad-CAM++)[11]. It is an interpreted method for rendering model output. Because a convolutional neural network uses a sliding window (convolution kernel) to extract features from an input image, the extracted feature map often has some correlation with the input image in pixel space. We can say that the information expressed in the weight of the feature map can reflect the importance of the respective feature for network prediction.

In recent years, deep learning security research has increasingly focused on the interpretability of hostile (adversarial) examples. Since adversarial examples are algorithmically constructed as modifications of "pure" data, the interpretability of adversarial examples should account for differences in the processing of "normal" and adversarial data. Why is the adversarial example such? In [13, 14, 15, 16], various methods for studying this aspect are proposed. One of the simplest and most easily interpretable approaches is to use salience maps, attribution maps, and class activation maps to identify differences between normal and adversarial examples. These methods avoid the need to change the model and use the local interpretability of the model to explain the input.

We propose [18] a locally interpretable algorithm that combines two interpretable methods: "class activation maps" and "model attribution maps". In the initial experimental phase, several model attribution methods were compared, such as integral gradient, significance map [17], and controlled backpropagation. Controlled

backpropagation has a higher definition due to the processing of the model backpropagation gradient values, while the contours of the adversarial example maps are blurred to a certain extent due to the introduced perturbations.

For adversarial examples, the area of data that the model pays attention to is more scattered[19]. Thus, a normal sample class activation attribution map has a sharper contour and more focused areas of interest, while an adversarial sample class activation attribution map has a fuzzy contour and more disparate regions of interest, showing several different regions of interest in a class activation attribution map.

5. Conclusion

The paper presents a new algorithm for detecting adversarial evasion attacks based on the interpretation of the process of processing adversarial examples. The detector uses two different approaches to interpretability: "attribution maps" and "class activation maps". Using both the "outline" of the object in the image and the "region of interest" of the model, the differences between normal and adversarial examples are explored. Using the ResNet architecture, the detector was trained on a dataset containing class activation attribution maps for both normal and adversarial instances, resulting in a CNN model for adversarial instance detection after training. Experiments were performed on the ImageNet dataset using a pre-trained ResNet50 model as a threat model. The results show that the detector handles various types of adversarial examples with small perturbation coefficients well and exhibits high tolerance. Thus, the proposed algorithm has been tested on real attacks, the implementation code has been made publicly available.

REFERENCES

1. Ilyushin, Eugene, Dmitry Namiot, and Ivan Chizhov. "Attacks on machine learning systems-common problems and methods." International Journal of Open Information Technologies 10.3 (2022): 17-22. (in Russian)
2. Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov. "The rationale for working on robust machine learning." International Journal of Open Information Technologies 9.11 (2021): 68-74. (in Russian)
3. Namiot, Dmitry, and Eugene Ilyushin. "On the robustness and security of Artificial Intelligence systems." International Journal of Open Information Technologies 10.9 (2022): 126-134. (in Russians)
4. Namiot, Dmitry. "Schemes of attacks on machine learning models." International Journal of Open Information Technologies 11.5 (2023): 68-86. (in Russian)
5. Kostyumov, Vasily. "A survey and systematization of evasion attacks in computer vision." International Journal of Open Information Technologies 10.10 (2022): 11-20. (in Russian)

6. Huayu, Li, and Namiot Dmitry. "A Survey of Adversarial Attacks and Defenses for image data on Deep Learning." International Journal of Open Information Technologies 10.5 (2022): 9-16.
7. Zhou, Shuai, et al. "Adversarial Attacks and Defenses in Deep Learning: From a Perspective of Cybersecurity." ACM Computing Surveys 55.8 (2022): 1-39.
8. Silva, Samuel Henrique, and Peyman Najafirad. "Opportunities and challenges in deep learning adversarial robustness: A survey." arXiv preprint arXiv:2007.00753 (2020).
9. Bai, Tao, et al. "Recent advances in adversarial training for adversarial robustness." arXiv preprint arXiv:2102.01356 (2021).
10. Molnar, Christoph. Interpretable machine learning. Lulu. com, 2020.
11. Zhou, Bolei, et al. "Learning deep features for discriminative localization." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
12. Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017.
13. J. Gu and V. Tresp, "Saliency methods for explaining adversarial attacks," arXiv preprint arXiv:1908.08413, 2019.
14. Mangla, Puneet, Vedant Singh, and Vineeth N. Balasubramanian. "On saliency maps and adversarial robustness." Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part II. Springer International Publishing, 2021.
15. Fidel, Gil, Ron Bitton, and Asaf Shabtai. "When explainability meets adversarial learning: Detecting adversarial examples using shap signatures." 2020 international joint conference on neural networks (IJCNN). IEEE, 2020.
16. Amosy, Ohad, and Gal Chechik. "Using explainabilty to detect adversarial attacks." (2020).
17. Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." arXiv preprint arXiv:1312.6034 (2013).
18. Evasion Attacks Detector https://github.com/lhy19961016/Master_Degreetree/main/Code Retrieved: Jul, 2023
19. Zhang, Haichao, and Jianyu Wang. "Defense against adversarial attacks using feature scattering-based adversarial training." Advances in Neural Information Processing Systems 32 (2019).
20. Vaccari, Ivan, et al. "eXplainable and reliable against adversarial machine learning in data analytics." IEEE Access 10 (2022): 83949-83970.

UDC: 004.85

On the Machine Learning Models Inversion Attack Detector

Junzhe Song[1] and Dmitry Namiot[1]

[1]Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, 119991, Russian Federation

songjz@smbu.edu.cn, dnamiot@gmail.com

Abstract

This article is devoted to the detection of adversarial attacks on machine learning models. In the most general case, adversarial attacks are special data changes at one of the stages of the machine learning pipeline, which are designed to either prevent the operation of the machine learning system, or vice versa, to achieve the desired result for the attacker. But there is also a form of attack aimed at extracting non-public information from machine learning models. These include model inversion attacks. These types of attacks pose a threat to the use of machine learning as a service (MLaaS). Machine learning models accumulate a lot of redundant information during training, and the possibility of exposing this data while using the model can come as an unpleasant surprise.

Keywords: machine learning, adversarial attacks, model inversion

1. Introduction

Machine learning systems (and, at least now, it is a synonym for artificial intelligence systems) depend on data. This tautological statement leads, in fact, to quite serious consequences. Changing the data then, generally speaking, changes the performance of the model. Purposeful data changes are attacks on machine learning models [1]. But the models themselves can be directly affected during attacks. For example, weights can change on the fly, malicious code can be loaded into weights, etc. Adversarial attacks, which are possible for any discriminant machine learning models, pose a great threat to machine learning systems, since they do not guarantee the results and quality of the system. And such guarantees are, for example, mandatory for the use of a machine learning (artificial intelligence) system in critical areas such as avionics, automatic driving, special applications, etc. [2, 3]. An attack directly on the model also carries additional risks of extracting private information stored in machine learning models [4]. Detection (determination of the

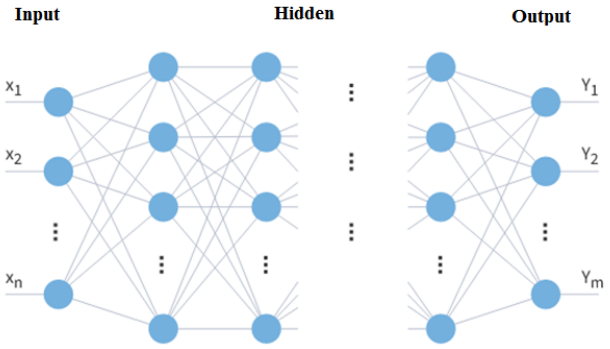


Fig. 1. Multilayer perceptron

fact of implementation) of attacks related to the extraction of information is the subject of this work. The article was written based on the results of the master's thesis, performed at Lomonosov Moscow State University.

The remainder of the article is structured as follows. In section 2, we focus on attacks on intellectual property. Section 3 presents the developed algorithm for detecting extraction attacks. Section 4 presents the conclusion.

2. On data extraction from machine learning models

In some classifiers, these types of attacks are also referred to as attacks aimed at stealing intellectual property. With the help of such attacks, one can, for example, restore the algorithm of the model or obtain various information about the training data. The American Standards Institute NIST in its glossary [5] defines 5 types of such attacks:

- Data Reconstruction
- Memorization
- Membership Inference
- Model Extraction
- Property Inference

The Model Extraction attack is perhaps the most understandable in its logic [14]. If we have the ability to interrogate the model, then we can accumulate a set of inputs and outputs <x, Y> and use this set as a training dataset to create a shadow model. This is one of the easiest ways to replicate the functionality of an existing model. In this regard, we can mention a multilayer perceptron, which just solves such problems (Fig. 1), by selecting hidden layers. We also note that all attacks of

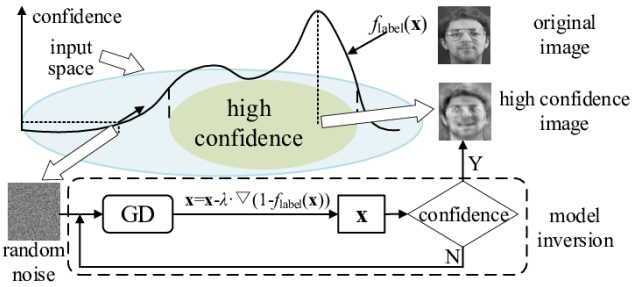


Fig. 2. Model inversion [8]

this class depend on the possibility of multiple polling of models. First of all, they are focused on attacks on MLaaS (machine learning as a service) systems [15].

Data reconstruction attacks should be recognized as the most serious in terms of access to private attributes. These attacks try to restore the input (training) data of the attacked model based on the results of its work [6]. Another name is model inversion attacks [7].

The idea is actually quite transparent - the data that was in the training set should be recognized better (the peak of the graph in Fig. 2) than those that were not in the set. Memorization attacks are a class of techniques that allow an attacker to extract training data from generative machine learning models such as language models [9]. Generalization and memorization in machine learning models are coupled, and neural networks can remember randomly selected datasets: deep learning models (in particular, generative models) often remember rare details about the training data that are completely unrelated to the task at hand. This "extra" data becomes the target of the attack. Membership Inference attacks are aimed at determining whether a particular record or data sample was part of the training data set [10]. As a rule, such attacks are adapted to be executed in the black box mode (Fig. 3).

In Property Inference attacks, the attacker tries to learn global information about the distribution of the training data. The goal is to disclose confidential information about the training sample (for example, dependence on some attributes, etc.) [11].

3. On the model inversion attack detection

The Fig. 4 shows a generic form of the model inversion attack, where the attacker wishes to infer sensitive information related to the class represented by *label*. Let $\tilde{f}_{label}(\mathrm{x})$ represents the output confidence that the input vector x matches the target class, represented by *label*. E.g., in the case of facial recognition, *label* presents the name (ID) of a person whose facial features are the goal of an attack. The algorithm receives a vector $\mathrm{x} = (x_0, x_1, .., x_n)$, where some of the values may be non-sensitive

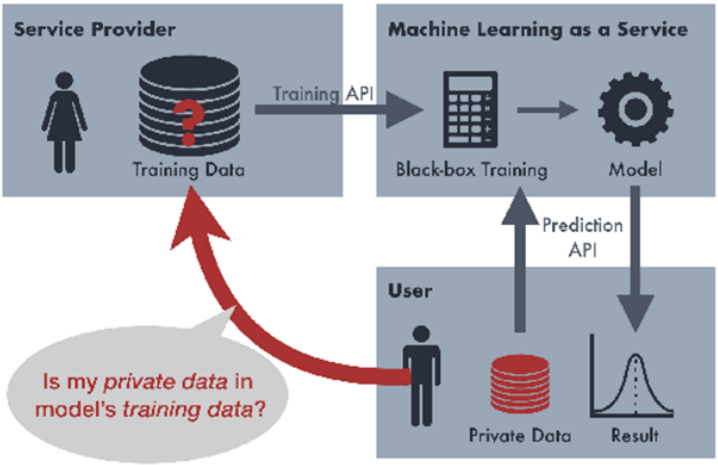


Fig. 3. Membership inference [11]

Algorithm 1 General Model Inversion Attack

Require: Initial vector x with non-sensitive features and random value for sensitive features
Ensure: Output vector x with inferred sensitive features
while $\tilde{f}_{label}(\mathbf{x}) <$ THRESHOLD **do**
$\mathbf{x} \leftarrow \mathbf{x} + \alpha \cdot \nabla \tilde{f}_{label}(\mathbf{x})$

Fig. 4. A typical model inversion attack [13]

data that the attacker could obtain about a target user and unknown sensitive data (they are initialized with any values). The goal of the attack is to maximize confidence (or minimize error) in the prediction according to the given threshold. This is achieved in a standard way for machine learning, using gradient $\nabla \tilde{f}_{label}(\mathrm{x})$ movement with some step α.

Our Model Inversion attack detection technique is to find the difference between the results of the (n)-th and (n + 1)-th iterations. Initially, we define a window in which ten image requests can be placed. When the user submits the second request, we start calculating the L2 norms. If the L2-norm value is less than 100 (this threshold may differ in different datasets), we can identify this query as a possible attack behavior. We then calculate the cosine similarity for each of these suspicious queries. If the cosine similarity consistently exceeds 0.9, this can be classified as a

```
Algorithm  Detector for Model Inversion Attack
function Detector(inputs:list, window:int, L2_max, cosine_max)
    L2_value = []
    cosine_value = []
    if len(inputs)>=2 then
        L2_value.append(L2_norm(inputs[-1],inputs[-2]))
        cosine_value.append(cosine_similarity(inputs[-1],inputs[-2]))
    end if
    if in L2_value has continuous window items >= L2_max then
        if in cosine_value has continuous window items >= cosine_max then
            print('This is Model Inversion Attack')
        end if
    end if
end function
```

Fig. 5. The proposed model inversion attack detector

model inversion attack given that the average user rarely repeats the same query for the same image more than ten times. The pseudo-code is presented in Fig. 5.

4. Conclusion

The paper presents a new algorithm for detecting inversion attacks based on the evaluation of successive requests. The idea of attack detection is that during an inversion attack, the attacker will sequentially refine the solution, and, accordingly, successive requests will be similar to each other. Detection of the fact of small regular changes in successive requests serves as a signal that an attack is being carried out. The proposed algorithm has been tested on real attacks, and the implementation code has been made publicly available [16].

REFERENCES

1. Ilyushin, Eugene, Dmitry Namiot, and Ivan Chizhov. "Attacks on machine learning systems-common problems and methods." International Journal of Open Information Technologies 10.3 (2022): 17-22. (in Russian)
2. Kostyumov, Vasily. "A survey and systematization of evasion attacks in computer vision." International Journal of Open Information Technologies 10.10 (2022): 11-20. (in Russian)
3. Namiot, Dmitry, and Eugene Ilyushin. "On the robustness and security of Artificial Intelligence systems." International Journal of Open Information Technologies 10.9 (2022): 126-134. (in Russians)

4. Namiot, Dmitry. "Schemes of attacks on machine learning models." International Journal of Open Information Technologies 11.5 (2023): 68-86. (in Russian)
5. White Paper NIST AI 100-2e2023 (Draft) Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations https://csrc.nist.gov/publications/detail/white-paper/2023/03/08/adversarial-machine-learning-taxonomy-and-terminology/draft Retrieved: Jul, 2023
6. Malekzadeh, Mohammad, and Deniz Gunduz. "Vicious Classifiers: Data Reconstruction Attack at Inference Time." arXiv preprint arXiv:2212.04223 (2022).
7. Song, J., Namiot, D. (2023). A Survey of the Implementations of Model Inversion Attacks. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds) Distributed Computer and Communication Networks. DCCN 2022. Communications in Computer and Information Science, vol 1748. Springer, Cham.
8. Zhang, Jiliang, et al. "Privacy threats and protection in machine learning." Proceedings of the 2020 on Great Lakes Symposium on VLSI. 2020.
9. Carlini, Nicholas, et al. "The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks." USENIX Security Symposium. Vol. 267. 2019.
10. Hisamoto, Sorami, Matt Post, and Kevin Duh. "Membership inference attacks on sequence-to-sequence models: Is my data in your machine translation system?." Transactions of the Association for Computational Linguistics 8 (2020): 49-63.
11. De Cristofaro, Emiliano. "An overview of privacy in machine learning." arXiv preprint arXiv:2005.08679 (2020).
12. Junzhe, S. A Survey of Model Inversion Attacks and Countermeasures / S. Junzhe, D. E. Namiot // Proceedings of the Institute for Systems Analysis Russian Academy of Sciences. – 2023. – Vol. 73, No. 1. – P. 82-93. – DOI 10.14357/20790279230110. – EDN MQODQW.
13. Alves, Tiago AO, Felipe MG França, and Sandip Kundu. "MLPrivacyGuard: Defeating confidence information based model inversion attacks on machine learning systems." Proceedings of the 2019 on Great Lakes Symposium on VLSI. 2019.
14. Gong, Xueluan, et al. "Model extraction attacks and defenses on cloud-based machine learning models." IEEE Communications Magazine 58.12 (2020): 83-89.
15. Miao, Yuantian, et al. "Machine learning–based cyber attacks targeting on controlled information: A survey." ACM Computing Surveys (CSUR) 54.7 (2021): 1-36.
16. MI detector Codes and results are available at https://github.com/UnReAlKiNg/detection- of-model-inversion-attack Retrieved: Jul, 2023

УДК: 519.872

Модель работы процессора в условиях конкуренции за вычислительный ресурс

И.Л.Лапатин, А.А.Назаров, С.В.Пауль

Национальный исследовательский Томский государственный университет,
пр. Ленина, 36, Томск, Россия

lapatin@mail.ru, nazarov.tsu@gmail.com, paulsv82@mail.ru

Аннотация

В статье рассматривается модель работы вычислительного процессора в виде системы массового обслуживания с неограниченным числом приборов и деградацией скорости обслуживания. Деградация скорости обслуживания означает функциональную зависимость интенсивности обслуживания поступающих запросов от общего числа запросов в системе. Введение в модель такой функции позволяет учитывать снижение производительности процессора при увеличении нагрузки на него. В результате модификации и применения метода асимптотического анализа была получена Гауссовская аппроксимация распределения вероятностей числа заявок в системе, что соответствует распределению одновременно выполняемых задач процессором.

Ключевые слова: конкуренция, функция деградации, теория массового обслуживания, метод асимптотического анализа

1. Введение

Физические вычислительные машины могут иметь различный масштаб от персонального компьютера с несколькими ядрами в процессоре до огромных серверов и Data-центров. И для всех этих случаев наблюдается явление снижения производительности при увеличении нагрузки на них. Под увеличением нагрузки подразумевается увеличения одновременно выполняемых задач. Освещение этой проблемы в литературе прежде всего связано с изучением работы облачных узлов [1], [2] и работы на них разного числа виртуальных машин. Моделирование скорости обслуживания таких систем должно включать зависимость от количества клиентов, работающих с облачным узлом. В данной работе мы предлагаем рассмотреть модель работы не облачного узла, где виртуальные машины могут менять режимы своей работы, а модель отдельного процессора, который

Исследование выполнено при поддержке Программы развития Томского государственного университета (Приоритет-2030)

выполняет посылаемые на него задачи и среди них происходит конкуренция за вычислительный ресурс.

Моделирование работы вычислительных машин можно проводить с помощью методов теории массового обслуживания. Это позволяет учитывать стохастическую природу исследуемого объекта, так как моменты запуска виртуальных машин в облаке, запуска очередного процесса на персональном компьютере происходят в случайные моменты времени, а продолжительность работы виртуальной машины или время выполнения задачи на компьютере является недетерминированной величиной. Кроме этого, результаты таких исследование позволяют делать вероятностные выводы (например, находить квантили) по изучаемым характеристикам.

Существенным усложнением модели для описания работы вычислительных машин является необходимость учитывать зависимость скорости обслуживания запросов от общего числа запросов в системе. Наиболее популярным типом моделей, зависящей от состояния скорости обслуживания, является ступенчато-убывающая. Такой подход к моделированию используется в работах [3], [4]. Другой метод заключается в использовании моделирования скоростей обслуживания, скоростей поступления и времени ожидания (до обслуживания) как зависимых случайных величин [5].

Мы предлагаем использовать так называемую функцию деградации скорости обслуживания, которая каждому значению числа заявок в системе ставит в соответствие коэффициент снижения скорости обслуживания. При таком подходе можно моделировать различные ситуации чувствительности скорости обслуживания от числа заявок в системе.

Заметим, что определение конкретного вида функции деградации является отдельной задачей и не входит в цели данной работы.

Для моделирования работы процессора будем использовать систему массового обслуживания с неограниченным числом приборов. Входящий поток описывает некоторый агрегированный поток задач, который поступает на процессор. Каждый занятый прибор будем отождествлять с выполняемой задачей. Чем больше приборов занято, тем больше задач одновременно выполняется. Для моделирования снижения скорости обслуживания будем использовать функцию деградации, которая фактически меняет интенсивность обслуживания в зависимости от числа заявок в системе. Таким образом, случайность времени выпонения запросов порождается разным объемом поступающих задач и переменной скоростью их выполнения. Исследование будем проводить методом асимптотического анализа.

2. Описание математической модели

Рассмотрим модель массового обслуживания с неограниченным числом приборов (Рисунок 1).

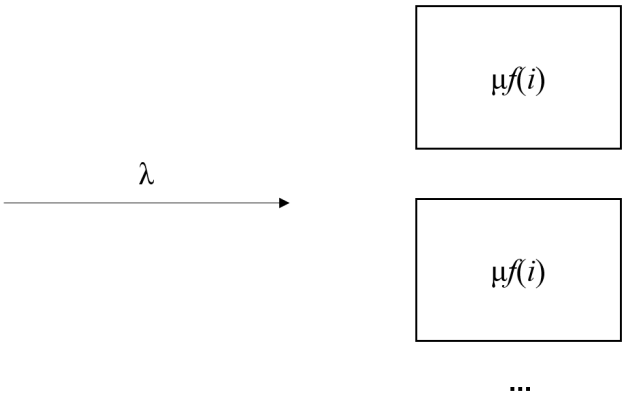


Рис. 1. Математическая модель

На вход системы поступает простейший поток заявок с параметром λ. Каждая заявка, которая поступает в систему, мгновенно начинает обслуживание. Время обслуживания распределено по экспоненциальному закону, но параметр распределения зависит от числа заявок в системе. Интенсивность обслуживания можно записать как $\mu f(i)$, где μ – интенсивность обслуживания единственной заявки в системе (без конкуренции), а $f(i)$ – функция деградации, значения которой являются безразмерным коэффициентом снижения интенсивности обслуживания в зависимости от числа заявок i в системе. Чем больше заявок одновременно находится в системе, тем ниже интенсивность их обслуживания.

Функционирование системы определятся Марковским случайным процессом $i(t)$ – число заявок в системе в момент времени t. Для распределения вероятностей $P(i,t) = P\{i(t) = i\}$ можно записать систему дифференциальных уравнений Колмогорова

$$\frac{\partial P(i,t)}{\partial t} = -(\lambda + i\mu f(i))P(i,t) + (i+1)\mu f(i+1)P(i+1,t) + \lambda P(i-1,t). \quad (1)$$

Запишем ее в стационарном виде

$$-(\lambda + i\mu f(i))P(i) + (i+1)\mu f(i+1)P(i+1) + \lambda P(i-1) = 0. \quad (2)$$

Систему (2) будем решать методом асимптотического анализа в предельном условии высокоинтенсивного входящего потока.

3. Метод асимптотического анализа

Предельное условие высокоинтенсивного входящего потока определяется бесконечно большим параметром $T \to \infty$ и равенствами

$$\lambda = \lambda_1 T, \quad f(i) = f_1\left(\frac{i}{T}\right). \tag{3}$$

Величину T будем называть параметром высокой интенсивности входящего потока. Параметр λ_1 – фиксированная величина. На практике интенсивность $\lambda_1 T$ рассматривается как единая величина с достаточно большими, но конечными значениями [6]. С учетом сделанных обозначений, систему (2) перепишем в виде

$$-\left(\lambda_1 + \frac{i}{T} f_1\left(\frac{i}{T}\right)\right) P(i) + \frac{i+1}{T} f_1\left(\frac{i+1}{T}\right) P(i+1) + \lambda_1 P(i-1) = 0. \tag{4}$$

Обозначим некоторую неотрицательную величину $\varepsilon = \frac{1}{T}$ и в системе (4) сделаем следующие замены

$$\frac{i}{T} = i\varepsilon = x, \quad P(i) = P_1(x, \varepsilon), \tag{5}$$

имеем

$$-\left(\lambda_1 + x f_1(x)\right) P_1(x, \varepsilon) + (x+\varepsilon) f_1(x+\varepsilon)) P_1(x+\varepsilon, \varepsilon) + \lambda_1 P_1(x-\varepsilon, \varepsilon) = 0. \tag{6}$$

Реализуя метод асимптотического анализа в предельном условии высокоинтенсивного входящего потока, решая систему уравнений (6), получим предельную гауссовскую плотность распределения вероятностей $i(t)$ числа занятых приборов в исследуемой системе, которая определяется равенством

$$p(x) = \frac{1}{\sqrt{2\pi\kappa_2}} e^{-\frac{(x-\kappa_1)^2}{2\kappa_2}} \tag{7}$$

где κ_1 является единственным решением уравнения

$$\lambda_1 - \kappa_1 f_1(\kappa_1) = 0. \tag{8}$$

велична κ_2 определяется равентсвом

$$\kappa_2 = \lambda_1 \left(\mu \frac{\partial}{\partial x}(x f_1(x))\right)^{-1}. \tag{9}$$

Для получения дискретного распределения вероятностей рассматриваемого процесса $i(t)$ будем использовать следующую формулу

$$p_{discret}(i) = \frac{p(i)}{\sum\limits_{i=0}^{\infty} p(i)}, \quad i \geq 0. \tag{10}$$

Допредельное гауссовское распределение вероятностей $p(i)$ определяется средним числом $\kappa_1 T$ приборов, занятых обслуживанием заявок в системе и дисперсией $\kappa_2 T$.

4. Численный пример

Рассмотрим следующие значения параметров модели: $\lambda = 50$ — среднее число задач, которые отправляются на процессор в единицу времени (например, в час), то есть интенсивность поступления задач; $\mu = 10$ (час)$^{-1}$ — величину, обратную среднему времени выполнения запроса при отстутсвии других задач на процессоре; параметр высокой интенсивнсти входящего потока $T = 100$; функцию деградации скорости обслуживания определим в виде неотрицательной убывающей функции

$$f(i) = \frac{1}{\sqrt{1+i}}. \tag{11}$$

На рисунке 2 представлена гауссовская аппроксимация распределения вероятностей числа занятых приборов в исследуемой системе с учетом функции деградации вида (11), заданное средним числом занятых приборов в системе $\kappa_1 T = 5$ и дисперсией $\kappa_2 T = 5$.

Получение результата в виде оценки распределения вероятностей числа заявок в системе позволяет делать выводы, например, о среднем числе одновременно выполняемых задач на процессоре или о числе задач, которые не будет превышено с некоторой вероятностью q. Подставляя полученные значения в функцию деградации, мы можем оценивать, какая производительность выполнения задач была в среднем или ниже какого уровня производительность не опускалась вероятностью q.

5. Заключение

В работе рассмотрена система массового обслуживания с неограниченным числом приборов и деградацией скорости обслуживания в зависимости от числа заявок в системе. Такая система может быть полезна для моделирования реальных систем, в которых рост числа клиентов приводит к снижению производительности для каждого клиента. В частности, в данной работе рассматривалась

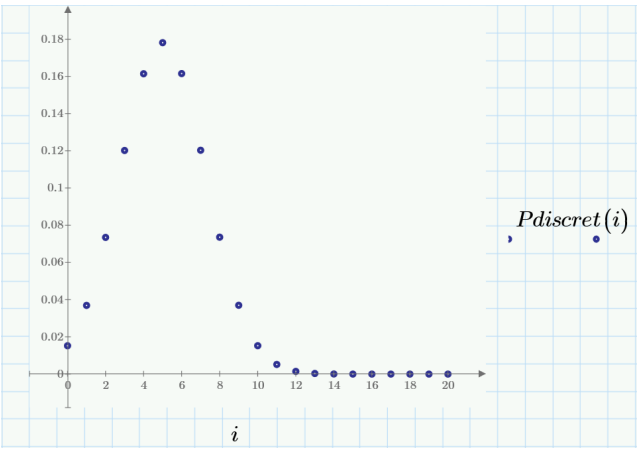


Рис. 2. Гауссовская аппроксимация распределения вероятностей числа занятых приборов в системе

модель работы процессора в условиях конкуренции задач за вычислительный ресурс. В результате исследования получена гауссовская аппроксимация распределения вероятностей числа занятых приборов в системе с учетом деградации скорости обслуживания.

Литература

1. Huber, N., Quast M. V., Brosig F., Hauck M., & Kounev S. A method for experimental analysis and modeling of virtualization performance overhead. In International Conference on Cloud Computing and Services Science. 2011. P. 353–370.
2. Bermejo B., Juiz C. A general method for evaluating the overhead when consolidating servers: performance degradation in virtual machines and containers. The Journal of Supercomputing. 2022. P. 1–28.
3. Liu X., Li S., Tong W. A queuing model considering resources sharing for cloud service performance. The Journal of Supercomputing. 2015. 71(11). P. 4042–4055.
4. Choudhary A., Chakravarthy S. R., Sharma D. C. Analysis of MAP/PH/1 Queueing System with Degrading Service Rate and Phase Type Vacation. Mathematics. 2021. 9(19). P. 2387.
5. Morozov E. A general multiserver state-dependent queueing system. Centre de Recerca Matemàtica. 2010. 927. – 17 p.
6. Моисеев А. Н., Назаров А.А. Бесконечнолинейные системы и сети массового обслуживания. – Томск : Изд-во НТЛ, 2015. – 240 с.

UDC: 004.7

Semi-orthogonal Precoder for Improving Throughput and Fairness in Downlink NOMA-MIMO Systems

I.A. Levitsky, S.A. Tutelian, A.A. Kureev , E.M. Khorov

Institute for Information Transmission Problems of the Russian Academy of Sciences, Moscow, Russia

[levitsky,tutelian,kureev,khorov]@wireless.iitp.ru

Abstract

In Non-Orthogonal Multiple Access (NOMA) systems, the base station can simultaneously transmit data from an antenna to a group of users. In this group, a user with higher channel gains performs Successive Interference Cancellation (SIC) to mitigate interference from the signal destined for the user with a weak channel. Typically, downlink NOMA-MIMO (Multiple-Input Multiple-Output) systems use precoders that form beams focused on users with higher channel gains of each group. This approach induces interference on the users with low gain, degrading throughputs and fairness in the network. This paper proposes a novel semi-orthogonal precoder that forms non-orthogonal beams for users within each group while maintaining orthogonality between the groups. Simulations show the efficiency of this approach in improving the geometric mean user throughput.

Keywords: Fairness, MIMO, NOMA, precoder construction, Zero Forcing

1. Introduction

Downlink Non-Orthogonal Multiple Access (NOMA) allows a base station (BS) to transmit data to several users, using the same time and frequency resources. It is achieved by superposing the messages intended for different users in the power domain. Users with higher channel gains employ the Successive Interference Cancellation (SIC) technique to effectively mitigate inter-user interference [1]. This idea can be extended to Multiple-Input Multiple-Output (MIMO) systems [2]. In NOMA-MIMO systems, the users are split into several groups, each of which contains users with highly-correlated channels. The number of groups may equal the number of transmitter antennas, which corresponds to the number of transmit beams. Each

The research has been carried out at IITP RAS and supported by the Russian Science Foundation (Grant No 21-19-00846, https://rscf.ru/en/project/21-19-00846/)

group is allocated a specific beam, and within each group, the users apply NOMA as described above. Many studies [3,4] consider a case with two users per group, namely the cell-edge user and the cell-center user. Generally, the BS builds a precoder that forms beams focused on the cell-center user to increase aggregated throughput performance. However, this approach leads to significant inter-group interference affecting the cell-edge user in the same group. Consequently, their throughput degrades, and resource allocation becomes unfair.

To address this challenge, the paper proposes a novel approach where the BS constructs a semi-orthogonal precoder that forms non-orthogonal beams for users within each group while maintaining orthogonality between the groups. With simulations, we show that this approach improves the geometric mean of user throughput in the downlink NOMA-MIMO system.

The rest of the paper is organized as follows. Section 2 describes the system model. Section 3 describes the semi-orthogonal precoder. In Section 4, we explain the simulation scenario and provide numerical results. Section 5 concludes the paper.

2. System model

Consider a downlink multiuser NOMA-MIMO system, where a BS communicates with a set of users $\hat{\mathcal{K}}$ with cardinality $|\hat{\mathcal{K}}| = \hat{K}$. The BS has M antennas, whereas each user has a single antenna. The BS intends to send data to all users, but it may choose a subset of users to service at a time. Thus, BS selects a *configuration*, which is a tuple $(\mathcal{K}, f)$ that includes a subset $\mathcal{K} \subset \hat{\mathcal{K}}$ of $|\mathcal{K}| = K$ users, and a grouping function f that subdivides $\mathcal{K}$ into G mutually non-overlapping subsets $\mathcal{K}_g$. The grouping function f takes an index of a user k and returns the group number g and the user's index i inside this group*:

$$f : \mathcal{K} \to (\mathcal{G}, \mathcal{K}_g), \;\; \mathcal{G} = \{1, ..., G\}, \;\; \mathcal{K}_g = \{1, ..., K_g\}. \tag{1}$$

For simplicity, let $K_g \leq 2$. $K_g > 2$ rarely notably improves the performance [5]. If $K_g = 2$, we refer to the users as cell-edge and cell-center ones. The system model can be easily adapted to include an arbitrary maximum number of users per group. The BS sends data to users in $\mathcal{K}$, encoded into the symbol s_k for the k-th user. The symbols s_k are independent and normalized. Thus, BS forms a vector $\mathbf{s} = [s_1, ...s_K]^T \in \mathbb{C}^{K\times 1}$.

To send data to the user k, the BS applies precoder $\mathbf{p}_k \in \mathbb{C}^{M\times 1}$. Hence, for all simultaneously served users, the whole precoder matrix is $\mathrm{P} = [\mathbf{p}_1, ..., \mathbf{p}_K] \in \mathbb{C}^{M\times K}$. The resulting transmitted signal $\mathbf{x} \in \mathbb{C}^{M\times 1}$ in a slot equals $\mathbf{x} = \mathrm{P}\mathbf{s}$, with the average

*Thanks to f, we can switch between global user indexing k and in-group indexing $f(k) = (g, i)$: $a_k \equiv a_{f(k)} \equiv a_{g,i}$, where a is any user property.

total power E_{max}. The received signal for each y_k is modeled as

$$y_k = \mathbf{h}_k^H \sum_{l=0}^{K} \mathbf{p}_k s_k + n_k, \tag{2}$$

where $\mathbf{h}_k \in \mathbb{C}^{M\times 1}$ represents a channel coefficient vector for user k. The noise component n_k is a circularly symmetric complex Gaussian random variable with zero mean and variance σ^2, and they are i.i.d. for all k. Without loss of generality, grouping f indexes users so the cell-edge user comes first: $||\mathbf{h}_{g,1}|| < ||\mathbf{h}_{g,2}||$.

To receive symbol s_k, the user k either performs SIC or treats the interference as noise. We assume an ideal SIC, where cell-center users can always decode data for cell-edge users. If a user i in a group g decodes symbol $s_{g,j}$, the signal-to-interference-plus-noise ratio (SINR) equals

$$\gamma_{g,(j\to i)} = \frac{|\mathbf{h}_{g,i}^H \mathbf{p}_{g,j}|^2}{|\mathbf{h}_{g,i}^H \mathbf{p}_{g,i}|^2 \cdot \mathbb{I}[j=1] + \sum_{k\notin\mathcal{K}_g} |\mathbf{h}_{g,i}^H \mathbf{p}_k|^2 + \sigma^2}, \tag{3}$$

where the first term of the denominator represents the intra-group interference (from the signal for the cell-center user), the second one is the inter-group interference, and the third one is the Gaussian noise variance. The value of $\gamma_{g,(j\to i)}$ shall be high enough to allow the cell-center user to decode the message for the cell-edge user and perform SIC successfully. The corresponding achievable rate for user i when decoding its data is $R_{g,i} = \log_2\left(1 + \gamma_{g,(i\to i)}\right)$.

3. Precoder Design

Let K users are selected, and groups $\mathcal{G} = \{1, ..., G\}$ are formed. Given the channel state information $\mathbf{h}_k$ for each k-th user, the precoder matrix is $\tilde{\mathrm{P}} = [\tilde{\mathbf{p}}_1, ..., \tilde{\mathbf{p}}_K]$, where $\tilde{\mathbf{p}}_k$ is the precoder vector for user k. To reallocate the power among the users, the precoder can be multiplied $\tilde{\mathrm{P}}$ by a diagonal matrix D: $\mathrm{P} = \tilde{\mathrm{P}}\mathrm{D}$. For simplicity in this short paper, we allocate power to a group g proportional to K_g. Inside each group g, we further search for the best power allocation coefficients $\alpha_{g,i}$, as described in Section 4.

3.1. Zero Forcing precoder. For non-NOMA systems, a widely used precoder is zero-forcing (ZF). Given the channel matrix $\mathrm{H} = [\mathbf{h}_1, ..., \mathbf{h}_K]$, the precoder equals: $\tilde{\mathrm{P}} = [\tilde{\mathbf{p}}_1, ..., \tilde{\mathbf{p}}_K] = \mathrm{ZF}(\mathrm{H}) \equiv \mathrm{H}(\mathrm{H}^H\mathrm{H})^{-1}$. The main property of the ZF precoder $\mathrm{ZF}([\mathbf{h}_1, ..., \mathbf{h}_i]) = [\mathbf{p}_1, ..., \mathbf{p}_i]$is that $\mathbf{h}_i^H \mathbf{p}_j = 0 \ \forall i \neq j$.

The downside of the ZF precoder is that the effective channel gain for some users is low if they have correlated channel vectors. Moreover, K cannot exceed the number of antennas M at the BS.

3.2. State-of-the-art NOMA-MIMO precoder. The widely-used state-of-the-art (SOTA) NOMA-MIMO precoder forms the precoder matrix based on the channel state information of the cell-center user in each two-user group. Let $\hat{\mathrm{H}} = [\mathbf{h}_{1,K_1}, ..., \mathbf{h}_{g,K_g}, ..., \mathbf{h}_{G,K_G}]$. Then, the precoder equals $\hat{\mathrm{P}} = [\hat{\mathbf{p}}_1, ..., \hat{\mathbf{p}}_g, ..., \hat{\mathbf{p}}_G] = \mathrm{ZF}(\hat{\mathrm{H}}) = \hat{\mathrm{H}}(\hat{\mathrm{H}}^H\hat{\mathrm{H}})^{-1}$, where $\hat{\mathbf{p}}_g$ is the precoder vector for the group g. As all users in the group get the same precoder vector, respective columns in $\hat{\mathrm{P}}_{M\times G}$ are duplicated to get $\tilde{\mathrm{P}}_{M\times K}$. Note that inside two-user groups, the SOTA precoder cancels inter-group interference on cell-center users only. The signal quality on cell-edge users deteriorates due to not only misaligned precoding but also inter-group interference.

3.3. Semi-orthogonal precoder. Proposed in this paper semi-orthogonal precoder is formed as follows. Consider a user k from a group $\mathcal{K}_g$ and all users from other groups, we build a precoding vector for this user k:

$$\tilde{\mathbf{p}}_k = \mathrm{ZF}([\mathbf{h}_k, \mathbf{h}_{l_1}, ..., \mathbf{h}_{l_{K-K_g}}])[1, 0, ..., 0]^T_{1\times K-K_g+1} \ \forall l_j \notin \mathcal{K}_g : k \in \mathcal{K}_g. \quad (4)$$

Performing this procedure for each user, we get $[\tilde{\mathbf{p}}_1, ..., \tilde{\mathbf{p}}_K] = \tilde{\mathrm{P}}$.

This precoder allows us to eliminate inter-group interference. Also, the cell-edge user receives a better signal than with the SOTA precoder. To construct this precoder, M shall not exceed $K - K_g + 1$.

Note that if all groups have only one user, the precoder becomes equal to ZF.

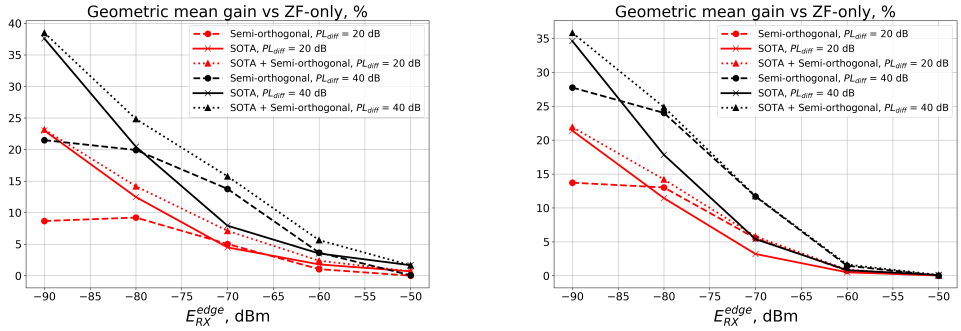
4. Numerical Results

To compare the performance of the proposed precoder against SOTA, we use simulation. Because of the paper size limitation, we consider only a simple scenario, with a BS with $M = 3$ or $M = 4$ antennas and $\hat{K} = 4$ single-antenna users. Two pairs of users are located on two straight lines that intersect at the BS with angle $\theta = 45°$ between lines.[†] For this scenario, we generate channel matrices using QuaDRiGa [6]. Channel noise variance $\sigma^2 = -94$ dBm. Two cell-center users have path loss PL_{center} and two cell-edge users have pathloss PL_{edge}. The pathloss difference between cell-edge and cell-center users in dB is $PL_{diff} = PL_{center} - PL_{edge}$.

We consider a slotted system. In every slot, the BS chooses the configuration $(\mathcal{K}, f)$ and the precoder P. There are several strategies for the precoder selection: ZF, SOTA, and semi-orthogonal precoder. Also, we consider a case, where every slot the BS selects the precoder that maximizes the geometric mean throughput of all users. The latter is denoted as SOTA+Semi-orthogonal.

In the series of experiments, we fix $PL_{center} = 80$ dB and change the maximum transmit power E_{max} and PL_{diff}. Figure 1 shows the relative gain of various strategies vs. ZF for various values of $E^{edge}_{RX} = E_{max} - PL_{center} - PL_{diff}$.

[†]In the extended version of the paper, we demonstrate results for more complicated scenarios.



(a) $M = 3$ (b) $M = 4$

Fig. 1. Gain in the geometric mean of throughput of all users compared to ZF precoder usage.

First, it can be seen that a higher pathloss difference between cell-edge and cell-center users gives a greater increase in throughput for precoders with NOMA (SOTA or semi-orthogonal), which is consistent with the prior general observation of NOMA behavior. However, if users are close to the BS, i.e., for high $E_{RX}^{edge} \gtrsim -60$ dBm, NOMA precoders give minimal gain compared to ZF. The reason is that power gains using ZF with perfect channel knowledge for users are big enough to outweigh performance drops associated with high channel correlation. On the other hand, with NOMA-based precoders, the interference from the cell-center signal is much higher than the Gaussian noise at cell-edge users, so if PL_{diff} is fixed, SINR stays the same even when channel gains increase.

For small $E_{RX}^{edge} < -80$ dBm, SOTA achieves a greater throughput gain compared to a semi-orthogonal precoder. It happens because cell-edge users have rather low channel gains, and interference reduction for cell-edge users with the semi-orthogonal precoder has almost no effect on the system performance in terms of the geometric mean of throughput. At the same time, SOTA obtains higher SINR values on cell-center users. As a result, it is easier for BS to allocate more power to cell-edge users.

For medium values -80 dBm $\lesssim E_{RX}^{edge} \lesssim -60$ dBm semi-orthogonal precoder shows better results compared to both ZF and SOTA strategies. In this case, the interference nulling for cell-edge users plays an essential role in the considered problem. Channel conditions of cell-edge users are insufficient for ZF to overcome correlation issues, and, on the other hand, they are sufficient to have a noticeable contribution to system performance. Note that the join usage of various precoders shows better

results than using them separately. This strategy is able to use the advantages of both NOMA-based precoders depending on the channel conditions.

5. Conclusion

Existing Downlink NOMA-MIMO systems suffer from inter-group interference, which is caused by applying beamforming on a specific user within each group. In contrast, proposed in this paper the semi-orthogonal precoder reduces this interference by forming non-orthogonal beams for users within each group while maintaining orthogonality between the groups. With simulations, we showed that this precoder outperforms existing ones. However, to achieve higher gains the joint usage of various precoders should be done. Also, we consider power allocation, as the direction of further improvement of the precoder.

REFERENCES

1. E. Khorov, A. Kureev, I. Levitsky, I. F. Akyildiz, Prototyping and experimental study of non-orthogonal multiple access in Wi-Fi networks, IEEE Network 34 (4) (2020) 210–217.
2. M. Zeng, A. Yadav, O. A. Dobre, H. V. Poor, A fair individual rate comparison between MIMO-NOMA and MIMO-OMA, in: 2017 IEEE Globecom Workshops (GC Wkshps), IEEE, 2017, pp. 1–5.
3. M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, H. V. Poor, On the sum rate of MIMO-NOMA and MIMO-OMA systems, IEEE Wireless communications letters 6 (4) (2017) 534–537.
4. Z. Ding, F. Adachi, H. V. Poor, The application of MIMO to non-orthogonal multiple access, IEEE Transactions on Wireless Communications 15 (1) (2015) 537–552.
5. F. Liu, M. Petrova, Performance of proportional fair scheduling for downlink PD-NOMA networks, IEEE Transactions on Wireless Communications 17 (10) (2018) 7027–7039.
6. S. Jaeckel, L. Raschkowski, K. Börner, L. Thiele, QuaDRiGa: A 3-D multi-cell channel model with time evolution for enabling virtual field trials, IEEE transactions on antennas and propagation 62 (6) (2014) 3242–3256.

UDC: 51-74

Minimization of peak age of information in LoRaWAN-based monitoring systems

D.K. Kim[1], A.M. Turlikov[2], N.V. Markovskaya[3]

[1]Narxoz University, Zhandossov str., 55, Almaty, Kazakhstan
[2,3]HSE University, Kantemirovskaya str., 3A, St. Petersburg, Russia

dmitriy.kim@narxoz.kz, atiurlikov@hse.ru, nmarkovskaya@hse.ru

Abstract

We study the communication network based on LoRaWAN technology. The communication network consists of N nodes and carries out one-way transmission of monitoring data to the base station (BS). The messages from each node are sent to the BS at random time intervals and independently of each other. If they have the same spreading factor (SF) and propagation time of messages from two or more nodes overlap, a collision occurs and none of the messages reaches the BS. We assume that the importance of information from each node is different, and the peak age of information (PAoI) is used to measure its freshness. We define the functional as the maximum PAoI among all nodes in the system, and our task is to find the parameters under which its minimization is achieved. We choose various message intensities to account for the different freshness of information from each node and formulate the problem of optimal SF allocation.

Keywords: LoRaWAN, Optimization problem, ALOHA, Peak Age of Information, SF allocation

1. Introduction

Long Range Wide-Area Networks (LoRaWAN) technologies might be used to monitor the results of measurements at objects located over a large area and enable low-cost data transmission and collection systems. There are three classes: Class A, B, and C in LoRaWAN [1]. Devices of Class A initiate transmission itself and send packets as they have been generated. We consider a low-cost monitoring system based on LoRaWAN devices of Class A with transmission at random intervals. If during the sending of a message from one node an another node transmits its message, a collision occurs and both messages are lost (ALOHA type of protocol).

[1]This research is funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP19680230).

[2,3]The work was prepared within the framework of the Basic Research Program at HSE University.

We assume that information from different nodes has various "importance". Our goal is to use different message sending intensities depending on the importance of the node and the allocation of the nodes themselves to different spreading factors (SF allocation) to optimize some functional that characterizes the communication system. For example, to monitor voltage fluctuations in the low-voltage (LV) electrical grid, the voltage in the nodes near the transformer changes "more slowly" than in the remote nodes. To assess the state of the LV grid by monitoring the voltage at the nodes, information from the nodes remote (according to the grid topology) from the transformer must be received more frequently than from other nodes. We can take this into account by varying the intensity of sending messages from different nodes: if the voltage in a node behaves more predictably, the intensity of sending messages may be low, and vice versa - nodes with large voltage fluctuations send their messages with a higher intensity.

At the same time, a very high total message sending intensity from all nodes can occupy the transmission medium and thus increase the number of message collisions. Then it is needed that the information received will be "fresh". The same can happen when the total message sending intensity is low, i.e. when all nodes rarely send up-to-date information. To account for the freshness of information, we use the notion of the Age of Information [2].

The age of information is a concept that reflects the freshness of the information [3, 4] and was defined as $\Delta(t) = t - u(t)$, i.e. the difference of the current time t instant and the time stamp of the received update $u(t)$. The more general form of updating delay, so called Cost of Update Delay (CoUD) (see [4, 5]), was suggested as a stochastic process that increases as a function of time between updates:

$$C(t) = f\left(t - u(t)\right),$$

where $f(\cdot)$ is a non-negative, monotonically increasing function. We use linear function

$$f(t) = \sigma^2 t \tag{1}$$

for some constant $\sigma^2 > 0$, which characterizes the "importance" of information of different nodes. A more tractable age metric as peak age of information (PAoI) (see Fig. 1) was introduced in works [2, 6].

The peak age-of-information (PAoI) metric (see [7]) is defined as:

$$PAoI = \lim_{M\to\infty} \frac{\sum_{m=1}^{M} A_m}{M}.$$

It is easy to see that if $A_1, A_2, A_3, \ldots$ are independent and identically distributed then

$$PAoI = \mathbf{E}A_1.$$

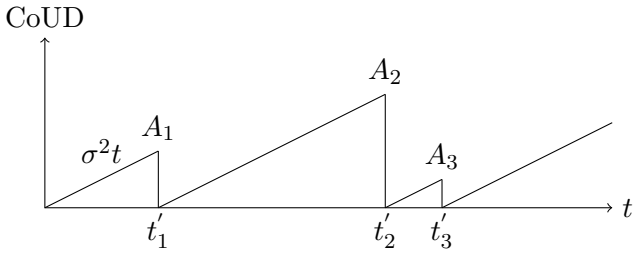


Fig. 1. CoUD trajectory with a linear function (1) , $t_1', t_2', t_3', \ldots$ – time instants of the information update, $A_1, A_2, A_3, \ldots$ – peak values of CoUD.

In [7], the authors consider close problem of optimizing the PAoI by controlling the arrival rate of update messages and derive properties of the optimal solution for the M/G/1 and M/G/1/1 models.

To take into account the "importance" of information from different nodes, we use different growth rates σ_i^2, $i = 1, 2, \ldots, N$, [8] for CoUD of each node. It means that each node has its own $PAoI_i$, $i = 1, 2, \ldots, N$.

Based on the concept of Age of Information it is possible to construct a lot of different functionals characterizing the communication system. In our work, we consider the functional:

$$MPAoI = \max\{PAoI_1, PAoI_2, \ldots, PAoI_N\},$$

which means maximal average peak value among all nodes in the system.

The idea of our work is to choose such parameters of the communication system at which the desired functional is minimized. We first find the message sending intensities from different nodes (section 2), and then partition the entire set of nodes into spreading factors (section 3) to further reduce the maximum PAoI among all nodes. In section 4, we obtain an estimate of the lower bound for our functional.

2. One SF: Problem formulation for minimizing of MPAoF

In this section we will assume that all nodes use the same SF with propagation time of one message $Q > 0$ and the same frequency [9].

We have N nodes and each of them sends messages with intensities $\lambda_i > 0$, $i = 1, 2, \ldots, N$. It means that intervals between messages of node i have exponential distribution with parameter $\lambda_i > 0$. We suppose that they are independent for each node.

Now let us formulate the optimization problem for one SF: to find $(\lambda_1^*, \lambda_2^*, \ldots, \lambda_N^*)$ at which

$$\max\{PAoI_1, PAoI_2, \ldots, PAoI_N\} \to \min. \tag{2}$$

If during the time interval $(-Q, Q)$ from the instant one node sends a message, the other nodes do not send messages, then the message is successfully delivered to the BS. Otherwise. a collision occurs and the messages are lost. This assumption is fulfilled in practice when all nodes are approximately the same distance from the BS. According to this distance, the spreading factor is chosen so that in the absence of collisions the messages are delivered successfully.

Based on these assumptions and by performing a series of arguments [10] we consider a simplified model for the total stream of successfully delivered messages

$$\Lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_N.$$

and use the properties of the Poisson process. According to this model the stream of successfully delivered messages from ith node might be defined as Poisson with intensity $\lambda_i p$, where $p = e^{-2Q\Lambda}$ and the total intensity of all successfully delivered messages is equal to

$$\Lambda \times p.$$

Put $e_{\lambda_i p}$ as an exponentially distributed random variable with intensity $\lambda_i p$. In this case, we can conclude that

$$PAoI_i = \mathbf{E}\left(\sigma_i^2 e_{\lambda_i p}\right) = \frac{\sigma_i^2}{\lambda_i p}.$$

The optimization problem (2) can be written in the form: to find $(\lambda_1^*, \lambda_2^*, \ldots, \lambda_N^*)$ at which

$$\max\left\{\frac{\sigma_1^2}{\lambda_1 p}, \frac{\sigma_2^2}{\lambda_2 p}, \ldots, \frac{\sigma_N^2}{\lambda_N p}\right\} \to \min.$$

The solution to the optimization problem might be achieved if for some minimal $K > 0$

$$\frac{\sigma_1^2}{\lambda_1 p} = \cdots = \frac{\sigma_i^2}{\lambda_i p} = \cdots = \frac{\sigma_N^2}{\lambda_N p} = K.$$

Then $\sigma_i^2 = K\lambda_i p$, $i = 1, 2, \ldots, N$, and

$$\sum_{i=1}^{N} \sigma_i^2 = K\Lambda p$$

or

$$K = \frac{\sum_{i=1}^{N} \sigma_i^2}{\Lambda p}.$$

It means that we have to find the maximal value of the intensity of the total stream of successfully delivered messages. Consider the optimization problem:

$$\Lambda \times e^{-2Q\Lambda} \to \max.$$

Lemma 1. The maximal value of the intensity of the total stream of successfully delivered messages

$$\Lambda^* = \frac{1}{2Q}. \tag{3}$$

Then we obtain that minimal $K > 0$ is equal to

$$K^* = 2eQ \sum_{i=1}^{N} \sigma_i^2 \tag{4}$$

and optimal intensity for ith node

$$\lambda_i^* = \frac{1}{2Q} \frac{\sigma_i^2}{\sum_{i=1}^{N} \sigma_i^2}.$$

3. SF allocation for MPAoI minimizing

The LoRa technology uses distributed spectrum modulation with 6 orthogonal spreading factors (SF= $7, 8, \ldots, 12$). Messages with different SF can be transmitted simultaneously. A smaller SF provides a higher data rate. A larger SF increases the receiver's sensitivity and therefore the range of the system.

There are many works (see, for example, [11, 12]), where the task of assigning spreading factors is considered without taking into account the specifics of the transmitted data.

We can reduce MPAoI by allocating nodes to different SF. We need to divide indices of all nodes $G = \{1, 2, \ldots, N\}$ into $1 \le k \le 6$ sets $G_{13-k}, G_{13-k+1}, \ldots G_{12}$, such that $G = G_{13-k} \cup \cdots \cup G_{12}$, $G_i \cap G_j = \emptyset$, $i \ne j$, where G_i – node indices with SF= i and G_j – node indices with SF= j. We assume that message transmission in each of the SF occurs independently of the nodes with other SF. The optimal allocation of nodes allows us to reduce the MPAoI value.

Let $Q_7, Q_8, \ldots, Q_{12}$ be propagation times for our SFs and we need to split the set G into k subsets so that

$$\max \left\{ 2eQ_{13-k} \sum_{i \in G_{13-k}} \sigma_i^2, \ldots, 2eQ_{12} \sum_{i \in G_{12}} \sigma_i^2 \right\} \to min.$$

If we define

$$\nu = \max_{13-k \le j \le 12} Q_j \sum_{i \in G_j} \sigma_i^2,$$

then

$$\nu \ge Q_j \sum_{i \in G_j} \sigma_i^2, \quad j = 13-k, 13-k+1, \ldots, 12,$$

and we can perform our task as well known a linear integer programming problem.

Let x be a vector from $N \times k$ elements

$$x = (x_1, x_2, \ldots, x_{kN}),\ x_i \in \{0, 1\},$$

where

$$x_i + x_{N+i} + \cdots + x_{(k-2)N+i} + x_{(k-1)N+i} = 1,\ i = 1, 2, \ldots, N.$$

Put

$$s_{13-k} = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2, 0, \ldots, 0),$$
$$s_{13-k+1} = (0, \ldots, 0, \sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2, 0, \ldots, 0),$$
$$\ldots,$$
$$s_{12} = (0, \ldots, 0, \sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2).$$

For brevity, we can rewrite $\sum_{i \in G_j} \sigma_i^2 = s_j x^T$, $j = 13-k, 13-k+1, \ldots, 12$, where x^T is a transposed vector x. It is to formulate the optimal SF allocation problem:

$$\nu \to \min, \tag{5}$$
$$\nu - Q_j s_j x^T \ge 0,\ j = 13-k, 13-k+1, \ldots, 12.$$
$$x_i \in \{0, 1\},\ i = 1, 2, \ldots, kN,$$
$$x_i + x_{N+i} + \cdots + x_{(k-2)N+i} + x_{(k-1)N+i} = 1,\ i = 1, 2, \ldots, N.$$

Example 1. Let $N = 30$ and $k = 3$. It means that we allocate 30 nodes among SF= 10, SF= 11 and SF= 12. Put $Q_{10} = 0.3707$ sec., $Q_{11} = 0.8233$ sec. and $Q_{12} = 1.4828$ sec. Let $\sigma_i^2 = \frac{i}{10}$, $i = 1, 2, \ldots, N$.

Then

$$G_{10} = \{2-8, 14-20, 22-26\},$$
$$G_{11} = \{9, 27-30\},$$
$$G_{12} = \{1, 10-13, 21\},$$
$$2e \max\left\{ Q_{10} \sum_{i \in G_{10}} \sigma_i^2, Q_{11} \sum_{i \in G_{11}} \sigma_i^2, Q_{12} \sum_{i \in G_{12}} \sigma_i^2 \right\} =$$
$$2e \max\{10.157, 10.126, 10.083\} = 55.22.$$

Remark that integer programming is a quite consuming procedure. Therefore, it is important to find a lower bound for our functional. In other words, if we obtain a solution satisfying the constraints of the optimization problem (5) and close enough to the lower bound, we can stop the computational process of searching for the optimal solution and, thus, significantly reduce the number of calculations.

4. Lower bound for MPAoI

Define $S = \sum_{i=1}^{N} \sigma_i^2$. Let

$$S_j = \sum_{i \in G_j} \sigma_i^2, \ j = 13-k, 13-k+1, \ldots, 12.$$

It is easy to find the lower bound if we consider the system of equations:

$$\begin{cases} S_{13-k} + S_{13-k+1} + \cdots + S_{12} = S, \\ Q_i S_i = Q_j S_j, \ i, j = 13-k, 13-k+1, \ldots, 12. \end{cases}$$

It is easy to prove the following lemma:

Lemma 2.

$$\max\left\{2eQ_{13-k}\sum_{i\in G_{13-k}}\sigma_i^2, \ldots, 2eQ_{12}\sum_{i\in G_{12}}\sigma_i^2\right\} \geq 2e \times \frac{\prod_{j=13-k}^{12} Q_j}{\sum_{i=13-k}^{12} \frac{\prod_{j=13-k}^{12} Q_j}{Q_i}} \times S.$$

Let us illustrate the lemma with an

Example 2. Under the conditions of the previous example for three SF:

$$2e \times \frac{Q_{10}Q_{11}Q_{12}}{Q_{10}Q_{11} + Q_{11}Q_{12} + Q_{10}Q_{12}} \times S = 2e \times 9.466 = 51.463.$$

5. Conclusion

We study communication system, which works based on LoRaWAN technology. The system consists of N nodes. Every node initiates transmission with some information and sends messages to the BS at random time intervals independently of each other. Every node belongs to one out of k SF, which has a different propagation time. If nodes have the same spreading factor and propagation time of messages from two or more nodes overlap, a collision occurs and none of the messages reaches the BS. According to our model, every node has different importance for the system. Such metric as the peak age of information (PAoI) with various growth rates is used to measure the freshness and importance of information from the nodes. We choose

the functional describing the entire system as the maximal PAoI among all nodes and investigate how to minimize it.

At first, we choose various message intensities to account for different freshness of information from each node and then formulate the problem of optimal SF node allocation. In this work, we used a simplified model that does not take into account the distance from the BS and the signal fading in the channel, as in the works [11, 12]. At the end, we give an example and also find the lower bound for our functional, which might be used to reduce calculations for optimization problem.

6. Author contribution

Dmitriy Kim developed the mathematical model, Andrey Turlikov formulated the problem, Natalya Markovskaya obtained numerical results of the examples.

REFERENCES

1. "What is LoRaWAN," https://lora-alliance.org/about-lorawan/, 2022
2. Yates R. D., Sun Y., Brown D. R., Kaul S. K., Modiano E. and UlukusS. Age of Information: An Introduction and Survey // IEEE Journal on Selected Areas in Communications. 2021. V. 39, no. 5. P. 1183–1210.
3. Chen X., Gatsis K., Hassani H. and Bidokhti S. S. Age of Information in Random Access Channels // IEEE Transactions on Information Theory. 2022. V. 68, no. 10. P. 6548-6568.
4. Yates S. K. R., Gruteser M. Real-time status: How often should one update // IEEE Int. Conf. Comp. Commun. (INFOCOM). 2012. P. 2731-2735.
5. Sun Y. , Uysal-Biyikoglu E., Yates R., Koksal C. E., Shroff N. B. Update or wait: How to keep your data fresh // 35th Annual IEEE Int. Conf. Comput. Commun. (INFOCOM). 2016. P. 1-9.
6. Costa M., Codreanu M., Ephremides A. Age of information with packet management // Proc. IEEE Int. Symp. Inf. Theory (ISIT), Jun. 2014. P 1583-1587.
7. Huang L., E. Modiano. Optimizing age-of-information in a multi-class queueing system // Proc. IEEE Int. Symp. Inf. Theory (ISIT). 2015. P 1681–1685.
8. Kosta A., Pappas N., Ephremides A. and Angelakis V. Age and Value of Information: Non-linear Age Case // 2017 IEEE International Symposium on Information Theory (ISIT), Aachen, Germany. 2017. P. 326–330.
9. Silva F. S. D., Neto E. P., Oliveira H., Rosário D., Cerqueira E., Both C., Zeadally S., Neto A. V. A Survey on Long-Range Wide-Area Network Technology Optimizations // IEEE Access. 2021. V.9. P 106079–106106.
10. Kim D., Georgiev G., Markovskaya N. A Model of Random Multiple Access in Unlicensed Spectrum Systems // 2022 Wave Electronics and its Application in Information and Telecommunication Systems, WECONF. 2022.

11. Bankov D., Khorov E., Lyakhov A. LoRaWAN Modeling and MCS Allocation to Satisfy Heterogeneous QoS Requirements // Sensors. 2019. V.19, no. 19:4204.
12. Garlisi D., Tinnirello I., Bianchi G., Cuomo F. Capture Aware Sequential Waterfilling for LoRaWAN Adaptive Data Rate // IEEE Transactions on Wireless Communications. 2021. V. 20, no. 3. P. 2019–2033.

УДК: 004.852

К прогнозированию качества радиоканала между БПЛА в рое с применением многослойной нейронной сети

А.Ю. Сырцов[1], Е.В. Бобрикова[1], И.С. Ярцева[1], В.С. Шоргин[2], Ю.В. Гайдамака[1,2]

[1]Российский университет дружбы народов, Российская Федерация, 117198, г.Москва, ул. Миклухо-Маклая, 6

[2]Федеральный исследовательский центр «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН), Российская Федерация, 119333, г. Москва, ул. Вавилова, д. 44-2

1032191341@rudn.ru, bobrikova-ev@rudn.ru, vshorgin@ipiran.ru, gaydamaka-yuv@rudn.ru, yartseva-is@rudn.ru

Аннотация

В работе предложена и исследована методика выбора схемы модуляции и кодирования для передачи данных в радиоканале между устройствами беспроводной сети, основанная на прогнозировании значения отношения сигнал-помеха-шум (SINR) в следующем временном интервале по последовательности известных значений этого отношения в прошлом, с использованием машинного обучения с применением многослойной нейронной сети. Методика может применяться для оценки теоретически достижимых характеристик точности и надежности механизмов поддержки топологии в автономных роях дронов.

Ключевые слова: SINR, машинное обучение, нейронная сеть, БПЛА, рой дронов, UAV swarm

1. Введение

В настоящее время особенности беспроводных сетей с использованием технологий 5G [1] хорошо изучены, [2], и внимание исследователей привлекают сети 6G, которые не только наследуют нерешенные проблемы сетей предыдущих поколений, но и ставят перед проектировщиками новые вопросы [3]. Кроме появления современных антенн для высоких диапазонов частот mmWave или THzWave [4], повысившихся требований к энерго- и спектральной эффективности [5], а также других важных аспектов более сложного планирования и развертывания площадок [1, 5], сохраняются проблемы маршрутизации, вызванные

подвижностью узлов и усилившиеся из-за необходимости соблюдения жестких требований к скорости передачи данных при предоставлении услуг. Как показано в [6, 7], при построении маршрутов по основанным на поиске кратчайших путей протоколам маршрутизации без учета качества радиоканала между узлами сети звено с самым низким качеством радиоканала ограничивает скорость передачи данных на всем маршруте, так что кратчайший маршрут может проигрывать другим с точки зрения межконцевой задержки и других показателей качества предоставления услуги. Поскольку в этом случае качество радиоканала начинает оказывать существенное влияние на выбор следующего узла при построении маршрута, задача оценки и прогнозирования характеристик радиоканала становится остроактуальной. Дополнительные трудности, включая динамическую перестройку маршрута, вносят макро- и микромобильность узлов сети, при этом первая приводит к повышению вероятности обрыва соединения из-за исчезновения узла из радиуса слышимости или перекрытия линии прямой видимости между приемо-передающими устройствами блокатором радиосигнала, а также из-за усиливающейся при сближении узлов интерференции, а вторая – к необходимости более частого запуска процедуры поиска луча и привлечения специальных процедур энергосбережения типа механизма прерывистого приема (discontinuous reception, DRX) [8]. Отметим, что при прогнозировании качества радиоканала для роя БПЛА каждый узел сети может обладать неполной информацией о своем местоположении, иметь ограничения на сохранение и пополнение заряда аккумуляторной батареи БПЛА, использовать принципы прерывистого приема, терять выравнивание луча с БПЛА-передатчиком [9] или покидать позицию в строю под воздействием внешних факторов, например, ветра. Для решения многопараметрических задач оптимизации отлично зарекомендовали себя методы искусственного интеллекта, в частности, аппарат многослойных нейронных сетей [10]. Отметим, что разработанная методика выбора схемы модуляции и кодирования для передачи данных в радиоканале между устройствами беспроводной сети может быть использована для сценариев с наземными мобильными приемо-передающими устройствами.

2. Системная модель

В настоящей работе продолжено исследование эффективности обмена данными в рое БПЛА при отсутствии информации о взаимном позиционировании [11], при этом для прогнозирования интерференции, как одного из основных показателей, напрямую влияющих на скорость передачи данных, применена многослойная нейронная сеть. Использована разработанная в [12] и предназначенная для оценки теоретически достижимых характеристик точности и надежности механизмов поддержки топологии в автономных роях дронов методика выбора

схемы модуляции и кодирования для передачи данных в радиоканале между двумя БПЛА, основанная на прогнозировании значения отношения сигнал/шум (Signal Interference plus Noise Ratio, SINR) в следующем временном интервале по последовательности известных значений этого отношения в прошлом. Предлагаемый в [12] способ назначения модуляционно-кодовой схемы БПЛА-передатчиком, основанный на прогнозировании значения SINR с применением машинного обучения, состоит из двух этапов. На первом этапе для заданной модели движения БПЛА-приемника строится и обучается двуслойная нейронная сеть для прогнозирования значения SINR на мобильном оборудовании пользователя на основе известных значений этого отношения в течении прошлых временных интервалов. На втором этапе по прогнозируемому значению SINR определяется модуляционно-кодовая схема, которая требуется для передачи данных БПЛА-приемнику при оказании услуги с соответствующим уровнем качества. Отметим, что задача оптимизации, решаемая при обучении нейронной сети, является многомерной, при этом параметры нейронной сети существенно зависят от модели движения БПЛА.

3. Математическая модель многослойной нейронной сети

В общем виде задача исследования состоит в настройке нейронной сети для прогнозирования по значениям SINR на n последовательных временных тактах значения SINR на $(n+t)$-м такте, $n \geq 1$, $t \geq 1$. Регрессия строится для объекта, который является последовательностью n тактов времени, признаками объекта служит соответствующая этим временным тактам последовательность значений SINR, измеренных на оборудовании БПЛА-приемника. В качестве инструмента прогнозирования используется многослойная нейронная сеть [13, 14] с двумя скрытыми слоями, содержащими m нейронов каждый. На нейроны входного слоя подаются n признаков объекта, единственный нейрон выходного слоя содержит прогнозируемое на $(n+1)$-м такте значение SINR. Сложность обучения модели многослойной нейронной сети многократно возрастает по сравнению со сложностью обучения однослойной нейросети, где достаточно единственной формулы градиента для обновления весов методом стохастического градиентного спуска. Поэтому исследование ограничилось двумя скрытыми слоями. Граф вычислений нейронной сети схематически показан на рис. 1.

Точность прогнозирования оценивалась с помощью классических метрик оценки качества в задачах регрессии - среднеквадратичной ошибки (Mean Squared Error, MSE), средней абсолютной ошибки (Mean Absolute Error, MAE), коэффициента детерминации R^2 (Coefficient of determination), средней абсолютной ошибки в процентах (Mean Absolute Percentage Error, MAPE). Задача настройки параметров нейронной сети решалась в двух численных экспериментах – с по-

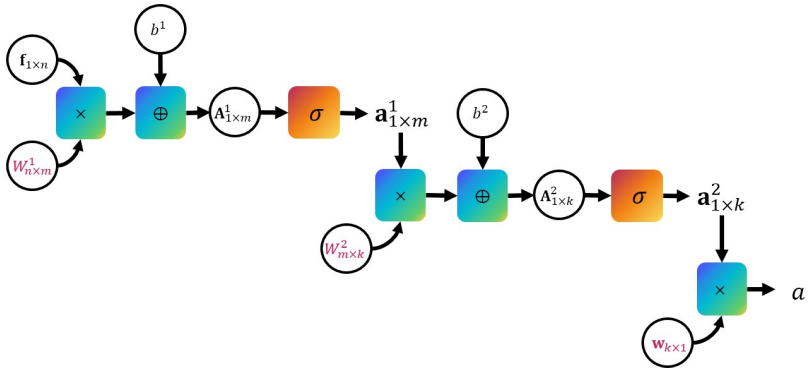


Рис. 1. Граф вычислений многослойной нейронной сети

стоянным числом нейронов в скрытых слоях и с изменением числа нейронов в скрытых слоях. При этом в каждом эксперименте число признаков менялось от 1 до n.

4. Численные эксперименты

В настоящем разделе представлены результаты двух численных экспериментов прогнозирования значения SINR для временных рядов переменной длины $n = 1, ..., 6$ с горизонтом прогнозирования $k = 1$, таким образом, значение SINR на следующем такте прогнозируется по известным значениям SINR на предыдущих n тактах.

В обоих численных экспериментах число n признаков во входном слое варьируется от 1 до 6. При этом в первом эксперименте число m нейронов скрытых слоев постоянно, $m = 6$, а во втором эксперименте $m = n$. В дальнейшем планируется провести исследование для переменного числа нейронов в скрытых слоях. Оценки точности прогнозирования для обоих численных экспериментов на выборке из 67 046 объектов. Траектории генерировались в соответствии с алгоритмом Grid Random Walk, согласно которому на каждом такте возможно перемещение в одну из 4 соседних ячеек, при этом значения вероятностей выбора ячейки определяются заданной моделью движения БПЛА-приемника. Вся выборка делилась на тренинговую X_{train} (75%) и тестовую X_{test} (25%), представлены ниже (см. табл. 1 и 2).

Анализ результатов показал, что попарные значения основных метрик оценивания точности прогнозирования для двух численных экспериментов отличаются друг от друга незначительно, при этом в обоих случаях наблюдается тенденция

	Компоненты вектора входных признаков					
	f_1	$f_1 f_2$	$f_1 f_2 f_3$	$f_1, ..., f_4$	$f_1, ..., f_5$	$f_1, ..., f_6$
$MSE(a, X_{test})$	0.01339	0.00340	0.00198	0.00225	0.00212	0.00331
$MAE(a, X_{test})$	0.07325	0.04375	0.02642	0.03343	0.02655	0.02976
$R^2(a, X_{test})$	0.99991	0.99997	0.99998	0.99998	0.99998	0.99997
$MAPE(a, X_{test})$	0.00354	0.00198	0.00138	0.00156	0.00138	0.00152

Таблица 1. Оценка точности прогнозирования для численного эксперимента 1

	Компоненты вектора входных признаков					
	f_1	$f_1 f_2$	$f_1 f_2 f_3$	$f_1, ..., f_4$	$f_1, ..., f_5$	$f_1, ..., f_6$
$MSE(a, X_{test})$	0.01325	0.03334	0.03198	0.01342	0.00142	0.00305
$MAE(a, X_{test})$	0.07429	0.12240	0.11749	0.07481	0.02647	0.03416
$R^2(a, X_{test})$	0.99991	0.99977	0.99978	0.99991	0.99999	0.99997
$MAPE(a, X_{test})$	0.00356	0.00568	0.00551	0.00359	0.00126	0.00162

Таблица 2. Оценка точности прогнозирования для численного эксперимента 2

роста качества прогнозирования с увеличением числа входных признаков. Заметим, что однослойная нейросеть в [15] показала противоположную тенденцию - прогнозирование на большем числе входных признаков приводило к снижению точности прогноза. Это послужило мотивацией для настоящего исследования.

5. Заключение

Дальнейшие исследования могут быть направлены на анализ горизонта прогнозирования с помощью метрик оценки качества прогнозирования и показателей качества соединения между БПЛА в рое для $k > 1$. Также интересной задачей является вопрос исследования длины временного ряда n с точки зрения временного лага, при наличии которого задача нейросетевого моделирования временных рядов существенно усложняется, так как появляется неопределенность выбора обучающих пар из имеющегося временного ряда. Авторы в дальнейшем планируют провести сравнение результатов используемого в статье прогноза алгоритмом нейронной сети с прогнозами других классических алгоритмов машинного обучения (решающие деревья, случайный лес и другими).

6. Благодарности

Исследование выполнено за счет гранта Российского научного фонда № 22-29-00694, https://rscf.ru/project/22-29-00694/.

ЛИТЕРАТУРА

1. 3GPP, “5G; NR; NR and NG-RAN Overall description; stage 2 (Release 15),” 3GPP TS 38.300 V15.11.0, 3GPP, November 2020.
2. Moltchanov D., Sopin E., Begishev V., Samuylov A., Koucheryavy Y., Samouylov K. A Tutorial on Mathematical Modeling of 5G/6G Millimeter Wave and Terahertz Cellular Systems || IEEE Communications Surveys Tutorials, 2022, vol. 24, no. 2, pp. 1072-1116, 2022, https://doi.org/10.1109/COMST.2022.3156207
3. Trommler, Kimberley Hafner, Matthias Kellerer, Prof Merz, Peter Schuster, Sigurd Urban, Josef Baeder, Uwe Gunzelmann, Dr Kornbichler, Andreas. (2022). Six Questions about 6G.
4. 3GPP, “Study on channel model for frequencies from 0.5 to 100 GHz (Release 14),” 3GPP TR 38.901 V14.1.1, July 2017.
5. 3GPP, «Advanced plans for 5G», https://www.3gpp.org/news-events/ 2210-advanced 5g, 3GPP, 2021.
6. Samuylov, A.; Moltchanov, D.; Kovalchukov, R.; Gaydamaka, A.; Pyattaev, A.; Koucheryavy, Y. GAR: Gradient assisted routing for topology self-organization in dynamic mesh networks. Comput. Commun. 2022, 190, 10–23. [CrossRef]
7. Молчанов Д. А. Сети 5G/6G: архитектура, технологии, методы анализа и расчета: монография / Молчанов Д. А., В. О. Бегишев, К. Е. Самуйлов, Е. А. Кучерявый. М.: Изд-во РУДН, 2022. 515 с.
8. Li Yu-N. R., Chen M., Xu J., Tian L., Huang K. Power Saving Techniques for 5G and Beyond. // IEEE Access – 2020 – P. 99, https://doi.org/10.1109/ACCESS.2020.3001180
9. O. Chukhno et al., "A Holistic Assessment of Directional Deafness in mmWave-Based Distributed 3D Networks,"in IEEE Transactions on Wireless Communications, 2022, vol. 21, no. 9, pp. 7491-7505; https://doi.org/10.1109/TWC.2022.3159086
10. Ming Xu, Wei Liu, Jinwei Xu, Yu Xia, Jing Mao, Cheng Xu, Shunren Hu, Daqing Huang Recurrent Neural Network Based Link Quality Prediction for Fluctuating Low Power Wireless Links Sensors 2022, 22(3), 1212; https://doi.org/10.3390/s22031212
11. Keyela, P., Yartseva, I.S., Gaidamaka, Y.V. (2022). Discrete Time Markov Chain for Drone’s Buffer Data Exchange in an Autonomous Swarm. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds) Distributed Computer and Communication Networks: Control, Computation, Communications. DCCN 2022. Lecture Notes in Computer Science, vol 13766 . Springer, Cham. https://doi.org/10.1007/978-3-031-23207-7_3

12. Ekaterina Bobrikova, Anna Platonova, Ekaterina Medvedeva et al. Using Neural Networks for Channel Quality Prediction in Wireless 5G Networks. In: Distributed Computer and Communication Networks: Control, Computation, Communications. DCCN 2022; 2022: pp. 132-143. Lecture Notes in Computer Science.
13. Vorontsov, K.: Matematicheskie metody obucheniya po precedentam (teoriya obucheniya mashin) [Mathematical teaching methods by precedents (machine learning theory)], http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf (2011)
14. Trask, A. W.: Grokking deep learning. Simon and Schuster (2019)
15. Bobrikova E.V., Platonova A.A., Medvedeva E.G. et al. A Maching learning approach for predicting SINR. In: Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь (DCCN-2022) = Distributed computer and communication networks : control, computation, communications (DCCN-2022) : материалы XXV Международной научной конференции. Россия, Москва, 26–30 сентября 2022 г; 2022: pp. 333-338

УДК: 519.872.4

Модель гетерогенной системы передачи данных с очередью и переключением каналов

С.М. Чижикова[1], Е.А. Пакулова[2], А.Н. Моисеев[1], С.П. Моисеева[1]

[1]Томский государственный университет, Ленина,36, Томск, Российская Федерация

[2]Южный федеральный университет,ул. Большая Садовая 105/42, Ростов-на-Дону, Российская Федерация

lana.cheese23@gmail.com, epakulova@sfedu.ru, moiseev.tsu@gmail.com, smoiseeva@mail.ru

Аннотация

В данной работе рассматривается гетерогенная многолинейная система массового обслуживания с ожиданием, особенность которой заключается в том, что для обработки и передачи данных предоставлены два гетерогенных ресурса разной интенсивности обслуживания и предоставляется возможность переподключения на более быстрый канал, при наличии там свободных единиц канального ресурса. Предложен рекуррентный алгоритм для нахождения стационарных вероятностей и проведен анализ характеристик качества обслуживания.

Ключевые слова: системы массового обслуживания с очередью, гетерогенная сеть передачи данных, переподключение каналов

1. Введение

Последние десятилетия активно развиваются сетевые технологии передачи данных. Появляется все больше услуг связи: видеоконференции и трансляции в режиме реального времени, многомодальные системы взаимодействия, системы телемедицины и телемониторинга, технологии Интернета вещей и пр. При этом все больше предъявляется требований к пропускной способности сети с одной стороны, с другой стороны необходимо выполнение требований качества обслуживания (Quality of Service, QoS) для различных типов трафика. В теории и практике выделяют экстенсивные и интенсивные методы обеспечения качества сервиса. Экстенсивные методы обеспечения качества сервиса предполагают добавление дополнительной пропускной способности за счет добавления

новых ресурсов. Интенсивные методы обеспечения качества сервиса предполагают повышение эффективности использования существующих ресурсов сети и включают различные группы методов [1]. Одной из таких групп являются методы многопоточной маршрутизации или методы многопоточной передачи данных.

Такие методы предполагают разделение на уровне приложения потока данных на несколько субпотоков, которые передаются через несколько транспортных соединений, каждый из которых использует свой маршрут. Реализуют передачу данных таких субпотоков многопоточные протоколы. В настоящее время предложено более 20 версий таких протоколов [2], однако наиболее популярными являются MPTCP [3], SCTP [4], MPQUIC [5].

Одной из основных задач при организации многопоточной передачи данных является распределение данных по субпотокам. В данной статье рассмотрена математическая модель распределения данных по субпотокам с переключением. За основу взят аппарат теории массового обслуживания (ТМО), как наиболее адекватный для описания процессов разделения ресурсов в сети.

2. Постановка задачи

В настоящей статье предлагается модель многолинейной системы обслуживания с ожиданием и переключением, особенность которой заключается в том, что для обработки и распределения данных на субпотоки предоставлены два гетерогенных ресурса конечного объема и разной интенсивности обслуживания. Под ресурсами будем понимать значения пропускной способности на интерфейсах доступных каналов связи, выраженной в единице канального ресурса (ЕКР). Предполагается, что каналы связи обладают гетерогенными характеристиками качества обслуживания (QoS). Также предполагается, что первый канал более быстрый по сравнению со вторым каналом связи. Таким образом, можно говорить об организации двухканальной системы передачи данных, так как информация приложения пользователя может быть передана по двум интерфейсам связи.

Каждый пакет распределяется на интерфейс быстрого канал связи, если последний обладает достаточным количеством ресурсов. То есть занимает прибор для обслуживания, если вимеется свободные "быстрые"приборы. В противном случае при недостаточным количестве свободного ресурса для обслуживания пакет обращается к интерфейсу второго более медленного канала. Если во втором канале имеются свободные приборы, тогда пакет попадает на обслуживание. В противоположном случае пакет попадает в очередь. Дисциплина обслуживания в очереди FIFO (First In – First Out). В случае если первый канал освободился и очередь пуста, то пакет с интерфейса второго канала связи перераспределяется

на обслуживание на интерфейс первого более быстрого канала. Таким образом, имеет место переключение.

Ставится задача определения вероятностных характеристик системы: вероятность моментального обслуживания «быстрым» каналом, вероятность моментального обслуживания «медленным» каналом, средняя длина очереди, среднее время пребывания заявки в очереди и системе.

3. Математическая модель с переключением на более быстрый канал

Рассмотрим систему массового обслуживания (СМО) с двумя блоками обслуживания различной интенсивности обслуживания на вход которой поступает пуассоновский поток пакетов с интенсивностью λ. Дисциплина обслуживания определяется следующим образом: если имеется достаточное количество свободных ЕКР у быстрого канала, то пакет обслуживается в течение случайного времени имеющего экспоненциальное распределение вероятностей с параметром μ_1. Если быстрый канал занят, то пакет поступает во второй канал и обслуживается там в течение случайного времени, имеющего экспоненциальное распределение вероятностей с параметром μ_2. Выполняется следующее соотношение $1/\mu_1 < 1/\mu_2$.

Используя символику Кендалла-Башарина в общем виде можно записать полученную СМО в виде $M|M^{(1,2)}|N + M|\infty$.

Схематическое изображение представлено на рис. 1

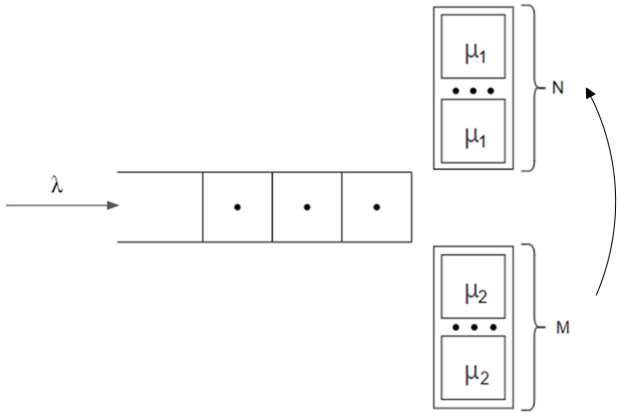


Рис. 1. СМО двухканальная система вида $\mathrm{M}|\mathrm{M}^{(1,2)}|N + M|\infty$

Состояние системы определим случайным процессом $i(t)$, где $i(t)$ – число заявок в системе в момент времени t.

Случайный процесс $i(t)$ изменения во времени состояний системы является цепью Маркова (рис. 2). Для которой граф вероятностей изменения состояний имеет вид:

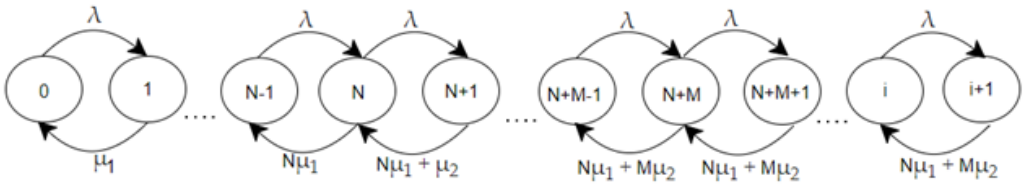


Рис. 2. Граф вероятностей системы для $\mathrm{M}|\mathrm{M}^{(1,2)}|N+M|\infty$

Для стационарных вероятностей числа пакетов в системе имеем системы линейных уравнений следующего вида:

$$\begin{cases} 0 = -\lambda \sqcap_0 + \mu_1 \sqcap_1 \\ 0 = \lambda \sqcap_0 - (\lambda + \mu_1) \sqcap_1 + 2\mu_1 \sqcap_2 \end{cases} \tag{1}$$

$$\begin{cases} 0 = -(\lambda + i\mu_1) \sqcap_i + \lambda \sqcap_{i-1} + \mu_1 \sqcap_{i+1} \\ .... \qquad\qquad i < N \\ 0 = -(\lambda + N\mu_1) \sqcap_N + \lambda \sqcap_{N-1} + (N\mu_1 + \mu_2) \sqcap_{N+1} \end{cases} \tag{2}$$

$$\begin{cases} 0 = -(\lambda + N\mu_1 + (i-N)\mu_2) \sqcap_i + \lambda \sqcap_{i-1} + \\ +(N\mu_1 + (i+1-N)\mu_2 \sqcap_{i+1} \\ .... \qquad\qquad N \le i < N+M \\ 0 = -(\lambda + N\mu_1 + M\mu_2) \sqcap_{N+M} + \lambda \sqcap_{N+M-1} + \\ +(N\mu_1 + M\mu_2) \sqcap_{N+M-1} \end{cases} \tag{3}$$

$$\begin{cases} 0 = -(\lambda + N\mu_1 + M\mu_2) \sqcap_i + \lambda \sqcap_{i-1} + (N\mu_1 + M\mu_2) \sqcap_{i+1} \\ .... \qquad\qquad i \ge N+M \end{cases} \tag{4}$$

Так как сумма слагаемых, начиная с $\sqcap_{N+M}$ представляет собой геометрическую прогрессию, то ряд сходиться при условии $\frac{\lambda}{N\mu_1 + M\mu_2} < 1 \Rightarrow \lambda < N\mu_1 + M\mu_2$.

Данное условие является условием существования стационарного режима в рассматриваемой системе.

Уравнения (1-4) являются основой для итерационного (рекуррентного) алгоритма с учетом условия нормировки. Положим $V_0 = a$, где a – некоторая произвольная положительная константа. Тогда

$$V_1 = \frac{\lambda}{\mu_1} V_0, V_2 = \frac{(\lambda + \mu_1)V_1 - \lambda V_0}{2\mu_1}.$$

Для i<N имеем

$$V_{i+1} = \frac{(\lambda + i\mu_1) \sqcap_i - \lambda V_{i-1}}{(i+1)\mu_1}.$$

$$V_{N+1} = \frac{(\lambda + N\mu_1)V_N - \lambda V_{N-1}}{N_1\mu_1 + \mu_2},$$

Для N<i<N+M

$$V_{i+1} = \frac{(\lambda + N\mu_1 + (i-N)\mu_2)V_i - \lambda V_{i-1}}{N\mu_1 + (i+1-N)\mu_2},$$

$$V_{N+M+1} = \frac{(\lambda + N\mu_1 + M\mu_2)V_{N+M} - \lambda V_{N+M-1}}{N\mu_1 + M\mu_2},$$

i> N+M

$$V_{i+1} = \frac{(\lambda + N\mu_1 + M\mu_2)V_i - \lambda V_{i-1}}{N\mu_1 + M\mu_2}.$$

Для нахождения финальных вероятностей воспользуемся условием нормировки $1 = \sum_{i=0}^{\infty} V_i$. Обозначим: $A = \sum_{i=0}^{\infty} V_i$.Далее нормируем и получаем финальное распределение $\sqcap_i = \frac{1}{A} V_i$.

В результате были получены выражения для следующих вероятностных характеристик системы: вероятность моментального обслуживания «быстрым» каналом, вероятность моментального обслуживания «медленным» каналом, средняя длина очереди, среднее время пребывания заявки в очереди и системе.

4. Численный пример

Для построенной модели проводился численный эксперимент с различными входными данными: интенсивность прибытия пакетов λ состаляет 2, интенсивность обслуживания для каналов $\mu_1 = 1$, $\mu_2 = 0.01$. Нагрузку системы обозначим через ρ. В результате были получены результаты, представленные в таблице 1.

Характеристики	**Входные параметры**		
	$\rho = 0.66$ $N = 3$ $M = 2$	$\rho = 0.98$ $N = 2$ $M = 4$	$\rho = 0.39$ $N = 5$ $M = 10$
Вероятность обслуживания «быстрым» каналом	0.56	0.03	0.94
Вероятность обслуживания «медленным» каналом	0.25	0.07	0.06
Вероятность попасть в очередь	0.19	0.90	5.5×10^{-6}
Среднее число заявок в системе	2.86	50.62	2.03
Средняя длина очереди	0.38	44.95	3.6×10^{-6}
Среднее время пребывания заявки в очереди	0.19	22.48	1.8×10^{-6}
Среднее время пребывания в системе	0.52	22.97	0.20

Таблица 1. Вероятностные характеристики СМО для систем двухканальной передачи передачи данных с переключением на более быстрый прибор

5. Заключение

В настоящей статье представлена математическая модель двухканальной системы передачи данных с ожиданием и переключением. С помощью разработанного рекуррентного алгоритма получено распределение вероятностей, и найдены вероятностные характеристики системы, которые можно использовать при проектировании реальных инфокоммуникационных систем.

Из полученных результатов численного эксперимента можно сделать вывод, что распределение данных с "медленного"канала на "быстрый"в режиме переключения приводит к увеличению очереди, и как следствие может привести к увеличению задержки передачи данных, что неприемлемо при передаче трафика, чуствительного к задержке. Однако такой режим может быть использован для

передачи эластичного трафика с целью балансировки нагрузки. Однако данный вопрос требует дальнейшего изучения.

ЛИТЕРАТУРА

1. Е.П. Степанов, Исследование методов многопоточной маршрутизации для обеспечения качества сетевого сервиса: дис. ... канд. тех. наук: 05.13.11: защищена 15.07.22. М., 2022, 144 с.
2. S. Habib, J. Qadir, A. Ali, D. Habib, M. Li, A. Sathiaseelan, The past, present, and future of transport-layer multipath, Journal of Network and Computer Applications, V. 75, 2016, pp. 236-258,
3. C. Raiciu, C. Paasch, S. Barre, A. Ford, M. Honda, F. Duchene, O. Bonaventure, and M. Handley, How hard can it be? designing and implementing a deployable multipath TCP. In Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation (NSDI'12). USENIX Association, USA, 2012, pp. 399–412.
4. R. Stewart and C. Metz, SCTP: new transport protocol for TCP/IP, in IEEE Internet Computing, vol. 5, no. 6, pp. 64-69, Nov.-Dec. 2001.
5. Q. De Coninck and O. Bonaventure, Multipath quic: Design and evaluation. Proceedings of the 13th international conference on emerging networking experiments and technologies, 2017, pp. 160–166.

УДК: 681.3.06 (075.32)

Анализ полноты и актуальности выходной информации в распределенных компьютерных и телекоммуникационных системах, обеспечивающих проведение избирательных кампаний

А.И. Костогрызов[1]

[1]Федеральное государственное учреждение Федеральный исследовательский центр "Информатика и управление" Российской академии наук, ул. Вавилова, д. 44, корп. 2., Москва, Российская Федерация

akostogr@mail.ru

Аннотация

Применительно к системам, обеспечивающим проведение избирательных кампаний, проанализировано качество выходной информации для избирателей. На основе применения авторских вероятностных моделей количественно оценено влияние используемых распределенных компьютерных и телекоммуникационных систем на такие свойства качества, как полнота и актуальность выходной информации о ходе и результатах выборов.

Ключевые слова: актуальность информации, вероятность, модель, полнота информации, система

1. Введение

При проведении избирательных кампаний сбор, обработка и передача информации базируется на широкомасштабном применении компьютерных и телекоммуникационных систем и сетей. Главным вопросом остается доверие к результатам выборов непосредственно в день голосования и сразу после него. Для избирателей это доверие формируется на основе выходной информации о ходе и результатах выборов. Требуемое при этом качество на всех иерархических уровнях выборной системы во многом достигается за счет оперативности доставки, полноты и достоверности выходной информации. Перед глазами – последний пример обеспечения качества выходной информации о ходе и результатах второго тура выборов президента Турции 28 мая 2023 года с гласным информированием избирателей в среднем каждые 20-40 минут. Это – частный положительный пример, но до него было множество других примеров, где недостаточное качество информации для избирателей служило источником социального взрыва.

Для более аргументированного ответа на вопрос о роли компьютерных и телекоммуникационных систем и сетей для обеспечения качества выходной информации о ходе и результатах выборов необходимо более детальное математическое моделирование. К сожалению, для приложений к выборным технологиям при наличии множества разнородных неопределенностей целенаправленных количественных оценок качества выходной информации практически нет (за некоторыми исключениями – см., например, [1]).

Именно количественному анализу таких свойств качества выходной информации, как полнота и актуальность применительно к системам, обеспечивающим проведение избирательных кампаний непосредственно в день выборов и при подсчете результатов выборов, посвящена настоящая работа.

2. Основные понятия, принятые положения и допущения

В работе приняты следующие положения и допущения, касающиеся достижения цели обеспечения оперативности доставки, полноты и достоверности выходной информации о ходе и результатах выборов.

Оперативность доставки информации измеряется временем обработки бюллетеней и доведения соответствующей информации о ходе и результатах выборов до избирателей. По сравнению с ручным пересчетом бюллетеней время обработки и доведения выходной информации до избирателей объективно сокращается на порядки – до нескольких минут. По этой причине целевые эффекты, связанные с повышением оперативности доставки информации за счет применения компьютерных и телекоммуникационных систем и сетей, представляются очевидными и в работе используются лишь в контексте формирования исходных данных для математического моделирования.

Все последующие определения адаптированы из ГОСТ Р 59341-2021 «Системная инженерия. Защита информации в процессе управления информацией системы», смысл других терминов (не раскрываемых ниже) сохранен согласно этому стандарту.

Под полнотой выходной информации понимается свойство предоставляемой избирателям информации отражать состояния всех впервые появляющихся фактов и объектов учета предметной области системы (например, о готовности избирательных комиссий, об открытии участков, о первых нарушениях в ходе выборов, о признании выборов состоявшимися, о первых результатах голосования и др.). Каждый случай неполноты выходной информации становится предметом особого внимания, ведь тем самым нарушается необходимая степень доверия избирателей к выборам. Здесь неопределенность заключается в наступлении моментов первого появления фактов и объектов учета предметной области системы,

времени подготовки, доведения и отражения соответствующей информации в базах данных (БД) системы.

Под достоверностью выходной информации в настоящей работе понимается свойство периодически обновляемой в БД и подлежащей доведению до избирателей выходной информации отражать реальные состояния хода выборов (по количеству проголосовавших, по количеству нарушений и пр.), промежуточных и окончательных результатов голосования (по количеству набираемых голосов), изменяемых во времени в ходе голосования и при подсчете количества бюллетеней за кандидатов. Достоверность выходной информации определяется корректностью обработки заполненных бюллетеней, безошибочностью при хранении и передаче соответствующей информации и сохранением ее актуальности на момент использования. Под актуальностью выходной информации понимается свойство обновляемой информации, корректно обработанной из заполненных бюллетеней и подлежащей доведению до избирателей, отражать текущее состояние хода выборов и результатов выборов с соблюдением согласованности с объективной реальностью. Рассогласование реальной и хранимой в БД информации вызвано устареванием информации в результате какого-либо значимого изменения до следующего обновления этого изменения в БД – см. рис. 1.

Актуальность выходной информации для избирателей обеспечивается путем достаточно быстрого и непротиворечивого отражения в БД объективно имевших место значимых изменений информации о ходе выборов и результатов выборов с учетом технических задержек в используемых компьютерных и телекоммуникационных системах и сетях. Оперирование неактуальной информацией в системе может быть воспринято избирателями как обман или необоснованные затяжки времени для подделки результатов голосования, вызвать недоверие к выборам и послужить источником социального взрыва. При этом вполне обоснованно можно полагать, что безошибочность входной информации из заполненных бюллетеней и корректность обработки бюллетеней обеспечиваются в результате применения на избирательных участках специальных комплексов автоматической обработки избирательных бюллетеней. Благодаря использованию самими избирателями этих комплексов осуществляется автоматический подсчет голосов, устраняются ошибки ручного подсчета голосов и предотвращаются попытки фальсификации итогов голосования. Также не без оснований полагается, что безошибочность выходной информации при ее хранении и передаче достигается применением современных компьютерных технологий обеспечения информационной безопасности. Неопределенность заключается в наступлении моментов значимого изменения состояния объектов учета в реальности, времени подготовки, доведения и отражения соответствующей информации в БД системы. С учетом вышеизложенного при использовании компьютерных и телекоммуника-

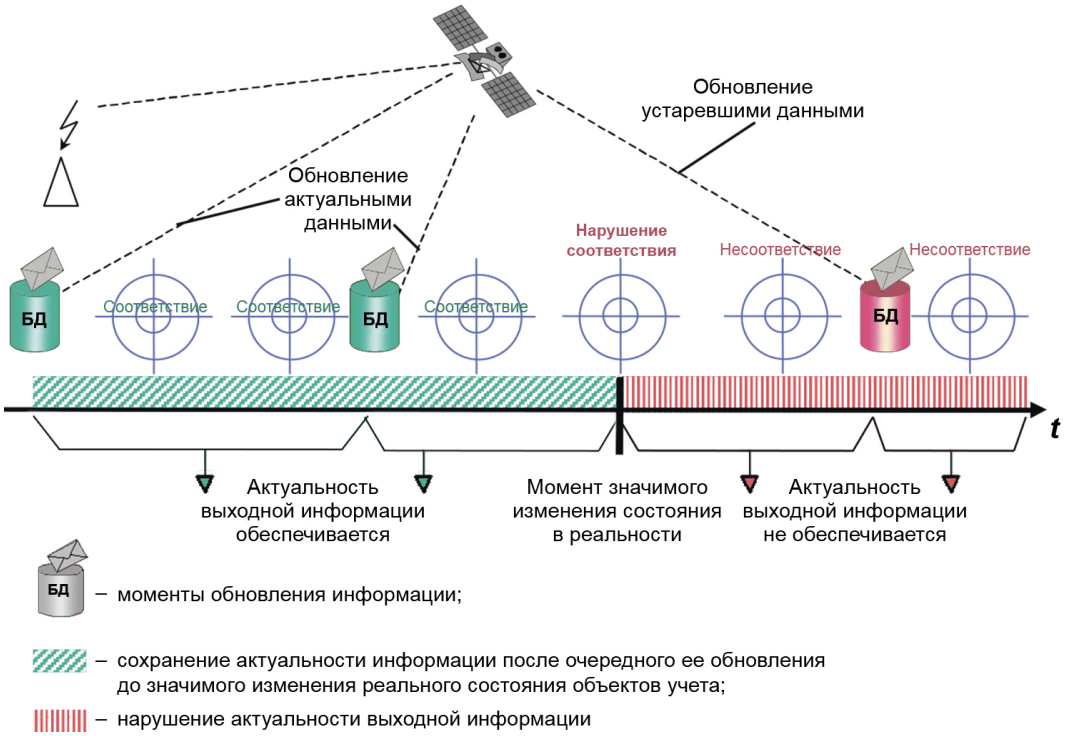


Рис. 1. Иллюстрация формирования актуальности выходной информации

ционных систем и сетей из свойств, характеризующих качество, как наиболее критичные выбраны свойства полноты и актуальности выходной информации. Понимая, что реальные распределенные компьютерные и телекоммуникационные системы, обеспечивающие проведение избирательных кампаний, имеют сложную сетевую структуру, в настоящей работе принято следующее допущение. В целях анализа полноты и актуальности выходной информации система, обеспечивающая проведение избирательных кампаний, рассматривается как черный ящик. Сетевая структура учитывается лишь путем формирования временных исходных данных для математического моделирования.

Далее оценим функциональные возможности некоторой автоматизированной системы (АС), структурно представляющей собой многоуровневую территориально-распределенную систему комплексов средств автоматизации центральной избирательной комиссии (ЦИК), десятков избирательных комиссий субъектов, сотен окружных, тысяч территориальных и десятков тысяч участковых избирательных комиссий. К примеру, в ГАС «Выборы» России насчитывается более 96 тысяч

участковых избирательных комиссий. Исследуемая АС может выступать как некая система-аналог для многих реальных систем, обеспечивающих проведение избирательных кампаний. Для расчетов используются модели для оценки полноты и актуальности информации, отраженные в авторских работах [1–4], а также в ГОСТ Р 59341–2021, в котором эти модели реализованы.

3. Оценка полноты выходной информации

Используемая «Модель для оценки полноты оперативного отражения в системе новых объектов и явлений» [1–4] позволяет оценить вероятность обеспечения полноты выходной информации в БД системы. Искомая вероятность вычисляется в предположении пуассоновского потока моментов появления новых фактов и объектов учета. В качестве исходных данных используется частота появления новых фактов и объектов учета в день выборов и при подсчете результатов выборов и среднее время подготовки, передачи и ввода новых объектов учета в БД.

При оценке полноты речь идет о новых фактах и объектах учета – это может быть информация о готовности республиканских, окружных, территориальных и участковых избирательных комиссий, об открытии участков, о нарушениях в ходе выборов, о признании выборов состоявшимися, о предварительных результатах голосования и др. Пусть в период выборов на одном избирательном участке в среднем 1 раз в 2 часа происходит новое важное событие, информация о котором должна быть доведена до сведения вышестоящей избирательной комиссии для принятия решения. В свою очередь, на более высокий уровень передается лишь 10 % всех сообщений, поступающих от подчиненных избирательных комиссий. Остальные 90% сообщений разрешаются в первой же более высокой инстанции. Пусть до внедрения АС на подготовку данных уходило в среднем 15 минут (печать поступивших документов на печатной машинке), на передачу (с помощью телеграмм, телефонограмм, факса) – 5 минут, прием и доведение до ЦИК – 2 минуты. В АС подготовка сократилась до 1 минуты, передача – до 3 секунд и доведения до БД ЦИК – 30 секунд.

Требуется оценить степень повышения полноты выходной информации в АС по сравнению с неавтоматизированным режимом.

Результаты моделирования показали, что вероятность обеспечения полноты выходной информации в БД АС повысится на уровне территориальных избирательных комиссий с 0.002 до 0.65, на уровне окружных избирательных комиссий с 10-6 до 0.34, на уровне избирательных комиссий субъектов с 0.003 до 0.66. Из-за большого количества избирательных участков информация на уровне ЦИК окажется неполной с вероятностью, близкой к 1. Неполнота будет иметь место по 3-7 объектам учета. Но избиратель привык к такой неполноте. Надо признать,

что на практике из-за многоминутных задержек полнота выходной информации о ходе выборов в вероятностном выражении будет объективно низкой даже с использованием высокопроизводительных систем и сетей.

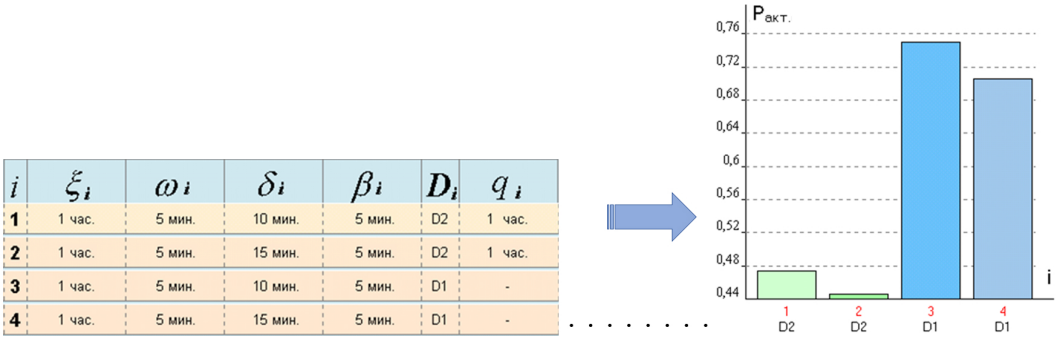
4. Оценка актуальности выходной информации

Используемая "Модель для оценки актуальности обновляемой информации" [1–4] позволяет оценить вероятность сохранения актуальности выходной информации в системе на момент ее предоставления избирателям при задаваемых: среднем времени между значимыми изменениями состояния объекта учета (ξ); среднем времени подготовки информации (ω); среднем времени передачи информации (δ); среднем времени ввода информации в БД (β); дисциплине обновления в информации в системе (D). $D = D_1$ означает, что сбор информации в системе происходит «сразу по происшествии значимого изменения» состояния объектов учета. $D = D_2$ означает, что сбор происходит вне явной зависимости от изменения состояний объектов учета. Для случая $D = D_2$ дополнительно задается среднее время (q) между обновлениями информации.

Для начала оценим актуальность информации о количестве проголосовавших, отображаемой на табло коллективного пользователя ЦИК АС в ходе выборов. Положим, что значимые для выборов изменения происходят 1 раз в час, регламент сбора информации от избирательных участков составляет также 1 раз в час, время подготовки данных в АС – 3 минуты, время передачи от избирательных участков до ЦИК с учетом обобщения в вышестоящих округах – 10-15 минут, а время ввода в БД с учетом контроля – 5 минут. Согласно результатам моделирования вероятность того, что на табло ЦИК отражается актуальная информация о количестве проголосовавших, равна 0.35-0.58 (т.к. на практике могут быть отступления от регламента сбора информации). Таким образом, собранные данные на момент отображения существенно устаревают. Может показаться удивительным, но, похоже, из-за отсутствия необходимости высокой точности количества проголосовавших более высокая актуальность и не требуется. Она требуется лишь в случае недобора необходимого минимума проголосовавших для признания выборов состоявшимися.

Дальнейший анализ проведен уже в приложении к информации о результатах выборов, когда скорость нужна не только для понимания «кто побеждает», но и для отсутствия подозрений о подтасовке результатов выборов. Автоматизация процесса подсчета и регистрации голосов позволяет обновлять информацию в режиме реального времени. Пусть на подготовку сообщения о промежуточных результатах уходит 5 минут, на передачу данных от избирательного участка через всю цепочку иерархии 10-15 минут, на контроль и ввод данных – 5 минут. Обобщенные оценки показывают (см. рис. 2), что регламентный сбор инфор-

мации (т.е. дисциплина $D_2, i = 1, 2$) существенно хуже по сравнению со сбором «сразу по изменении» (дисциплина $D_1, i = 3, 4$). Повышения актуальность с достижимого уровня 0.70-0.75 до уровня 0.8 можно добиться, если от участковых избирательных комиссий результаты будут направлены в оригинале по иерархии и одновременно – непосредственно в ЦИК. Так будет улучшена «прозрачность» выборов, доверие к их результатам будет повышено, избиратели смогут в режиме реального времени наблюдать актуальную информацию.



i	ξ_i	ω_i	δ_i	β_i	D_i	q_i
1	1 час.	5 мин.	10 мин.	5 мин.	D2	1 час.
2	1 час.	5 мин.	15 мин.	5 мин.	D2	1 час.
3	1 час.	5 мин.	10 мин.	5 мин.	D1	-
4	1 час.	5 мин.	15 мин.	5 мин.	D1	-

Рис. 2. Исходные данные и обобщенные результаты оценки актуальности

5. Заключение

Проведенный анализ влияния распределенных компьютерных и телекоммуникационных систем на качество выходной информации в системах, обеспечивающих проведение избирательных кампаний, показал следующее:

1) по сравнению с неавтоматизированным голосованием вероятность обеспечения полноты выходной информации в БД повысится до уровня 0.34-0.66. Из-за большого количества избирательных участков информация на уровне ЦИК окажется неполной с вероятностью, близкой к 1 (т.е. в масштабах системы отображаемая информация оказывается заведомо неполной, неполнота будет иметь место по 3-7 объектам учета);

2) отображаемая на табло ЦИК информация о количестве проголосовавших в период между обновлениями оказывается актуальной с вероятностью 0.35-0.58 (на практике более высокая актуальность и не требуется);

3) вероятность сохранения актуальности выходной информации на табло ЦИК о результатах выборов («кто побеждает») может достигать уровня 0.70-0.75. Повышения актуальности до уровня 0.8 можно добиться, если от участковых избирательных комиссий результаты будут направлены в оригинале по иерархии и одновременно – непосредственно в ЦИК.

ЛИТЕРАТУРА

1. Костогрызов А.И., Степанов П.В. Инновационное управление качеством и рисками в жизненном цикле систем – М.: Изд. "Вооружение, политика, конверсия", 2008. – 404с.
2. Kostogryzov A., Nistratov G., Nistratov A. Some Applicable Methods to Analyze and Optimize System Processes in Quality Management. Total Quality Management and Six Sigma, InTech, 2012, pp. 127-196
3. Kostogryzov A., Korolev V. Probabilistic methods for cognitive solving some problems of artificial intelligence systems. Probability, combinatorics and control. IntechOpen, 2020, pp. 3-34. https://www.intechopen.com/books/probability-combinatorics-and-control
4. Kostogryzov A., Makhutov N., Nistratov A., Reznikov G. Probabilistic predictive modeling for complex system risk assessments. Time Series Analysis - New Insights. IntechOpen, 2023, pp. 73-105. https://www.intechopen.com/chapters/83570

UDC: 621.391

Numerical Evaluation of the Optimal Precoder Design for Mobile Users in MISO System

Alexander A. Kalachikov[1]

[1]Siberian State University of Telecommunications and Information Sciences, Kirova st. 86, Novosibirsk, Russia

330rts@gmail.com

Abstract

In paper the numerical investigation of the impact of delayed channel state information (CSI) due to user movement and caused channel aging on the performance of multiuser precoder in downlink multiple-input sngle-output (MISO) system. The CSI of time variant wireless channel is obtained by the least square (LS) channel estimation. We consider Zero Forcing (ZF) algorithm and numerical optimization based solution of calculating precoder vectors maximizimg sum rate of multiuser system. For numerical simulation the QUADRIGA channel model reflecting the real propagation conditions for moving users is used. The obtained performance of multiuser Zero Forcing and optimization based precoder in spatially correlated channel are compared based on the empirical cumulative density function of the sum rate of multiple users.

Keywords: Multiuser precoding, QUADRIGA 3GPP channel model, ZF precoding, optimal precoding desing, MMSE channel estimation

1. Introduction

Massive multiple-input multiple-output (MIMO) systems enchance the capacity of multi-user MIMO systems by using beamforming over a transmit antenna array on the base station achieving spatial multiplexing gain [1].

Most of the performance gain of massive MIMO depends heavily on the accurate channel state information at the base station (BSs). For mobile users the channel impulse response is time varying and the coherence time is reduced.

The mismatch between the channel coefficients obtained by the channel estimation and used for precoding and the actual channel coefficients refers to channel aging [2].

It was observed that the moderate-mobility scenario at 30 km/h leads to as much as 50 of the performance reduction compared to low-mobility scenario[3].

The publication has been prepared according to the state order of Mintsifry Rossii No. 071-03-2023-001.

Therefore, the study of channel aging effects is crucial for complex system simulation in scenarios with moving users.

In this paper, we investigate the effect of channel aging on a precoder performance in multiuser downlink MISO system with vehicle users in a more realistic scattering scenario with spatially correlated channels.

2. SYSTEM MODEL

2.1. System Model. In the typical scenario [4] the Base Station with massive MISO equipped N_T transmit and receive antennas serves K users UE each having one antenna. The N_T antennas at the BS are defined as two-dimensional (2D) uniform rectangular array (URA) .

During the channel estimation perion users transmit pilot sequences of length τ symbols. The channel remains constant during this period. In data transmission period $(T - \tau)$ OFDM symbols the channel varies from symbol to symbol.

In matrix form the MU-MISO channel matrix for time index n and subcarrier index s is composed as $\mathbf{H}_{n,s} = [\mathbf{h}_{1,n,s} \dots \mathbf{h}_{K,n,s}]^T$ and the received vector is defined as

$$\mathbf{y}_{n,s} = \mathbf{H}_{n,s}^T \mathbf{x}_{n,s} + n_{n,s} \tag{1}$$

By using precoding (beamforming) the received signal for user k is defined as

$$y_k = \mathbf{h}_k^T \mathbf{w}_k s_k + \sum_{j \in S, j \neq k} \mathbf{h}_k^T \mathbf{w}_j s_j + \mathbf{n}_k, \ for\ k = 1, \dots, K \tag{2}$$

where the sum term corresponds to the interference from other users .

2.2. Channel Aging. Due to movement of the UEs the temporal variations in the propagation environment arise which affect the channel coefficient in a resource slots. The time-variant channel vector for the k-th user at time n $\mathbf{h}_k[n]$ can be modeled as a function of its initial state $\mathbf{h}_k[0]$ and an innovation component [5] as

$$\mathbf{h}_k[n] = \rho_k[n]\mathbf{h}_k[0] + \overline{\rho}_k[n]\mathbf{g}_k[n] \tag{3}$$

where time 0 corresponds to the last symbol transmitted in the channel estimation period, $\mathbf{g}_k[n]$ represents the independent innovation component at the time instant n, $\rho_k[n]$ represents the temporal correlation coefficient of channel vector between the channel realizations at time 0 and n.

2.3. Channel Estimation. The Least Square (LS) estimation is aplied to the received signal for each user on the pilot positions p of SRS sequence to obtain the esimate of the channel coefficients $\tilde{\mathbf{h}}_k[n]$. The LS estimate is computed by division of received symbols on corresponding values of pilot sequence. The estimate obtained

at pilot symbol period is used as the initial state $\mathbf{h}_k[0]$ to compute estimates of the channels $\hat{\mathbf{h}}_k[n]$ at other symbols of this resource slot and period of pilot insertion in resourse slot depend on the channel aging effect [6] .

2.4. ZF Precoding. The ZF precoding vector $\mathbf{w}_k$ of the k-th UE is orthogonal to the transpose conjugate channel vectors of all other UEs $\mathbf{h}_k^H \mathbf{w}_j = 0 \; for \; j \neq k$. The ZF precoding matrix W is composed of all precoding vectors $\mathbf{w}_j$ and is computed as the pseudoinverse of the channel matrix of the selected users as $\mathbf{W} = \mathbf{H}(\mathbf{H}^H\mathbf{H})^{-1}$.

The sum rate is the sum of the rates achieved by the UEs on all subcarriers and is depend on the signal to interference noise ratio (SINR) of each UE. The SINR at the $k-th$ UE on one subcarrier is

$$SINR_k = \frac{|\mathbf{h}_k^T \mathbf{w}_k|^2}{\sum_{j \neq k} |\mathbf{h}_k^T \mathbf{w}_j|^2 + K\sigma^2/P} \tag{4}$$

The achieved multiuser sum rate is determined as

$$R_{BF} = \sum_{k=1}^{K} (log_2(1 + SINR_k)), \text{ bits/s/Hz} \tag{5}$$

This metric is used to evaluate the spectral efficiency of considered precoding algorithms under channel aging effect [7].

2.5. Optimization based precoding design. The objective is to design precoding vectors $\mathbf{w}_1, ..., \mathbf{w}_K$ that maximize the multiuser sum rate R_{BF} depending of individual user $SINR_k$ under the BS tramsmit power constraint.

The cost function R_{BF} depends on the SINRs which are non-convex functions of the precoding vectors $\mathbf{w}_1, ..., \mathbf{w}_K$. The design of transmit precoder vectors can be transformed as the problem of minimizing the total transmit power subject to SINR constraints γ_k at each of the K receivers. The values of γ_k are the minimal acceptable SINR for the kth user. The corresponding minimization problem is formulated as follows

$$\min_{w_1,...,w_K \in C^{N_t}} \sum_{k=1}^{K} \|w_k\|^2 \tag{6}$$

$$\text{s.t. } \mathrm{SINR_k} \geq \gamma_\mathrm{k} \tag{7}$$

The solution of this problem gives the precoding vectors that achieves the required SINR using the minimum of power. The problem can be reformulated as a convex problem.

The reformulated SINR constraint is a second order cone constraint [8], [9].

3. NUMERICAL RESULTS

In this section the performance of the precoding algorithms are presented by computer simulation. The performances of ZF precoding and optimization based design are evaluated by a system level simulation in terms of the average spectral efficiency based on the QUADRIGA channel model [8], [9].

The simulation scenario is 3GPP 38.901 UMa NLOS. According to the UMa scenario the users are uniformly distributed around the BS in the area of 500 m from the transmitter. The path loss model and shadow fading are disabled and the narrowband channel vectors for selected subcarrier are normalized to unit power. The BS is equipped with 16 antennas. The number of randomly selected users is set as K = 4 and each user is equipped with single antenna. For each of users the linear movement track with specified user speed is defined. The lengths of all the simulated tracks are equals 250 meters which gives approximately 800 snapshots of the channel impulse responses.

3.1. Sum rate performance of of ZF precoding and optimal precoder design. This section provides the numerical results to observe the impact of channel aging on the sum rate of MISO system. The comparison of precoder performances is presented using the cumulative distribution functions (CDFs) of sum rate. The SOCP optimization problem is solved numerically using the convex optimization software CVX [10].

The impact of imperfect CSI is presented for the speed of the users 30 and 60 km/h for SNR value equals 8 dB. The performance curves of ZF precoder and optimization based precoder for stationary scenario (0 km/h) are presented as a benchmark. Fig.2 provides an comparison of precoder performances of 4 users and for the speed of the users 30 km/h.

At low mobility of v = 30 kmph the performance loss is slightly remarkable for ZF and optimization based precoders. The median SE for ZF precoder reduces to 4.6 bps in comparison with 6 bps=Hz of the benchmark. The median SE for optimization based precoder reduces to 7.8 bps in comparison with 8.2 bps=Hz of the static users. The SE of the optimization based precoder outperforms the ZF precoder. The sum-rate of optimization based precoder is about 1.8 times of that of the ZF precoder at SNR=8 dB for the 30 kmph.

With the increase of users speed the performance loss of ZF precoder becomes more notable. At medium mobility of v = 60 kmph the performance loss is remarkable for ZF and optimization based precoders. The median SE for ZF precoder reduces to 3 bps in comparison with 4 bps=Hz of the benchmark. The median SE for optimization based precoder reduces to 6 bps in comparison with 8 bps=Hz of the static users. The SE of the optimization based precoder outperforms the ZF precoder.

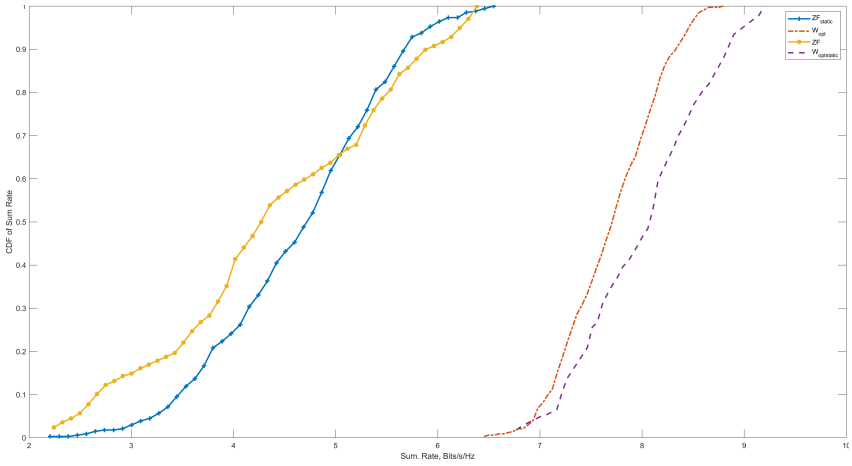


Fig. 1. CDF of Sum Rate speed 30 kmph.

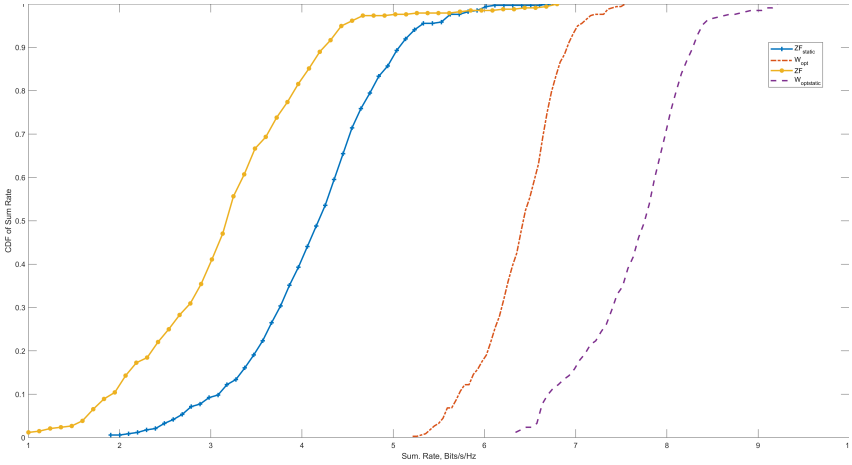


Fig. 2. CDF of Sum Rate speed 60 kmph.

The sum-rate of optimization based precoder is about 2 times of that of the ZF precoder at SNR=8 dB for the 60 kmph.

4. Conclusion

The results show that the performance gain of precoder based on SOCP is more significant than ZF precoder in scenario with larger speed of users.The performance gain of the optimization based precoder precoders compared to ZF become more significant when SNR increases for both of speed values.

REFERENCES

1. E. Castaneda, A. Silva, A. Gameiro, and M. Kountouris, "An overview on resource allocation techniques for multi-user MIMO systems," IEEE Communications Surveys and Tutorials, vol. 19, no. 1, pp. 239-284, 2017.
2. Truong, K.T.; Heath, R.W. Effects of channel aging in massive MIMO systems. J. Commun. Netw. 2013, 15, 338-351.
3. Yin, H.;Wang, H.; Liu, Y.; Gesbert, D. Addressing the Curse of Mobility in Massive MIMO With Prony-Based Angular-Delay Domain Channel Predictions. IEEE J. Sel. Areas Commun. 2020, 38, 2903-2917.
4. 3GPP, NR; Physical channels and modulation, 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.211, 10, version 16.3.0.
5. R. Chopra, C. R. Murthy, H. A. Suraweera, and E. G. Larsson, Performance analysis of FDD massive MIMO systems under channel aging, IEEE Trans. Wireless Commun., vol. 17, no. 2, pp. 1094-1108, Feb. 2018.
6. L. H. Nguyen, R. Rheinschmitt, T. Wild, and S. Brink, Limits of channel estimation and signal combining for multipoint cellular radio, in Proc. 8th Int. Symp. Wireless Communication Systems, 2011, pp. 176-180.
7. J. Zheng, J. Zhang, E. Bjornson, and B. Ai, Impact of channel aging on cell-free massive MIMO over spatially correlated channels, IEEE Trans. Wireless Commun., vol. 20, no. 10, pp. 6451-6466, 2021.
8. M. Bengtsson and B. Ottersten, Optimal and suboptimal transmit beamforming, in Handbook of Antennas in Wireless Communications, L. C. Godara, Ed. CRC Press, 2001.
9. W. Yu and T. Lan, Transmitter optimization for the multi-antenna downlink with per-antenna power constraints, IEEE Trans. Signal Process., vol. 55, no. 6, pp. 2646-2660, 2007.
10. S. Jaeckel, L. Raschkowski, K. Boerner and L. Thiele, F. Burkhardt, E. Eberlein, "QuaDRiGa - Quasi Deterministic Radio Channel Generator," User Manual and Documentation. Tech. Rep. v2.2.0, Fraunhofer Heinrich Hertz Institute (2019).
11. S. Jaeckel, L. Raschkowski, K. Boerner and L. Thiele, QuaDRiGa: A 3-D Multicell Channel Model with Time Evolution for Enabling Virtual Field Trials, IEEE Transactions on Antennas Propagation, 2014.
12. M. Grant, S. Boyd CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx.

UDC: 519.217.2

Mathematical model for analyzing all-optical switch performance metrics in transient mode

K.A. Vytovtov[1], E.A. Barabanova[1], G.K. Vytovtov[2], N.A. Antonov[3]

[1] V. A. Trapeznikov Institute of Control Sciences of RAS, 65 Profsoyuznaya Street, Moscow, Russia
[2] Astrakhan State Technical University, 16 Tatishchev Street, Astrakhan, Russia
[3] MIREA - Russian Technological University, 78 Vernadsky Avenue, Moscow, Russia

vytovtov_konstan@mail.ru, elizavetaalexb@yandex.ru

Abstract

A transient behavior of all-optical switch using a single-channel finite buffer queuing system with Poisson input flow and exponential service time distribution model is considered in this paper. The apparatus of the Laplace transform was used to solve the Kolmogorov equations system. Analytical expressions for the probability of losses and the throughput of all-optical switch in transient mode are found.

Keywords: all-optical switch, transient mode, queueing systems, Laplace transform, state probabilities, throughput

1. Introduction

With the development of new-generation optical networks, there has been an increased interest in their modeling. In most cases, mathematical tools from the theory of queuing systems [1, 2] are employed for this purpose. Depending on the parameters of optical network devices, various types of queuing systems (QS) can be used to describe their operation [3]. Performance metrics of QS, such as buffer size, number of requests in the system, loss probability, and waiting time of requests in the queue, can be described both in steady-state mode [1-3] and in transient mode [4–10]. In the steady-state, also known as the stationary mode, the system's characteristics do not depend on time. The transient mode is achieved after a certain period from the start of the system's operation or its restart and continues for some time before transitioning to the steady-state mode. Such modes also occur when

The reported study was funded by Russian Science Foundation, project number 23-29-00795, https://rscf.ru/en/project/23-29-00795/.

there is a sudden increase in the flow of requests in the system or an abrupt failure of a servicing device.

It should be noted that a significant portion of research on QS and the search for their performance metrics are conducted in the steady-state mode [1,3]. However, as the capacity of next-generation communication networks increases and the transition to all-optical systems occurs [2], the duration of the transition mode becomes comparable to the duration of an optical packet. Therefore developing a model for calculating packet loss probability and other performance metrics of all-optical switching devices in non-stationary conditions will allow communication operators to formulate accurate and precise requirements for the technical specifications of switching equipment, leading to improved a quality of provided services.

Evaluating the duration of the transition mode based on input flow parameters and servicing device characteristics enables the optimization of system operation during the transition mode, minimizing its impact on overall system performance, and is a critical problem in the design of communication systems.

The analytical expressions of the queuing systems transient characteristics is essential for analyzing operation of queuing systems and are necessary for solving synthesis problems for such systems.

This study investigates the model of the all-optical switch with a finite buffer and analyzes its characteristics in the transition mode using the analytical method based on Laplace transforms. The state diagram of the system is presented, and the problem formulation is described. The analytical method for solving Kolmogorov equations which describe the operation of the considered system in the transition mode is used.

2. The statement problem

In this study, the model of all-optical switch in the transition mode is considered. The model represents a queuing system $M/M/1/N$, where the input is a simple arrival process with intensity λ, and the service time of requests follows an exponential distribution with parameter μ. When the device is busy, a request enters the buffer if there is available space. If all N places in the buffer are occupied, the request exits the system unprocessed. The state diagram of the analyzed queuing system is presented in Fig. 1.

In Fig.1 S_0 is the initial state at which the system has no orders; S_1 is one order is processed by the server and the buffer is free; S_2 is the state when one order is processed by the server and one request is in the buffer; S_n is the state when one request is processed by the server and n request is in the buffer; S_{n+1} is the state

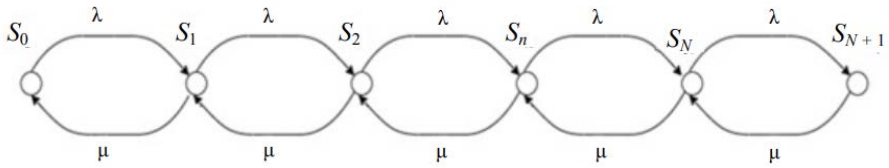


Fig. 1. Graph of states of $M/M/1/N$ system

when one request is processed by the server, the buffer is full, and the newly received request is lost.

The Kolmogorov equations describing the operation of this system have the form

$$\begin{cases} \dfrac{dP_0(t)}{dt} = -\lambda P_0(t) + \mu P_1(t), n = 0 \\ \dfrac{dP_n(t)}{dt} = \lambda P_{n-1}(t) - (\lambda + \mu) P_n(t) + (n+1)\mu P_{n+1}(t), \quad 1 \le n \le N \\ \dfrac{dP_{N+1}(t)}{dt} = \lambda P_N(t) - \mu P_{N+1}(t) \end{cases} \tag{1}$$

where $P_n(t)$ is the probability of the system being in state n at time t, corresponding to the probability of having n packets in the system. Thus, $P_n(t)$ is the probability of having no requests in the system, $P_{n-1}(t)$ is the probability of having N–1 packets in the system, and $P_{n+1}(t)$ is the probability of the buffer being fully occupied, resulting in the loss of the next arriving packet, which corresponds to the packet loss probability. The development of an analytical model for an optical switch with a limited buffer is required to analyze the state probabilities of its operation and its performance characteristics in both steady-state and transient regimes.

3. Description of the Analytical Method

To solve the system of equations (1), we will use the Laplace transform technique [1], which allows us to transform the system of differential equations (1) into a system of linear algebraic equations and significantly simplify its solution. To begin, let's express the system (1) in matrix form:

$$\frac{d\overrightarrow{P}(t)}{dt} = \mathbf{B}\overrightarrow{P}(t) \tag{2}$$

where $\overrightarrow{P}(t)$ is the vector of state probabilities; $\mathbf{B}$ is the matrix of coefficients for the differential equations of the system (1). After applying the Laplace transform to the

system (2) $\int_0^\infty e^{-st}\frac{d\overrightarrow{P}(t)}{dt} = \int_0^\infty e^{-st}\mathbf{B}\overrightarrow{P}(t)dt$ and utilizing the property of differentiation $\frac{d\overrightarrow{P}(t)}{dt} \to s\overrightarrow{P}(s) - \overrightarrow{P}(0)$, where s is a complex variable; $\overrightarrow{P}(s)$ is Laplace domain representation of the vector of state probabilities; $\overrightarrow{P}(0)$ is the vector of initial conditions, we rewrite (2) as follows:

$$s\overrightarrow{P}(s) - \overrightarrow{P}(0) = \mathbf{B}P(s)$$

Thus, we obtain a non-homogeneous system of linear algebraic equations

$$(\mathbf{B} - s\mathbf{I})\overrightarrow{P}(s) = -\overrightarrow{P}(0) \tag{3}$$

where I – the identity diagonal matrix. Let's express (3) as

$$\mathbf{A}P(s) = -\overrightarrow{P}(0) \tag{4}$$

where $\mathbf{A} = (\mathbf{B} - s\mathbf{I})$.

By solving the system (4) using Cramer's method, we obtain a vector of Laplace domain representations of the state probabilities $\overrightarrow{P}(s)$, where the k-th element is given by

$$P_k(s) = \frac{\Delta_k(s)}{\Delta(s)},$$

Where Δ is the determinant of the matrix $\mathbf{A}$; $\Delta_k(s)$ is the determinant of the matrix $\mathbf{A}_k$, obtained by replacing the k-th column of matrix $\mathbf{A}$ with the vector of initial condition $\overrightarrow{P}(0)$.

Therefore, the solutions of the system (4) are operator expressions for the probabilities of k-th states of the system $P_k(s)$, represented as fractions, where the numerators and denominators contain power polynomials $\Delta_k(s)$ and Δ respectively. To transition to the original domain $P_k(t)$ we will use the expansion formula, according to which the final expressions for the probabilities of k-th states of the system in the transition regime will be as follows:

$$P_k(t) = \sum_{i=1}^{n} \frac{\Delta_k(s_i)}{\frac{d\Delta(s_i)}{ds}} e^{s_i t} \tag{5}$$

where n is the total number of states of the system.

4. Performance Characteristics of Optical Switch in Transition mode

Let's examine the key performance characteristics of the optical switch in the transition mode.

4.1. Packet Loss Probability. The packet loss probability is equal to the probability of the system being in state when the servicing device is busy, and the switch buffer is completely filled. In this state, an incoming packet is discarded. Thus, the expression for the packet loss probability is given by

$$P_{loss}(t) = \sum_{i=1}^{n} \frac{\Delta_n(s_i)}{\frac{d\Delta(s_i)}{ds}} e^{s_i t} \tag{6}$$

4.2. Throughput Capacity of an Optical Switch in Transition mode. As the optical switch processes incoming packets in all system states except $P_loss(t)$, either by immediately transmitting them to the servicing device, as in state $P_0()t$, or by temporarily holding them in the buffer when the switch is in one of the k-th $P_k(t)$ states $k = \overline{1, n-1}$, and since the sum of the probabilities of all states equals one, the throughput capacity of the optical switch in the transition regime is determined by $A(t) = [1 - P_{loss}]\lambda$, where λ is the arrival rate of packets. The final expression for the throughput capacity of the optical switch in the transition regime, taking into account (6), is given by

$$A(t) = \left[1 - \sum_{i=1}^{n} \frac{\Delta_n(s_i)}{\frac{d\Delta(s_i)}{ds}} e^{s_i t}\right] \lambda$$

4.3. Transition time. Transition time refers to the period from the start of the transition regime until it reaches a steady-state, i.e., a state where its characteristics can be considered constant and independent of time. To determine the transition time, it is initially necessary to calculate the time constant using the formula

$$\tau = max\{\, \tau_i, i = \overline{1, n+2}\} \; = max\left\{ \frac{1}{|\alpha_i|, i = \overline{1, n+2}} \right\} \tag{7}$$

where α is the real part of the complex variable Δ_s.

Expression (7) signifies that out of all the roots of the polynomial $\Delta(s)$ it is necessary to select a nonzero root of the equation $\Delta(s) = 0$, where $s_i = \alpha_i + j\beta_i$ with the smallest real part α_i, $j\beta_i$ – imaginary part. The smallest α_i will determine the largest value of the time constant [10]. Having the time constant, the transition time can be determined using the formula $t_{tr} = (3 \div 5)\tau$.

5. Conclusion

This study has conducted an analysis of the transient mode of operation of an optical switch, represented as a model of a single-line queuing system with a limited buffer, Poisson input flow, and exponential service time distribution. The Laplace

transform technique was utilized to solve the Kolmogorov equations and determine the state probabilities of the system in the transient mode. The proposed analytical model enables the investigation of transient performance characteristics of an optical network switch.

REFERENCES

1. Dudin, A.N., Klimenok, V.I., Vishnevsky, V.M. The Theory of Queuing Systems with Correlated Flows. Springer: Berlin/Heidelberg, Germany. 2020. 410p.
2. Barabanova E., Vytovtov K., Vishnevsky V., Khafizov I. Analysis of Functioning Photonic Switches in Next-Generation Networks Using Queueing Theory and Simulation Modeling // Communications in Computer and Information Science.- 2023.- vol. 1748.- P. 356–369.
3. Rohit Singh Tomar, Dr.R.K.Shrivastav. Three Phases of Service For A Single Server Queueing System Subject To Server Breakdown And Bernoulli Vacation // International Journal of Mathematics Trends and Technology (IJMTT).- 2020.- vol. 66 (5).- P. 124–136.
4. Neelam Singla, Garg P.C. Transient and Numerical Solution of a Feedback Queueing System with Correlated Departures // American Journal of Numerical Analysis.- 2014. vol. 2 (1).- P. 20–28.
5. Shyam Sundar Sah, Ram Prasad Ghimire. Transient Analysis of Queueing Model // Journal of the Institute of Engineering.- 2015. vol. 11 (1).- P. 165-171.
6. Kempa Wojciech M., Paprocka Iwona. Transient behavior of a queueing model with hyper-exponentially distributed processing times and finite buffer capacity // Sensors.- 2022.- vol. 22(24). 9909. https://doi.org/10.3390/s22249909
7. Shyam S. S., Ram P. G. Transient analysis of queueing model // Journal of the Institute of Engineering.- 2015.- vol. 11 (1). -P. 165-171. DOI: 10.3126/jie.v11i1.14711.
8. Rakesh K., Bhavneet S. S. Transient numerical analysis of a queueing model with correlated reneging, balking and feedback // Reliability: Theory & Applications.- 2019, no. 4 (55).- P. 46-54.
9. Kaczynski W. H., Leemis L. M., Drew J. H. Transient queueing analysis // INFORMS Journal on Computing.- 2012, vol. 24, no. 1.- P. 10–28.
10. Vishnevsky V.M., Vytovtov K.A., Barabanova E.A., Semenova O.V. Transient Behavior of the MAP/M/1/N Queuing System // Mathematics, 2021.- 9(20). 2559.

UDC: 519.217.1

Mathematical models for reliability analysis of all-optical switches

E.A. Barabanova[1], K.A. Vytovtov[1], A.N. Fedorovskaya[2]

[1]V. A. Trapeznikov Institute of Control Sciences of RAS, 65 Profsoyuznaya Street, Moscow, Russia

[2]Astrakhan State Technical University, 16 Tatishchev Street, Astrakhan, Russia

elizavetaalexb@yandex.ru, vytovtov_konstan@mail.ru, an_fedorovskaya@mail.ru

Abstract

The reliability of all-optical switches based on new approach is analyzed. The proposed approach allows to take into account three main criteria influencing on switch functioning such as switch architecture, switch technology and method of switching control. The mathematical expressions of the reliability functions and the mean time to failure of well-known all-optical switches have been obtained. The numerical results for reliability functions of Clos, Banyan, Dual and crossbar switches based on electro-optical basic elements are presented.

Keywords: all-optical switches, a reliability function, a reliability block diagram

1. Introduction

All-optical networks are the perspective new generation networks with low latency and high bandwidth. The main element of an all-optical network is an all-optical switch which is characterized by such parameters as switching speed, insertion loss, crosstalk, scheme complexity, throughput and reliability [1-3].

To date, there are a number of approaches that allow calculating the reliability metrics of switching systems such as reliability function and mean time to failure based on analysis of their architecture [4,5]. For example, the 3D model of the all-optical MEMS-switch and his operating mode block diagram is used for investigating reliability of the optical switch in [4]. The reliability metrics of multistage switching systems have been calculated in [5].

It should be noted that for the accurate evaluating the all-optical switch reliability metrics beside of the switch architecture it must be taking into account the other important features such a method of switching scheme controlling and the type of its basic switching elements.

The publication has been prepared with the support of Russian Science Foundation according to the research project No. 23-29-00795, https://rscf.ru/en/project/23-29-00795/.

In this paper the new mathematical models for reliability analysis of all-optical switches are presented. The authors propose the approach taking into account the impact of several characteristics of all-optical switches on their reliability metrics at once. Such characteristics are a switch architecture, a switching control method, and a switch technology.

2. The reliability block diagrams of all-optical switches

First of all, to calculating and analyzing the reliability of all-optical switches, it is necessary to develop their reliability block diagrams. In this work we consider well-known switching schemes that can be used for constructing all-optical switches such as Clos [6], Banyan [7], Dual [8] and crossbar switches [5]. The reliability block diagrams of these schemes are presented in Figure 1, Figure 2, and Figure 3.

The reliability block diagram of Clos scheme consists of basic switching elements of three stages and control device connected in series. The m basic switching elements of central stage are connected in parallel. They are necessary for providing strictly non-blocking switching function but also can be considered as redundancy blocks. The minimum number of such elements is determined by using following rule: $m = 2n - 1$, where $n = (\sqrt{N/2})$ is the optimal number of inputs of the first stage switching blocks [6]. As Banyan and Dual schemes are decentralized control ones their reliability block diagrams do not include the control device block and contain only M basic switching elements connected in series, where M is equal to the number of switch stages. For the Banyan scheme $M = \log_2 N$ [6] and for Dual one $M = 0,5 \log_2 N$, where N is the number of switch inputs. The reliability block diagram of crossbar switch consists of $2N - 1$ basic elements which are necessary for the switching function.

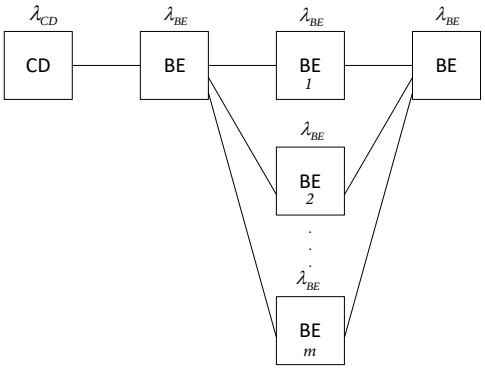


Fig. 1. The reliability block diagram of a Close switch

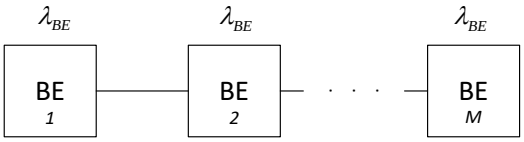


Fig. 2. The reliability block diagram of a Banyan and a Dual switches

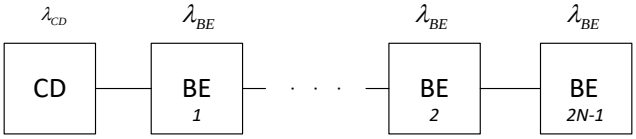


Fig. 3. The reliability block diagram of a crossbar switch

3. The reliability metrics

Using the developed reliability block diagrams as well as the failure rates of different types of optical switches [5] it is possible to calculate a reliability function $R(t)$ and a mean time to failure $MTTF$ (Table 1).

Let us consider the simplest exponential failure model when failure rates are constant for all times. For this case the reliability function of the switching element and the control device can be calculated using formulas: $R(t) = exp(-\lambda_{BE}t)$ and $R(t) = exp(-\lambda_{CD}t)$, where λ_{BE} and λ_{CD} are failure rates of the base element and the control device correspondingly. As the reliability block diagram of the Clos scheme includes m redundancy switching elements in the intermediate stage and two switching elements in the input and in the output stages the final expression of the Clos switch reliability function can be written as following

$$R(t) = (1 - (1 - exp(-\lambda_{BE}t))^m) \cdot exp(-(2\lambda_{BE} + \lambda_{CD})t) \tag{1}$$

The mean time to failure $MTTF$ of all-optical switches can be calculated as

$$MTTF = \int_0^\infty R(t)\, dt \tag{2}$$

Analogously the reliability functions of the other all-optical switching systems have been obtained and presented in table.1.

At the next stage of reliability calculation, the failure rates of switching and control elements must be found. These rates depend on fabric technology and reliability metrics of electron processors [3]. The results of numerical calculation of

Type of switch	$R(t)$	$MTTF$
Close	$(1 - (1 - exp(-\lambda_{BE}t))^{\sqrt{2N}-1}) \times \times exp(-(2\lambda_{BE} + \lambda_{CD})t)$	$\frac{1}{2\lambda_{BE} + \lambda_{CD}} -$ $-\int_0^\infty (1 - exp(-\lambda_{BE}t))^{\sqrt{2N}-1} \times \times exp((2\lambda_{BE} + \lambda_{CD})t)dt$
Dual	$exp(-0.5\lambda_{BE} \cdot \log_2 N \cdot t)$	$\frac{1}{0.5 \log_2 N \cdot \lambda_{BE}}$
Banyan	$exp(-\lambda_{BE} \cdot \log_2 N \cdot t)$	$\frac{1}{log_2 N \cdot \lambda_{BE}}$
Crossbar	$exp(-((2N-1) \cdot \lambda_{BE} + \lambda_{CD})t)$	$\frac{1}{(2N-1) \cdot (\lambda_{BE} + \lambda_{CD})}$

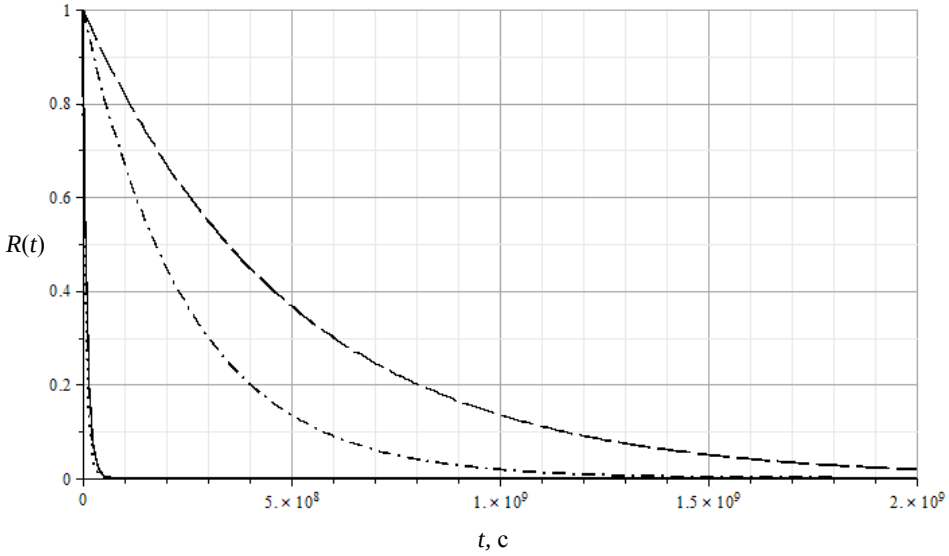
Table 1. The reliability metrics of all-optical switches

the reliability functions for the four described above switch architectures based on electro-optical basic switching elements are presented in Fig.4.

The proposed approach and the developed mathematical models make it possible to compare the reliability of all-optical switches based on different switch-fabric technologies. In this case, it is possible to obtain such a combination of architecture/type of basic elements, in which a switch built according to a less reliable architecture, but using more reliable basic elements, will have higher reliability metrics.

4. Conclusion

The main aim of this work is investigations of the all-optical switch reliability as well as comparison of well-known all-optical switches reliability metrics. The proposed approach is based on the derivation of mathematical expressions of reliability metrics taking into account the reliability block diagrams, the control method, the operation switch algorithms and the type of basic elements. This approach can be



Dotted line - crossbar switch; solid line - Close switch; dash-dotted line - Banyan switch; dashed line - Dual switch

Fig. 4. Comparison of reliability functions of all-optical switches based on electro-optical basic elements

used to obtain mathematical models of the reliability of the other switching circuits of all-optical switches, such as Benes, Spanker schemes, and others.

5. Acknowledgments

The study was supported by a grant from the Russian Science Foundation No. 23-29-00795.

REFERENCES

1. Maier, Martin. Optical switching networks. Cambridge University Press. 2008. 343p.
2. E. A. Barabanova, K. A. Vytovtov, V. M. Vishnevsky, A. V. Dvorkovich and V. F. Shurshev. Investigating the reliability of all-optical switches in transient mode // Journal of Physics: Conference Series. – 2021. vol. 2091.-012039.
3. Xiaohua Ma, G. S. Kuo. Optical switching technology comparison: optical mems vs. other technologies // IEEE Optical Communications.-2003.- 41(11).-P. 16-23.

4. Hasanov M.H., Agayev N.B., Atayev N.A., Fataliyev V.M. (2021) A new generation of controlled optic switch // T-Comm, vol. 15, no.3, - P. 64-68. (in Russian).
5. Fathollah Bistouni, Mohsen Jahanshahi. Scalable crossbar network: a non-blocking interconnection network for large-scale systems // J Supercomput. -2015.- Vol. 71.- P. 697–728.
6. C. Clos. A study of non-blocking Switching Networks // The Bell System Technical Journal.-1953.-P.406-424.
7. M. Zulfin, Suherman, Maksum Pinem, Rahmad Fauzi, M. Razali. Reducing Crosspoints on Multistage Switching by Using Batcher Banyan Switches //Proceedings of The 2nd International Conference On Advance And Scientific Innovation, ICASI 2019, 18 July, Banda Aceh, Indonesia.
8. E. A. Barabanova, K. A. Vytovtov, V. M. Vishnevsky, V. S. Podlazov. High-capacity strictly non-blocking optical switches based on new dual principle // Journal of Physics: Conference Series. – 2021. vol. 2091.-012040.
9. A. Birolini. Reliability Engineering. Theory and Practice. – 3.ed.-Berlin: Springer, 1999.

UDC: 621.3.019.34

Investigation of tethered unmanned high-altitude platform reliability

V.M. Vishnevsky[1], E.A. Barabanova[1], K.A. Vytovtov[1], G.K. Vytovtov[2]

[1] V. A. Trapeznikov Institute of Control Sciences of RAS, 65 Profsoyuznaya Street, Moscow, Russia

[2] Astrakhan State Technical University, 16 Tatishchev Street, Astrakhan, Russia

elizavetaalexb@yandex.ru, vytovtov_konstan@mail.box, an_fedorovskaya@mail.ru

Abstract

The reliability of the tethered unmanned high-altitude platform as a single complex system is investigated in this paper. The reliability block diagram of the platform has been developed. It consists of the main subsystems and its elements connected in series. The method of calculating the reliability metrics of the tethered unmanned high-altitude platform such as the reliability function, and the mean time to failure is presented. The example of the reliability investigation method application and the example of communication subsystem reliability metrics calculation is presented. The approach of fiber-optical cable failure rate calculation is considered.

Keywords: reliability function, mean time to failure, failure rate

1. Introduction

Currently, tethered high-altitude unmanned telecommunication platforms of long-term operation are widely used for solving a class of very important problems, including long-term communication and remote video surveillance for critical infrastructure sites for a long time [1].Thus the main parameter of such drones is reliability, which is primarily determined by the reliability indicators of its main elements.

In some of previous works k out-of-n type models were used to study the reliability of engines of tethered drones in the case when k out of n engines failed [2,3]. The other important problem is to calculate the reliability of the local hybrid navigation system consists of optical and radio subsystems in the cases of bad weather conditions or lifting and landing drone time moments. The reliability analysis of such navigation systems is performed using Markov models considered in transient mode [4].

The publication has been prepared with the support of Russian Science Foundation, according to the research project No. 22-49-02023

In a number of works, the investigating the autonomous quadrocopters reliability was carried out using the fault tree analysis method [5-8].Unlike autonomous drones tethered drones consist of more complex power, communication and control subsystems so the methods of autonomous drone reliability calculation cannot be applied to tethered ones.

The purpose of this work is investigating the reliability of tethered drone as a one complex system, and calculating its reliability indicators such as the reliability function, and the mean time to failure.

2. The reliability block diagram of the tethered unmanned high-altitude platform

The approach of constructing the reliability block diagram of the tethered unmanned high-altitude platform is used in this work to investigate the reliability of tethered drone (Fig.1).

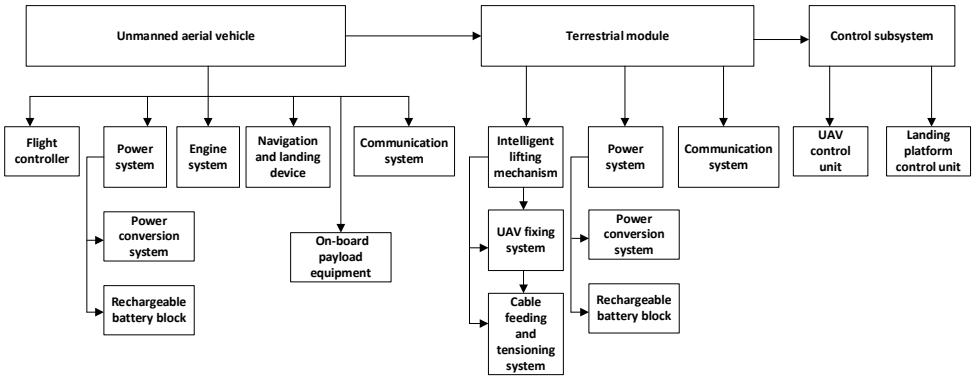


Fig. 1. The reliability block diagram of the tethered unmanned high-altitude platform

The main tethered drone subsystems are the unmanned aerial vehicle subsystem, the Terrestrial module, and the control subsystem. As shown in Fig. 1, the subsystem of the unmanned aerial vehicle includes the following main components such as the flight controller, the power system, the electric engine system, the navigation and landing device, and the communication system. The terrestrial module includes the intelligent lifting mechanism subsystem, a power subsystem, and a communication system. The tethered unmanned aerial platform control subsystem includes the unmanned aerial vehicle (UAV) control unit, and the landing platform control unit.

The ground-to-aircraft transmission of high-power energy system (power system) consists of the power conversion system and the rechargeable battery block. The

intelligent lifting mechanism subsystem consists of the intelligent fixing and the cable feeding and tensioning systems. Since all elements of the reliability block diagram are connected in series, and also taking into account the exponential distribution of failure rates of tethered unmanned aerial platform components, it is possible to calculate the failure rate of the entire system using the formula [9]:

$$\Lambda = \sum_{i=1}^{n} \lambda_i, \tag{1}$$

where n is the number of system elements.

The failure rates of system elements can be calculated by knowing the mean time to failure of their radio electronic components and the methodology for calculating their reliability [9]. The failure rates of complex subsystems can be found using the algorithms of calculating the reliability metrics of engines [2], and navigation system [11].

The failure rates of the system elements calculated in accordance with [10] are presented in Table 1.

Subsystems of the tethered unmanned platform	Failure rates λ_i, 1/h
Flight controller	$0.8963 \cdot 10^{-6}$
Power conversion system	$7,5276 \cdot 10^{-5}$
Rechargeable battery block	$6.8325 \cdot 10^{-5}$
Engine system	$2.5734 \cdot 10^{-6}$
Navigation and landing device	$1.4576 \cdot 10^{-5}$
Communication system	$9,5935 \cdot 10^{-6}$
UAV fixing system	$5,7834 \cdot 10^{-5}$
Cable feeding and tensioning system	$4.4529 \cdot 10^{-4}$
UAV control unit	$2.7623 \cdot 10^{-6}$
Landing platform control unit	$3.5623 \cdot 10^{-5}$

Table 1. Reliability metrics of tethered high-altitude unmanned platform subsystems

Using λ_i from the table 1 and substituting these values into the expression (1) we obtain $\Lambda = 7.1275$ 1/h.

The reliability function of the tethered unmanned high-altitude platform can be calculated as:

$$R(t) = e^{-\Lambda t} \tag{2}$$

$$\int_0^\infty R(t)\mathrm{d}t = \frac{1}{\Lambda} \tag{3}$$

3. The algorithm of communication system reliability calculation

For example, consider the method of calculating the reliability of a communication system for an unmanned aerial platform.

Let us assume that the equipment presented in Table 2 is used to build the communication system.

Communication elements	Type of equipment	$MTTF$, h	λ, 1/h
Ethernet switch	SW-70202	144890	$6,9018 \cdot 10^{-6}$
Fiber transceiver	SFP-S1SC13	445890	$2,2427 \cdot 10^{-6}$

Table 2. Reliability metrics of communication system elements

To calculate the failure rate of an optical cable λ_e using for providing communication from the terrestrial module to the unmanned aerial vehicle the following mathematical model is used [9]:

$$\lambda_e = [\lambda_{b.1} \cdot m \cdot K_{T1} + \lambda_{b.2} \cdot K_{T2}] \cdot L_K \cdot K_O + \lambda_{b.3} \cdot m \cdot K_{T1} \cdot K_{RG1}, \tag{4}$$

where $m = 2$ is the number of optical fibers in the cable; $\lambda_{b.1} = 2.33 \cdot 10^{-15}$ 1/h·m is the basic failure rate of optical fibers in the process of their operating time referred to one meter of cable type length; $\lambda_{b.2} = 6.86 \cdot 10^{-15}$ 1/h·m is the basic rate of sudden failures of the cable structure in the process of their operating time referred to one meter of the cable type length; $\lambda_{b.3} = 1.21 \cdot 10^{-11}$ 1/h·m is the basic rate of gradual failures of fiber-optical cables in the process of their operating time; $L_K =$ 110 meters is the length of the fiber-optical cable; $K_O = 4$ is the coefficient of rigidity of operating conditions for a given group of equipment [10].

The value of the suitability criterion KG_1 is calculated by the formula:

$$KG_1 = \frac{d}{K_{T4}} \tag{5}$$

Here $d = 13$ dBm is the maximum allowable value of the attenuation coefficient in the fibre-optic cable. The value of d is determined by the sensitivity of the receiving optical module SFP. $K_{T4} = 1,3$ is the temperature coefficient characterizing the maximum reversible changes of the fiber-optical cable attenuation coefficient in the range of negative operating temperatures [10]. Therefore $KG_1 = \frac{13}{1,3} = 10$.

The values of temperature coefficients K_{T1} and K_{T2} are determined by the formulas:

$$K_{T1} = exp[-K_{E1}(\frac{1}{T_{eq}} - \frac{1}{298})], \tag{6}$$

$$K_{T2} = exp[-K_{E2}(\frac{1}{T_{eq}} - \frac{1}{298})], \tag{7}$$

where $K_{E1} = 18,06 \cdot 10^3$ is the coefficient that depends on the activation energy of degradation processes [10] for claddings of optical fibers; $K_{E2} = 8,05 \cdot 10^3$ is the coefficient depending on the activation energy of degradation processes [10] for fiber-optical sheaths; T_{eq} is the equivalent operating temperature of components which can be calculated by the formula:

$$T_{eq} = (\frac{1}{T_{max}} + \frac{1}{K_E} ln \frac{(\sum_{i=1}^{n} t_i}{\sum_{i=1}^{n} t_i{}^* \cdot t_{T_{max}}})^{-1}, \tag{8}$$

where

$$t_i{}^* = t_i\{exp[-K_E(\frac{1}{T_i} - \frac{1}{T_{max}})]\}. \tag{9}$$

Here t_i is the total time interval of the communication system element at a temperature T_i; $T_{max} = 313$ K is the maximum operating temperature; $t_{T_{max}}$ is the total time interval of the component operation at the maximum operating temperature.

The example of the distribution of a tethered platform operating time at a certain temperature during the year is presented in Table 3. Here it is assumed that during the year the tethered unmanned aerial platform operates 3456 hours and 18 hours of them at the maximum operating temperature.

Taking into account the valuers of operating time presented in Table 3 and expressions (6)-(9), we have obtained the following coefficients: $K_{T1} = 1821,4$; $K_{T2} = 1393,7$. After substituting K_{T1} and K_{T2} into (4), we get $\lambda_e = 4,49 \cdot 10^{-7}$ 1/h. Then, according to (1) - (3) and taking into account the series connection of elements in the reliability block diagram (Fig.1), we obtain the reliability metrics of the communication subsystem for one year of operation: $\Lambda = 95.935 \cdot 10^{-7}$ 1/h; $MTTF = 104237$ h; $R(t) = 0.92$.

4. Conclusion

This paper proposes the methodology and the mathematical models for calculating the tethered high-altitude unmanned platform reliability as a complex single system. The approach is based on the constructing of the tethered high-altitude unmanned

T_i,K	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
233	-	18	-	-	-	-	-	-	-	-	-	-
238	18	36	-	-	-	-	-	-	-	-	-	-
243	36	72	-	-	-	-	-	-	-	-	-	18
248	72	72	-	-	-	-	-	-	-	-	-	36
253	72	36	-	-	-	-	-	-	-	-	-	72
258	36	36	18	-	-	-	-	-	-	-	18	72
263	36	18	36	18	-	-	-	-	-	18	36	36
268	18	-	72	36	-	-	-	-	18	36	72	36
273	-	-	72	72	18	-	-	-	36	72	72	18
278	-	-	36	72	36	18	-	18	72	72	36	-
293	-	-	36	36	72	36	18	36	72	36	36	-
288	-	-	18	36	72	72	36	72	36	36	18	-
293	-	-	-	18	36	72	72	72	36	18	-	-
298	-	-	-	-	36	36	72	36	18	-	-	-
303	-	-	-	-	18	36	36	36	-	-	-	-
308	-	-	-	-	-	18	36	18	-	-	-	-
313	-	-	-	-	-	-	18	-	-	-	-	-

Table 3. Operating time of a tethered unmanned aerial platform at a given temperature T_i during the month t_i, h

platform reliability block diagram, which is presented for the first time in this paper. The example of calculating the reliability metrics of the tethered unmanned platform communication subsystem is considered.

REFERENCES

1. Vishnevsky V. M., Efrosinin D. V., Krishnamoorthy A. Principles of construction of mobile and stationary tethered high-altitude unmanned telecommunication platforms of long-term operation // Communications in Computer and Information Science (Cham: Springer). 2018. V. 919. P. 561—569.
2. Kozyrev D. V., Phuong N. D., Houankpo N. G. K., Sokolov A. Reliability evaluation of a hexacopter-based flight module of a tethered unmanned high-altitude platform // Communications in Computer and Information Science (Cham: Springer). 2019. V. 1141. P. 646-–656.
3. Vishnevsky V., Selvamuthu D., Rykov V., Kozyrev D., Ivanova N. Reliability modeling of a flight module of a tethered high-altitude telecommunication platform

// 2022 International Conference on Information, Control, and Communication Technologies (ICCT). 2022. https://ieeexplore.ieee.org/document/9976764.

4. Vishnevsky V.M., Vytovtov K. A., Barabanova E. A., Buzdin V. E. Mathematical model for reliability indicators calculation of tethered UAV hybrid navigation system // Journal of Physics: Conference Series 2091. 2021.
5. Kazem Imani, Amirhossein Gholami, Mahdi BagherianDehaghi. Reliability calculation with error tree analysis and breakdown effect analysis for a quadcopter power distribution system // Maintenance, Reliability and Condition monitoring. 2022. V. 2. P. 45–57.
6. Beatriz Juliana de Oliveira Martins Franco, Luiz Carlos Sandoval Góes. Failure analysis methods in unmanned aerial vehicle (UAV) applications // 19th International Congress of Mechanical Engineering. 2007.
7. Enrico Petritoli, Fabio Leccese, and Lorenzo Ciani. Reliability and Maintenance Analysis of Unmanned Aerial Vehicles, `https://doi.org/10.3390/s18093171`
8. Mahmood Shafiee, Zeyu Zhou, Luyao Mei, FatemeDinmohammadi, Jackson Karama and David Flynn. Unmanned Aerial Drones for Inspection of Offshore Wind Turbines: A Mission-Critical Failure Analysis, `https://doi.org/10.3390/robotics10010026`
9. A. Birolini. Reliability Engineering. Theory and Practice. Berlin: Springer, 1999.
10. Reliability of electrical devices. Manual. – M.: MDRF, 2006. (in Russian)
11. Vishnevsky V. M., Vytovtov K. A., Barabanova E. A., Buzdin V. E. and Frolov S. A. Local hybrid navigation system of tethered high-altitude platform // Lecture Notes in Computer Science (Cham: Springer). 2021. V. 13144. P. 67—79.

UDC: 004.896

Machine Learning-Based Models for the Compressibility Factor of Natural Gas

Olga Kochueva [1] and Ruslan Akhmetzianov[1]

[1]National University of Oil and Gas "Gubkin University", 65 Leninsky Prospekt, Moscow 119991, Russia

kochueva.o@gubkin.ru

Abstract

Rational technical solutions related to the design and operation of pipeline systems are based on a large volume of hydraulic calculations, the performance and accuracy of which depends on many factors. One of them is the correct calculation of the compressibility coefficient, which is included in the gas equation of state to represent the real properties of the gas and depends on pressure, temperature and component composition of the gas. We develop and analyze the models for calculating the compressibility factor trained on a large amount of data calculated with AGA-8 equation of state. The presented models can be applied to replace the original model when analyzing the development of risk situations and searching for optimal gas transportation modes, in software designed for staff training on computer simulators, such approach is called surrogate modeling.

Keywords: Machine learning, surrogate modelling, compressibility factor

1. Introduction

Hydraulic calculations are the primary tool for rational technical decisions related to the design and operation of pipeline systems. The compressibility factor is introduced into the gas equation of state to account for its real properties and depends on the pressure, temperature, and gas component composition. Most explicit and implicit approximation dependencies have been developed based on the accumulated significant volume of experimental data. The calculation of the compressibility factor with regard to gas fraction composition can be performed according to the equation of state (in recent years, the work of most researchers is based on GERG-2008, AGA8 [1], [2] , AGA10), where the problem is reduced to an iterative method of solving a nonlinear equation. The time of calculations plays a significant role in non-stationary modeling of gas transfer modes in main pipelines as

well as determining optimal mode of gas transmission and parameters identification of a gas transportation system. The work aimed to build approximations to calculate the compressibility factor using machine learning algorithms and to analyze the quality of the resulting models.

In recent decades, finding the best computational method for the compressibility factor has been a hot research issue. In [3] the relationship in explicit form, constructed on the basis of a set of experimental data (3038 records) is presented. The correlation is built as a fraction, the numerator and denominator of which are functions containing the reduced pressures and temperature in various degrees and their logarithms, there are 20 coefficients in the formula. The paper [4] presents the formula that is a modification of the implicit relation [5]. The resulting model contains 19 coefficients, and the polynomial functions of reduced temperature and pressure and exponential functions of temperature are used in the calculation. A feature of [6] is the division of the range of reduced pressure values into 2 subsets, a separate model is built for each range. Due to this it became feasible to reach a higher accuracy of the model. The authors used the group method of data handling (GMDH), the proposed models do not include logarithmic or exponential functions and can be easily used in programming of flow-meters. The paper [7] introduces a model based on a multidimensional nonlinear relationship, built on a sample of 6988 values obtained from the digitized Standing–Katz diagram. The authors give a formula that is the quotient of two polynomials, the formula has 20 coefficients, a mean absolute percentage error (MAPE) about 1.5% is indicated. To calculate the compressibility factor in [8], an artificial neural network with two hidden layers is proposed, it was trained on 4158 experimental data entries. The papers [9] and [10] presented correlations to calculate the Z-factor explicitly based on genetic programming or Symbolic Regression (SR) method. These two separate studies considered different ranges of pressure and temperature, presented different correlations and reported accuracy of the models (MAPE) of 2.5% and 0.03%, respectively). However, i it is difficult to compare the results of all mentioned models since the testing was carried out on a disparate data. In many works, authors usually present cross-plots determining the ratio between the obtained and experimental values of the compressibility factor, they also provide the determination coefficient values and the average absolute error, but the majority of them do not conduct a complete analysis of the models including the study of the derivatives of the compressibility factor with respect to pressure and temperature. Our study fills this gap.

2. Generating a data set

The aim of the work was to obtain and analyze models for the compressibility factor, which could replace the iterative procedure proposed in [2], for the ranges

of pressure, temperature and gas compositions, typical for the main gas pipelines modes. Such approximations, provided that the deviation of the compressibility factor calculated by them will not exceed 0.05-0.1% of the result obtained by the procedure [2], can be used when carrying out calculations of non-stationary modes of gas transfer, when searching for the optimum conditions, in simulators - everywhere where the speed of calculation is important. The authors didn't pursue the goal to obtain universally applicable models, thus we specified a temperature in the range from 273 to 333 K and two pressure ranges: $P_1 \in [3.5 - 5.6]$ MPa and $P_2 \in [5.5 - 7.5]$ MPa for main pipelines of different capacities. The gas compositions as a mixture of methane (from 90% to 97%), hydrocarbons from C_2 to C_6 plus nitrogen, carbon dioxide, and helium were taken to generate initial data. Since many components are contained in natural gas in small amounts, the study planned to analyze the possibility of reducing the number of input variables without compromising the model quality, therefore the molar mass of the gas was added to the input parameters.

For different gas compositions, pressure values were set with a step of 0.02 MPa, temperature with a step of 2K, and the values of the compressibility factor were calculated with the method [2]. Thus, initial data for model training and testing were generated with the number of entries 24722 for the range P_1 and 23028 for the range P_2. It should be noted that the specified ranges are not a limitation for the proposed technique.

3. Results and discussion

We generated 4 models, based on Random Forest (RF) algorithm [11], Group Method Data Handling (GMDH) [12], Symbolic Regression (SR) [13], artificial neural network (ANN). Two of them - SR and GMDH present the approximation in explicit form. Symbolic Regression is not as widely used machine learning algorithm as RF or ANN, its advantage is computational efficiency, the possibility to control the complexity of the obtained correlations, the fact that dependence can be obtained in explicit form. A number of publications ([14], [15], [16]) emphasize the effectiveness of its use to solve a number of practical problems. The detailed description of models built with SR method is presented in [10]. The developed models for calculating the compressibility factor have R^2=0.99, MAE=0.0002-0.001, MAPE=0.02-0.1%. The model based on GMDH algorithm has the best metrics, the models obtained by other machine learning methods lose slightly in accuracy, although their MAPEs satisfy the requirement to be less than 0.1 %.

The hypotheses of constancy of the mathematical expectation, the normality of the distribution, the absence of autocorrelation, and the constancy of the variance were tested for the residuals of each model. We used ANOVA for testing the hypotheses of constancy of the mathematical expectation, Goldfeld–Quandt test for checking

homoscedasticity, Pearson χ^2 test for the hypothesis about normal distribution of the residuals and plotting autocorrelation function (ACF). For all models, conclusions were drawn about the constancy of the mathematical expectation of the residuals, the normality of the distribution, the absence of autocorrelation, and the constancy of the variance.

We compared the derivatives of the compressibility factor with respect to pressure and temperature calculated by each model with the derivative numerically calculated with method [2]. Models built with SR and GMDH methods demonstrate good correspondence of the derivatives of the compressibility factor both with respect to pressure and to temperature with the derivatives of compressibility factor calculated with method [2]. Figure 1 on the left presents graph of $\partial Z/\partial T$ for $P = 6.84$ MPa for SR model.

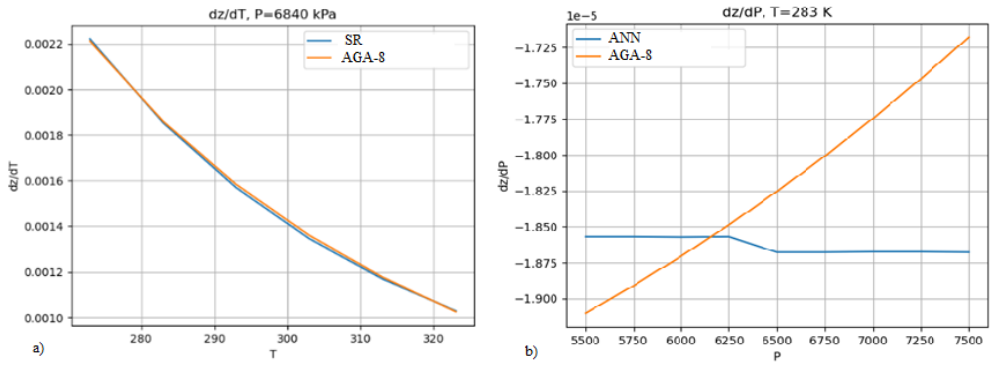


Fig. 1. Derivatives of the compressibility factor with respect to pressure for SR method (on the left) and with respect to temperature for ANN (on the right).

The model built with GMDH method showed good performance, its mean absolute percentage error for $\partial Z/\partial P$ when $T = 283$K is about 1.4%, and for $\partial Z/\partial T$ when $P = 6.48$MPa is less 1%. Models built with RF and ANN methods demonstrate poor agreement with the derivatives of compressibility factor calculated with method [2]. Figure 1 on the right shows a graph of $\partial Z/\partial P$ for $T = 283$K for ANN model.

We have tested the models developed for the interval P_2 for pressure $P < 5.5$ MPa and $P > 7.5$ MPa. The predictive performance of the GMDH model and SR model for pressure values beyond the intervals where the models were trained, turned out to be of acceptable quality. ANN and RF models had poor performance for the interval $P > 7.5$ MPa and $P < 5.5$ MPa.

One of the main objectives of our study was to develop the computationally effective approximation of the compressibility coefficient, which can be used in models

of unsteady gas flow, in solving optimization problems, etc. MAPE and computation time for different machine learning algorithms are shown in Figure 2. Calculation time is given in seconds to calculate 100 values of the compressibility coefficient for the samples from the test set. The last bar corresponds to the method [2], since it is the source of the dataset, its error is assumed to be 0. The best computational time belongs to multiple linear regression (MLR), symbolic regression (genetic algorithm (GA)) and GMDH.

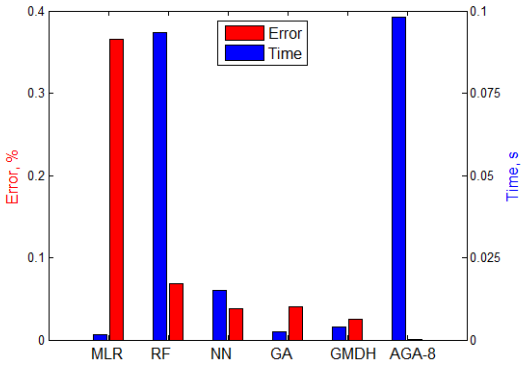


Fig. 2. MAPE and computation time for different machine learning algorithms.

4. Conclusion

The developed surrogate models for calculating the compressibility factor have R^2=0.99, MAE=0.0002-0.001, MAPE=0.02-0.1%. The error of the calculation methods used to obtain the initial data for model building does not exceed 0.2%. The results indicate that the goal of the work has been achieved, and approximations have been obtained that can successfully replace computationally demanding iterative procedures without loss of accuracy. The best results including the shape of the derivatives of the compressibility factor with respect to pressure and temperature, showed models based on GMDH method and symbolic regression. The resulting models are easy to integrate into existing software and in higher-level models.

REFERENCES

1. ISO 12213-2:2006, Natural Gas — Calculation of Compression Factor Switzerland, Geneva, ISO, 2006
2. Repository for the supplementary files to AGA 8 NIST USA, `https://pages.nist.gov/AGA8/` (accessed on 10.06.2023)

3. Azizi N., Behbahani R., Isazadeh M.A. An efficient correlation for calculating compressibility factor of natural gases.// J. Nat. Gas Chem. 2010. V. 19. P. 642 – 645.
4. Kareem L.A., Iwalewa T.M., Al-Marhoun M. New explicit correlation for the compressibility factor of natural gas: linearized z-factor isotherms //J Petrol Explor Prod Technol. 2016. V. 6. P. 481 – 492.
5. Hall K.R., Yarborough L. A new equation-of-state for Z-factor calculations //Oil Gas J. 19 P. 73. V.71, P. 82 – 92.
6. Luan Lin, Shiyang Li, Sihao Sun, Yaqi Yuan, Ming Yang, A novel efficient model for gas compressibility factor based on GMDH network // Flow Measurement and Instrumentation. 2020. V. 71. 101677.
7. Wang Y., Ye J, Wu Sh. An accurate correlation for calculating natural gas compressibility factors under a wide range of pressure conditions // Energy Reports. 2022. V.8(2). P. 130 – 137.
8. Azizi N., Rezakazemi M., Zarei M.M. An intelligent approach to predict gas compressibility factor using neural network model // Neural Computing and Applications. 2019. V.31(1). P. 55 – 64.
9. Towfighi S. An empirical equation for the gas compressibility factor // Z. Pet. Sci. Technol. 2020. V.38. P. 24 – 27.
10. Kochueva O., Zadorozhnyy V. Analysis of approximations of the gas compressibility factor derived from genetic algorithms // E3S Web Conf. Mathematical Models and Methods of the Analysis and Optimal Synthesis of the Developing Pipeline and Hydraulic Systems. 2023. V.397. 01005.
11. Breiman, L. Random forests. // Mach. Learn. 2001. V.45. P. 5 – 32.
12. Madala H.R., Ivakhnenko O.G. Inductive Learning Algorithms for Complex Systems Modeling. Boca Raton: CRC Press, 1994, ISBN 978-0849344381
13. Koza J.R. Genetic programming: on the programming of computers by means of natural selection. The MIT Press, Cambridge, 1992.
14. Kochueva O., Nikolskii K. Data Analysis and Symbolic Regression Models for Predicting CO and NOx Emissions from Gas Turbines // Computation. 2021. V.9. 139.
15. Kochueva O. Razrabotka modelej prognozirovaniya vybrosov oksidov ugleroda i azota gazovyh turbin na osnove geneticheskih algoritmov //Delovoj zhurnal Neftegaz.RU. 2022. V. 5-6 (125-126). P. 14 – 20. (in Russian).
16. Praks P., Lampart M., Praksová R., Brkić D., Kozubek T., Najser J. Selection of Appropriate Symbolic Regression Models Using Statistical and Dynamic System Criteria: Example of Waste Gasification // Axioms. 2022. V. 11. 463.

UDC: 519.873

On Reliability Function of a k-out-of-n Model in Case of Quick Recovery of its Components

Rykov V.[1,2,3] and Ivanova N.[1,4]

[1]Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St., Moscow, 117198, Russian Federation
[2]Gubkin Russian State Oil and Gas University, 65 Leninsky Prospekt, Moscow, 119991, Russia
[3]Institute for Transmission Information Problems (named after A.A. Kharkevich) RAS, Bolshoy Karetny, 19, GSP-4, Moscow, Russia
[4]V.A.Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, 65 Profsoyuznaya street, Moscow, 117997, Russia

vladimir_rykov@mail.ru, nm_ivanova@bk.ru

Abstract

In some previous works, closed-form representations for reliability characteristics of k-out-of-n model with an exponential distribution of components lifetime and arbitrary distribution of their repair time have been found. In that investigation, the markovization method was applied for calculation of the main reliability characteristics. In the recent paper, reliability function is considered for a repairable k-out-of-n model in the same assumptions about life and repair time distributions. The problem of its sensitivity in case of quick recovery of system's components is discussed on an example of 2-out-of-n model. It is shown that the reliability function in scale of its mean time to failure is asymptotically exponential.

Keywords: k-out-of-n model, reliability and sensitivity analysis, arbitrary repair time distribution, quick recovery

1. Introduction

Many real technical systems can be described with a k-out-of-n model. A high reliability of any system is achieved through redundancy and "quick" recovery of its failing components. The k-out-of-n model is an example of a redundant system that is of interest from both theoretical and practical points of view [1].

A number of papers have been devoted to the study of various redundant systems under "quick" recovery. In particular, B. Gnedenko [2, 3] considered a double redundant system with a Poisson failure flow and an arbitrarily distributed repair

time with the help of the theory of regenerative random processes. A. Solov'ev [4] investigated a system of cold redundancy with arbitrarily distributed life and repair time, using the theory of analytic functions.

This raises the issue of convergence rate of system lifetime distribution to the exponential one. Some works by I. Kovalenko and V. Kalashnikov [5, 6] are dedicated to this problem.

This paper continues a series of investigations [7, 8, 9] of k-out-of-n models and is devoted to calculating the reliability function of this model, as well as to the analysis of convergence rate of the reliability function to the exponential one under "quick" recovery of its components.

2. Problem Setting and Notation

Consider a k-out-of-n model with an arbitrary distribution for its component-wise repair. Such a system consists of n components and fails due to any k components' failure. Suppose that

- the system works till it first enters state k, which is the state of the system failure;
- failed system's components are repaired by a single repair facility;
- component's failure arises according to a Poisson flow with intensity α;
- repair times of components are independent and their common cumulative distribution function (c.d.f.) $B(t)$ is absolutely continuous with probability density function (p.d.f.) $b(t) = B'(t)$, mean $b = \int\limits_0^\infty (1 - B(x))dx < \infty$ and its Laplace transform (LT) $\tilde{b}(s) = \int\limits_0^\infty e^{-sx} b(x)dx$;
- the system's states space $E = \{0, 1, ..., k\}$ means the number of failed components.

To study considered k-out-of-n model, use the so-called markovization method based on introduction of supplementary variables [10]. Thus, consider a two-dimensional process $Z = \{Z(t) = (J(t), X(t)),\ t \geq 0\}$, where the first component $J(t)$ represents the number of failed components at time t, and the supplementary variable $X(t)$ means elapsed repair time, that is the time spent by the repair facility for restoration of the component being repaired.

Due to the supplementary variable, the process Z is a Markov one. Denote its micro-state p.d.f.'s with respect to the supplementary variable in domain $0 \leq x \leq t < \infty$ by

$$\pi_j(t; x)dx = \mathbf{P}\{J(t) = j,\ x < X(t) \leq x + dx\}\ \ (j = \overline{1, k})$$

and appropriate macro-state probabilities for $t \geq 0$ by

$$\pi_j(t) = \mathbf{P}\{J(t) = j\} = \int_0^t \pi_j(t;x)dx \ \ (j = \overline{1,k}).$$

In this paper, we deal with the system reliability function $R(t)$,

$$R(t) = \mathbf{P}\{T > t\},$$

where $T = \inf\{t : \ J(t) = k\}$ is the system lifetime.

3. Main Results

For the study of reliability function of the considered model, consider the process Z with absorption state k. Its transition graph is presented in Fig. 1. Here, we denote $\lambda_i = (n-i)\alpha, \ (i = \overline{0,k-1})$ the system failure intensity when i components of n fail, $\beta(x) = \frac{b(x)}{1-B(x)}$ is a conditional repair density of elements, given elapsed repair time is x.

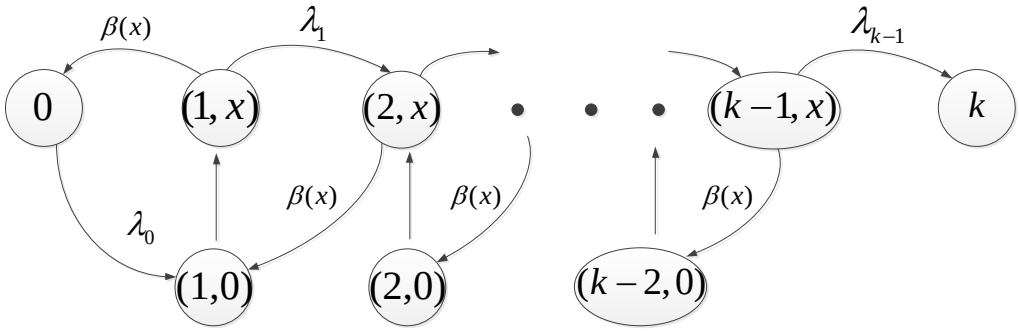


Fig. 1. Transition graph of the process Z with absorption

In [7] the Kolmogorov system of equations for the process micro-state probabilities have been found and the algorithm of its solution based on the method of characteristics was proposed to use. To find an analytic solution of this system consider further a 2-out-of-n model. The system of Kolmogorov forward partial differential equations for process Z with absorption state $k = 2$ in the scope $0 \leq x < t < \infty$ has

the following form

$$\begin{aligned} \frac{d}{dt}\pi_0(t) &= -\lambda_0\pi_0(t) + \int_0^t \beta(x)\pi_1(t,x)dx, \\ \left(\frac{\partial}{\partial t} + \frac{\partial}{\partial x}\right)\pi_1(t;x) &= -(\lambda_1 + \beta(x))\pi_1(t,x), \\ \frac{d}{dt}\pi_2(t) &= \lambda_1 \int_0^t \pi_1(t,x)dx, \end{aligned} \tag{1}$$

jointly with initial

$$\pi_0(0) = 1, \tag{2}$$

and boundary conditions

$$\pi_1(t,0) = \lambda_0\pi_0(t). \tag{3}$$

The reliability function is connected with the probability of absorbing state by the expression,

$$R(t) = \mathbf{P}\{T > t\} = 1 - \mathbf{P}\{T \le t\} = 1 - \pi_k(t) = 1 - \pi_2(t), \tag{4}$$

which is its closed-form representation.

According to the algorithm, proposed in [7], obtain Laplace transform (LT) of reliability function of a 2-out-of-n model,

$$\tilde{R}(s) = \frac{s + \lambda_0(1 - \tilde{b}(s+\lambda_1)) + \lambda_1}{(s+\lambda_1)(s + \lambda_0(1 - \tilde{b}(s+\lambda_1)))}, \tag{5}$$

whence it is easy to obtain mean system lifetime,

$$\mathbb{E}[T] = \tilde{R}(0) = \frac{1}{\lambda_1} + \frac{1}{\lambda_0(1 - \tilde{b}(\lambda_1))}.$$

The final form of the reliability function can be obtained by passing to the inverse LT in the expression (5). However, this step can be done only in some special cases of component's repair time distribution, for example for distributions with fractionally linear Laplace transform.

For an arbitrary distribution of repair time, we can only find two real roots of the denominator of equation (5). One of them evidently equal to $s = \lambda_1$ and the second one can be found from the equation $s + \lambda_0 = \lambda_0\tilde{b}(s + \lambda_1)$, that holds from fig.2.

Let the component repair time has an exponential distribution with parameter β, that leads to $\tilde{b}(s) = \dfrac{\beta}{s+\beta}$. Then, the following statement holds.

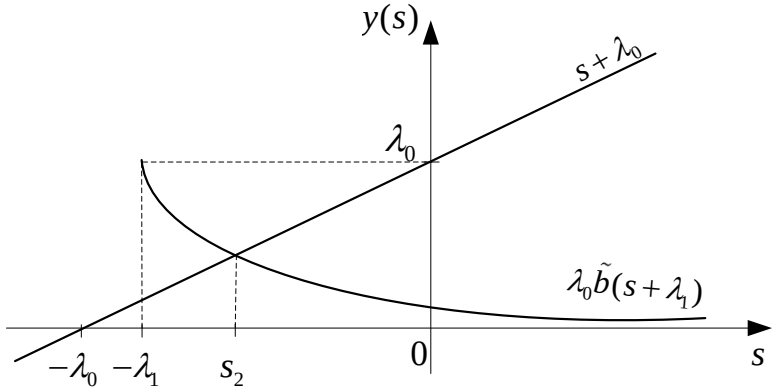


Fig. 2. Graphical representation of the root s_2

Statement. The reliability function $R(t)$ and the mean system lifetime $\mathbb{E}[T]$ of a 2-out-of-n model with exponentially distributed life and repair time take the forms,

$$R(t) = 1 - \frac{n(n-1)\alpha^2}{s_1 - s_2} \cdot \left[\frac{1 - e^{s_2 t}}{s_2} - \frac{1 - e^{s_1 t}}{s_1} \right], \quad \mathbb{E}[T] = \frac{\beta + \alpha(2n-1)}{n(n-1)\alpha^2},$$

where

$$s_{1,2} = \frac{1}{2} \left[-\alpha(2n-1) - \beta \pm \sqrt{\alpha^2 + 2\alpha\beta(2n-1) + \beta^2} \right].$$

For a particular case of the 2-out-of-n model, one can estimate and prove the convergence of the reliability function to the exponential one on the scale of its mean system lifetime. Denote $\rho = \frac{\beta}{\alpha}$ relative speed of component's recovery.

Theorem 1. As $\rho = \frac{\beta}{\alpha} \to \infty$ there is a uniform convergence of the reliability function of a 2-out-of-n model on the scale of its mean system lifetime: $\hat{R}(t) \to e^{-t}$, and the rate of convergence is of the order ε:

$$|\hat{R}(t) - e^{-t}| < \varepsilon,$$

where $\varepsilon = \dfrac{n(n-1)}{(\rho + 2n - 1)^2}$.

4. Conclusion

In this paper, the reliability function of k-out-of-n model with arbitrary distribution of component repair time is considered on an example of 2-out-of-n model.

The exact formulas for the reliability function and the mean system lifetime are obtained in the case of an exponential distribution of the repair time. For this model, the convergence of the reliability function on the scale of its mean lifetime to the exponential one is studied, and an estimate of convergence rate is obtained.

REFERENCES

1. K. Trivedi, Probability and Statistics with Reliability, Queuing and Computer Science Applications, 2nd ed., John Wiley Sons, USA, 2016.
2. B. Gnedenko, On cold double redundant system, Izv. AN SSSR. Texn. Cybern. (4) (1964) 3–12.
3. B. Gnedenko, On duplication with renewal, Izv. AN SSSR. Texn. Cybern. (5) (1964) 111–118.
4. A. Solov'ev, Asymptotic distribution of the lifetime of a duplicated element, Izv. AN SSSR. Texn. Cybern. (5) (1964) 119–121.
5. I. Kovalenko, On the asymptotic consolidation of random processes, Kibernetika 6 (1980) 87–95.
6. V. Kalashnikov, Geometric Sums: Bounds for Rare Events with Applications: Risk Analysis, Reliability, Queueing, Springer, Dordrecht, 1997. doi:10.1007/978-94-017-1693-2.
7. V. Rykov, D. Kozyrev, A. Filimonov, N. Ivanova, On reliability function of a k-out-of-n system with general repair time distribution, Probability in the Engineering and Informational Sciences 35 (2021) 885–902. doi:10.1017/S0269964820000285.
8. V. Rykov, N. Ivanova, Reliability and sensitivity analysis of a repairable k-out-of-n:f system with general life- and repair times distributions, in: In proceedings of the 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference. Edited by Piero Baraldi, Francesco Di Maio and Enrico Zio, 2020. doi:10.3850/978-981-14-8593-0 5750-cd.
9. V. Rykov, N. Ivanova, D. Kozyrev, T. Milovanova, On reliability function of a k-out-of-n system with decreasing residual lifetime of surviving components after their failures, Mathematics 10 (2022). doi:10.3390/math10224243.
10. D. Cox, The analysis of non-markovian stochastic processes by the inclusion of supplementary variables, Mathematical Proceedings of the Cambridge Philosophical Society 51 (1955) 433–441. doi:10.1017/S0305004100030437.

UDC: 519.21

About quasi-renewal processes and quasi-regenerative processes

G.A. Zverkina [1,2]

[1]V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, 65 Profsoyuznaya Street, Moscow 117997, Russia

[2]Institute for Information Transmission Problems, of Russian Academy of Sciences (A. A. Kharkevich Institute), Bolshoy Karetny per. 19, build.1, Moscow 127051, Russia

Abstract

The study of the behaviour of stochastic processes in the queueing theory and related fields is often based on the study of the behaviour of regenerative and Markov piecewise linear processes. Here we represent the concept of quasi-renewal and quasi-regenerative processes that are found in applications and discuss these concepts.

Keywords: Renewal theory, Regenerative processes, Queueing theory, Coupling method, Lorden's inequality, Quasi-renewal and quasi-regenerative processes

1. Introduction

Regenerative processes are extremely popular in applied research, in particular, in queuing theory and related fields.

Definition 1 (According to the text [1])**.** $\mathbf{x}_t$ is regenerative with respect to $\mathfrak{Z}$ over $\mathcal{A}$ if for all $\mathbf{z} \in \mathfrak{Z}$, $A \in \mathcal{A}$, there is a function $\phi_A(\cdot)$, depending only on A (and not on $\mathbf{z}$) such that

$$\mathbf{P}\{\mathbf{x}_t \in A | \mathbf{z};\, n_t > 0; T_{n_t}; \mathbf{x}_s, s \leqslant T_{n_t}\} = \Phi_A(t - T_{n_t}))$$

is a valid representation of the conditional probability, for some $\{t_i\}$. We write $\mathbf{x}_t$ is $\mathcal{R}(\mathfrak{Z}, \mathcal{A}, \{t_i\})$. ▷

This is an old definition. Now the definition of a regenerative process is based on the concept of a Markov moment and is longer (see, e.g., [2, 3]). Regenerative processes have useful properties for use in various models of applied problems – see, e.g., [4].

A significant number of publications are devoted to the analysis of the asymptotic behaviour of regenerative processes ([5, 6, 7] et al.).

Since the convergence rate of the regenerative process distribution is bounded from above by the rate of convergence of the backward renewal time distribution of the embedded renewal process, this rate can be estimated, for example, in the same way as in paper [8].

However, the behaviour of some complex models in queuing theory and in related fields is described by non-regenerative Markov processes – see, e.g., [9, 10] et al. In fact, in these papers, processes that are closely related to non-regenerative processes are studied.

2. Quasi-renewal and quasi-regenerative processes

Recall that the renewal process is a counting process $N_s \stackrel{\text{def}}{=} \sum\limits_{i=1}^{\infty} \mathbf{1}\left(\sum\limits_{j=1}^{i} \xi_j < s\right)$ that is specified using the renewal time ξ_j distribution function (d.f.) $F(s)$. Here random variables (r.v.'s) ξ_j are equally distributed and mutually independent; their distribution can be arbitrary and non-singular. The distribution of r.v. can also be specified using the intensity $\lambda(s) \stackrel{\text{def}}{=} \frac{F'(s)}{1-F(s)}$ (or failure rate in reliability theory); $F(t) = 1 - \exp\left(-\int\limits_0^t \lambda(s)\,\mathrm{d}\,s\right)$. For non-continuous distribution functions, we use the generalized intensity $\lambda(s) \stackrel{\text{def}}{=} \dfrac{f(s)}{1-F(s)} - \sum\limits_i \delta(s-a_i)\ln\big(F(a_i+0) - F(a_i-0)\big)$, where $\{a_i\}$ — is the set of all points of discontinuity of a function $F(s)$, and $\delta(\cdot)$ is a standard δ-function (see [11]).

Thus, the renewal process can be defined by the generalized intensity function.

Definition 2 (Quasi-renewal process)**.** The process $\widetilde{N}_t \stackrel{\text{def}}{=} \sum\limits_{i=1}^{\infty} \mathbf{1}\left\{\sum\limits_{s=1}^{i} \widetilde{\xi}_s \leqslant t\right\}$ is named quasi-renewal process if the r.v.'s $\widetilde{\xi}_i$ are defined by their generalized intensity $\lambda_i(s)$ which satisfy the following

Conditions:

1. The (generalized) measurable non-negative functions $\varphi(s)$ and $Q(s)$ exist such that for all $s \geqslant 0$, $\varphi(s) \leqslant \lambda_j(s) \leqslant Q(s)$; **2.** $\int\limits_0^{\infty} \varphi(s)\,\mathrm{d}\,s = \infty$; **3.** $Q(0) < (1-\varepsilon)\delta(0)$ for some $\varepsilon > 0$. ▷

A one-dimensional quasi-renewal process is impossible; it can only be a component of a more complex process. Obviously, the quasi-recovery times given by the compound intensities $\lambda(s)$ are dependent. However, this dependence is not arbitrary.

This dependence is in some sense "weak". Indeed, the intensity of the quasi-recovery is the sum of the function $\varphi(s)$ and some additional variable intensity.

It is easy to see that if two independent r.v. ξ and η are given by the intensities $\lambda(s)$ and $\mu(s)$, then the intensity of the r.v. $\zeta = \min\{\xi, \eta\}$ equals to $\nu(s) = \lambda(s)+\mu(s)$. Accordingly, the flow of moments of the quasi-renewal process can be perceived as a certain fixed flow given by the usual renewal process, to which an additional flow of dependent events is added.

If, in addition to the Definition 2, condition $\int\limits_0^\infty x^{k-1} \exp\left(-\int\limits_0^x \varphi(s)\,\mathrm{d}\,s\right)\,\mathrm{d}\,x = M_k < \infty$ for some $k \geqslant 2$ is added, and the intensity $\lambda(s)$ determines the non-lattice distribution, then the generalized Lorden's inequality holds for the quasi-renewal process.

Theorem 1 (Generalized Lorden's inequality ([12])). If conditions 1–3 are satisfied, and $\ell \leqslant k-1$, then for the backward renewal time $B_t \stackrel{\text{def}}{=\!=} t - \sum\limits_{i-1}^{N_t} \widetilde{\xi}_j$ of the quasi-renewal process the following inequality is true inequalities for the ℓ-th moment of B_t: $\mathbb{E}\,(B_t)^\ell \leqslant \mathbb{E}\,\eta^\ell + \frac{\mathbb{E}\,\eta^{\ell+1}}{(\ell+1)\mathbb{E}\,\zeta} \stackrel{\text{def}}{=\!=} \Xi_\ell$, where $\mathbf{P}\{\eta \leqslant x\} = 1 - \exp\left(-\int\limits_0^x \varphi(t)\,\mathrm{d}\,t\right) = \Phi(x)$; $\mathbf{P}\{\zeta \leqslant x\} = 1 - \exp\left(-\int\limits_0^x Q(t)\,\mathrm{d}\,t\right) = \widehat{\Phi}(x)$.

Definition 3 (Quasi-regenerative process)**.** The process $(X_t,\, t \geqslant 0)$ on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, with a measurable state space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ is quasi-regenerative, if on other probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbf{P}})$ there exists regenerative Markov process $\widetilde{X}_t$ with the same measurable state space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ such that for all $t \geqslant 0$, the distribution $\mathcal{P}_t$ of X_t is equal to the distribution $\widetilde{\mathcal{P}}_t$ of the process $\widetilde{X}_t$. ▷

Since the marginal distributions of processes X_t and $\widetilde{X}_t$ coincide, the quasi-regenerative process has the same properties as its regenerative "copy".

Some quasi-regenerative processes have been studied in the papers [9, 10].

The question arises: how to find out if the stochastic process under study is quasi-regenerative?

3. Main result

Previously studied quasi-regenerative processes were non-one-dimensional.

Therefore, we assume that the process X_t under study is multidimensional ($X_t = \{X_t^{(1)}, X_t^{(2)}, \ldots, X_t^{(2)}\}$), and its components are dependent processes. We suspect that this process is quasi-regenerative.

Each component $X_t^{(i)}$ of X_t has a state space $\mathcal{X}^{(i)}$ with a σ-algebra $\mathcal{B}(\mathcal{X}^{(i)})$.

Let us fix some one element $x^{(i)}$ from the set $\mathcal{X}^{(i)}$.

Consider the times $t_j(x^{(i)})$ (and $t_j(x^{(i)}) < t_{j+1}(x^{(i)})$) such that $X^{(i)}_{t_j(x^{(i)})} = x^{(i)}$, and for all $\tau \neq t_j(x^{(i)})$, $X^{(i)}_\tau \neq x^{(i)}$.

If the set $\{t_j(x^{(i)})\}$ is infinite, then the sequence $T^{(i)}(x^{(i)}) = \{t_1(x^{(i)}), t_2(x^{(i)}), \ldots\}$ can be regarded as a sequence of Markov moments.

Let us call this sequence $x^{(i)}$-sequence for sub-process $x^{(i)}$.

If this sequence forms a quasi-renewal process, this process is called an embedded quasi-renewal process.

Remark 1. For a regenerative process X_t with state space $\mathcal{X}$, there is such $x \in \mathcal{X}$ such that the corresponding embedded x-sequence forms a classical renewal process. ▷

Theorem 2. If all sub-processes $X^{(i)}_t$ of the process $X_t = \{X^{(1)}_t, X^{(2)}_t, \ldots, X^{(2)}_t\}$ have embedded quasi-renewal processes, then process X_t is a quasi-regenerative process. ▷

Proof. The proof of the Theorem 2 is based on the coupling method, and generalization of Lorden's inequality given above.

This proof is similar to the proof in paper [13], but it is very long and therefore not given here.

We will demonstrate the scheme of the proof using the simplest example. ■

Example 1. Consider two dependent quasi-renewal processes with quasi-renewal periods ξ_i and η_i satisfying **conditions 1, 2, 3** from the Definition 2, and, in addition, the following **assumptions** hold:
a. $\int\limits_0^\infty x^{k-1} \exp\left(-\int\limits_0^x \varphi(s)\,\mathrm{d}\,s\right) \mathrm{d}\,x = M_k < \infty$ for some $k \geqslant 2$; **b.** There exists the constant $T \geqslant 0$ such that $\varphi(s) > 0$ a.s. for all $s > T$. Condition **1** holds: the d.f.'s of ξ_i and η_i are bounded by the functions $\Phi(x) = 1 - \exp\left(-\int\limits_0^x Q(u)\,\mathrm{d}\,u\right)$ and $\widehat{\Phi}(x) = \exp\left(-\int\limits_0^x \varphi(u)\,\mathrm{d}\,u\right)$, and their distribution density can be bounded from below by the function $\psi(s) = \varphi(s)\exp\left(-\int\limits_0^s Q(u)\,\mathrm{d}\,u\right)$. Condition **2** guarantees the presence of a nonzero variance of the r.v.'s ξ_i and η_i. Assumption **a** holds: there exists $\mathbb{E}\,\xi_j^k < \infty$, $k \geqslant 2$. Assumption **b** holds the density of the d.f. of ξ_i positive a.s. for time $> T$, i.e. we may consider delayed switching for the considered quasi-renewal processes. Also, assumption **b** holds the distribution of ξ_i are non-lattice.

We use

Lemma 1 (Basic Coupling Lemma ([14, 15])). Let $f_i(s)$ be the distribution density of r.v. θ_i $(i = 1, 2)$. And let $\int\limits_{-\infty}^{\infty} \min(f_1(s), f_2(s))\,\mathrm{d}\,s = \varkappa > 0$. Then on some probability space there exists two random variables ϑ_i such that $\vartheta_i \stackrel{\mathcal{D}}{=} \theta_i$, and $\mathbf{P}\{\vartheta_1 = \vartheta_2\} \geqslant \varkappa$. ▷

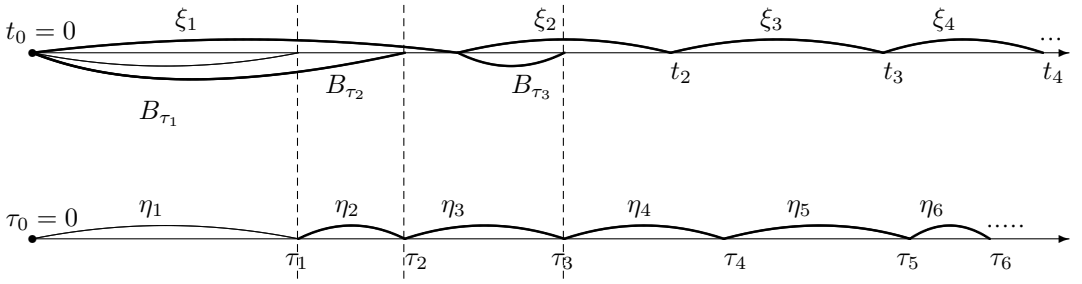


Fig. 1. Visualization of two quasi-renewal processes.

The Fig.1 shows the implementation of two quasi-renewal processes. At the time τ_1, by the generalized Lorden's inequality 1: $\mathbb{E}\,B_{\tau_1} \leqslant \Xi_1$. Thus, by Markov inequality, for some fixed $A > \Xi_1$, $\mathbf{P}\{B_{\tau_1} \leqslant A\} \geqslant 1 - \frac{\Xi_1}{A} \stackrel{\text{def}}{=} p_0$. If $B_{\tau_1} \leqslant A$, then (see Lemma 1) with probability greater then $\varkappa(A) \stackrel{\text{def}}{=} \inf\limits_{a \leqslant A} \int\limits_0^{\infty} \min\{\psi(s+a), \psi(s)\}\,\mathrm{d}\,s = p_1$ we can prolong the periods ξ_1 and η_2 by such a way, that their ends coincide. If they do not coincide, we repeat this procedure at the time τ_2, etc. Thus, with probability greater than $p_0 p_1$ the coupling time T is the time τ_{i+1}, and T is geometrical (conditional) sum of periods $\eta_1, \eta_2, \ldots, \eta_{i+1}$. This sum can be bounded, and an upper bound for $\mathbb{E}\,T^k$ can be calculated. So, we can then estimate the convergence rate of the distribution of this doubled quasi-renewal process. We skip here these calculations. ▷

4. Conclusion

After proving that some process in the queuing theory and in related fields is quasi-regenerative, we can use the technique of obtaining an upper bound on the rate of its convergence to a stationary distribution.

The author is grateful to one of the anonymous reviewers for valuable advice.

REFERENCES

1. Smith, W. L. Regenerative Stochastic Processes // Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences,1955, Vol. 232(1188), P. 6–31. doi:10.1098/rspa.1955.0198

2. Rykov V.V., Kozyrev D.V., Fundamentals of queuing theory. NIC INFRA-M, 2022.
3. Asmussen, S. Regenerative Processes // Applied Probability and Queues. Stochastic Modelling and Applied Probability, 2003, Vol. 51. P. 168—185. doi:10.1007/0-387-21525-5_6. ISBN 978-0-387-00211-8.
4. Vlasiou, M. (2011). *Regenerative Processes* // Cochran, J. J. (Editor-in-Chief). Wiley Encyclopedia of Operations Research and Management Science, Wiley InterScience (Online service), Hoboken, N.J.: J. Wiley, 2010.
5. A. A. Borovkov. Stochastic processes in queueing theory. Springer, 1976.
6. A. A. Borovkov. Asymptotic methods in queuing theory. Wiley, 1984.
7. H. Thorisson. Coupling, Stationarity, and Regeneration, Springer, New York, 2000.
8. Zverkina G. Coupling method for backward renewal process and Lorden's inequality // Distributed Computer and Communication Networks: 20th International Conference, DCCN 2017, Moscow, Russia, September 25–29, 2017, Proceedings, Springer International Publishing, 2017, P. 368–379.
9. Veretennikov A. *On Polynomial Recurrence for Reliability System with a Warm Reserve* // Markov Processes and Related Fields. 2019. Vol. 25. P. 745–761.
10. Zverkina G. A System with Warm Standby // Computer Networks (Proceedings of the 26th International Conference (CN 2019, Kamień Śląski, Poland). Cham: Springer, 2019. P. 387-399.
11. Kalimulina E., Zverkina G. On generalized intensity function and its application to the backward renewal time estimation for renewal processes // Proceedings of the 5th International Conference on Stochastic Methods (ICSM-5, 2020). M.: RUDN, 2020. P. 306-310.
12. Kalimulina E., Zverkina G. On some generalization of Lorden's inequality for renewal processes // arXiv.org. Cornell: Cornell university library 2019, Vol. 1, P. 1–5.
13. Zverkina G. Ergodicity and Polynomial Convergence Rate of Generalized Markov Modulated Poisson Processes // Proceedings of the 23rd International Conference on Distributed Computer and Communication Networks: Control, Computation, Communications (DCCN-2020, Moscow). Cham: Springer, 2021. Vol.1337. P. 367-381.
14. Kato, K. Coupling Lemma and Its Application to The Security Analysis of Quantum Key Distribution // Tamagawa University Quantum ICT Research Institute Bulletin 2014, Vol. 4 No 1, P. 23–30.
15. Veretennikov, A. and Butkovsky, O.A. On asymptotics for Vaserstein coupling of Markov chains // Stochastic Processes and their Applications 2013, Vol. 123(9), P. 3518–3541.

УДК: 004.81

Задача анализа вероятностных характеристик системы интегрированного доступа и транзита*

В.С. Феоктистов[1], Д.И. Николаев[1], Ю.В. Гайдамака[1,2], К.Е. Самуйлов[1,2]

[1]Кафедра прикладной информатики и теории вероятностей, Российский университет дружбы народов, ул. Миклухо-Маклая, д.6, Москва, 117198, Россия
[2]Федеральный исследовательский центр «Информатика и управление» РАН, ул. Вавилова, д. 44-2, Москва, 119333, Россия

1032192939@pfur.ru, {1032201198, gaydamaka-yuv, samuylov-ke}@rudn.ru

Аннотация

В представленном исследовании анализируются вероятностные характеристики интегрированной системы доступа и обратной связи, разработанной консорциумом 3GPP для беспроводных сетей миллиметрового диапазона. Данная технология решает проблемы развертывания мобильных базовых станций, в том числе на беспилотных летательных аппаратах, в густонаселенных городских районах с высокой плотностью застройки и наличием как мобильных, так и стационарных блокираторов радиосигнала. Исследование включает в себя разработку модели и сценария для внедрения IAB, имитационной модели на языке GPSS и аналитической модели фрагмента системы, представленную в виде системы массового обслуживания на основе поллинга. В качестве основных показателей эффективности выбраны средние сквозные задержки и среднее число пакетов в сети и на отдельных маршрутах.

Ключевые слова: mmWave, IAB, GPSS, мобильная сеть, технология интегрированного доступа и транзита, система массового обслуживания, дисциплина поллинга

1. Введение

Технология интегрированного доступа и транзита (IAB, Integrated Access and Backhaul) в сетях 5G представляет инновационное решение, которое объединяет функции доступа и транзита данных в одном сетевом узле, используя один и тот же радиоресурс. Это снижает необходимость устанавливать и поддерживать отдельные устройства для связи между базовыми станциями (БС). В сетях 5G

*Исследование профинансировано Российским научным фондом (РНФ), проект № 22-29-00694, https://rscf.ru/en/project/22-29-00694/.

технология IAB использует беспроводную обратную связь, чтобы уменьшить зависимость от волоконно-оптических соединений и упростить установки. Кроме того, IAB хорошо зарекомендовала себя при развертывании ретрансляционных станций на беспилотных летательных аппаратах, когда возможности свободного полета используются для поиска оптимальной связи пользователя и базовой станции, местоположения беспилотника и распределения мощности. Это позволяет гибко развертывать сети 5G, включая подвижные БС, которые адаптируются к изменяющимся условиям в городской среде.

Для достижения своих целей технология IAB повторно использует существующие функции и интерфейсы, определенные для доступа в сетях 5G. В частности, архитектура IAB использует функции MT (Mobile-Termination), gNB-DU, gNB-CU, UPF (User Plane Function), AMF (Access and Mobility Management Function) и SMF (Session Management Function), а также соответствующие интерфейсы NR Uu (между MT и gNB), F1, NG, X2 и N4 [1]. Для исследования выбрана архитектура 1a, описанная в техническом отчете 3GPP TR 38.874 [1].

В данной работе применен подход, заключающийся в использовании методов теории массового обслуживания, имитационного моделирования и измерений на программном комплексе, разработанном на языке GPSS [2]. В рамках исследования построена модель системы массового обслуживания с дисциплиной поллинга.

Используемая в работе литература раскрывает детали технологий 5G и IAB [1, 3]; разъясняет принципы работы языка GPSS [2]; дает информацию о беспроводных сетях [4], статистику об использовании интернет-трафика [5] и теорию о поллинговых системах [6].

2. Системная модель

Для анализа технологии выбрана модель, состоящая из одного IAB-донора и IAB-узла, а также двух групп пользовательских устройств: UEs-1 и UEs-2, причем IAB-донор имеет проводное соединение с внешней сетью.

Особенностью исследования является использование полудуплексного соединения с разделением режимов на доступ и транспорт. В режиме транспорта, передача пакетов осуществляется между IAB-узлами и IAB-донором, а в режиме доступа – между пользовательскими устройствами и базовыми станциями, к которым они подключены.

Для корректной и бесперебойной работы вводится фиксированная политика управления трафиком внутри сети IAB. Для исследуемой сети выбрана следующая последовательность действий: (1) пакеты перемещаются от IAB-донора к IAB-узлу; (2) пакеты перемещаются от IAB-донора к группе UEs-1 и от IAB-узла к группе UEs-2; (3) пакеты перемещаются от группы UEs-1 к IAB-донору

и от группы UEs-2 к IAB-узлу; (4) пакеты перемещаются от IAB-узла к IAB-донору; (5) пакеты перемещаются от IAB-донора к группе UEs-1 и от IAB-узла к группе UEs-2; (6) пакеты перемещаются от IAB-донора к IAB-узлу; (7) пакеты перемещаются от IAB-донора к группе UEs-1 и от IAB-узла к группе UEs-2.

Для проведения численного эксперимента выбраны следующие параметры модели: время моделирования – 10 сек; продолжительность одного такта – 1 мс; длина служебного пакета – 64 байта; длина пользовательского пакета – распределение длины пользовательских IP пакетов со скачками в областях 64, 1400, 1500 байт; пропускная способность каналов связи между БС – 9460,3 Мбит/с; пропускная способность нисходящих каналов связи для UEs – 6464,5 Мбит/с; пропускная способность восходящих каналов связи для UEs – 3547,6 Мбит/с; вероятность успешной доставки пакетов между БС – 0,95; вероятность успешной доставки пакетов между БС и UEs – 0,9; поступление пакетов – пуассоновский поток пакетов; интенсивность поступления пакетов для UEs (DL, Downlink) – 39.000 пакетов/с; интенсивность генерации пакетов UEs (UL, Uplink) - 1.000 пакетов/с.

3. Поллинговая модель

Будем называть граничными узлы сети IAB, не имеющие дочерних узлов. На них в очередь Q_1 (нисходящий канал) поступают заявки только от родительского узла, а в очередь Q_2 (восходящий канал) — от абонентских устройств (UE). Построим аналитическую модель для граничных узлов.

Принимая во внимание полудуплексный режим передачи данных в сети IAB, для описания процесса обслуживания пакетов данных граничным узлом предлагается модель поллинга с исчерпывающей дисциплиной обслуживания и дополнительными условиями. Процесс работы узла делится на 3 стадии: прогулка, во время которой заявки поступают в обе очереди; обслуживание первой очереди; обслуживание второй очереди. Таким образом, получается следующий цикл: прогулка — очередь Q_1 — очередь Q_2 — прогулка, причем переключение от первой очереди ко второй происходит мгновенно.

Обозначим λ_1, λ_2 — интенсивности поступления заявок в первую и вторую очереди; μ_1, μ_2 — интенсивности обслуживания заявок в первой и второй очередях; s_1^{-1}, s_2^{-1} — интенсивности переключения прибора от второй к первой и от первой ко второй очередям соответственно (средние времена переключения s_1 и s_2); r_1, r_2 — максимальные длины первой и второй очередей (емкость накопителей).

В нашей системе переключение от первой очереди ко второй мгновенно, то есть $s_2 = 0$, тогда переобозначим $s_1 = s, s_1^{-1} = s^{-1}$. Основную особенность

системы можно сформулировать следующим образом:

$$\lambda_i = \begin{cases} \lambda_i, & q = 0 \text{ (прогулка прибора)} \\ 0, & q = 1,2 \text{ (обслуживание очереди } Q_1 \text{ или } Q_2) \end{cases}, i = 1,2 \qquad (1)$$

Функционирование такой системы можно описать с помощью случайного процесса (СП):

$$\mathbf{X}(t) = \{(q(t), n_1(t), n_2(t)), t \geq 0\},$$

где $q(t) \in \{0,1,2\}$ — состояние прибора ($q = 0$ — поступление заявок, $q = 1,2$ — обслуживание первой и второй очередей соответственно), $n_1(t) \in \{0,1,\ldots,r_1\}$ — число заявок в первой очереди, $n_2(t) \in \{0,1,\ldots,r_2\}$ — число заявок во второй очереди в момент t.

Тогда пространство состояний представимо в следующем виде (2):

$$\begin{aligned} \mathbb{X} = \{&(0,n_1,n_2) : n_1 \in \{0,1,\ldots,r_1\}, n_2 \in \{0,1,\ldots,r_2\}, \\ &(1,n_1,n_2) : n_1 \in \{1,\ldots,r_1\}, n_2 \in \{0,1,\ldots,r_2\}, \\ &(2,0,n_2) : n_2 \in \{1,\ldots,r_2\}\} \end{aligned} \qquad (2)$$

В состоянии прогулки может быть любое число заявок в каждом из накопителей. В состоянии обслуживания в первой очереди должна быть хотя бы одна заявка, иначе прибор мгновенно переключится ко второй очереди. В состоянии обслуживания второй очереди в первой очереди нет заявок, но есть хотя бы одна во второй, так как в выбранном нами порядке обслуживания очередей сначала полностью обслуживается первая очередь, а затем вторая.

Пространство состояний также можно представить, как объединение трех непересекающихся подпространств:

$$\mathbb{X} = \mathbb{X}^{Vacation} \bigcup \mathbb{X}^{(1)} \bigcup \mathbb{X}^{(2)},$$

при этом подпространство состояний прогулки

$$\mathbb{X}^{Vacation} = \{(0,n_1,n_2) \in \mathbb{X} : n_1 \in \{0,1,\ldots,r_1\}, n_2 \in \{0,1,\ldots,r_2\}\},$$

подпространство состояний обслуживания первой очереди

$$\mathbb{X}^{(1)} = \{(1,n_1,n_2) \in \mathbb{X} : n_1 \in \{1,\ldots,r_1\}, n_2 \in \{0,1,\ldots,r_2\}\},$$

подпространство состояний обслуживания второй очереди

$$\mathbb{X}^{(2)} = \{(2,0,n_2) \in \mathbb{X} : n_2 \in \{1,\ldots,r_2\}\}.$$

Число состояний можно вычислить по следующей формуле (3):

$$|\mathbb{X}| = (r_1 + 1)(r_2 + 1) + r_1(r_2 + 1) + r_2. \tag{3}$$

Пусть $p(q, n_1, n_2) = \lim\limits_{t\to\infty} P\{\mathbf{X}(t) = (q(t), n_1(t), n_2(t))\}, (q, n_1, n_2) \in \mathbb{X}$ — стационарное распределение вероятностей СП $\mathbf{X}(t)$.

Для нахождения стационарных вероятностей необходимо решить систему уравнений равновесия (СУР)

$$\begin{cases} \mu_1 p(1,1,0) + \mu_2 p(2,0,1) = (\lambda_1 + \lambda_2) p(0,0,0), \\ \lambda_1 p(0,i-1,0) = (u(r_1 - i)\lambda_1 + \lambda_2 + s^{-1}) p(0,i,0), i = 1,...,r_1, \\ \lambda_2 p(0,0,i-1) = (\lambda_1 + u(r_2 - i)\lambda_2 + s^{-1}) p(0,0,i), i = 1,...,r_2, \\ \lambda_1 p(0,i-1,j) + \lambda_2 p(0,i,j-1) = (u(r_1 - i)\lambda_1 + u(r_2 - j)\lambda_2 + s^{-1}) p(0,i,j), \\ i = 1,...,r_1, j = 1,...,r_2, \\ s^{-1} p(0,i,j) + u(r_1 - i)\mu_1 p(1,i+1,j) = \mu_1 p(1,i,j), i = 1,...,r_1, j = 0,...,r_2, \\ s^{-1} p(0,0,j) + \mu_1 p(1,1,j) + u(r_2 - j)\mu_2 p(2,0,j+1) = \mu_2 p(2,0,j), j = 1,...,r_2, \end{cases} \tag{4}$$

Зная стационарное распределение вероятностей, можно получить интересующие нас метрики: вероятность нахождения в состоянии прогулки, вероятность нахождения в состоянии обслуживания первой очереди, вероятность нахождения в состоянии обслуживания второй очереди, вероятность блокировки, среднее число заявок в первой очереди, среднее число заявок во второй очереди.

4. Численный эксперимент

Поскольку модель IAB с полудуплексным режимом без упрощающих предположений рассматриваемой математической модели с дисциплиной поллинга представляет собой сложную систему, для нее целесообразно разработать имитационную модель. В качестве языка имитационного моделирования выбран GPSS. Обработка и визуализация данных осуществлялась с использованием языка программирования Python версии 3.7 и библиотек matplotlib, pandas, numpy, scipy и PIL.

Под сквозной задержкой в сети IAB далее понимается длина интервала времени от момента поступления последнего бита пакета на IAB-донор от внешней сети или момента поступления сгенерированного пользовательским устройством пакета в очередь для последующей его передачи на узел сети до момента поступления последнего бита успешно переданного пакета на пользовательское устройство или IAB-донор. Она может быть оценена с использованием подхода

декомпозиции и агрегации как сумма задержек в элементах сети и на соединяющих их звеньях. Задержка пакета в каждом элементе сети IAB оценивается как время пребывания заявки в системе массового обслуживания, а время передачи по звену моделируется с учётом быстродействия канала.

Результаты численного эксперимента представлены на четырех графиках. На рисунке 1 наблюдаем, что средняя сквозная задержка по сети, полученная с помощью имитационной модели, практически совпадает со средней сквозной задержкой, полученной по формуле Литтла через среднее число пакетов в сети. Относительная погрешность составляет менее 1,5%. При этом, как видно из рисунка 2, среднее число пакетов по маршрутам и по всей сети почти линейно зависит от продолжительности такта.

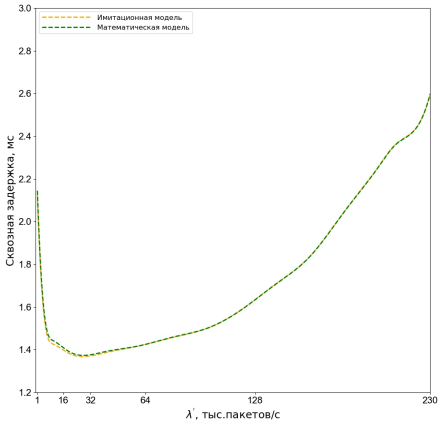


Рис. 1. Математическая и модельная средняя сквозная задержка ($\lambda_{\text{UEs-1}} = \lambda_{\text{UEs-2}} = \lambda'$).

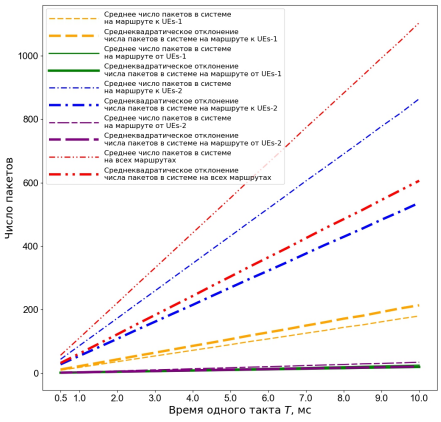


Рис. 2. Среднее число пакетов на маршрутах сети.

При увеличении интенсивности потока пакетов, поступающих на IAB-донор из внешней сети, среднее число пакетов по всей IAB-сети и по маршрутам передачи пакетов к группам пользовательских устройств растет экспоненциально, в то время как среднеквадратичное отклонение имеет более сложный вид (рис. 3). Отметим, что система достигает состояния перегрузки для $\lambda' \approx 240000$ пакетов/с, что соответствует примерно 600 UEs на каждую базовую станцию и 1200 UEs на всю исследуемую сеть. Более того, до значения интенсивности $\lambda' \approx 240000$ пакетов/с средние сквозные задержки в системе не превышают 4

мс, определенные стандартами 3GPP для предоставления услуг eMBB [7]. Этот результат следует учитывать при планировании сетей IAB с ретрансляционными станциями.

Продолжая анализ, на основе рисунка 4 приходим к выводу, что только по направлениям передачи пакетов от IAB-донора к группе пользовательских устройств UEs-1 и от IAB-узла к группе UEs-2 очереди достигают состояния перегрузки при указанной интенсивности $\lambda^{'}$.

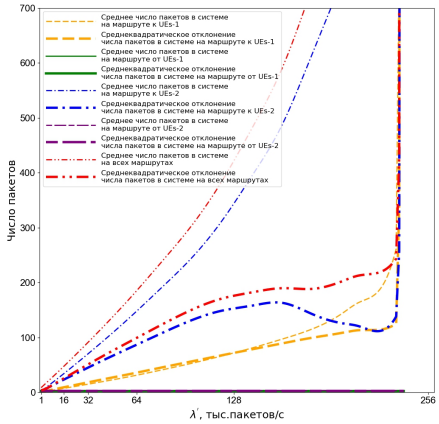


Рис. 3. Среднее число пакетов на маршрутах сети ($\lambda_{\text{UEs-1}} = \lambda_{\text{UEs-2}} = \lambda^{'}$).

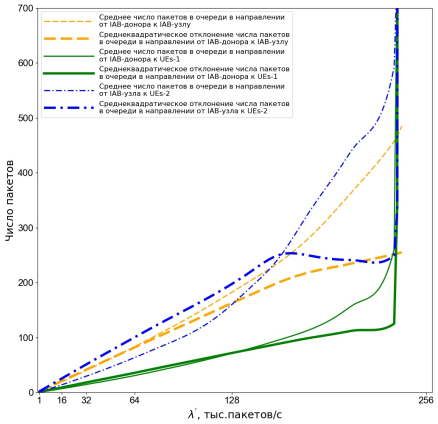


Рис. 4. Среднее число пакетов в трех очередях ($\lambda_{\text{UEs-1}} = \lambda_{\text{UEs-2}} = \lambda^{'}$).

5. Заключение

В работе построена системная модель и сценарии использования системы интегрированного доступа и транзита (IAB) в сетях 5G, создана программа имитационного моделирования на языке GPSS и программа для обработки данных и визуализации результатов на языке Python версии 3.7, подготовлены исходные данные для численного эксперимента.

Проведенный численный анализ для предложенной сети с фиксированной политикой управления показывает, что технология IAB может обслуживать до 1200 одновременно подключенных пользовательских устройств при умеренных нагрузках на IAB-сеть. Сквозные задержки в IAB-сети остаются относительно небольшими до достижения состояния перегрузки, что обеспечивает требуемое качество обслуживания пользователей. Небольшое увеличение сквозной задерж-

ки по сравнению с традиционными сетями 5G незаметно для пользователей и компенсируется увеличением покрытия сети при меньших затратах на развертывание оборудования.

В дальнейшем планируется провести дальнейшие исследования, включающие моделирование сети с несколькими узлами IAB и адаптированное управление трафиком.

ЛИТЕРАТУРА

1. 3GPP TR 38.874 v16.0.0. NR; Study on Integrated Access and Backhaul. 2018-12. // 3GPP Portal : [Электронный ресурс]. — URL: https://portal.3gpp.org (дата обращения: 17.02.2023).
2. С. И. Матюшенко, Д. А. Пяткина, Р. В. Разумчик. Моделирование систем массового обслуживания в среде GPSS World : учебное пособие. — 2-е изд., перераб. и доп. изд. — М.: Российский университет дружбы народов, 2022. — 94 с.
3. 3GPP TR 38.801 v14.0.0 Study on new radio access technology: Radio access architecture and interfaces. 2017-03. // 3GPP Portal : [Электронный ресурс]. — URL: https://portal.3gpp.org (дата обращения: 17.02.2023)
4. Beard Cory and Stallings William. Wireless communication networks and systems. — Hoboken: Pearson Higher Education, Inc., 2015. — 642 с.
5. Cisco Annual Internet Report (2018–2023) White Paper // CISCO : [Электронный ресурс]. — URL: https://www.cisco.com (дата обращения: 17.02.2023).
6. Vishnevskiy V. M., Semenova O. V. Polling systems: Theory and application in the broadband wireless networks. — M.: Technosphere, 2007. — 312 p.
7. 3GPP TR 38.913 v17.0.0 Study on scenarios and requirements for next generation access technologies. 2022-04. // 3GPP Portal : [Электронный ресурс]. — URL: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2996 (дата обращения: 20.07.2023)

UDC: 519.21

Maximum Likelihood Estimation of the Dead Time Distribution Parameter in Recurrent Semi-Synchronous Doubly Stochastic Events Flow

Anna Vetkina[1] and Ludmila Nezhel'skaya[1]

[1]National Research Tomsk State University, 36 Lenina Ave., Tomsk 634050, Russia

anyavetkina@gmail.com, ludne@mail.ru

Abstract

In this study, recurrent semi-synchronous events flow with two states is described. Operation of the flow is considered with random non-prolonging dead time that has uniform distribution. Maximum likelihood estimation of a distribution parameter of the dead time duration is provided.

Keywords: semi-synchronous events flow, random dead time, maximum likelihood, parameter estimation

1. Introduction

Currently, the information message flows in different information and computational networks are best described by doubly stochastic events flows [1, 2]. These types of flows are characterized by both random occurrence times of events and their intensity. In general, doubly stochastic events flows are correlated flows [3].

Events flows that involve double randomness (stochasticity) can be divided into two classes: the first class consists of flows whose process (intensity) is a continuous random process; the second class consists of flows whose process (intensity) is a piecewise-constant random process with a finite (arbitrary) number of states. The latter are called MC-flows (Markov chain) or MAP-flows (Markovian Arrival Process) and are most commonly used in solving applied problems.

In practical information systems, it is often impossible to observe all events in a flow due to the dead time of recording devices [4]. This dead time occurs when recorded events cause other events to be lost or inaccessible for observation. The dead time of recording devices is influenced by various factors and can be categorized as non-prolonging or prolonging. The duration of unobservability may be deterministic or random with a specific distribution law. In case of random dead time, its maximum duration is limited by an upper bound. Therefore, it is reasonable to consider the

distribution of dead time as uniform over some interval when treating it as a random variable.

The object of this study is a recurrent semi-synchronous doubly stochastic events flow with an intensity that is a piecewise-constant random process with two states. The flow operates with non-prolonging dead time that duration is random and uniformly distributed.

In order to identify lost events caused by dead time factor, it is necessary to estimate the value of the duration of the dead time. The maximum likelihood (ML) method is a widely used approach for determining unknown parameters in various fields of study. In this work, 1) an algorithm to find the ML estimations for the parameter of the dead time distribution is applied to the observed semi-synchronous events flows; 2) for the investigation of obtained estimations, a numerical study of constructed likelihood function in the domain of the unknown variable is provided.

2. Mathematical model

This research focuses on a semi-synchronous doubly stochastic events flow that accompanying process (intensity) is a piecewise-constant random process $\lambda(t)$ with two states S_1 and S_2. It is assumed that the i-th state of the process S_i occurs when $\lambda(t) = \lambda_i$, and during that time a Poisson events flow with intensity λ_i occurs, $i = 1, 2$. The transition from the state S_1 of the $\lambda(t)$ process to the state S_2 is only possible at the occurrence of an event (synchronicity property of the flow), and this transition occurs with probability p (the $\lambda(t)$ process remains in first state with probability $1 - p$). The transition from the state S_2 to the state S_1 can occur at any time not related to the occurrence of an event (asynchronicity property of the flow). The duration of the $\lambda(t)$ process being in the second state is a random variable distributed according to an exponential law $F(t) = 1 - e^{-\alpha_2 t}$, $t \geqslant 0$, where α_2 is the intensity of the transition from the state S_2 to the state S_1. Hence, we have a semi-synchronous doubly stochastic events flow. Under these assumptions, $\lambda(t)$ is a hidden Markov process ($\lambda(t)$ is a fundamentally unobservable process; only the moments of events occurrence in the flow are observable).

After each recorded event at time t_k, a period of dead time of random duration occurs, which is generated by this event, so that other events in the original flow that occur during this dead time period are not observable and do not cause its prolongation. It is assumed that the random duration of dead time is uniformly distributed with probability density $p(T) = 1/T^*$, where T is the value of the duration of dead time, $0 \leqslant T \leqslant T^*$.

We consider the semi-synchronous events flow, when $p = 1$, i.e. a flow that instantly transitions to the second state at each occurrence of an event in the first state. When this restriction is satisfied, the original semi-synchronous events flow,

operating under conditions of deterministic dead time, becomes a recurrent flow:

$$p(\tau_1, \tau_2 \mid T) = p(\tau_1 \mid T)p(\tau_2 \mid T),\ \tau_1 \geqslant T, \tau_2 \geqslant T$$

where $p(\tau_i \mid T)$, $i = 1, 2$, is the probability density of the duration values between adjacent events in the observed flow, $p(\tau_1, \tau_2 \mid T)$ is the joint probability density [5].

A possible case is shown in Fig. 1, where S_1 and S_2 are the states of the random process $\lambda(t)$; the time axis $(0, t)$ is the axis of moments of occurrence of observable events at times $t_1, t_2, \ldots$; the time axis $(0, t^{(i)})$, $i = 1, 2$, is the axis of occurrence of events in the i-th state of the $\lambda(t)$ process, on which the values of the duration of dead times generated by observable events in the flow are also indicated; white circles denote observable events, black ones denote unobservable events, hatching denotes periods of dead time; the trajectory of the $\lambda(t)$ process is attached to the time axis $(0, t^{(1)})$.

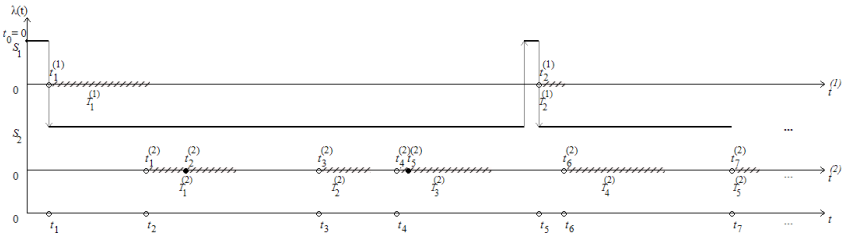


Fig. 1. Formation of the observed recurrent events flow

Note that the stationary mode of operation of the flow is considered, i.e. we assume that the flow operates infinitely, and its probabilistic characteristics do not depend on time.

For the semi-synchronous events flow with $p = 1$ we consider two cases of the ratio of parameters of this flow: a) general case, when $\lambda_1 - \lambda_2 - \alpha_2 \neq 0$; b) special case, when $\lambda_1 - \lambda_2 - \alpha_2 = 0$.

The objectives of this study are:

1) To estimate the parameter of the uniform distribution of non-prolonging dead time duration T^* based on a sample of moments of observed event occurrence $t_1, t_2, \ldots, t_n$ on the time interval $(0, T_m)$, where T_m is the observation time of the flow ($t_n < T_m$), using the maximum likelihood method.
2) To investigate the estimation $\hat{T^*}$ for the general and special cases of the considered flow. For that, to conduct statistical experiments to determine the properties of the obtained estimates.

3. Approximate ML estimate of the parameter T^*

The maximum likelihood method is used to estimate the unknown parameter T^* of the uniform distribution of non-extendable dead time duration in the observed flow. This method consists in maximizing a likelihood function to estimate the parameter so that, under the assumed statistical model, the observed data is most probable.

Let we have n values of time intervals between observed events: $\tau_1, \tau_2, \ldots, \tau_n$ that are measured on the time interval $(t_0, t]$. Then we have the form of the likelihood function:

$$L(T^* \mid \tau_1, \tau_2 \ldots, \tau_n) = \prod_{k=1}^{n} p(\tau_k \mid T^*),\ T^* > 0, \tag{1}$$

where $p(\tau_k \mid T^*)$ is a probability density of the duration of the interval between neighboring events in the observed flow with $\tau = \tau_k$ (τ_k is a measurement); T^* is a variable value ($T^* > 0$). a) For the general case, we have:

$$p_1(\tau \mid T^*) = \frac{1}{T^*}\left\{1 - e^{-\lambda_1\tau} - e^{-(\lambda_1+\alpha_2)\tau} + e^{-(\lambda_2+\alpha_2)\tau}\right\},\ 0 \le \tau < T^*,$$
$$p_2(\tau \mid T^*) = \frac{1}{T^*}\left\{e^{-\lambda_1\tau}\left[-1 + C_1\, e^{\lambda_1 T^*} + C_2\, e^{-\alpha_2 T^*}\right] + \right.$$
$$\left. + e^{-(\lambda_2+\alpha_2)\tau}\left[-1 + C_1 e^{-(\lambda_1-\lambda_2)T^*} + C_2 e^{(\lambda_2+\alpha_2)T^*}\right]\right\},\ \tau \ge T^*,$$

where $C_1 = \frac{-\alpha_2(\lambda_2+\alpha_2)}{(\lambda_1+\alpha_2)(\lambda_1-\lambda_2-\alpha_2)}$, $C_2 = \frac{\lambda_1(\lambda_1-\lambda_2)}{(\lambda_1+\alpha_2)(\lambda_1-\lambda_2-\alpha_2)}$, $C_1 + C_2 = 1$, $\lambda_1 - \lambda_2 - \alpha_2 \neq 0$.

b) For the special case, we have:

$$p_1(\tau \mid T^*) = \frac{1}{T^*}\left\{1 - 2e^{-\lambda_1\tau} + e^{-(\lambda_1+\alpha_2)\tau}\right\},\ 0 \le \tau < T^*,$$
$$p_2(\tau \mid T^*) = \frac{1}{T^*}e^{-\lambda_1\tau}\left\{-2 + \left[1 + \frac{\lambda_1\alpha_2}{\lambda_1+\alpha_2}(\tau - T^*)\right]e^{\lambda_1 T^*} + \right.$$
$$\left. + \left[1 - \frac{\lambda_1\alpha_2}{\lambda_1+\alpha_2}(\tau - T^*)\right]e^{-\alpha_2 T^*}\right\},\ \tau \ge T^*,$$

where $\lambda_1 - \lambda_2 - \alpha_2 = 0$.

Since the values τ_k, $k = \overline{1,\ n}$, are independent, we can order them: $0 < \tau^{(1)} < \ldots < \tau^{(n)} < \infty$. Then we can rewrite the expression (1) in the formula given below

$$L(T^*|\tau^{(1)}, \tau^{(2)}, \ldots, \tau^{(n)}) = \prod_{k=1}^{i} p_1(\tau^{(k)}|T^*) \prod_{k=i+1}^{n} p_2(\tau^{(k)}|T^*),$$
$$\tau^{(i)} \le T^* < \tau^{(i+1)},\ i = \overline{1,\ n}, \tag{2}$$

assuming $\tau^{(0)} = 0$, $\tau^{(n+1)} = \infty$.

Previously in the paper [6] the following theorem for the recurrent events flow with $p = 1$ was proven.

Theorem 1. The function $p\left(\tau^{(k)}|T^*\right)$ of the variable T^* ($T^* > 0$) in both general and special cases reaches its global maximum at a point $T^* = \tau^{(k)}$.

The result of the theorem indicates that the likelihood function (2) is an increasing function at the interval $\left[0,\, \tau^{(1)}\right]$ and reaches its local maximum at a point $T^* = \tau^{(1)}$, and it is an decreasing function at the interval $\left[\tau^{(n)},\, \infty\right)$ and reaches its local maximum at a point $T^* = \tau^{(n)}$. Therefore, to find the global maximum of the likelihood function (2), it is necessary to investigate the interval $\left[\tau^{(1)},\, \tau^{(n)}\right]$ of change of the variable T^* ($\tau^{(1)} \le T^* \le \tau^{(n)}$).

Since the function $p\left(\tau^{(k)}|T^*\right)$ ($T^* > 0$) at the point $T^* = \tau^{(k)}$, $k = \overline{1,\, n}$, reaches its global maximum, the point $T^* = \tau^{(k)}$, $k = \overline{1,\, n}$, is considered a point suspicious of the local maximum of the likelihood function (2), and the algorithm for finding the value of the approximate MP estimate of the parameter T^* is as follows: 1) the values of the likelihood function (2) are calculated at the points $T^* = \tau^{(k)}$, $k = \overline{1,\, n}$; 2) the maximum value of the function (2) is found on the set of these points; 3) the value $\hat{T}^*$ that provide the maximum value of the function (2) at the previous step of the algorithm is selected as the value of the approximate ML estimate of the parameter T^*.

It is not analytically possible to find out explicitly the behavior of the likelihood function (2) over the entire segment $\left[\tau^{(1)},\, \tau^{(n)}\right]$ of variable T^* change. Instead, we can numerically investigate the behavior of the likelihood function at some intermediate points of the segment $\left[\tau^{(1)},\, \tau^{(n)}\right]$ by dividing it evenly into several parts with a given sampling step ΔT^*.

The present study focuses on examining and improving the approximations derived from the maximum likelihood approach explained earlier by calculating all the values of the likelihood function on the segment $\left[\tau^{(1)},\, \tau^{(n)}\right]$ of variable T^* change with a given sampling step ΔT^* and comparing its maximum value with the maximum found from the values of the likelihood function at the points $T^* = \tau^{(k)}$, $k = \overline{1,\, n}$. Therefore, we find out the behavior of the likelihood function (2) over the entire interval of variable T^* change (with a given accuracy of calculations), and not only at the points where the global maximum of the function $p\left(\tau^{(k)}|T^*\right)$ is reached.

For that purpose, numerous experiments were provided on a simulation model of the observed flow in both general and special cases. Approximate and refined ML estimations of the parameter T^* were obtained numerically according to the two algorithms presented above. Results of the experiments indicate the high accuracy of the ML estimates $\hat{T}^*$ calculated by methodology of estimating the unknown

parameter of the model by the maximum likelihood method, since they received insignificant deviations from the approximate estimates and refined ones.

4. Conclusion

In this paper, we consider a recurrent semi-synchronous doubly stochastic events flow with $p = 1$ with uniformly distributed random dead time in the general case, when $\lambda_1 - \lambda_2 - \alpha_2 \neq 0$ and special case, when $\lambda_1 - \lambda_2 - \alpha_2 = 0$. An explicit form of the likelihood function is given for estimating the parameter T^* of a uniform distribution of the duration of dead time. The procedure for constructing the MP estimate $\hat{T^*}$ is described.

REFERENCES

1. Cox D.R. The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables. Proceedings of the Cambridge Philosophical Society. 1955. V. 51(3). P. 433–441.
2. Kingman Y.F.C. On doubly stochastic Poisson process. Proceedings of the Cambridge Philosophical Society. 1964. V. 60(4). P. 923–930.
3. Vishnevsky V.M., Dudin A.N., Klimenok, V.N. Stokhasticheskie sistemy s korrelirovannymi potokami. Teoriya i primenenie v telekommunikatsionnykh setyakh [Stochastic Systems with Correlated Flows. Theory and Application in Telecommunication Networks]. 2018. Moscow: Tekhnosfera (In Russian).
4. Apanasovich V.V., Kolyada A.A., Chernyavsky A.F. Statisticheskiy analiz sluchaynykh potokov v fizicheskom eksperimente [The statistical analysis of series of random events in physical experiment]. Minsk: Universitetskoe. 1988. (In Russian).
5. Nezhelskaya L.A. Otsenka sostoyanij i parametrov dvazhdy stohasticheskih potokov sobytij. Dissertation for the degree of Doctor of Physical and Mathematical Sciences. Tomsk. 2016. P. 341. (In Russian).
6. Gortsev A.M., Vetkina A.V. An application of the maximum likelihood estimation for the parameter of uniform distribution of the duration of unextendable dead time in recurrent alternating semi-synchronous flow. Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitelnaja tehnika i informatika // Tomsk State University Journal of Control and Computer Science. 2023. V. 62. P. 36–49. (In Russian).

Научное электронное издание

Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь (DCCN-2023)

МАТЕРИАЛЫ XXVI МЕЖДУНАРОДНОЙ НАУЧНОЙ КОНФЕРЕНЦИИ

(25–29 СЕНТЯБРЯ 2023 г., МОСКВА)

Под общей редакцией д.т.н. В.М. Вишневского, д.т.н. К.Е. Самуйлова

Составитель: к.ф.-м.н. **Козырев** Дмитрий Владимирович

Локальное электронное издание
Номер госрегистрации в НТЦ "Информрегистр" 0322303901

Минимальные системные требования:
Pentium 4; 1,3 ГГц и выше; Windows 7/8 и выше; Acrobat Reader 4.0
Дата подписания к использованию 10.11.2023
1 электронно-оптический диск (CD-R), 26,1 Мб, Тираж 100 экз.

Федеральное государственное бюджетное учреждение науки
Институт проблем управления им. В.А. Трапезникова
Российской академии наук
117997, Россия, Москва,
ул. Профсоюзная, д. 65
www.ipu.ru