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-2020)

МАТЕРИАЛЫ XXIII МЕЖДУНАРОДНОЙ НАУЧНОЙ КОНФЕРЕНЦИИ
(14–18 СЕНТЯБРЯ 2020 г., МОСКВА, РОССИЯ)

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

НАУЧНОЕ ЭЛЕКТРОННОЕ ИЗДАНИЕ

Москва
ИПУ РАН
2020

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-2020)

PROCEEDINGS OF THE XXIII INTERNATIONAL SCIENTIFIC CONFERENCE
(September 14–18, 2020, Moscow, Russia)

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

MOSCOW
ISC RAS
2020

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-2020) = Distributed computer and communication networks: control, computation, communications (DCCN-2020) [Электронный ресурс] **:** материалы XXIII Междунар. научн. конфер, 14–18 сент. 2020 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: 44,9 Мб). – М.: ИПУ РАН, 2020. – 1 электрон. опт. диск (CD-R). – Систем. требования: Pentium 4; 1,3 ГГц и выше; Internet Explorer; Acrobat Reader 4.0 или выше. – Загл. с экрана. – ISBN 978-5-91450-248-2. – № госрегистрации 0322002892.

В научном электронном издании представлены материалы XXIII 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
- Большие данные
- 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-2020, 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.

Текст научного электронного издания
воспроизводится в том виде, в котором представлен авторами

Утверждено к изданию Программным комитетом конференции

ISBN 978-5-91450-248-2 **© ИПУ РАН, 2020**

Содержание / Contents

UDC: 519.2

Statistical Analysis of the End-to-End Delay of Packet Transfers in a Peer-to-Peer Network

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

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

[2]Fakultät WIAI, Otto-Friedrich-Universität, An der Weberei 5, D-96047 Bamberg, Germany

markovic@ipu.rssi.ru, nat.markovich@gmail.com, udo.krieger@ieee.org

Abstract

The paper is devoted to the statistical analysis of the end-to-end (E2E) delay of packet transfers between source and destination nodes in a peer-to-peer (P2P) overlay network. The E2E delay is determined by the sum of a random number of per-hop (p-h) delays along the links of a considered overlay path. We propose to use the maximum of the p-h delays instead of the sum. Based on recent analytic results derived from extreme-value theory we show that such sums and maxima corresponding to different paths may have the same tail and extremal indexes. These indexes determine the heaviness of the distribution tail and the dependence of extremes, respectively. Using the extremal index we identify limit distributions of the maxima of the E2E delays and the maxima of p-h delays at a path among all source-destination paths. The distributions are used to identify quality-of-service (QoS) metrics of a P2P model like the packet missing probability and the corresponding playback delay as well as the equivalent capacity.

Keywords: End-to-end delay, per-hop delay, tail index, extremal index, packet missing probability, playback delay, equivalent capacity, quality-of-service

1. Introduction

The identification of the distribution of E2E delays arising between source and destination nodes in a P2P-overlay network constitutes an important problem of telecommunication due to live TV and video-on-demand applications. The delay of information transmission through the network and, hence, the playback delay that is the lag between the generation of the packet and its playout deadline have a big impact on the quality of service and experience. As the E2E delay can be represented as a sum of a random number of p-h delays, its distribution depends

on the distributions of the random length of the overlay path between source and destination and the p-h delays. The latter are determined by the structure of the overlay network.

In [2], [5] the relation between the distribution of the packet delay and the packet missing probability has been considered. The distribution of the E2E delay of the ith path $D_i(D) = \sum_{j=1}^{L_i(D)} X_{i,j}$ is required. $\{X_{i,j} : 1 \le j \le L_i(D)\}$ are the p-h delays of this overlay path i from the source S to the destination node D with a random length $L_i(D)$. The paths between S and D are schematically shown in Fig.1. Its randomness is caused by the random number of nodes and links of the paths due to the dynamics of the network over time. The exceedance of the packet delay over the playback deadline b is considered as one of the main reasons to miss a packet. Then the part of the missing probability is the following $P_m(b) = P\{D_i(D) > b\}$. The exceedance of the rate over the equivalent capacity of the channel is considered in [5] as the second reason to loose packets. One of the objectives of the paper is to identify the missing probability under more general assumptions than in [2], [5] in view of the last results obtained in [6]. Namely, as in [11] it was assumed that $\{X_{i,j}\}$ are independent and identically distributed (i.i.d.) random variables (r.v.s) with light or heavy tails depending on the overlay structure, and the number of nodes N in the network and $L_i(D)$ are stationary distributed. The mutual dependence or independence of $X_{i,j}$ and $L_i(D)$ and the assumption which tail of these r.v.s is heavier are principal to identify the distribution of the sum, see for instance [3].

We assume that $\{X_{i,j}\}$ are not necessarily i.i.d.. This assumption is realistic since paths may be overlapping as in Fig. 1. We assume that $\{X_{i,j}\}$ are stationary distributed at links located at the same distance with regard to the number of links from S. The random path length $L_i(D)$ is assumed to be stationary distributed, but its mutual independence on the p-h delay is omitted. Another objective is to find the relation between the local dependence (cluster) structure and the distributions of the E2E delay and a maximal p-h delay at a path. This allows us to generalize the probability $P_m(b)$ uniformly to all paths of lengths $\{L_i\}$ and to obtain $P\{\max_i D_i(D) > b\}$. Our achievements are based on the results of extreme-value theory obtained in [6].

In [6] it is derived that the tail index (TI) and extremal index (EI) of the asymptotic distributions of sums and maxima of random length sequences may be the same. One may conclude that the sum and maxima of p-h delays at the paths have the same heaviness of the distribution tail and the same dependence structure. This implies that the distribution of the E2E delay may be approximated by the distribution of the maximum of the p-h delays at the source-destination path. As the E2E delays can be made available in practice easier than the p-h delays, this allows us to approximate the distribution of the p-h delays at the most heavy-tailed link. Moreover, one can use the maximum distribution to determine the packet missing probability.

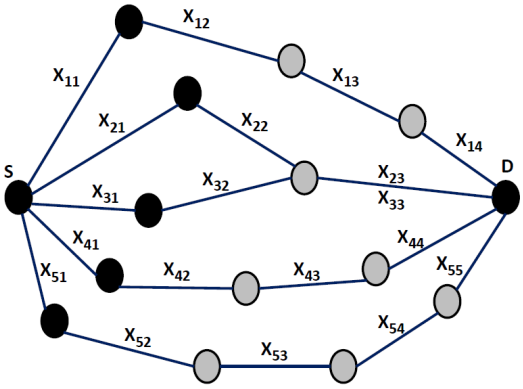


Fig. 1. Paths of random length between source node S and destination node D with the per-hop time delays $\{X_{i,j}\}$ of packet transmissions on the ith path between these nodes; the nodes in black between S and D indicate those ones with a distance of one link from S.

The paper is organized as follows. Section 2 contains a survey of related results. Our main results are presented in Section 3 with parts concerning the stochastic model and its nonparametric estimation. The exposition is finalized by some conclusions.

2. Related Work

Let the links be enumerated from the source node S. We assume that the p-h delay $X_{i,j}$, $i, j \geq 1$, at link j of the path i is regularly varying distributed. Then

$$P\{X_{i,j} > x\} \quad = \quad \ell_j(x)x^{-k_j} \tag{1}$$

holds with the TI k_j and a slowly varying function $\ell_j(x)$, i.e. $\lim_{x\to\infty} \ell_j(tx)/\ell_j(x) = 1$ for any $t > 0$. Positive constants and logarithms give examples of $\ell_j(x)$. The links with the same number j are assumed to be stationary distributed and their distribution may be different from the distribution of the links with another number.
The EI θ is called the local dependence measure having in mind that consecutive exceedances over a high threshold u occur usually in clusters. Such clusters of exceedances are caused by the dependence in stochastic sequences. The clustering can be intensified by heavy distribution tails.

Definition 1. [4] The stationary sequence of r.v.s $\{X_n\}_{n\geq 1}$ with cumulative distribution function (cdf) $F(x)$ and $M_n = \max\{X_1, ..., X_n\}$ is said to have the EI $\theta \in [0, 1]$ if for each $0 < \tau < \infty$ there is a sequence of real numbers $u_n = u_n(\tau)$ such that it holds

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

The EI has the following relation to the distribution of the maximum:

$$P\{M_n \le u_n\} \quad = \quad P^{n\theta}\{X_1 \le u_n\} + o(1) = F^{n\theta}(u_n) + o(1), \quad n \to \infty.$$

It holds $\theta = 1$ if the r.v.s $\{X_n\}$ are i.i.d.. The converse is incorrect. An EI that is close to zero implies a kind of a strong dependence.
In order to use the results in [6], we assume that the p-h delays $\{X_{i,j} : i \ge 1\}$ are stationary distributed as (1) and have their TI $k_j > 0$ and EI $\theta_j \in [0, 1]$, and that among all sets of the links there exists a unique set with the minimal TI. Without loss of generality, this can be the set of first links $\{X_{i,1} : i \ge 1\}$ from the source node S with TI equal to k_1. Such set has the heaviest distribution tail. Other sets have TIs larger than k_1 and, hence, they are not so heavy-tailed distributed. Despite some of such $\{k_j\}_{j\ge 2}$ may be equal, the corresponding distributions of the link sets may be not the same if the slowly varying functions $\ell_j(x)$, $j \ge 2$ in (1) are different.
In [6] the EI and the TI of sums and maxima of random sequences of random lengths $\{L_n\}$ were considered. It is proved that the sequences of sums and maxima

$$\begin{aligned} X_n(z, L_n) &= z_1 X_{n,1} + z_2 X_{n,2} + ... + z_{L_n} X_{n,L_n}, \\ X_n^*(z, L_n) &= \max(z_1 X_{n,1}, z_2 X_{n,2}, ..., z_{L_n} X_{n,L_n}) \end{aligned}$$

with positive constants $z_1, ..., z_{L_n}$ have a distribution (1) with the same k_1 and θ_1. Since the E2E delays constitute random sums of a random number of terms, the mentioned result relates to our problem. L_n is geometrically distributed irrespective of the distributions of the packet transmission rates and E2E delays and depending only on the levels of their quantiles (Theorem 1, [5]). This geometric model fits the result in [6]. In case that some paths include a node with light-tailed distributed p-h delay and (or) the distribution of the p-h delays at some link away from the source contains a mixture of light- and heavy-tailed distributions, the basic statistical result developed in [6] is still valid.

3. Statistical Analysis of the End-to-End Delay

Let $n \ge 1$ be the number of possible paths constructed by nodes of the P2P overlay network. Since the P2P network may be dynamically changed in time, the number of nodes available for the packet transmission is changing and n is random. We can neglect its randomness considering the approach as conditional one, since n is proportional to the number of nodes in the network and the latter can be large. The theoretical result in [6] assumes that n is deterministic and tends to infinity.
Let us consider the double-indexed array of the p-h delays $\mathbb{X} = (X_{i,j} : i, j \ge 1)$. The "row index" i corresponds to the p-h delays belonging to the same path i between the source S and destination D, and the "column index" j corresponds to the p-h

delays arising at the jth link enumerated from the source node. All p-h delays relate to the same source-destination pair (S, D). We consider the corresponding matrix

$$\mathbb{X} = \begin{pmatrix} X_{1,1} & X_{1,2} & 0 & ...0 & X_{1,L_1} \\ X_{2,1} & X_{2,2} & X_{2,3} & ...0 & X_{2,L_2} \\ ... & ... & ... & ... & ... \\ X_{n,1} & X_{n,2} & X_{n,3} & ...X_{n,L_n-1} & X_{n,L_n} \end{pmatrix}$$
$$\begin{pmatrix} k_1, & k_2, & k_3, & ..., & k_{L_n-1}, & k_{L_n} \\ \theta_1, & \theta_2, & \theta_3, & ..., & \theta_{L_n-1}, & \theta_{L_n} \end{pmatrix}$$

where the first and last columns corresponding to the one-hop links to the source and destination nodes are full and internal columns are completed by zeros up to the maximal dimension, let's say L_n. We assume the most general case: the columns can be dependent, and each column may consist of dependent p-h delays, and the distribution of each column is stationary with the positive TI k_j and its local dependence structure is described by the EI θ_j. For any location of zeros in the matrix, the minimal TI (and the corresponding EI) of the internal columns taken together is determined by the distribution of the most heavy-tailed distributed element. The sum $D_i = \sum_{j=1}^{L_i} z_j X_{i,j}$ and maximum $M_i = \max_{j=1,...,L_i}\{z_j X_{i,j}\}$ of weighted elements of the ith string set are the weighted E2E delay and the longest weighted p-h delay at the ith path. The weights $\{z_i\}$ may reflect a priority which can be proportional to capacities of links or impact on the scheduling of the peer selection determining the path. In the simplest case, $\{z_i\}$ are all equal to one.
Suppose without loss of generality $k_1 < k$, where $k = \lim_{n\to\infty} \inf_{2\le j\le l_n} k_j$, $l_n = [n^\chi]$, $0 < \chi < (k - k_1)/(k_1(k + 1))$ holds. According to Theorem 4 in [6] it follows*

$$P\{M_i > x\} \sim P\{D_i > x\} \sim x^{-k_1}, \qquad x \to \infty. \tag{3}$$

This means that the most heavy-tailed distributed column of the p-h delays determines the distributions of the E2E delay and the maximal p-h delay at the ith path. Instead of the E2E delays, one can consider the maximal p-h delay at each path since they have the same heaviness of tail, i.e. the same distribution up to slowly varying functions. This allows us to model the distribution of the p-h delays since the E2E delays can be easily gathered as statistics in practice, rather than the p-h delays.
The EI of M_i and D_i is equal to θ_1 corresponding to k_1. Then the maxima of the sequences $\{D_i\}$ and $\{M_i\}$, $i = 1, ..., n$, have the same limiting distributions. It holds

$$\lim_{n\to\infty} P\{M_n^s \le u_n\} = \lim_{n\to\infty} P\{M_n^m \le u_n\} = e^{-\tau\theta_1} \tag{4}$$

*The symbol $\sim$ means asymptotically equal to or $f(x) \sim g(x) \iff f(x)/g(x) \to 1$ as $x \to a$, $x \in M$, where the functions $f(x)$ and $g(x)$ are defined on some set M and a is a limit point of M.

by (2) with $\lim_{n\to\infty} nP\{M_n > u_n\} = \lim_{n\to\infty} nP\{D_n > u_n\} = \tau$, where

$$M_n^s = \max\{D_1, ..., D_n\}, \quad M_n^m = \max\{M_1, ..., M_n\},$$

and $\{u_n\}$ is an increasing sequence of thresholds. In [6] u_n is selected in such a way that $\tau = (z_1/y)^{k_1}$ with a constant $y > 0$ holds, namely, $u_n = yn^{1/k_1}\ell_1^\sharp(n)$, where $\ell_1^\sharp(n)$ is a slowly varying function.
Regarding the transmission rates of the packets we can argue in the same way. Following [5], each node is a bottleneck and it may upload an own superimposed flow coming from other nodes. Then the ith packet is associated with the sequence of rates $\{R_{i,1}, R_{i,2}, ..., R_{i,L_i}\}$ corresponding to the links of the ith path. We denote $R_{i,j} = Y_i/Z_{i,j}$, where Y_i is the packet length and $Z_{i,j}$ is the inter-arrival time between the ith packet and the previous (or next) one arriving at the jth node. Considering the matrix $\mathbb{X}$ above one can substitute $X_{i,j}$ by $R_{i,j}$ assuming that the columns of the rates have the TIs $\{k_i^*\}$ and EIs $\{\theta_i^*\}$ and a unique minimal TI k_1^* exists as for the p-h delays. Then we obtain (4) with corresponding replacements.
The probability of the successful transmission P_{st} of n packets over their n paths is determined by

$$P_{st} \quad = \quad P\{M_n^{m*} \le u_n^*\} + P\{M_n^m \le b_n\},$$

where $M_n^{m*} = \max\{M_1^*, ..., M_n^*\}$ and $M_i^* = \max_{j=1,...,L_i}\{z_j R_{i,j}\}$ are the maximal transmission rates of the packet over n paths and over the path i, respectively. The excess of the rate over the equivalent channel capacity u_n^* may cause the miss of packets [5]. By (4) $P\{M_n^m \le b_n\}$ is the probability that the longest p-h delay M_n over n paths is less than the playback delay b_n. The sequences $\{u_n^*\}$ and $\{b_n\}$ are determined to be increasing as $n \to \infty$ in the same way as $\{u_n\}$ in [6], i.e. $u_n^* = yn^{1/k_1^*}$ and $b_n = yn^{1/k_1}$ omitting slowly varying functions. Then it holds

$$P_{st} \approx e^{-\tau^*\theta_1^*} + e^{-\tau\theta_1} = e^{-(z_1/y)^{k_1^*}\theta_1^*} + e^{-(z_1/y)^{k_1}\theta_1}, \ y > 0$$

for sufficiently large n, where y is selected in such a way to keep $P_{st} < 1$. The approximate probability to loose at least one packet during the transmission over n paths is $P_m = 1 - P_{st}$. Taking $P_m = \eta$, where $\eta \in (0, 1)$ is small, one can find y.
The biggest problem of the approach is to detect whether the unique column with the smallest TI k_1 exists or not. The discrimination tests of close distribution tails built by only higher order statistics can be used, [9], [10]. The application of the test to each pair of columns to discriminate the heaviest tail consistently may constitute a calculation problem that is out of scope of this paper. Many proposed network architectures place nodes with large upload capacities close to the source S, [2]. One

may expect the smallest capacities and rates at the last link before the destination node D. This may lead to the heaviest distribution tail of the p-h delays and the smallest TI at the last link. Thus, one can estimate and compare the TIs and EIs of the p-h delays (the rates) at the internal part and the last column of the matrix $\mathbb{X}$ above and find the minimal k_1 (k_1^*). To estimate the TI one can recommend estimators designed for dependent data and based on sums and maxima of non-intersecting data blocks, [7], [8]. Among the nonparametric estimators of the EI, the blocks, runs and intervals ones are the most popular, [1]. The first two require a tuning parameter and the threshold u. The intervals estimator needs only u.

4. Conclusions and Open Problems

The performance analysis of the data transfer along paths of random lengths in P2P-overlay networks subject to QoS constraints is considered. The distribution of the E2E packet transfer delay between source and destination nodes is modeled. The E2E delay is determined by the sum of a random number of per-hop (p-h) delays along the links of an overlay path. Based on [6] and assuming that the p-h delays are regularly varying distributed it is shown that the sums and maxima of the p-h delays corresponding to different paths of random lengths may have the same TI and EI. These indexes determine the heaviness of the tail of the delay distribution and the dependence indicator that measures the cluster tendency (i.e. how extreme values arise by groups of observations). Using the EI the limit distributions of the maxima of the E2E and p-h delays over all source-destination paths are identified. For real-time applications with stringent E2E delay constraints, the latter distributions are used to identify QoS metrics like the packet missing probability, the corresponding playback delay and the required equivalent capacity to transfer the flows of data.

The proposed approach requires the verification and comparison of the TIs of p-h delays to find the set of links whose delays have the heaviest tail. Regarding modern network architectures one can expect that the last link before the destination node has the heaviest distribution tail. The known tests allow us to compare pairs of samples in the columns of $\mathbb{X}$ regarding the similarity of their distributions.

The described asymptotic results are valid for sufficiently high thresholds that are in our context the playback delay and the equivalent capacity of the channel. Our results provide the basis for an improved control scheme regarding the optimal selection of transport paths in a P2P-overlay network subject to QoS constraints on the E2E delay and packet loss metrics. Regarding the application of a P2P-overlay concept in 5G networks, we may consider the deployment of a blockchain functionality on top of an underlying network of mining peers that are validating transactions of IoT data processing or the use of P2P video streaming as important examples. In the case of such real-time applications, we are looking for short playback delays,

but they may lead to a large packet missing probability. The derived asymptotic performance analysis models of the E2E transfer delay provide a tendency with an increasing probability of successful packet transmission as both the playback delay and the equivalent capacity increase. But these performance analysis models require an adjustment for short playback delays and not high, realistic capacities. Our future studies will focus on these analysis and design issues of modern teletraffic theory.

Acknowledgments

The first author was partly supported by Russian Foundation for Basic Research (grant 19-01-00090).

REFERENCES

1. Beirlant J., Goegebeur Y., Teugels J., Segers J. Statistics of Extremes: Theory and Applications. Wiley, Chichester, West Sussex, 2004.
2. Dán G., Fodor V. Delay Asymptotics and Scalability for Peer-to-Peer Live Streaming // IEEE Trans. Parallel Distrib. 2009. V. 20(10). P. 1499–1511.
3. Jessen A. H., Mikosch T. Regularly varying functions // Publ. Inst. Math. (Beograd) (N.S.). 2006. V. 80. P. 171–192.
4. Leadbetter M.R., Lingren G., and Rootzen H. Extremes and Related Properties of Random Sequence and Processes, Ch. 3. Springer, New York, 1983.
5. Markovich N.M. Quality assessment of the packet transport of peer-to-peer video traffic in high-speed networks // Performance Evaluation. 2013. V. 70. P. 28–44.
6. Markovich N.M., Rodionov I.V. Maxima and sums of non-stationary random length sequences // Extremes. 2020. V. 23. P. 451–464.
7. Markovich N. M., Vaičiulis M. Modification of Moment-Based Tail Index Estimator: Sums versus Maxima. In: Bertail P. et al. (eds), Nonparametric Statistics. ISNPS 2016. Springer Proceedings in Mathematics & Statistics. V. 250. P. 85–102. Springer, Cham, 2018.
8. McElroy T. and Politis D.N. Moment-Based Tail Index Estimation // J. Statist. Plan. Infer. 2007. V. 137. P. 1389–1406.
9. Rodionov I.V. On discrimination between classes of distribution tails // Probl. Inform. Transm. 2018. V. 54(2). P. 124–138.
10. Rodionov I.V. Discrimination of close hypotheses about the distribution tails using higher order statistics // Theory of Probability and its Applications. 2019. V. 63(3). P. 364–380.
11. Shih M.F., Hero A.O. Unicast-based inference of network link delay distributions using mixed finite mixture models // IEEE Trans. Signal Process. 2003. V. 51(8). P. 2219–2228.

UDC: 519.872

Space merging approach to analysis of queuing system with heterogeneous servers and N-policy

Agassi Melikov[1], Sevinc Aliyeva[2], Mammed Shahmaliyev[3]

[1]Institute of Control Systems, National Academy of Science of Azerbaijan
[2]Baku State University
[3]National Aviation Academy of Azerbaijan, Azerbaijan

agassi.melikov@gmail.com, s@aliyeva.info, mamed.shahmaliyev@gmail.com

Abstract

In this paper, we propose an approximate method to investigate the Markovian queuing system with two separate pools of heterogeneous servers (HS) under N-policy. It is assumed that fast servers (F-servers) remain awake all the time while slow servers (S-servers) will go to sleep independently when number of calls in the buffer less than some threshold. At the end epoch of a sleep period, if the number of the calls gathered in the system buffer reaches or exceeds a given threshold, the corresponding S-server will wake up independently; otherwise, the S-server will begin another sleep period. An approximate method is applied under the condition that the sleep rates is essentially less than both arrival intensity of calls and their service intensity. The joint probability distribution of the number of calls in system and number of busy S-servers is determined by simple computational procedures. Illustrative numerical examples show the high accuracy of the proposed approximate method.

Keywords: queuing system, heterogeneous servers, N-policy, space merging, calculation algorithm

1. Introduction

Because of the intensive development of computer technology, cloud data centers use servers of various capacities. Therefore, for their correct mathematical analysis models of queuing systems with heterogeneous servers (QSwHS) are used. Moreover, such kind of models might be used to study systems where arrived calls are processed by people rather than by machines.

It seems that first serious paper devoted to QSwHS is [1]. In [1], a Markovian system with infinite queue and randomized call admission control (CAC) scheme was investigated. Algrotihm to calculation of steady-state probabilities as well as formulas to determine the mean number of calls in system are proposed.

In QSwHS, problems of determining optimal CAC schemes are important. In [2] it is proven that for QSwHS with two servers the following CAC is optimal to minimize

mean number of calls in the system: the fast server (F-server) always works if there is at least one call in the system, and slow server (S-server) is only involved when queue length reachs a cetain threshold value (N). This CAC usually called N-policy as well. Later this result is stated for models with unreliable servers [3]; generalization of N-policy to models with more than two servers has been proposed in [4-7].

Recently in [8] Markovian QSwHS is applied to study energy consumption problem in cloud data center. In the indicated paper authors propose a clustered Virtual Machine (VM) allocation strategy based on a sleep-mode with a wake-up threshold. The VMs are clustered into two pools, namely, pool of F-servers and pool of S-servers; it is assumed that F-servers remain awake at all times, while S-servers asynchronously go to sleep under a light workload. After a sleep time expires, the S-server will resume processing calls only if the number of waiting calls reaches the wake-up threshold. In other words, in [8], QSwHS with two separate pools of servers and N-policy is investigated. To study constucted two-dimensional Markov chain (2D MC) matrix-geometric method by Neuts [9] is used.

Note that in literature models of QSwHS in case of identical calls are studied in detail. However, models of QSwHS with calls of different kinds represents some practical and scientific interests. Recently models of buffer-less QSwHS and QSwHS with buffers and jump priorities were investigated in papers [10] and [11], respectively. Note that the indicated papers contain review of available papers devoted to QSwHS.

This paper in spirit is close to [8] and here, to improve readability, we keep its notation. In this paper, we consider alternate (approximate) method to investigate the model that proposed in [8]. Proposed here approach based on the space-merging algorithm (SMA) to calculation of the steady-state probabilities of 2D MC in [10, 11].

The paper is organized as follows. In Section 2 the model of the QSwHS under study is described. Algorithm to calculate the elements of generating matrix (Q-matrix) and explicit formulas for performance measures are given in Section 3. In Section 4 SMA to approximately calculation of steady-state probabilities and performance measures is developed. Results of numerical experiments are demonstrated in Section 5. Concluding remarks are given in Section 6.

2. The model

The investigated system contains two pools of heterogeneous servers. Pool I contains F-servers and their number is equal to c; pool II contains S-servers and their number is equal to d. The input flow is Poisson one with rate λ. The service rates of calls in Pool I and in Pool II are assumed to be exponentially distributed with parameters μ_1 and μ_2, respectively and $\mu_1 > \mu_2$.

In the system, the following service mechanism is used. The F-servers remain awake all the time while S-servers will go to sleep independently when number of calls in the buffer less than a given wake-up threshold N. At the end epoch of a sleep period, if the number of the calls gathered in the system buffer reaches or exceeds

wake-up threshold, the corresponding S-server will wake up independently. Otherwise, the S-server will begin another sleep period. The sleep time length is assumed to follow an exponential distribution with parameterθ ($\theta > 0$).

Rules for transitions from awake state to sleep state and from sleep state to awake state are defined as follows.

· For a busy S-server, the state transition from awake state to sleep state occurs only at the instant when a call either in Pool I or Pool II is completely processed. When a call is completely processed in F-server, if the number of call in system is zero and there is at least one call being processed in S-server, one of the calls being processed in Pool II will migrate Pool I, and then the evacuated S-server will go to sleep. When a call is completely processed in S-server, if the number of call in system is zero, the evacuated S-server will go to sleep directly.

· For a sleeping S-server, the state transition from sleep state to awake state occurs only at the end epoch of a sleep period. When a sleep timer expires, if the number of call in system is equal to or greater than the wake-up threshold N, the corresponding S-server will wake up to process the first call waiting in the system buffer; otherwise, a new sleep timer will be started and the S-server will begin another sleep period.

The problem is finding the joint probability distribution of the number of calls in system and number of busy S-servers. Determination of the indicated probability distribution allows calculate the desired QoS metrics as well.

3. Exact method

State of the system is defined by the two-dimensional (2D) vector (i, j), where i is the total number of calls in the system, $i = 0, 1, ...$, and j indicates the number of busy S-channels, $j = 0, 1, ..., d$. Based on the distribution function of the random variables involved in the formation of the model, we determine that the given system is described by the two-dimensional Markov chain (2D MC). The state space of this 2D MC is defined as following

$$E = \bigcup_{i=0}^{d} E_i \,,\ E_{i_1} \bigcap E_{i_2} = \emptyset \ ,\ \text{if } i_1 \neq i_2 \ , \tag{1}$$

where $E_0 = \{(i,\, 0) \,:\, i \geq 0\}$, $E_k = \{(i,\, k) \,:\, i \geq c + k\}$, $k = \overline{1,\, d}$.

The transition rate from the initial state $(i_1, j_1) \in E$ to the final state $(i_2, j_2) \in E$ is denoted as $q((i_1, j_1), (i_2, j_2))$. The combination of these quantities forms Q-matrix of given 2D MC and are determined from the following relations.

For the initial state of the type $(i_1,\, 0)$:

$$q\left((i_1,\, 0)\,, (i_2, j_2)\right) = \begin{cases} \lambda\,, & \text{if } (i_2, j_2) = (i_1 + 1,\, 0)\ , \\ \min\left(i_1,\, c\right)\mu_1, & \text{if } (i_2, j_2) = (i_1 - 1,\, 0)\ , \\ d\theta, & \text{if } i_1 \geq c + N,\ (i_2, j_2) = (i_1,\, 1)\ , \\ 0 & \text{in other cases.} \end{cases} \tag{2}$$

For the initial state of the type $(i_1,\, j_1)$, $i_1 \geq c + j_1$:

$$q\left((i_1, j_1), (i_2, j_2)\right) = \begin{cases} \lambda\,, & \text{if } (i_2, j_2) = (i_1 + 1, j_1)\,, \\ c\mu_1 + j_1\mu_2, & \text{if } i_1 > c + j_1\,,\ (i_2, j_2) = (i_1 - 1, j_1)\,, \\ c\mu_1 + j_1\mu_2, & \text{if } i_1 = c + j_1\,,\ (i_2, j_2) = (i_1 - 1, j_1 - 1)\,, \\ (d - j_1)\,\theta\,, & \text{if } i_1 \geq c + N + j_1\,,\ (i_2, j_2) = (i_1, j_1 + 1)\,, \\ 0 & \text{in other cases.} \end{cases} \quad (3)$$

As we seen from (2) and (3) the given infinite 2D MC represent level dependent quasi birth-death (LDQBD) process. Below the ergodicity condition of this 2D MC is established.

Let $p(i, j)$ means the steady-state probability of state $(i, j) \in E$. It is easy to derive desired QoS metrics via steady-state probabilities. We consider following QoS metrics.

· The average number of busy F-channels (N_{av}^F) is given by

$$N_{av}^F = \sum_{i=1}^{c-1} ip\,(i,\, 0)\ + c\left\{\sum_{i=c}^{\infty} p\,(i,\, 0) + \sum_{i=c+1}^{c+d}\sum_{j=1}^{i-c} p\,(i, j) + \sum_{i=c+d+1}^{\infty}\sum_{j=1}^{d} p\,(i,\, j)\right\}\ . \quad (4)$$

· The average number of busy S-channels (N_{av}^S) is given by

$$N_{av}^S = \sum_{j=1}^{d} j \sum_{i=c+j}^{\infty} p\,(i,\, j)\ . \quad (5)$$

· The average number of the calls in system (L_s) is expressed as

$$L_s = \sum_{i=1}^{c} ip\,(i,\, 0) + \sum_{i=c+1}^{c+d} i\sum_{j=0}^{i-c} p\,(i,\, j) + \sum_{i=c+d+1}^{\infty} i\sum_{j=0}^{d} p\,(i,\, j)\ . \quad (6)$$

· The average sojourn time of the calls in system (W_s) is calculated by Little' formula, i.e.

$$W_s = \frac{1}{\lambda} L_s\,. \quad (7)$$

4. Approximate method

Here we propose an effective and simple numerical method for approximate analysis of the investigated system. It based on space merging approach to calculate the stationary distribution of 2D MC. For the correct use of this method, we assume that $\lambda >> \theta$ and $c\mu_1 + d\mu_2 >> \theta$.

In accordance to our assumption the transition intensities between states in the classes E_i , $i = \overline{0, d}$,are too large than intensities between states from different classes. By using this fact consider the following splitting of state space E:

$$E = \bigcup_{j=0}^{d} E_j\ ,\ E_{j_1} \bigcap E_{j_2} = \emptyset \ \text{ if } j_1 \neq j_2\ , \quad (8)$$

where $E_0 = \{(i,\,0) \,:\, i \geq 0\}$, $E_j = \{(i,\,j) \,:\, i \geq c+j\}$, $j = \overline{1,\,d}$.

Based of the splitting (8) the merge function on the state space E is determined as follows:

$$U\left((i,j)\right) =< j > \text{ if } (i,j) \in E_j \,,$$

where $< j >$ is a merge state, which includes all the states of the E_j , $j = \overline{0,d}$. Let $\Omega = \left\{< j >:\, j = \overline{0,\,d}\right\}$.

The approximate values of steady-state probabilities of the initial model are defined as follows:

$$\tilde{p}\,(i,j) \approx \rho_j\,(i)\,\pi\,(< j >) \,, \tag{9}$$

where $\rho_j\,(i)$ denotes the state probability of state (i,j) within the splitting model with state spaceE_j, and $\pi\,(< j >)$ is the probability of the merge state $< j >\in \Omega$.

From splitting scheme (8) it is clear that all the splitting models are one-dimensional birth and death processes (1D BDP), so that in the class of states E_j the second component is constant. Therefore, in the splitting model with state space E_j microstate $(i,j) \in E$ can be represent by scalar i.

From (2) we get that probabilities $\rho_0\,(i)$ coincide with the steady-state probabilities of the classical $M/M/c/\infty$ system with individual server utilization, $\nu_0 = \lambda/c\mu_1$, i.e. if $\nu_0 < 1$ we have

$$\rho_0\,(i) = \begin{cases} \frac{(c\nu_0)^i}{i!}\rho_0\,(0) \,, & 0 \leq i \leq c\,, \\ \frac{\nu_0^i c^c}{c!}\rho_0\,(0) \,, & i > c\,, \end{cases} \tag{10}$$

where $\rho_0\,(0) = \left(\sum_{i=0}^{c-1} \frac{(c\nu_0)^i}{i!} + \frac{(c\nu_0)^c}{c!} \cdot \frac{1}{1-\nu_0}\right)^{-1}$.

From (3) we get that probabilities $\rho_j\,(i)$, $j = \overline{1,d}$, coincide with steady-state probabilities of the classical $M/M/1/\infty$ with load ν_j where $\nu_j = \lambda/\left(c\mu_1 + j\mu_2\right)$, i.e. $\nu_j < 1$ we have

$$\rho_j\,(i) = \nu_j^{i-c-j}\,(1-\nu_j) \,,\ i = c+j,\, c+j+1\,, \ldots \tag{11}$$

Note. Since conditions $\nu_j < 1$ should be satisfied for each $j = \overline{0,d}$, so we obtain ergodicity condition of the initial model, i.e. $\nu_0 < 1$. This condition does not depend on number of S-servers (d) as well as their setup-time $\left(\theta^{-1}\right)$. This fact is expected one since according to our assumption, the sleep period of S-servers is very large with comparison of arrival rate and total service intensity, and so the ergodicity condition is determined only by the intensity of service of F-servers.

The transition intensity from the merge state $< j_1 >$ to other merge state $< j_2 >$ is denoted as $q\,(< j_1 >, < j_2 >)$, $< j_1 >, < j_2 >\in \Omega$. These quantities are calculated as follows:

$$q\,(< j_1 >, < j_2 >) = \sum_{\substack{(i_1,j_1) \in E_{j_1} \\ (i_2,j_2) \in E_{j_2}}} q\left((i_1,j_1)\,,\,(i_2,j_2)\right)\rho_{j_1}\,(i_1) \,. \tag{12}$$

After certain algebras on the bases of (2), (3) and (10)-(12) we obtain:

$$q\left(< j_1 >, < j_2 >\right) = \begin{cases} \alpha\left(j_1\right), & j_2 = j_1 + 1, \\ \beta\left(j_1\right), & j_2 = j_1 - 1, \\ 0 & \text{in other cases,} \end{cases} \tag{13}$$

where $\alpha\left(j_1\right) = \left(d - j_1\right)\theta\left(1 - \sum_{i=\chi(j_1)}^{c+N+j_1-1} \rho_{j_1}\left(i\right)\right)$, $\chi\left(j_1\right) = \begin{cases} c + j_1, & j_1 > 0, \\ 0, & j_1 = 0, \end{cases}$ $j_1 = \overline{0, d-1}$;

$$\beta\left(j_1\right) = \left(c\mu_1 + j_1\mu_2\right)\rho_{j_1}\left(c + j_1\right),\ j_1 = \overline{1, d}.$$

From (13) we conclude that the probabilities of the merging states $\pi\left(< j >\right), < j > \in \Omega$, are calculated as the state probabilities of 1-D BDP. In other words,

$$\pi\left(< j >\right) = \prod_{i=1}^{j} \frac{\alpha\left(i-1\right)}{\beta\left(i\right)} \pi\left(< 0 >\right),\ j = 1, ..., d, \tag{14}$$

where $\pi\left(< 0 >\right)$ is derived from normalizing condition, i.e. $\sum_{j=0}^{d} \pi\left(< j >\right) = 1$.

Finally, taking into account the relations (10), (11), (14) from (9) we calculate the steady-state probabilities of the initial 2D MC. After certain algebras, we obtain the following approximate formulas for calculating the desired performance measures of the system:

$$N_{av}^{F} \approx c\left(1 - \pi\left(< 0 >\right)\right) + \left\{\sum_{i=1}^{c-1} i\rho_0\left(i\right) + c\left(1 - \sum_{i=0}^{c-1} \rho_0\left(i\right)\right)\right\} \pi\left(< 0 >\right); \tag{15}$$

$$N_{av}^{S} \approx \sum_{j=1}^{d} j\pi\left(< j >\right); \tag{16}$$

$$\begin{aligned} L_s \approx\ & \pi\left(< 0 >\right) \sum_{i=1}^{c} i\rho_0\left(i\right) + \sum_{i=1}^{d} \left(c + i\right) \sum_{j=0}^{i} \rho_j\left(c + i\right) \pi\left(< j >\right) + \\ & + \sum_{i=1}^{\infty} \left(c + d + i\right) \sum_{j=0}^{d} \rho_j\left(c + d + i\right) \pi\left(< j >\right). \end{aligned} \tag{17}$$

The approximate value of average sojourn time of the calls in system is calculated from (17) by using (7).

5. Numerical Results

Numerical experiments have two goals: firstly, to analyze the accuracy of the state probabilities and performance measures calculated by the developed approximate formulas; secondly, to perform some numerical experiments to study behavior of the performance measures (4)-(7) with respect to sleep parameter and number of F-servers.

Regarding the first goal, note that, analytical analysis of the accuracy of developed formulas is impossible. Therefore, comparative analysis of the results via numerical

experiments is used. The accuracy of calculating the approximate values of stationary probabilities is estimated using the following norms: cosine similarity $\|N\|_1$; Jaccard coefficient $\|N\|_2$; Euclidean distance $\|N\|_3$

In numerical experiments, the values of system parameters are same as in [8]. Some results of the comparative analysis of the both steady-state probabilities and performance measures calculations for the exact and approximate approaches are given in table.

(c,d)	(N,θ)	(μ_1,μ_2)	Norms			N_{av}^F		N_{av}^S		W_s	
			$\|N\|_1$	$\|N\|_2$	$\|N\|_3$	EV	AV	EV	AV	EV	AV
(45, 5)	(10, 0.5)	(0.2, 0.1)	1	0.97	0.004	34.989	35.073	0.022	0.008	5.032	5.050
		(0.3, 0.2)	1	1	4E-08	23.333	23.333	1.7E-07	2E-08	3.333	3.333
		(0.4, 0.3)	1	1	3E-08	17.500	17.500	4.2E-10	6E-09	2.500	2.500
	(11, 1.0)	(0.2, 0.1)	1	0.97	0.005	34.989	35.117	0.023	0.013	5.032	5.059
		(0.3, 0.2)	1	1	3E-08	23.333	23.333	1.4E-07	3E-08	3.333	3.333
		(0.4, 0.3)	1	1	3E-08	17.500	17.500	2.9E-10	1E-08	2.500	2.500
	(12, 1.5)	(0.2, 0.1)	1	0.96	0.006	34.990	35.138	0.021	0.016	5.032	5.064
		(0.3, 0.2)	1	1	2E-08	23.333	23.333	9.8E-08	2E-08	3.333	3.333
		(0.4, 0.3)	1	1	3E-08	17.500	17.500	1.6E-10	2E-08	2.500	2.500
(42, 8)	(10, 0.5)	(0.2, 0.1)	0.99	0.80	0.031	34.931	35.688	0.138	0.133	5.097	5.298
		(0.3, 0.2)	1	1	8E-07	23.333	23.333	3.6E-06	7E-07	3.333	3.333
		(0.4, 0.3)	1	1	3E-08	17.500	17.500	2.2E-10	1E-08	2.500	2.500
	(11, 1.0)	(0.2, 0.1)	0.98	0.68	0.054	34.932	36.306	0.137	0.301	5.097	5.455
		(0.3, 0.2)	1	1	6E-07	23.333	23.333	2.9E-06	7E-07	3.333	3.333
		(0.4, 0.3)	1	1	3E-08	17.500	17.500	1.5E-10	2E-08	2.500	2.500
	(12, 1.5)	(0.2, 0.1)	0.96	0.61	0.070	34.937	36.738	0.326	0.439	5.099	5.566
		(0.3, 0.2)	1	1	5E-07	23.333	23.333	2.1E-06	6E-07	3.333	3.333
		(0.4, 0.3)	1	1	3E-08	17.500	17.500	8.2E-11	3E-08	2.500	2.500
(44, 6)	(10, 0.5)	(0.2, 0.1)	1	0.92	0.012	34.973	35.284	0.053	0.039	5.045	5.116
		(0.3, 0.2)	1	1	2E-07	23.333	23.333	6.6E-07	1E-07	3.333	3.333
		(0.4, 0.3)	1	1	3E-08	17.500	17.500	2.4E-11	1E-08	2.500	2.500
	(11, 1.0)	(0.2, 0.1)	1	0.89	0.019	34.974	35.503	0.051	0.077	5.045	5.163
		(0.3, 0.2)	1	1	1E-07	23.333	23.333	5.2E-07	1E-07	3.333	3.333
		(0.4, 0.3)	1	1	3E-08	17.500	17.500	1.5E-11	2E-08	2.500	2.500
	(12, 1.5)	(0.2, 0.1)	1	0.86	0.0228	34.977	35.631	0.046	0.101	5.046	5.192
		(0.3, 0.2)	1	1	8E-08	23.333	23.333	3.4E-07	1E-07	3.333	3.333
		(0.4, 0.3)	1	1	4E-08	17.500	17.500	7.9E-10	3E-08	2.500	2.500

Table 1. Estimation of calculation accuracy for steady-state probabilities and performance measures; EV – exact value, AV – approximate value.

From the table we conclude that the developed approximate formulas to calculate the steady-state probabilities has high accuracy, because the values of both norms $\|N\|_1$ and $\|N\|_2$ are very close to 1 while Euclidean distance $\|N\|_3$ is negligible. It is important to note that our results completely coincide with the results obtained in [8]. For the given initial data the analysis of accuracy of the calculation of system performance measures (4)-(7) was performed as well. It should be noted that the performance measures are almost the same when using exact and approximate approaches (see table). At the same time, the complexity of the proposed algorithm for calculating the steady-state probabilities and performance measures is significantly lower than the algorithm proposed in [8].

6. Conclusion

In this paper, we have analyzed a queuing system with heterogeneous servers in which N-policy is used. Fast servers are active for all the time while slow servers will

wake-up if the number of the calls in the buffer reaches or exceeds a given threshold; S-servers go to sleep asynchronously when number of calls in the buffer less than indicated threshold. An approximate method based on space merging approach is applied to calculate the steady-state probabilities of constructed two-dimensional Markov chain with infinite state space as well as performance measures of investigated system. Illustrative numerical examples show the high accuracy of the proposed approximate method.

REFERENCES

1. Gumbel H. Waiting Lines with Heterogeneous Servers. // Operations Research. 1960. V. 8. Issue 4. P. 504-511.
2. Larsen R.L., Agrawala A.K. Control of Heterogeneous Two-Server Exponential Queuing System. // IEEE Transactions on Software Engineering. 1983. Vol. 9. Issue 4. P. 522-526.
3. Efrosinin D. Sztrik J., Farkhadov M., Stepanova N. Reliability Analysis of Two-Server Heterogeneous Queuing System with Threshold Control Policy. // Communications in Computer and Information Sciences. 2017. Vol. 800. P. 13-27.
4. Viniotis I., Ephremides A. Extension of the Optimality of a Threshold Policy in Heterogeneous Multi-Server Queuing Systems. // IEEE Transactions on Automatic Control. 1988. Vol. 33. P. 104-109.
5. Rosberg Z., Makowski A.M. Optimal Routing to Parallel Heterogeneous Servers – Small Arrival Rates. // IEEE Transactions on Automatic Control. 1990. Vol. 35. Issue 7. P. 789-796.
6. Rykov V.V. Monotone Control of Queuing Systems with Heterogeneous Servers. // Queuing Systems. 2001. Vol. 37. P. 391-403.
7. Rykov V.V., Efrosinin D. On the Slow Server Problem. // Automation and Remote Control. 2009. Vol. 70. Issue 12. P. 2013-2023.
8. Jin S., Qie X., Zhao W., Yue W., Takahash Y. A Clustered Virtual Machine Allocation Strategy based on a Sleep-Mode with Wake-Up Threshold in a Cloud Environment. // Annals of Operations Research. https://doi.org/10.1007/s10479-019-03339-3
9. Neuts M.F. Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach. // Baltimore: John Hopkins University Press. 1981. 332 .
10. Melikov A.Z., Ponomarenko L.A.,Mekhbaliyeva E.V. Analysis of Models of Systems with Heterogeneous Servers. // Cybernetics and System Analysis. 2020. Vol. 56. Issue 1. P. 89-99.
11. Melikov A.Z., Mekhbaliyeva E.V. Analysis and Optimization of System with Heterogeneous Servers and Jump Priorities. // Journal of Computer and Systems Sciences International. 2019. Vol. 58. Issue 5. P. 718-735.

UDC: 519.21

On a Total Amount of Occupied Resource in the System with Parallel Service and Renewal Arrival Process

E. Chernyshova[1], E. Lisovskaya[2], S. Moiseeva[1], M. Pagano[3]

[1]Tomsk State University, Lenina ave. 36, Tomsk, 634050, Russia
[2]Peoples' Friendship University of Russia,
Miklukho-Maklaya str. 6, Moscow, 117198, Russia
[3]University of Pisa, Via Caruso 16, I-56122, Pisa, Italy

chernishova@stud.tsu.ru, lisovskaya-eyu@rudn.ru, smoiseeva@mail.ru,
michele.pagano@iet.unipi.it

Abstract

Resource queueing systems has emerged as a powerful tool to calculate the capacity of a base station that serves customers using video services. In this scenario, two types of resources are used: uplink and downlink bandwidth. This means that if there is not enough bandwidth to meet the requirements, the client will not be able to connect. Network designers set the task of determining the optimal value of the cell capacity in order to minimize the loss of connection and downtime of radio resources due to which the operator may suffer losses. To address such issues, this paper carries out an analytical study of a resource queue with random resource requirements. In more detail, the paper deals with a renewal arrival process and the parallel service of each customer in two infinite-server blocks. Balance equations are solved under the asymptotic condition of the high intensity of the arrival process and it is obtained that the two-dimensional stationary probability distribution of the amount of occupied resources in both server blocks is approximately two-dimensional Gaussian. Parameters of the distribution are also derived.

Keywords: service system, random resources, renewal arrival process.

1. Introduction

Nowadays, research on queueing systems is very popular, as the demand for high-quality communication is constantly growing. To describe complex processes, the modulated and renewal arrival processes are used, thanks to their flexibility and relative simplicity. Typically, these processes are investigated by numerical methods and simulation, while analytical results are obtained only for special cases.

The development of new models and methods of data storage is also the goal of queueing theory. Using resource systems, i.e., queueing system in which customers require random amounts of different resources, it becomes possible to make an overall assessment of the occupied resource amounts and prevent overloading the system with a limited number of physical resource blocks [1, 4]. In this paper, using the asymptotic analysis method and dynamic screening method [2], we analytically investigate a system with a renewal arrival process and parallel service of customers.

2. Mathematical model

Consider a queueing system consisting of two blocks [5], each of them has an unlimited number of servers and resources. The customers enter in the system according to a renewal arrival process, given by the CDF of the interarrival times $A(z) = P\{\zeta < z\}$. Each arrival customer is duplicated, then it goes into each of the blocks and instantly takes any free server where it is serviced for a random time ξ_i, $i = 1, 2$. The CDF of the service time in the first block is $B_1(\tau) = P\{\xi_1 < \tau\}$, in the second is $B_2(\tau) = P\{\xi_2 < \tau\}$. The CDFs of the random amount of resources in the first and second blocks are $G_1(y) = P\{\nu_1 < y\}$ and $G_2(y) = P\{\nu_2 < y\}$, respectively [3]. Upon completion of service, the customer leaves the system, frees the server and all used resources. The occupied resources amount and the service time are independent of each other.

Denote the total amount of occupied resources on the first and second blocks at time t by $V_1(t)$ and $V_2(t)$, respectively. Our goal is to obtain the probability distribution $P(v_1, v_2, t) = P\{V_1(t) < v_1, V_2(t) < v_2\}$ of the stochastic two-dimensional process $\{V_1(t), V_2(t)\}$. This process is not Markovian, and for further study we will be using the dynamic screening method, and adding the component $z(t)$ as the residual time from time t to the time of the next event in the flow. To do this, we introduce the functions $S_i(t) = 1 - B_i(T - t)$, which give the probability that the customer arrived in the system at moment t in the block $i = 1, 2$ will not finish the service at an arbitrary time T in the future, provided that at the initial moment t_0 the system was empty. After applying the dynamic screening method and introducing the partial characteristic function

$$h(z, u_1, u_2, t) = \int_0^\infty e^{ju_1v_1} \int_0^\infty e^{ju_2v_2} P(z, dv_1, dv_2, t),$$

we obtain the system of differential equations for the characteristic function of the distribution:

$$\frac{\partial h(z,u_1,u_2,t)}{\partial t} = \frac{\partial h(z,u_1,u_2,t)}{\partial z} + \frac{\partial h(0,u_1,u_2,t)}{\partial z}\Big[A(z) - 1 + \\ + A(z)S_1(t)(G_1^*(u_1) - 1) + A(z)S_2(t)(G_2^*(u_2) - 1) + \\ + A(z)S_1(t)S_2(t)(G_1^*(u_1) - 1)(G_2^*(u_2) - 1)\Big], \quad (1)$$

where $G_i^*(u_i) = \int\limits_0^\infty e^{ju_iy} dG_i(y)$, $i = \overline{1,2}$.

3. Asymptotic Analysis

We will find the solution of equation (1) by the asymptotic analysis method under the condition of the increasing intensity of arrivals. Let $A(N,z) = P\{N\zeta < z\}$ be the CDF of interarrival times, where N is a high-intensity parameter (theoretically, $N \to \infty$).

The equation (1) will change as:

$$\frac{1}{N}\frac{\partial h(z,u_1,u_2,t)}{\partial t} = \frac{\partial h(z,u_1,u_2,t)}{\partial z} + \frac{\partial h(0,u_1,u_2,t)}{\partial z}\Big[A(z) - 1 + \\ + A(z)S_1(t)(G_1^*(u_1) - 1) + A(z)S_2(t)(G_2^*(u_2) - 1) + \\ + A(z)S_1(t)S_2(t)(G_1^*(u_1) - 1)(G_2^*(u_2) - 1)\Big]. \quad (2)$$

Theorem 1. The first-order asymptotic partial characteristic function of the stochastic process $\{z(t), V_1(t), V_2(t)\}$ has the form:

$$h(z,u_1,u_2,t) \approx R(z)\exp\left\{ju_1N\lambda a_1\int\limits_{t_0}^{t} S_1(\tau)d\tau + ju_2N\lambda a_2\int\limits_{t_0}^{t} S_2(\tau)d\tau\right\},$$

where $R(z)$ is the stationary probability distribution of the stochastic process $z(t)$; $\lambda = \left[\int_0^\infty (1 - A(z))dz\right]^{-1}$; a_1 and a_2 are means of the occupied resource amounts.

Proof. Let us make the following substitutions in equation (2):

$$\varepsilon = \frac{1}{N}, \quad u_1 = \varepsilon x_1, \quad u_2 = \varepsilon x_2, \quad h(u_1,u_2,t) = f_1(x_1,x_2,t,\varepsilon).$$

We obtain

$$\varepsilon\frac{\partial f_1(z,x_1,x_2,t,\varepsilon)}{\partial t}=\frac{\partial f_1(z,x_1,x_2,t,\varepsilon)}{\partial z}+\frac{\partial f_1(0,x_1,x_2,t,\varepsilon)}{\partial z}\Big[A(z)-1+ \\ +A(z)S_1(t)(G_1^*(\varepsilon x_1)-1)+A(z)S_2(t)(G_2^*(\varepsilon x_2)-1)+ \\ +A(z)S_1(t)S_2(t)(G_1^*(\varepsilon x_1)-1)(G_2^*(\varepsilon x_2)-1)\Big]. \tag{3}$$

Step 1. Let us find asymptotic solution $f_1(z,x_1,x_2,t)=\lim\limits_{\varepsilon\to 0} f_1(z,x_1,x_2,t,\varepsilon)$ of equation (3) when $\varepsilon\to 0$:

$$\frac{\partial f_1(z,x_1,x_2,t)}{\partial z}+\frac{\partial f_1(0,x_1,x_2,t)}{\partial z}[A(z)-1]=0$$

and assume that:

$$f_1(z,x_1,x_2,t)=R(z)\Phi_1(x_1,x_2,t), \tag{4}$$

where $\Phi_1(x_1,x_2,t)$ is a scalar differentiable function satisfying the condition $\Phi_1(x_1,x_2,t_0)=1$.

Step 2. In equation (3), we make the transition to the limit $z\to\infty$:

$$\varepsilon\frac{\partial f_1(\infty,x_1,x_2,t,\varepsilon)}{\partial t}=\frac{\partial f_1(0,x_1,x_2,t,\varepsilon)}{\partial z}\Big[S_1(t)(G_1^*(\varepsilon x_1)-1)+ \\ +S_2(t)(G_2^*(\varepsilon x_2)-1)+S_1(t)S_2(t)(G_1^*(\varepsilon x_1)-1)(G_2^*(\varepsilon x_2)-1)\Big]. \tag{5}$$

We substitute (4) into (5) and use the first order exponent expansion in the form: $G_i^*(\varepsilon x)=\int\limits_0^\infty e^{j\varepsilon x_i y}dG_i(y)=\int\limits_0^\infty(1+j\varepsilon x_i y+O(\varepsilon^2))dG_i(y)=1+j\varepsilon x_i a_i+O(\varepsilon^2)$; further we divide everything by ε and perform the passage to the limit as $\varepsilon\to 0$. Notice that $R'(0)=\lambda$:

$$\frac{\partial\Phi_1(x_1,x_2,t)}{\partial t}=\Phi_1(x_1,x_2,t)[jx_1\lambda a_1 S_1(t)+jx_2\lambda a_2 S_2(t)]. \tag{6}$$

The solution of the differential equation (6) is

$$\Phi_1(x_1,x_2,t)=\exp\left\{ju_1\lambda a_1\int\limits_{t_0}^{t}S_1(\tau)d\tau+ju_2\lambda a_2\int\limits_{t_0}^{t}S_2(\tau)d\tau\right\}$$

that leads to the following asymptotic approximation equality for $\varepsilon\to 0$:

$$h(z,u_1,u_2,t)\approx R(z)\exp\left\{ju_1N\lambda a_1\int\limits_{t_0}^{t}S_1(\tau)d\tau+ju_2N\lambda a_2\int\limits_{t_0}^{t}S_2(\tau)d\tau\right\}.$$

The theorem is proved. ■

Theorem 2. The second-order asymptotic partial characteristic function of the stochastic process $\{z(t), V_1(t), V_2(t)\}$ has the form:

$$\begin{aligned} h(z,u_1,u_2,t) \approx R(z)\exp\Bigg\{ & ju_1N\lambda a_1\int\limits_{t_0}^{t} S_1(\tau)d\tau + ju_2N\lambda a_2\int\limits_{t_0}^{t} S_2(\tau)d\tau + \\ & + \frac{(ju_1)^2}{2}\left(N\lambda\alpha_1\int\limits_{t_0}^{t} S_1(\tau)d\tau + N\kappa a_1^2\int\limits_{t_0}^{t} S_1^2(\tau)d\tau\right) + \\ & + \frac{(ju_2)^2}{2}\left(N\lambda\alpha_2\int\limits_{t_0}^{t} S_2(\tau)d\tau + N\kappa a_2^2\int\limits_{t_0}^{t} S_2^2(\tau)d\tau\right) + \\ & + ju_1ju_2N(\lambda+\kappa)a_1a_2\int\limits_{t_0}^{t} S_1(\tau)S_2(\tau)d\tau\Bigg\}, \end{aligned} \tag{7}$$

where α_1 and α_2 are the second raw moments of the occupied resources.

Proof. Represent the function $h(z,u_1,u_2,t)$ as:

$$h(z,u_1,u_2,t) = h_2(z,u_1,u_2,t)\exp\left\{ju_1N\lambda a_1\int\limits_{t_0}^{t} S_1(\tau)d\tau + ju_2N\lambda a_2\int\limits_{t_0}^{t} S_2(\tau)d\tau\right\}.$$

We obtain the equation regarding $h_2(z,u_1,u_2,t)$:

$$\begin{aligned} & \frac{1}{N}\frac{\partial h_2(z,u_1,u_2,t)}{\partial t} + h_2(z,u_1,u_2,t)(ju_1\lambda a_1S_1(t) + ju_2\lambda a_2S_2(t)) = \\ & = \frac{\partial h_2(z,u_1,u_2,t)}{\partial z} + \frac{\partial h_2(0,u_1,u_2,t)}{\partial z}\Big[A(z) - 1 + \\ & + A(z)S_1(t)(G_1^*(u_1) - 1) + A(z)S_2(t)(G_2^*(u_2) - 1) + \\ & + A(z)S_1(t)S_2(t)(G_1^*(u_1) - 1)(G_2^*(u_2) - 1)\Big]. \end{aligned} \tag{8}$$

Let us make the following substitutions in equation (8)

$$\varepsilon^2 = \frac{1}{N}, \quad u_1 = \varepsilon x_1, \quad u_2 = \varepsilon x_2, \quad h_2(z,u_1,u_2,t) = f_2(z,x_1,x_2,t,\varepsilon).$$

We obtain

$$\begin{aligned}\varepsilon^2\frac{\partial f_2(z,x_1,x_2,t,\varepsilon)}{\partial t}+f_2(z,x_1,x_2,t,\varepsilon)(j\varepsilon x_1\lambda a_1S_1(t)+j\varepsilon x_2\lambda a_2S_2(t))=\\ =\frac{\partial f_1(z,x_1,x_2,t,\varepsilon)}{\partial z}+\frac{\partial f_1(0,x_1,x_2,t,\varepsilon)}{\partial z}\Big[A(z)-1+\\ +A(z)S_1(t)(G_1^*(\varepsilon x_1)-1)+A(z)S_2(t)(G_2^*(\varepsilon x_2)-1)+\\ +A(z)S_1(t)S_2(t)(G_1^*(\varepsilon x_1)-1)(G_2^*(\varepsilon x_2)-1)\Big].\end{aligned}\tag{9}$$

Step 1. Let us find asymptotic solution $f_2(z,x_1,x_2,t)=\lim\limits_{\varepsilon\to 0}f_2(z,x_1,x_2,t,\varepsilon)$ of equation (9) when $\varepsilon\to 0$:

$$\frac{\partial f_2(z,x_1,x_2,t)}{\partial z}+\frac{\partial f_2(0,x_1,x_2,t)}{\partial z}[A(z)-1]=0.$$

We will find the function $f_2(z,x_1,x_2,t)$ as:

$$f_2(z,x_1,x_2,t)=R(z)\Phi_2(x_1,x_2,t),\tag{10}$$

where $\Phi_2(x_1,x_2,t)$ is a scalar differentiable function satisfying the condition $\Phi_2(x_1,x_2,t_0)=1$.

Step 2. We write the solution of equation (9) as power expansion

$$\begin{aligned}f_2(z,x_1,x_2,t,\varepsilon)=\Phi_2(x_1,x_2,t)\big[R(z)+\\ +(j\varepsilon x_1\lambda a_1S_1(t)+j\varepsilon x_2\lambda a_2S_2(t))f(z)+O(\varepsilon^2)\big],\end{aligned}\tag{11}$$

where $f(z)$ is some differentiable function.

We substitute (11) into (9) and get a differential equation for an unknown function $f(z)$:

$$f(z)=f'(0)\int\limits_0^z(1-A(x))dx+\int\limits_0^z(R(x)-A(x))dx.$$

Step 3. In equation (9), we make the transition to the limit $z\to\infty$. The function $f_2(z,x_1,x_2,t)$ is monotonically increasing and bounded above on z, then:

$$\lim_{z\to\infty}\frac{f_2(z,x_1,x_2,t,\varepsilon)}{\partial z}=0.$$

We use the second-order exponent expansion in the form:

$$G_i^*(\varepsilon x)=\int\limits_0^\infty e^{j\varepsilon x_iy}dG(y)=1+j\varepsilon x_ia_i+\frac{(j\varepsilon x_i)^2}{2}\alpha_i+O(\varepsilon^3).\tag{12}$$

We substitute (11) into (9), using (12), dividing everything by ε^2, for $z \to \infty$ and $\varepsilon \to 0$, we obtain the equation for $\Phi_2(x_1, x_2, t)$:

$$\frac{\partial \Phi_2(x_1,x_2,t)}{\partial t} = \Phi_2(x_1,x_2,t)\left[\frac{(jx_1)^2}{2}(\lambda\alpha_1 S_1(t) + \kappa a_1^2 S_1^2(t)) + \right.$$
$$\left. + \frac{(jx_2)^2}{2}(\lambda\alpha_2 S_2(t) + \kappa a_2^2 S_2^2(t)) + jx_1 jx_2 a_1 a_1 (\lambda+\kappa) S_1(t) S_2(t)\right],$$

where $\kappa = 2f'(0) - 2f(\infty)$. The solution of the differential equation is

$$\Phi_2(x_1,x_2,t) = \exp\left\{\frac{(jx_1)^2}{2}\left(\lambda\alpha_1 \int_{t_0}^{t} S_1(\tau)d\tau + \kappa a_1^2 \int_{t_0}^{t} S_1^2(\tau)d\tau\right) + \right.$$
$$+ \frac{(jx_2)^2}{2}\left(\lambda\alpha_2 \int_{t_0}^{t} S_2(\tau)d\tau + \kappa a_2^2 \int_{t_0}^{t} S_2^2(\tau)d\tau\right) + \qquad (13)$$
$$\left. + jx_1 jx_2 (\lambda+\kappa) a_1 a_2 \int_{t_0}^{t} S_1(\tau) S_2(\tau) d\tau\right\}.$$

We substitute (13) in (10), then following the reverse substitutions, we write the approximate asymptotic equality regarding $h(z, u_1, u_2, t)$, which coincides with (7).

The theorem is proved. ■

4. Characteristic function of the stationary probability distribution

The asymptotic characteristic function of the stationary probability distribution of the stochastic process $\{V_1(t), V_2(t)\}$ has the form:

$$h(u_1,u_2) \approx \exp\left\{ju_1 N\lambda a_1 b_1 + ju_2 N\lambda a_2 b_2 + \frac{(ju_1)^2}{2}(N\lambda\alpha_1 b_1 + N\kappa a_1^2\beta_1) + \right.$$
$$\left. + \frac{(ju_2)^2}{2}(N\lambda\alpha_2 b_2 + N\kappa a_2^2\beta_2) + ju_1 ju_2 N a_1 a_2 (\lambda+\kappa) b_{12}\right\},$$

where

$$b_i = \int_0^\infty (1 - B_i(\tau))d\tau, \quad \beta_i = \int_0^\infty (1 - B_i(\tau))^2 d\tau, \quad i = 1, 2,$$

$$b_{12} = \int_0^\infty (1 - B_1(\tau))(1 - B_2(\tau))d\tau.$$

Hence, the means vector and the covariance matrix of the resulting Gaussian approximation have the form:

$$\mathbf{a} = N\lambda \left[a_1 b_1 \quad a_2 b_2 \right],$$

$$\mathbf{K} = N \begin{bmatrix} \lambda\alpha_1 b_1 + \kappa a_1^2 \beta_1 & a_1 a_2 (\lambda + \kappa) b_{12} \\ a_1 a_2 (\lambda + \kappa) b_{12} & \lambda\alpha_2 b_2 + \kappa a_2^2 \beta_2 \end{bmatrix}.$$

5. Conclusion

The two-dimensional stochastic process, which describes the total amounts of occupied resources on the blocks of the system with parallel service, has been analysed. In more detail, the probability distribution of the process can be approximated by a two-dimensional Gaussian distribution and its approximation parameters were found.

6. Acknowledgment

The publication has been prepared with the support of the "RUDN University Program 5-100".

REFERENCES

1. K. Ageev, A. Garibyan, A. Golskaya, Yu. Gaidamaka, E. Sopin, K. Samouylov, L. M. Correia, Modelling of Virtual Radio Resources Slicing in 5G Networks, Communications in Computer and Information Science, 1109, 150–161.
2. E. Lisovskaya, M. Pagano, On the Application of Dynamic Screening Method to Resource Queueing System with Infinite Servers, Applied Probability and Stochastic Processes. Infosys Science Foundation Series, 2020, 179–198.
3. E. Lisovskaya, E. Pankratova, Yu. Gaidamaka, S. Moiseeva, M. Pagano, Heterogeneous Queueing System MAP/GI$^{(n)}/\infty$ with Random Customers' Capacities, Lecture Notes in Computer Science, 2019, 11965, 315–329.
4. V. Naumov, K. Samouylov, N. Yarkina, E. Sopin, S. Andreev and A. Samuylov, LTE performance analysis using queuing systems with finite resources and random requirements, 2015 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), Brno, 2015, 100–103.
5. A. Nazarov, S. Moiseeva, Asymptotic analysis method in queueing theory, NTL, Tomsk, Russia. 2006 (in Russian).

UDC: 519.213

Resource QS with the Requests Duplication at the Second Phase and Renewal Arrival Process

A. Galileyskaya[1], E. Lisovskaya[2], M. Pagano[3], S. Moiseeva[1]

[1] National Research Tomsk State University,
36 Lenina Ave., Tomsk, Russian Federation

[2] Peoples' Friendship University of Russia (RUDN University),
6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation

[3]Department of Information Engineering, University of Pisa,
Via Caruso 16, I-56122, Pisa, Italy

n.galileyskaya@bk.ru, lisovskaya-eyu@rudn.ru michele.pagano@iet.unipi.it,
smoiseeva@mail.ru

Abstract

In this paper, a resource queueing system with renewal arrival process, arbitrary service time distribution and requests duplication at the second phase is considered. In more detail, we apply the dynamic screening method to obtain the asymptotic expression for the stationary probability distribution of the total amount of occupied resource in the system. Finally, we verify the goodness of the obtained Gaussian approximation by means of discrete event simulation.

Keywords: *queueing system, arbitrary service time, copying of requirement, characteristic function, asymptotic analysis method*

1. Introduction

The standardization of 3GPP is moving forward rapidly with research into wideband waveforms, as well as with the adoption of the new 5G New Radio (NR) access to unlicensed spectrum (NR-U) [2]. One of the basic architectures for NR-U includes carrier aggregation with an NR license and a secondary carrier in an unlicensed spectrum, which in general allows traffic to be unloaded unhindered in scenarios where data transfer speeds of several gigabits are required. While today's NR-U research focuses on physical and protocol aspects, the performance of NR-U unloading mechanisms at the system level has not been thoroughly investigated [1]. This paper develops an analytical model in the form of a queuing system that takes into account the characteristics of the session dynamics in the millimeter wave

(mmWave) including random requirements to the spectrum [6] and the result obtained here will allow us to estimate the required cell capacity from the point of view of the probability of a session reset and the system resource utilization coefficient [3].

2. Mathematical model

Consider a tandem queueing system with an infinite number of servers and arbitrary service time. Customers arrive according to a renewal process, described by the distribution function $A(x)$ of the interarrival time. We will assume that each request is characterized by some random volume.

Each arriving customer immediately occupies the first free server on the first phase and requires resources. Service time distribution is $B_0(\tau)$ and volume distribution is given by probability function $G_0(y)$. When the service in the first phase ends, the customer is duplicated and served by the next two blocks on the second phase in parallel. In more detail, the application takes a random amount of a certain resource with the distribution function $G_1(\tau)$ in first block of the second phase and $G_2(\tau)$ in the second block of the second phase, respectively. The service times on the first and the second blocks of the second phases don't depend on each other and have arbitrary distribution functions $B_1(y)$ and $B_2(y)$. When the service is completed in the first and the second blocks, the customer leaves the system. Resource amounts and service times are mutually independent and do not depend on the epochs of customer arrivals. Figure 1 shows the structure of the system.

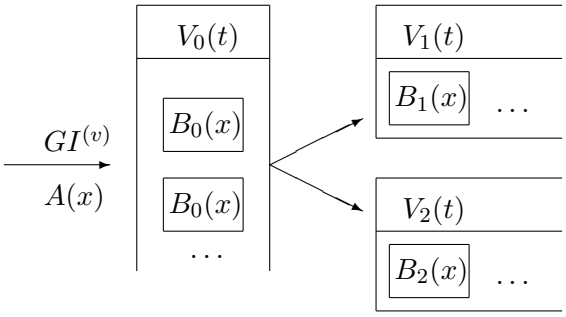


Fig. 1. Queueing system with the customers copying at the second phase and renewal arrival process

Denote by $V_i(t)$ the total volume of requirements in the i-th phase at time t, $(i = 0, 1, 2)$. Our goal is to derive the probabilistic characterization of the 3-dimensional process $\{V_0(t), V_1(t), V_2(t)\}$. This process, in general, is not Markovian and, therefore, we use the dynamic screening method for its investigation.

Consider four time axes that are presented in the Figure 2. Let axis GI shows the epochs of customers' arrivals, axis 0,1 and 2 will correspond to the first, second and third screened flow respectively.

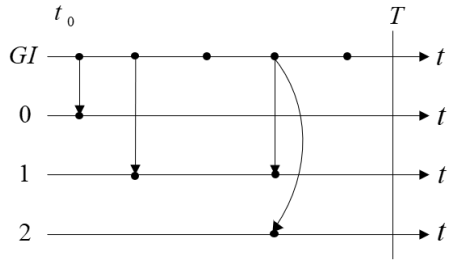


Fig. 2. Screening of the customers arrivals

We introduce the functions (dynamic probability) $S_0(t), S_1(t), S_2(t), S_{12}(t)$, the values of which lie in the range $[0, 1]$ and satisfy the property $S_0(t) + S_1(t) + S_2(t) + S_{12}(t) \leq 1$.

The arrival process event can be screened only on one of the axes 0, 1, 2, or on axes 1 and 2 simultaneously. Let the system be empty at moment t_0, and let us fix some arbitrary moment T in the future. $S_0(t)$ represents the probability that a customer arriving at the time t will be serviced in the $0-$unit by moment T. It is easy to show that $S_0(t) = 1 - B_0(T - t)$ for $t_0 \leq t \leq T$. The probability that the customer that arrived at time $t > t_0$ by time T will finish service in the $0-$ and $2-$ units, but not in the $1-$unit (i.e. will be screened in the axis 1) is equal to

$$S_1(t) = (B_2 * B_0)(T - t) - \int_0^{T-t} B_1(T - t - x)B_2(T - t - x)dB_0(x).$$

The probability that the customer that arrived at time $t > t_0$ by time T will finish service in the $0-$ and $1-$ units, but not in the $2-$unit (i.e. will be screened in the axis 2) is equal to

$$S_2(t) = (B_1 * B_0)(T - t) - \int_0^{T-t} B_1(T - t - x)B_2(T - t - x)dB_0(x).$$

Finally, the probability that a customer arriving in the system at time t, will finish service in the $0-$unit, but not in the $1-$ and $2-$ units (i.e. will be screened in the axes 1 and 2), equals

$$S_{12}(t) = B_0(T - t) - (B_1 * B_0)(T - t) - (B_2 * B_0)(T - t) +$$

$$+\int_0^{T-t} B_1(T-t-x)B_2(T-t-x)dB_0(x).$$

Denote by $W_i(t)$ the total amount of resources screened on axis $i = 0, 1, 2$. It is easy to prove that

$$P\{V_0(T) < x_0, V_1(T) < x_1, V_2(T) < x_2\} =$$
$$= P\{W_0(T) < x_0, W_1(T) < x_1, W_2(T) < x_2\}, \quad (1)$$

for $x_i > 0 (i = 0, 1, 2)$. We use equality (1) to investigate the process $\{V_0(t), V_1(t), V_2(t)\}$ via the analysis of the process $\{W_0(t), W_1(t), W_2(t)\}$.

3. Integro-Differential equations

Let us consider the four dimensional Markovian process $\{z(t), W_0(t), W_1(t), W_2(t)\}$, where $z(t)$ is the residual time from t to the next arrival. Denoting the probability distribution of this process by

$$P\{z(t) < z, W_0(t) < x_0, W_1(t) < x_1, W_2(t) < x_2\} = P(z, x_0, x_1, x_2, t).$$

and taking into account the formula of total probability, we can write the following system of Kolmogorov integro-differential equations

$$\frac{\partial P(z,x_0,x_1,x_2,t)}{\partial t} = \frac{\partial P(z,x_0,x_1,x_2,t)}{\partial z} + \frac{\partial P(0,x_0,x_1,x_2,t)}{\partial z}(A(z)-1) +$$
$$A(z)\left[S_0(t)\left(\int_0^{x_0}\frac{\partial P(0,x_0-y,x_1,x_2,t)}{\partial z}dG_0(y) - \frac{\partial P(0,x_0,x_1,x_2,t)}{\partial z}\right) +\right.$$
$$S_1(t)\left(\int_0^{x_1}\frac{\partial P(0,x_0,x_1-y,x_2,t)}{\partial z}dG_1(y) - \frac{\partial P(0,x_0,x_1,x_2,t)}{\partial z}\right) +$$
$$S_2(t)\left(\int_0^{x_2}\frac{\partial P(0,x_0,x_1,x_2-y,t)}{\partial z}dG_2(y) - \frac{\partial P(0,x_0,x_1,x_2,t)}{\partial z}\right) +$$
$$\left.S_{12}(t)\left(\int_0^{x_1}\int_0^{x_2}\frac{\partial P(0,x_0,x_1-y_1,x_2-y_2,t)}{\partial z}dG_2(y_2)dG_1(y_1) - \frac{\partial P(0,x_0,x_1,x_2,t)}{\partial z}\right)\right]$$

with the initial condition

$$P(z, x_0, x_1, x_2, t_0) = \begin{cases} R(z), x_0 = x_1 = x_2 = 0, \\ 0, \text{ otherwise}, \end{cases}$$

where $R(z)$ denotes the stationary probability distribution of the random variable, which is determined by equality

$$R(z) = \lambda \int_0^z (1 - A(x))dx, \quad \lambda = \left[\int_0^\infty (1 - A(z))dz \right]^{-1}.$$

We introduce the partial characteristic function

$$h(z, v_0, v_1, v_2, t) = \int_0^\infty e^{jv_0x_0} \int_0^\infty e^{jv_1x_1} \int_0^\infty e^{jv_2x_2} P(z, dx_0, dx_1, dx_2, t),$$

where $j = \sqrt{-1}$ is the imaginary unit. Then, we can write

$$\frac{\partial h(z, v_0, v_1, v_2, t)}{\partial t} = \frac{\partial h(z, v_0, v_1, v_2, t)}{\partial z} +$$

$$\frac{\partial h(0, v_0, v_1, v_2, t)}{\partial z} \{A(z) - 1 + A(z) [S_0(t) (G_0^*(v_0) - 1) + \quad (2)$$

$$S_1(t) (G_1^*(v_1) - 1) + S_2(t) (G_2^*(v_2) - 1) + S_{12}(t) (G_1^*(v_1)G_2^*(v_2) - 1)]\},$$

where

$$G^*(v) = \int_0^\infty e^{jvy} dG(y),$$

with the initial condition

$$h(z, v_0, v_1, v_2, t_0) = R(z). \quad (3)$$

4. Gaussian Approximation

A direct solution to equation (2) is impossible to find. Therefore, to solve problem (2) – (3), we use the method of asymptotic analysis under the condition of infinitely growing arrival rate. We write the distribution function of the lengths of the intervals between the moments of receipt of applications in the system in the form $A(Nz)$, where $N \to \infty$ is a parameter of high flow intensity [4, 5].

Then, the equation (2) takes the form

$$\frac{1}{N}\frac{\partial h(z,v_0,v_1,v_2,t)}{\partial t} = \frac{\partial h(z,v_0,v_1,v_2,t)}{\partial z} + \frac{\partial h(0,v_0,v_1,v_2,t)}{\partial z}\{A(z)-1+A(z)[S_0(t)(G_0^*(v_0)-1)+ S_1(t)(G_1^*(v_1)-1)+S_2(t)(G_2^*(v_2)-1)+S_{12}(t)(G_1^*(v_1)G_2^*(v_2)-1)]\}, \quad (4)$$

Theorem 1. The joint stationary probability distribution of the total resource amount in the system $GI^{(v)}/GI/\infty$ is asymptotically three dimensional Gaussian with mean:

$$\mathbf{a} = N\lambda\left[a_1^{(0)}b_0 \quad a_1^{(1)}b_1 \quad a_1^{(2)}b_2\right],$$

where $a_1^{(i)}, i=0,1,2$ are the means of resource requirements for a single customer and

$$b_0 = \int_0^\infty (1-B_0(\tau))d\tau, b_1 = \int_0^\infty (B_0(\tau)-(B_1*B_0)(\tau))d\tau, b_2 = \int_0^\infty (B_0(\tau)-(B_2*B_0)(\tau))d\tau,$$

and covariance matrix

$$\mathbf{K} = N\begin{bmatrix} \lambda a_2^{(0)}b_0+\kappa\left(a_1^{(0)}\right)^2\beta_0 & \kappa a_1^{(0)}a_1^{(1)}\beta_{01} & \kappa a_1^{(0)}a_1^{(2)}\beta_{02} \\ \kappa a_1^{(0)}a_1^{(1)}\beta_{01} & \lambda a_2^{(1)}b_1+\kappa\left(a_1^{(1)}\right)^2\beta_1 & \lambda a_1^{(1)}a_1^{(2)}b_{12}+\kappa a_1^{(1)}a_1^{(2)}\beta_{12} \\ \kappa a_1^{(0)}a_1^{(2)}\beta_{02} & \lambda a_1^{(1)}a_1^{(2)}b_{12}+\kappa a_1^{(1)}a_1^{(2)}\beta_{12} & \lambda a_2^{(2)}b_2+\kappa\left(a_1^{(2)}\right)^2\beta_2 \end{bmatrix},$$

where $a_2^{(i)}, i=0,1,2$ are the second raw moments of resource requirements for a single customer and

$$\beta_0 = \int_0^\infty (1-B_0(\tau))^2 d\tau, \beta_1 = \int_0^\infty (B_0(\tau)-(B_1*B_0)(\tau))^2 d\tau, \beta_2 = \int_0^\infty (B_0(\tau)-(B_2*B_0)(\tau))^2 d\tau,$$

$$\beta_{01} = \int_0^\infty (1-B_0(\tau))(B_0(\tau)-(B_1*B_0)(\tau))d\tau, \beta_{02} = \int_0^\infty (1-B_0(\tau))(B_0(\tau)-(B_2*B_0)(\tau))d\tau,$$

$$\beta_{12} = \int_0^\infty (B_0(\tau)-(B_1*B_0)(\tau))(B_0(\tau)-(B_2*B_0)(\tau))d\tau,$$

$$b_{12} = \int_0^\infty \left(B_0(\tau)-(B_1*B_0)(\tau)-(B_2*B_0)(\tau)+\int_0^\tau B_1(\tau-x)B_2(\tau-x)dB_0(x)\right)d\tau.$$

5. Numerical Example

We assume that the input renewal process is characterized by the following distribution function

$$A(Nz) = \begin{cases} 0, z < 0.5/N, \\ Nz - 0.5, z \in [0.5/N; 1.5/N], \\ 1, z > 1.5/N, \end{cases}$$

and service times have gamma distribution with parameters

$$\alpha_0 = \beta_0 = 0.5; \alpha_1 = \beta_1 = 1.5; \alpha_2 = \beta_2 = 2.5.$$

Moreover, the resources have a uniform distribution in the range: $[0; 3]$ for the first phase, $[0; 2]$ for the first block of the second phase and $[0; 1]$ for the second block of the second phase.

Table 1 presents the Kolmogorov distances between the asymptotic and empirical distributions of the total amount of resources occupied in three blocks. The approximation accuracy increases with incoming process intensity N, which is also illustrated by Figure 3.

N	1	5	10	50
Δ_0	0.303	0.036	0.021	0.009
Δ_1	0.312	0.038	0.021	0.009
Δ_2	0.303	0.035	0.019	0.008

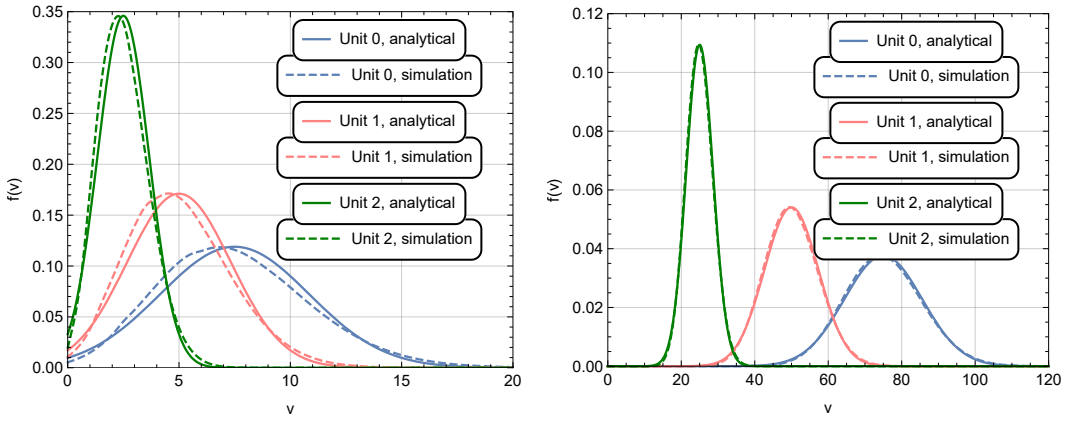
Table 1. Kolmogorov distances

6. Conclusion

In this paper, we presented the analysis of resource queueing system $GI^{(v)}/GI/\infty$ with renewal arrival process and arbitrary service time. In more detail, we applied the dynamic screening method to obtain the asymptotic expression for the joint stationary probability distribution of the total volume in the three blocks of the system and showed that it is three-dimensional Gaussian. Numerical experiments and simulations allow us to determine the applicability area of the asymptotic result for each phase of the system.

7. Acknowledgment

The publication has been prepared with the support of the “RUDN University Program 5-100”.



(a) $N = 5$ (b) $N = 50$

Fig. 3. Distribution of the total volume of the occupied resource for each phase of the system

REFERENCES

1. K. Ganesan, P. B. Mallick, J. Löhr, D. Karampatsis and A. Kunz, 5G V2X Architecture and Radio Aspects, 2019 IEEE Conference on Standards for Communications and Networking (CSCN), GRANADA, Spain, 2019, 1–6.
2. S. Lagen et al., New Radio Beam-Based Access to Unlicensed Spectrum: Design Challenges and Solutions, IEEE Communications Surveys and Tutorials, 2020, 22(1), 8–37.
3. X. Lu et al., Integrated Use of Licensed- and Unlicensed-Band mmWave Radio Technology in 5G and Beyond, IEEE Access, 2019, 7, 24376–24391.
4. A. Moiseev, A. Nazarov, Asymptotic analysis of the infinite-server queueing system with high-rate semi-markov arrivals, 2014 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2014, 507–513.
5. A. Moiseev, A. Nazarov, Queueing network with hight-rate arrivals, European Journals of Operational Research, 2016, 161–168.
6. V.A. Naumov, K.E. Samuilov, Analysis of Networks of the Resource Queuing Systems, Autom. Remote Control, 2018, 79, 822–829.

UDC: 123.456

Melanoma detection with deep neural networks

Eugene Yu. Shchetinin[1], Leonid A. Sevastianov[2,3], Edik A. Ayrjan[3,4], Anastasia V. Demidova[2]

[1]Financial University, Government of the Russian Federation, Moscow, Russian Federation
[2]Peoples' Friendship University of Russia (RUDN University), Moscow, Russian Federation
[3]Joint Institute for Nuclear Research, Dubna, Russian Federation
[4]Dubna State University, Dubna, Russian Federation

riviera-molto@mail.ru, sevastianov-la@rudn.ru, ayrjan@jinr.ru, demidova-av@rudn.ru

Abstract

In this paper, an approach to solving the problem of detecting skin malignancies, namely, melanoma, based on the analysis of dermoscopic images using the methods of deep learning. For this purpose, a deep convolutional neural network architecture was developed, which was applied to the processing of dermoscopic images of various skin lesions contained in the HAM10000 data set. The studied data was previously processed to eliminate noise, contamination, and change the size and format of images. In addition, since the disease classes are unbalanced, a number of transformations were performed to balance them. The data obtained in this way were divided into two classes: Melanoma and Benign. Computer experiments on the use of a built deep neural network on the data obtained in this way have shown that the proposed approach provides an accuracy of 91% on the test sample, which exceeds similar results obtained by other algorithms.

Keywords: melanoma, classification, deep learning, convolutional neural networks

1. Introduction

Melanoma is a deadly form of skin cancer that is often undiagnosed or misdiagnosed as a benign skin lesion. Its early and accurate detection is extremely important, because the lives of patients depend on it. In their practice, doctors are accustomed to rely on their professional experience and evaluate the injuries of each patient

The publication has been prepared with the support of the. RUDN University Program 5-100" and funded by Russian Foundation for Basic Research (RFBR) according to the research project No 19-01-00645.

based on a personal examination. However, such a system for detecting skin lesions is time-consuming, since it requires magnification and illumination of skin images to improve the clarity of pigment spots [1-3]. In addition, the manual Dermoscopy procedure is more prone to errors, requires many years of experience in complex situations, and a huge number of visual studies of similarities and differences between various skin lesions [4,5].

Clinical studies allow us to obtain an accuracy of the diagnosis of melanoma from 65 to 80%, which was a good result for some time [6,7]. However, modern research claims that the use of dermoscopic images in diagnosis significantly increases the accuracy of diagnosis of skin lesions. However, the visual differences between melanoma and benign skin lesions can be very small, making diagnosis difficult even for an expert doctor. Recent advances in the use of artificial intelligence methods in the analysis of medical images have allowed us to consider the development of intelligent medical diagnostics systems based on visualization as a very promising direction that will help the doctor in making more effective decisions about the health of patients and making a diagnosis at an early stage and in adverse conditions [8]. In this paper, we investigate an approach to solving the problem of classification of skin diseases, namely, detection of melanoma, based on deep learning methods. For this purpose, the architecture of a deep convolutional neural network was developed, which was applied to the processing of dermoscopic images of various skin lesions contained in the set of dermoscopic images HAM10000 [9]. The studied data was previously processed to eliminate noise, contamination, and change the size and format of images. In addition, since the disease classes are unbalanced, a number of transformations were performed to balance them. The data obtained in this way were divided into two classes: Melanoma and Benign. Computer experiments on the use of a built deep neural network on the data obtained in this way have shown that the proposed approach provides an accuracy of 91% on the test sample, which exceeds similar results obtained by other algorithms.

2. Review of modern achievements in the field of computer processing of dermatoscopic images

Most classical methods in the field of melanoma classification rely on manual selection of features such as the type of lesion (primary morphology), lesion configuration (secondary morphology), color, distribution, shape, texture, and uneven borders of the pigment spot [10] then, after extracting the main characteristics of images, machine learning methods such as the K-nearest neighbor (k-NN) algorithm, logistic regression, decision trees, and others are used to solve the classification problem [11]. Modern computer research on the diagnosis of skin diseases in order to detect melanoma actively implements deep learning methods and is aimed at improving

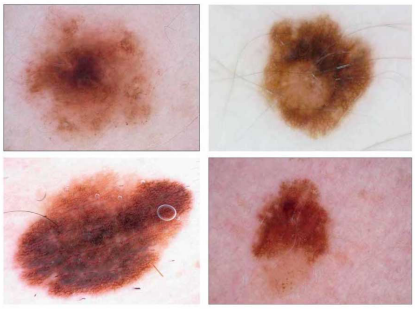
existing and developing new models of deep neural networks, primarily convolutional neural networks (CNN) [12-13]. Esfahani et al. [14] proposed a CNN architecture for the diagnosis of melanoma, where clinical images were pre-processed in such a way as to reduce the illumination of the image. Research results have shown that the proposed method is able to diagnose cases of melanoma in 70% of cases. Mahbod et al. showed that convolutional neural networks are superior to traditional machine learning methods [16]. The authors proposed a hybrid automated computerized method for classifying skin diseases using three pre-trained deep networks (AlexNet, VGG16, ResNet-18) to extract the features. The features extracted in this way are then used to train the support vector machine on 150 images from the ISIC 2017 dataset. Chelebi et al. [17] proposed an ensemble of threshold methods for determining the boundaries of skin lesions.

Heckler et al. [18] applied deep learning methods to classify histopathological diagnosis of melanoma and compared the result with qualified histopathologists. Esteva et al. [21] implemented a pre-trained INCEPTIONV3 network to classify nine classes, where they used a labeled set of dermatological data that has 3374 dermoscopic images, 129,450 clinical images, and reaches an accuracy of 72%. Harangi and co-authors [22] used the ensemble method DCNN (deep convolutional neural network), where they combined the result of four different architectures, improving the accuracy of melanoma classification on the 2017 ISBI dataset. XI and Bovik [23] proposed a method for segmentation of skin lesions in which the CNN model is combined with a genetic algorithm. In [24], a method for segmentation of skin lesions on dermoscopic images from the ISIC 2017 data set and their classification of various types of skin cancer using deep neural network models Mask R-CNN and U-net is proposed. The proposed method consists of preprocessing and segmentation using a hybrid learning algorithm. The goal of the first stage is to remove noise using the filtering method. In the second stage, images are segmented based on the clustering method. In [25], it was proposed to use the deep convolutional network ResNet50 for recognition of melanoma.

3. Data description and their pre-processing

In clinical dermatology, there are relatively few data sets with digital images of skin lesions. Most of these sets are too small and/or not publicly available, which creates an additional barrier to research in this area. Examples of such dermatology-related image datasets are: the Dermofit image Library [26] - a dataset containing 10 different classes, including 1,300 high-quality images of skin lesions collected worldwide. Dermnet [27] - the website-enabled Atlas of skin diseases contains more than 23,000 skin images divided into 23 classes. In early 2016, the international biomedical imaging Symposium (ISBI) published a set of data for analyzing skin

Table 1. Image examples from ISIC archive

Melanoma	Benign
	

lesions for early detection of melanoma [28]. In order to support the training of clinical dermatologists and the development of new information technologies, the International society for skin imaging (ISIC) has developed an international repository of dermoscopic images, known as the ISIC archive. This data set contains pigmented skin lesions obtained using standard Dermoscopy. Every year, ISIC adds new images to its archive and promotes the task of implementing computer methods for detecting melanoma and other skin diseases. For example, the HAM10000 data set created within this organization served as data for the ISIC-2018 Challenge [29]. In 2019, the number of samples already numbered more than 25,000 images for Dermoscopy, available for training in 7 different categories. in table.1 examples of images from the studied data are given. The lesion classes in the HAM10000 dataset are listed below.

1. nv: Melanocytic nevi-benign neoplasms of melanocytes [6705 images];
2. Mel: Melanoma-malignancy [1113 images];
3. bkl: Benign keratosis - a common class that includes seborrheic keratosis, solar lentigo, and lichen-squamous as keratosis [1099 images];
4. bcc: Basal cell carcinoma is a common variant of epithelial skin cancer that rarely metastasizes, but grows if untreated [514 images];
5. akiec: Actinic keratosis and intraepithelial carcinoma are common non-invasive variants of squamous cell carcinoma [327 images];
6. Vasc: Vascular lesions of the skin of cherry angiomas to angiokeratoma and pyogenic granulomas [142 images];
7. Df: Dermatofibroma-a benign skin lesion [115 images].

4. Building a deep neural network model and computer experiments

In this paper, a model of the CNN deep convolutional neural network was constructed to analyze dermoscopic images and detect melanoma in them [26]. The CNN model is initialized as a sequence of layers using the Sequential class. Next, a convolutional Conv2D layer was added, with the input parameters of the feature map $inputshape = (32, 32, 3)$, where 32 is the size of the spatial features of the input map, and 3 is the number of color channels (in this case, the color of the image in RGB format). Convolution is defined by the following parameters: the size of templates extracted from input data is (3x3); the depth of the output feature map is the number of filters calculated by the convolution. In this model, the first convolutional layer outputs a feature map of size (30,30,32) and calculates 32 filters based on input data. Each of these 32 output channels contains a 30x30 value grid-a map of filter responses to input data that defines the response of this filter template for different sections of input data. The last parameter of the convolutional layer is the activation function that we use to activate neurons in the neural network, specifically 'relu'. In the third step, the pooling layer (MaxPooling2D) with the map (3x3) was used. The main purpose of using this layer is to reduce the number of coefficients in the feature map for processing, as well as to implement spatial filter hierarchies by creating successive convolution layers for viewing larger and larger Windows. Then create a vector for the fully connected flatten () layer. The Flatten layer serves as a link between the data received by the network and the output vector, converting the multidimensional output of the preceding layer into a one-dimensional (vector). In the last step, a fully connected layer is built – the density layer. The density function has 2 parameters - the number of nodes for the output layer(128) and the 'relu' activation function. The output layer has 1 node where the 'sigmoid ' activation function is used. Next, you need to compile the model and optimize the weight coefficients and loss function for evaluating the model. The Adam algorithm is selected as the optimizer for our model. The loss function is *binarycrossentropy*, since we have two classes (1 or 0: Benign (Benign) or Malignant (Melanoma) tumors).

The HAM10000 data set used in this work contains 10013 images, pre-divided into 8000 training and 2000 test images. Computer experiments in research on the identification of melanoma was conducted in two stages. In the first stage, all images were placed as containing benign Benign and malignant Melanoma skin lesions. Then, training and test samples were created for each class. Due to the unbalanced classes of the studied data set, the test and training data sets were reduced by reducing the number of images in separate classes, as well as using special algorithms [27].

The quality of the constructed neural network model was evaluated using the basic accuracy, sensitivity, and Precision and Recall metrics. Table 1 shows a

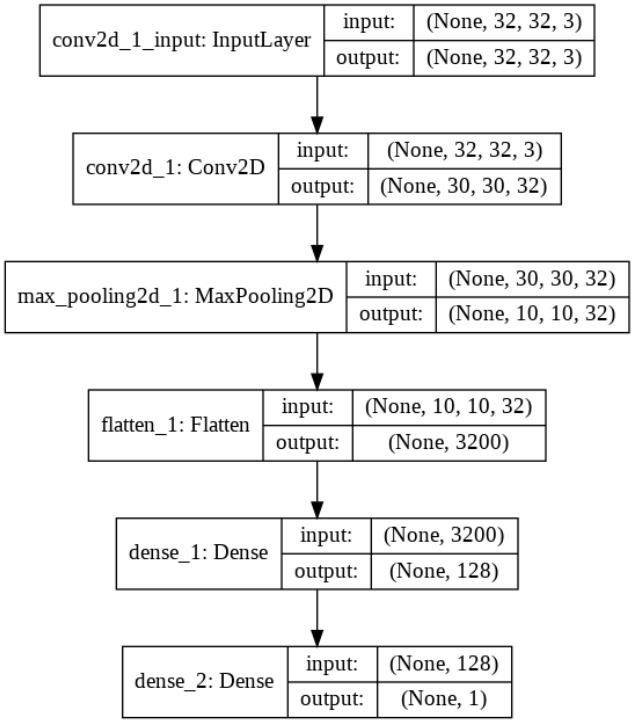


Fig. 1. The architecture of the deep convolutional network

matrix of inaccuracies (confusion matrix) obtained as a result of the neural network operation on a test sample. From the Table 1 we get accuracy=91.4%, $Precision = TP/(TP + FP) = 92\%$, $Recall = TP/(TP + FN) = 0.901\%$. All experiments were performed using a computer equipped with a core i5 processor, 8 GB SDRAM, and an NVIDIA GeForce 920M graphics card. Karas and TensorFlow were used to develop a neural network model and train it [28].

Table 2. Confusion Matrix for proposed DCNN

Matrix		Actual class	
		Melanoma	Non-Melanoma
Prediction Class	Melanoma	69	6
	Non-Melanoma	7	68

5. Conclusion

The paper deals with the problem of classification of skin diseases, primarily melanoma, and its recognition among other diagnosed skin lesions. To solve this problem, a deep convolutional neural network was implemented, which was then applied to solve the problem using the example of a set of HAM10000 digital images provided by ISIC (International Skin Imaging Collaboration). The proposed method includes various methods of pre - processing images in order to obtain a more informative and balanced training and control samples. The model accuracy in the validation sample was 91%. Further research will focus on improving the CNN architecture to improve accuracy, get more images to train our model, and apply other deep neural network architectures. We will also make efforts to make this model available for use as an application for mobile devices in telemedicine systems.

REFERENCES

1. Siegel R. L., Miller K. D., Jemal A. Cancer statistics 2019. CA: a cancer journal for clinicians, 2019, 69(1). P. 7-34.
2. American Cancer Society, Cancer Facts and Figures 2019.
3. R. M.- Cancer and undefined. An overview of skin cancers, Wiley Online Libr. 1995.
4. Rogers, H. W., Weinstock, M. A., Feldman, S. R., Coldiron, B. M. (2015). Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the US population, 2012. JAMA dermatology, 151(10), 1081-1086.
5. Gandhi, S. A., Kampp, J. (2015). Skin cancer epidemiology, detection, and manage-ment. Medical Clinics, 99(6), 1323-1335.
6. Kittler, H., Pehamberger, H., Wolff, K., Binder, M. (2002). Diagnostic accuracy of dermoscopy. The lancet oncology, 3(3), 159-165.
7. S. Dreiseitl, L. Ohno-Machado, H. Kittler, S. Vinterbo, H. Billhardt, and M. Binder, A comparison of machine learning methods for the diagnosis of pigmented skin lesions, Journal of biomedical informatics, vol. 34, no. 1, pp. 28–36, 2001.
8. https://biomedicalimaging.org/2016/.
9. https://www.isic-archive.com/.
10. Salerni, G., Teran, T., Puig, S., Malvehy, J., Zalaudek, I., Argenziano, G. (2013). Meta-analysis of digital dermoscopy follow-up of melanocytic skin lesions: a study on behalf of the International Dermoscopy Society. Journal of the European Academy of Dermatology and Venereology, 27(7), 805-814.
11. Esteva, A., Kuprel, B., Thrun, S. (2015). Deep networks for early stage skin disease and skin cancer classification. Project Report, Stanford University.

12. Harangi, B. (2018). Skin lesion classification with ensembles of deep convolutional neural networks. Journal of biomedical informatics, 86, 25-32.
13. Yuan, Y., Chao, M., Lo, Y. C. (2017). Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance. IEEE transactions on medical imaging, 36(9), 1876-1886.
14. Nasr-Esfahani E., Samavi S., Karimi N., Soroushmehr S. M. R.(2016). Melanoma detection by analysis of clinical images using convolu-tional neural network. In 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) pp. 1373-1376.
15. Giotis, I., Molders, N., Land, S., Biehl, M., Jonkman, M. F., Petkov, N. MED-NODE: a computer-assisted melanoma diagnosis system using non-dermoscopic images. Expert systems with applications, 42(19), 6578-6585, 2015.
16. Mahbod, A., Schaefer, G., Wang, C., Ecker, R., Ellinge, I. (2019). Skin lesion classifi-cation using hybrid deep neural networks. In ICASSP 2019-2019 IEEE International Confer-ence on Acoustics, Speech and Signal Processing (ICASSP) pp. 1229-1233.
17. M. Emre Celebi, Q. Wen, S. Hwang, H. Iyatomi, G. Schaefer, Lesion Border Detection in Dermoscopy Images Using Ensembles of Thresholding Methods, Ski. Res. Technol., vol. 19, no. 1, pp. 252–258, Feb. 2013.
18. Hekler, A., Utikal, J. S., Enk, A. H., Berking, C., Klode, J., Schadendorf, D., (2019). Pathologist-level classification of histopathological melanoma images with deep neural net-works. European Journal of Cancer, 115, 79-83.
19. Esteva A., Kuprel B., Novoa R. A., Ko J., Swetter S. M., Blau H. M. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115.
20. Harangi, B. (2018). Skin lesion classification with ensembles of deep convolutional neural networks. Journal of biomedical informatics, 86, 25-32.
21. Xie F., Bovik A. C. (2013). Automatic segmentation of dermoscopy images using self-generating neural networks seeded by genetic algorithm. Pattern Recognition, 46(3), 1012-1019.
22. Codella N. C., Nguyen Q. B., Pankanti S., Gutman D. A., Helba, B. (2017). Deep learning ensembles for melanoma recognition in dermoscopy im-ages. IBM Journal of Research and Development, 61(4/5), 5-1.
23. Yu L., Chen H., Dou Q., Qin J. (2016). Automated melanoma recog-nition in dermoscopy images via very deep residual networks. IEEE transactions on medical imaging, 36(4), 994-1004.
24. Dermofit image library, https://licensing.eri.ed.ac.uk/i/software/dermofit-image library.html.
25. Dermnet - skin disease atlas, http://www.dermnet.com/

26. Shollet F. Deep Learning with Python, Manning Publications Co. 2017.
27. L. A. Sevastianov, E. Yu. Shchetinin, On methods for imroving the accuracy of multiclass classification on imbalanced data, Informatics and Applications, 2020, Volume 14, Issue 1, pp. 63-70.
28. https://keras.io.

UDC: 519.7

Paralinguistic model for emotions recognition with deep neural networks

Eugene Yu. Shchetinin[1], Leonid A. Sevastianov[2,3], Dmitry S. Kulyabov[2,3], Edik A. Ayrjan[3,4], Anastasia V. Demidova[2]

[1]Financial University, Government of the Russian Federation, Moscow, Russian Federation
[2]Peoples' Friendship University of Russia (RUDN University), Moscow, Russian Federation
[3]Joint Institute for Nuclear Research, Dubna, Russian Federation
[4]Dubna State University, Dubna, Russian Federation

riviera-molto@mail.ru, sevastianov-la@rudn.ru, kulyabov-ds@rudn.ru, ayrjan@jinr.ru, demidova-av@rudn.ru

Abstract

In this paper the computer paralinguistic model for emotions recognition based on deep neural networks is proposed. The main stages of its construction were studied and relevant models of the deep convolutional networks and recurrent networks with LSTM memory cell were used. Intensive computer experiments on the emotions recognition from human speech with proposed model were conducted. As the data for research and testing of our model RAVDESS dataset of audio recordings was selected. The results showed a high efficiency of the explored model, and the accuracy estimates for some classes of emotions were reached 90%.

Keywords: emotions recognition, paralinguistic model, convolutional network, ResNet18, recurrent network, BLSTM model, RAVDESS

1. Introduction

Paralinguistics is a field of linguistics that studies various nonverbal aspects of speech, such as emotions, intonation, pronunciation, and other characteristics of the human voice [1]. Computer paralinguistics is one of the most relevant and dynamically developing areas of modern speech technologies, and the recognition of emotions in human speech is the most popular part of them [2, 3]. Computer classification of emotions sets the task of extracting some features from emotional

The publication has been prepared with the support of the. RUDN University Program 5-100" and funded by Russian Foundation for Basic Research (RFBR) according to the research project No 19-01-00645.

speech of a person based on audio recordings, video recordings of people who uttered this statement, and other modalities. Various paralinguistic models are used to evaluate the physical parameters of the voice, such as pitch, intensity, formants, and harmonics, to determine emotions. Such classifiers are used in the development of emotional intelligence systems, security systems, biometric research, telemedicine, mobile assistants, and others.

The complexity of this problem needs to determine such features that are sufficiently resistant to voice anomaly and noise, while maintaining all the main characteristics and features of the voice. Also, the model used must take into account the dynamics of features over time for effective analysis of changes in the voice. Most often, the method of feature extraction based on a sliding window is used to solve these problems. This method solves the problem of data normalization and prevents the model from overfitting.

We developed the paralinguistic emotion classification model in next stages: collecting information to form training and test samples, selecting features from the information that the model will be trained on, selecting the model and its architecture, configuring hyperparameters, training the model, and validating the model against new data.

The most common methods for modeling and classifying emotions are the mixtures of Gaussian distributions (GMM), Hidden Markov Models (HMM), Support Vector Machines (SVM) and artificial neural networks. With the advent of deep learning methods and the creation of deep neural networks (DNN), research in the field of computer analysis of emotions has acquired a qualitatively new direction of development. In this paper, we propose a computer paralinguistic model of emotion recognition based on the ensemble of the convolutional neural network of the ResNet18 architecture and the bidirectional recurrent neural network with an LSTM memory cell. On the basis of RAVDESS audio recordings, computer experiments on the classification of emotions using the proposed model and a comparative analysis of the results obtained with other models of neural networks, as well as the most effective machine learning algorithms, were conducted. The analysis of the results showed the advantages of the developed paralinguistic model in solving the problem, as well as the use of deep learning methods.

2. The development of the computer paralinguistic model of the emotions recognition

The following stages of building the computer paralinguistic model of the human emotions are suggested:

Database selection. The database combines sets of audio recordings intended for training and testing the model, sets of tags depending on the task being solved, as well as accompanying documentation.

Preliminary data processing. The purpose of preprocessing is to eliminate as much as possible the influence of external factors on the audio recording – recording quality, external noise, differences in the sensitivity of recording equipment, and so on. Typical types of preprocessing are filtering noise by frequency, cropping the audio recording by purity, reverberation, and normalization by audio recording by volume.

Allocation of low-level descriptors (LLD). At this stage, you can directly select features from the audio recording. This happens using a sliding window algorithm, usually 10-30 milliseconds wide. Window functions of various types are used: rectangular ones for selecting features based on the time distribution of the signal, and smooth ones for selecting spectral and frequency features. The main types of acoustic LLD include: intonation (tone, frequency, etc.), intensity (energy, Teagerfunction), linear kepstralny coefficients (LPCC), mel-features – mel-spectrograms and Mel-kepstralny coefficients (MFCCs), formants – amplitudes and so on, harmonic signs (the ratio of harmonics to noise, noise to harmonics, and so on [5].

Hierarchical feature selection. At this stage, attributes are selected from existing low-level descriptors. The purpose of this step is to reduce the data dimension and convert feature vectors of potentially unknown length to a scalar value. Thus, the analysis moves to the area of "super-segmentation", which experimentally gives a higher accuracy of models in the problems of paralinguistics.

Dimension compression. At this stage, the feature space is transformed so as to reduce covariances outside the main diagonal of the covariance matrix, often this is achieved by shifting and rotating the source space. Dimension compression techniques include principal component analysis (PCA) and linear discriminant analysis (LDA).

Feature selection. After reducing the dimension of the feature space by PCA or LDA methods, it is transformed by eliminating unnecessary features. For the selection of features, the criterion is needed – often criteria based on entropy and information growth are used for this purpose – the Gini criterion, the Shannon entropy, the Akaike information criterion, and so on. You can also use the value of the error functional of the trained model.

The choice of the model parameters. At this stage, the parameters of the trained paralinguistic model are fine-tuned. For neural networks, this process will consist in choosing the correct network topology – the number and organization of layers, the presence of network normalization mechanisms-Dropout, and so on, the number of neurons in hidden layers, the network initialization option, choosing an optimization

algorithm and setting it – the learning rate, choosing the value of annealing, and so on.

Model training. At this stage, the model is trained to solve the problem. The model uses the labels of the training set in order to find dependencies in the data and to learn how to extrapolate them to a new data.

3. Data description and pre-processing

3.1. Data description.

In this paper the RAVDESS data set (Ryerson Audio-Visual Database of Emotional Speech and Song) was used for research and development of the emotion recognition model. It was prepared by the Department of psychology and the Department of computer science and information technology at Ryerson University in Toronto, Canada, specifically for various paralinguistic studies [6]. The data set consists of video and audio recordings made by 24 professional actors from Toronto, Canada, whose task was to express a particular emotion on the record. The RAVDESS database consists of 7356 records. A total of 104 different recordings were made by each actor, which were later divided into three types – audio recording only, video recording only, and video recording with a sound track present. Thus, there are 312 different files for each of the actors.

3.2. Features selection and engeeniring.

To train machine learning algorithms and deep neural networks to recognize emotions, the audio recordings l must be pre-processed in such a way as to extract the main characteristic features of certain emotions. Let's look at the main ones:

Tone, volume, and frequency are the main characteristics of an audio signal. Tone refers to the pitch of the sound inside the window, and directly depends on the frequency of the signal. Sound volume is defined as the sound pressure level. The frequency of the signal expresses the number of vibrations of the sound wave per second.

Zero-crossing rate (ZR). ZR is a measure that expresses the number of times the audio signal graph crosses zero within a given window. ZR, as a feature, allows you to classify different types of content on audio recordings well – the values of this indicator for human conversation and, for example, for music differ significantly. However, this feature is subject to a strong influence of noise on audio recordings, so its use requires preliminary data cleaning.

Linear predictive coefficients (LPC). LPC is a method that allows to predict the value of the next window based on the value of previous windows.

The following characteristics are calculated based on the spectrogram of the sound wave. The spectrogram shows the dependence of the signal power density on

time, and allows you to evaluate the signal in the context of different frequencies. The original signal is decomposed into a spectrum using the Fourier transform.

Spectral centroid. This characteristic of the spectrum shows where its center of mass is, and another interpretation is the median of the spectrum values.

Spectral contrast. Spectral contrast shows the differences between spectral peaks and spectral troughs at a particular time, in the context of each frequency of the spectrum. It is a good measure of the range of the spectrum at a given time.

Spectral flatness. This characteristic is also called the tonality coefficient or Wiener entropy. This indicator characterizes the content of pure tone in the audio signal, or, if the values are too high, the noise content.

In addition, chromatograms are widely used. Chromagrams show the distribution of the signal by notes within the height class. Chromagrams are often used in tasks of classifying an audio signal by musical genres, but they can also be used in tasks of determining gender and age from an audio recording.

Mel-coefficients. Mels are the units of the measurement of audibility of a sound signal, calculated from the physiological features of the structure of the human ear. The mel-spectrogram is a spectrogram obtained by decomposing the signal using the Fourier transform, translated into mel. On this basis it is possible to calculate Mel-cepstral coefficients. Kepstr are the values obtained from the logarithm of the original spectrogram by the inverse Fourier transform. Mel-cepstral coefficients express the value of sound energy falling on each kepstr. A total of 24 Mel-kepstral coefficients are calculated, and usually the first 13 are sufficient for speech recognition or classification tasks [7].

4. Basic models of deep neural networks used in emotion recognition

Recurrent neural networks (RNN) are the group of deep neural network models used in sequence processing. This allowed to determine flexible long-term dependencies in the data, which is especially important in the context of analyzing human speech. To do this, the RNN computational graph contains loops that reflect the influence of previous information from the event sequence on the current information. However, it was found that despite the ability to model long-term dependencies, in practice, models of recurrent neural networks do not implement the requirements and suffer from problems with gradient descent [8]. To preserve the context for long periods of time and solve the problem of gradient attenuation, a special neural network architecture called “long Short-Term Memory” (LSTM) was developed[9].

An LSTM module is a memory cell that has multiple inputs and outputs that allow us to add or remove information about the state of the cell. Adding or removing information is controlled by the gates. To control the state of a cell, the LSTM

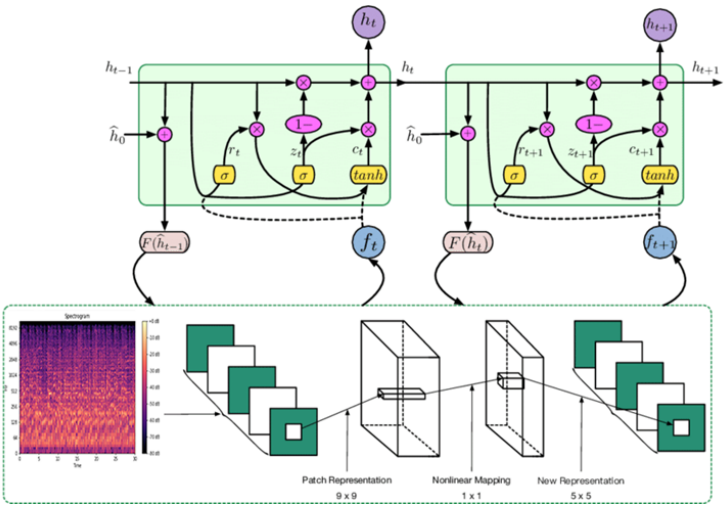


Fig. 1. Ensemble model architechture

contains three such gates. These are sigmoid layers (rectangles inside the RNN cell) that output numbers between 0 and 1, describing how much information should be skipped. A zero value means that we don't skip anything, while a one value means that we skip all the information.

In this form our neural network model only stores past information, since it processes the sequence in only one direction. To eliminate this disadvantage, a model of a bidirectional recurrent neural network with an LSTM memory cell was proposed [10]. Bidirectional LSTM networks work in both directions, combining the output of two hidden LSTM layers that transmit information in opposite directions — one in the course of time, the other against it, and thus simultaneously receiving data from past and future states. In this paper, we propose a paralinguistic model of emotion recognition using the BLSTM architecture and a deep convolutional network as an ensemble. See her graph on Fig.1.

5. Computer experiments for emotion recognition algorithms training

For computer experiments, only the audio part of the RAVDESS set was taken, containing 1440 3-second audio recordings made by 24 actors. Audio recordings are equally divided into 8 classes according to the emotions expressed in them: neutral, calm, upset, joy, irritation, fear, disgust and surprise. Each emotion was recorded with two types of intensity – medium and high, and two takes were performed for each recording. The Librosa module for the Python programming language was used

to highlight features. The original audio recordings were normalized in volume and cleared from noise that went beyond the amplitude range from 300 to 3400 Hz. Then, using a fast Fourier transform with window width settings of 93 milliseconds, and window overlap of 46.5 milliseconds, the audio recordings were decomposed into the frequency spectrum. The following features were identified from this spectrum: Mel Cepstral-coefficients 1-24, Delta Cepstral Mel-coefficient second-order, Delta Mel Cepstral coefficient, Mel Cepstral coefficient mean Mel Cepstral-coefficient standard deviation, Chromagram, Chromagram is normalized energy, the tone characteristics of the centroid, spectral contrast, zero crossing rate, spectral centroid, spectral bandwidth, spectral flatness, spectral rolloff, RMS.

During the experiment, were trained on 8 different models: logistic regression, support vector machine (SVM), decision tree (DT), random forest (RF), gradient boosting XGBoost, convolutional neural network CNN (ResNet18), recurrent neural network RNN (BLSTM), ensemble of convolutional and recurrent networks Stacked CNN-RNN. The results of computations showed, that the models based on neural network algorithms have much higher accuracy of emotion recognition and classification than linear algorithms or algorithms based on decision trees and XGB. Also, the BLSTM network showed slightly higher accuracy 78%, which is possible due to the use of long-term memory modules in this architecture, which allow two-way work with the context of the processed information. The accuracy of networks ensemble was found about 81%.

Also, for this database, computer experiments were conducted to classify the gender of the actor, and, in addition, the positivity (negativity) of the expressed emotions. In these cases, the classification accuracy was 91.4% and 93.7%, respectively. It is obvious that the reduction of emotion classes or their binarization leads to the expected significant increase in the accuracy of classification.

6. Results discussion and conclusion

In this paper the studies of RAVDES data set containing human emotional speech have been conducted, and the models of deep neural networks for classifying emotions have been proposed. The article presents the paralinguistic model based on ensemble of convolutional network ResNet18 and BLSTM neural network for classifying human emotions by voice. A comparative analysis of the results of using various models of neural networks and machine learning algorithms has shown the advantage of the architecture of recurrent neural networks.

Based on the results of the research, the following conclusions can be drawn. In fact, the results obtained in the work can be assessed as good, given that only audio recordings were used. Obviously, speech alone is not enough to accurately classify emotions, but you also need to use video recordings, facial expressions, gestures,

and other additional data sources to improve the quality of recognition. In many ways, the success of the algorithm depends on the quality of the training database. It should be representative of all types of emotions delivered by the experts, and preferably in equal proportions [11]. For this purpose, it is necessary to expand existing databases by creating new records, for example, using generative neural networks [12]. Finally, the correctness of the data markup plays an important role, since the ambiguity of the markup also reduces the efficiency of the algorithm. A large number of emotion classes drastically reduces the quality of recognition. For example, it turned out that the best results are obtained by algorithms that allocate only 2 emotions.

The architecture of an ensemble of convolutional and recurrent neural networks allows to combine the advantages of both types of networks and achieve even greater accuracy than when applying these types of networks separately. Further analysis of the research conducted on various models of deep neural networks showed that all subsequent variations in their parameters did not lead to a significant increase in recognition accuracy. In our opinion, this may indicate the need to develop new models and further develop the architecture of deep neural networks.

REFERENCES

1. Rabiner L., Juang B. Fundamental of Speech Recognition. Englewood Cliffs: Prentice-Hall N.J.,1993.
2. Schuller B. The Computational Paralinguistics Challenge, IEEE Signal Processing Magazine, 2012, Vol. 29, 4, P. 1264-1281.
3. Schuller B., Batliner A. Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing, Wiley, 2013.
4. Steidl S. Automatic classification of emotion-related user states in spontaneous children's speech, Logos Verlag, Berlin, 2009.
5. Batliner A., Schuller B. Computational Paralinguistics. Emotion, Affect and Personality in Speech and Language Processing. John Wiley and Sons Limited. 2015.
6. Livingstone S.R., Russo F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE, 2018 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391
7. Hasan R., Jamil M., Rabbani G., Rahman S. Speaker identification using mel frequency cepstral coefficients / 3rd International Conference on Electrical and Computer Engineering, 2004. P. 28–30.

8. Hochreiter S., Bengio Y., Frasconi P. Gradient flow in recurrent nets: the difficulty of learning long term dependencies, in: A Field Guide to Dynamical Recurrent Neural Networks, Kremer and Kolen, Eds. IEEE Press, 2001.
9. Hochreiter S., Schmidhuber J. Long Short-Term Memory, Neural Computation, 9(8), P.1735-1780, 1997.
10. Schuster M., Paliwal Kuldip K. Bidirectional Recurrent Neural Networks, IEEE transactions on signal processing, v. 45, 11, 1997.
11. L. A. Sevastianov, E. Yu. Shchetinin, On methods for imroving the accuracy of multiclass classification on imbalanced data, Informatics and Applications, 2020, Volume 14, Issue 1, pp. 63-70.
12. Yafeng Niu, Dongsheng Zou A breakthrough in speech emotion recognition deep retinal convolutional neural networks, https://arxiv.org/abs/1707.09917.

UDC: 519.245

A simulation approach to reliability assessment of a redundant system with arbitrary distributions of uptime and repair time of its elements

H.G.K. Houankpo[1], D.V. Kozyrev[1,2], E. Nibasumba[1], Mouale M.N.B.[1], I.A. Sergeeva[1]

[1]Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation

[2]V.A.Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, 65 Profsoyuznaya street, Moscow, 117997, Russia

gibsonhouankpo@yahoo.fr, kozyrev-dv@rudn.ru, ema.patiri2015@yandex.ru, bmouale@mail.ru, sergeevair@mail.ru

Abstract

With the rapid development and spread of computer networks and information technology, researchers face new sophisticated and complex challenges of both applied and theoretical nature in investigating the reliability and availability of networks and data transmission systems.

In current paper, we study the system-level reliability of a multiple redundant system using the simulation approach. Also, we obtain the values of the relative repair rate at which the desired level of reliability is achieved, present plots of the system uptime probability and plots of the empirical distribution function and the empirical reliability function. Software implementation of the developed simulation algorithms was carried out on the basis of the R language.

Keywords: Simulation, reliability assessment, redundant systems, relative repair rate, probability of failure-free system operation, sensitivity.

Notations

a1 – average time to failure of an element,
b1 – average repair time of a failed element,
N – number of elements in the system,
T – maximum model run time,
NG – number of trajectories plots,
"GI" – general independent distribution.

1. Introduction

Currently, simulation is effectively used for the tasks of modeling network information systems, developing mathematical methods, information technologies, including

the development of new calculation models, for analyzing the functioning of computer networks, teletraffic modeling, etc.

Previously in [1], it was shown that explicit analytical expressions for the stationary distribution of the considered system are not always obtainable. The simulation model developed in this work allowed to investigate the reliability of the system, defined as the stationary probability of failure-free operation of the system, as well as to calculate the reliability characteristics of the system; also numerical research and analysis charting shown that this dependence becomes vanishingly small under a "fast" recovery, that is, with the growth of the relative repair rate ρ.

Recently, the functioning of various aspects of modern society has become critically dependent on communication networks [2, 3]. With the migration of critical communications tools, it has become vital to ensure the reliability and accessibility of data networks and systems. A number of previous studies [4–8] have focused on analyzing the reliability of various complex telecommunications systems. In particular, a study was conducted on the reliability of cold-standby data transmission systems. Paper [9] focused on reliability analysis of a combined power plant running on a gas turbine engine. In a series of works by Enrico Zio et al. [10–12] the Monte Carlo simulation method was applied to reliability assessment and risk analysis of multi-state physics systems. The aim of [13] was to develop a model for studying system reliability and analyzing the sensitivity of system availability. In [14], a simulation method was considered to simulate the reliability of a task by a complex system by modeling a task cyclogram, modeling a run-time profile and a method of dynamic reliability modeling. In [15], modeling and estimation methods were presented that allow temperature optimization of the reliability of a multiprocessor system on a chip for specific applications.

The current paper summarizes the results of previous studies of the authors in the case of cold standby of the system $\langle GI_N/GI/1\rangle$ with an arbitrary distribution function (DF) of uptime and an arbitrary DF of repair time of its elements. The aim of the work is to conduct simulation to find the value of the coefficient ρ (relative repair rate), at which a given level of reliability is achieved and to graph the dependence of the probability of failure-free operation of the system on the relative repair rate. The results of calculating the reliability estimate for different input distributions are presented.

2. Problem Statement and Model Description

As a simulation model of a redundant data transmission system consisting of N different types of data transmission channels, we consider a repairable multiple cold standby system $\langle GI_N/GI/1\rangle$ with one repair device, with an arbitrary distribution function (DF) of uptime and an arbitrary DF of repair time of its elements.

In this paper, we consider the dependence of the probability of failure-free operation of the system $\langle GI_N/GI/1\rangle$ on the relative repair rate. The task is to develop a simulation model for calculating the steady-state probabilities of the system, to find the stationary probability of failure-free operation of the system for some special cases of distributions and assess the reliability of the system, for $N = 3$.

2.1. Simulation model for calculating steady-state probabilities of $\langle GI_N/GI/1\rangle$ system. Let's define the following states of the simulated system:

- State 0: One (main) element works, $N - 1$ are in a cold standby;
- State 1: One element failed and is being repaired, one – works, $N - 2$ are in a cold standby;
- State 2: Two elements have failed, one is being repaired, the other is waiting for its turn for repair, one – works, $N - 3$ are in a cold standby;
- State N: All the items have failed, one is being repaired, the rest are waiting their turn for repair.

To describe the reliability modeling algorithm for the $\langle GI_N/GI/1\rangle$ system we introduce the following variables:

- double t - simulation clock; changes in case of failure or repair of the system's elements;
- int i, j - system state variables; when an event occurs, the transition from i to j takes place;
- double $t_{nextfail}$ – service variable, which stores the time until the next element failure;
- double $t_{nextrepair}$ – service variable, which stores the time until the next repair of the failed element;
- int k - counter of iterations of the main loop.

For clarity, the simulation model is presented graphically in Figure 1 in the form of a flowchart. The criterion for stopping the main cycle of the simulation model is to achieve the maximum model execution time T.

Table 1 shows the values of the coefficient $\rho = \frac{a_1}{b_1}$ — the relative repair rate (i.e. the ratio of the average uptime of the main element to the average repair time of the failed element), at which the specified level of stationary reliability $1 - \pi_3 = 0.9; 0.99; 0.999$. To analyze and compare the results, the following distributions were chosen: Exponential (M), Weibull-Gnedenko (WB), Lognormal (LN).

We consider particular cases of the model at $\rho = 25$; $N = 3$; $NG = 100$; $T = 1000$; where $b_1 = 1$; T_1 - system uptime; T_2 - repair time of a failed element.

A sufficiently high level of the system's reliability is achieved with a relatively small excess of the average values of the uptime by the repair time, except when the uptime of the system elements is distributed according to the Weibull-Gnedenko law.

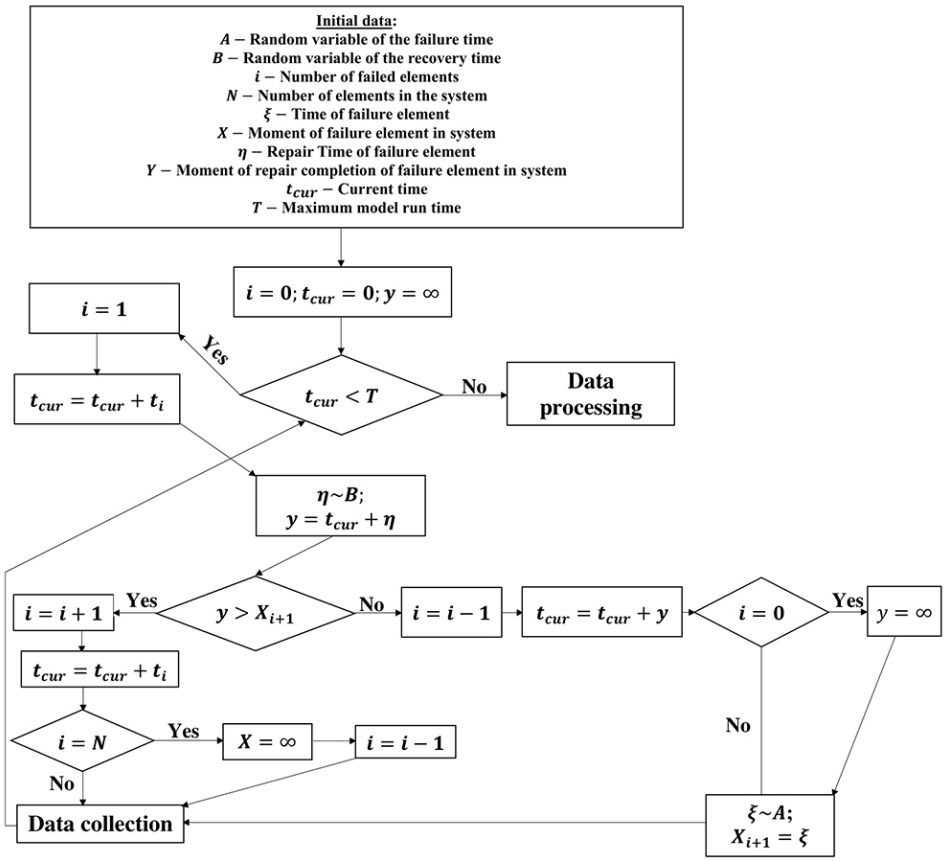


Fig. 1. Flowchart of the simulation model for estimating stationary probabilities.

Table 1. Values of the relative repair rate, at which a given level of the system's stationary reliability is achieved.

T_2 / T_1	$M(1/b_1)$			$WB(W)$			$LN(sig)$		
	0.9	0.99	0.999	0.9	0.99	0.999	0.9	0.99	0.999
$M(1/a_1)$	1.6	4.2	9.1	1.5	4.7	11.3	1.6	4 .4	9.7
$WB(W)$	3.2	12.2	25	2.9	11.6	25	3.3	11.9	25
$LN(sig)$	1.6	3.9	7.6	1.6	4.5	9.7	1.6	4	7.2

Figure 2 shows graphs of the probability of system uptime.

The obtained results demonstrate a high asymptotic insensitivity of the stationary reliability of the system. It can be seen that the differences between the curves

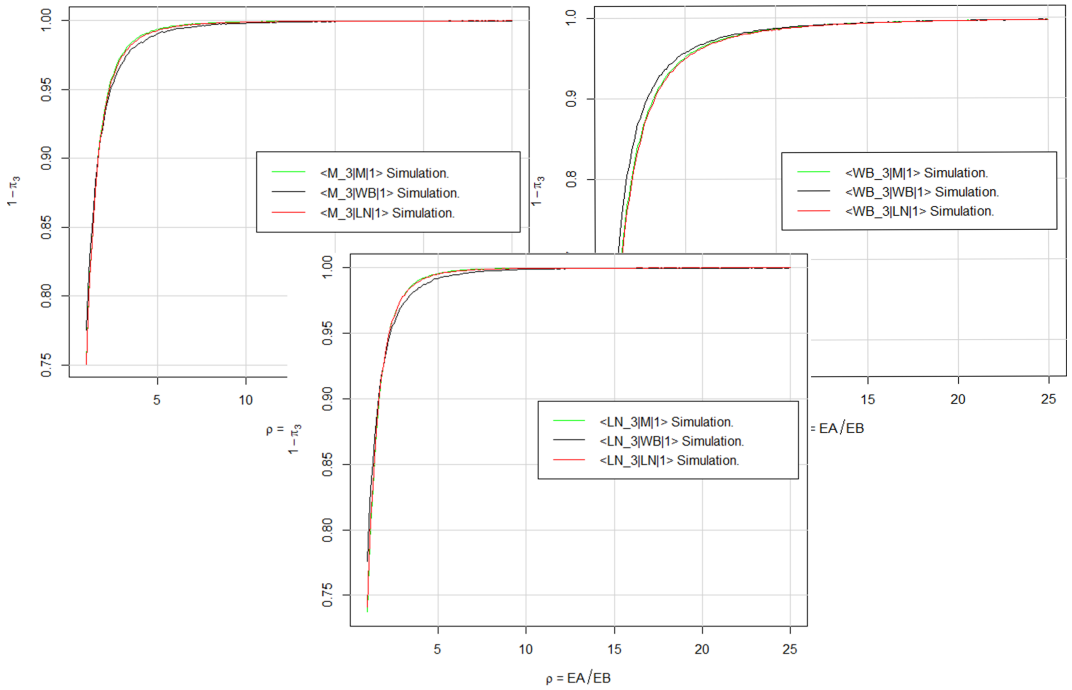


Fig. 2. Graphs of the probability of system uptime versus relative recovery rate for the systems $\langle M_3/GI/1\rangle$, $\langle WB_3/GI/1\rangle$ and $\langle LN_3/GI/1\rangle$.

during "fast" recovery become vanishingly small for all the considered distributions of the repair time of the system elements. For example, already starting from the value $\rho = 10$, all the curves are almost indistinguishable.

2.2. Simulation model for assessment of the $\langle GI_N/GI/1\rangle$ system reliability. In this case, the system stops functioning after all N elements have failed, and the maximum model run time T is equal to ∞.

For clarity, the simulation model is presented graphically in Figure 3.

Table 2 shows the values of the reliability estimates of the system (estimates of the mean time to failure of the system) with the time spent on modeling. The same distributions were chosen: Exponential, Weibull-Gnedenko, Lognormal.

We consider particular cases of the model at $\rho = 25$; $N = 3$; $NG = 10000$; where $b_1 = 1$; T_1 - system uptime; T_2 - repair time of a failed element.

As it can be seen from Table 2, the most reliable model is a model with a lognormal distribution of uptime and an exponential distribution of the repair time of a failed element.

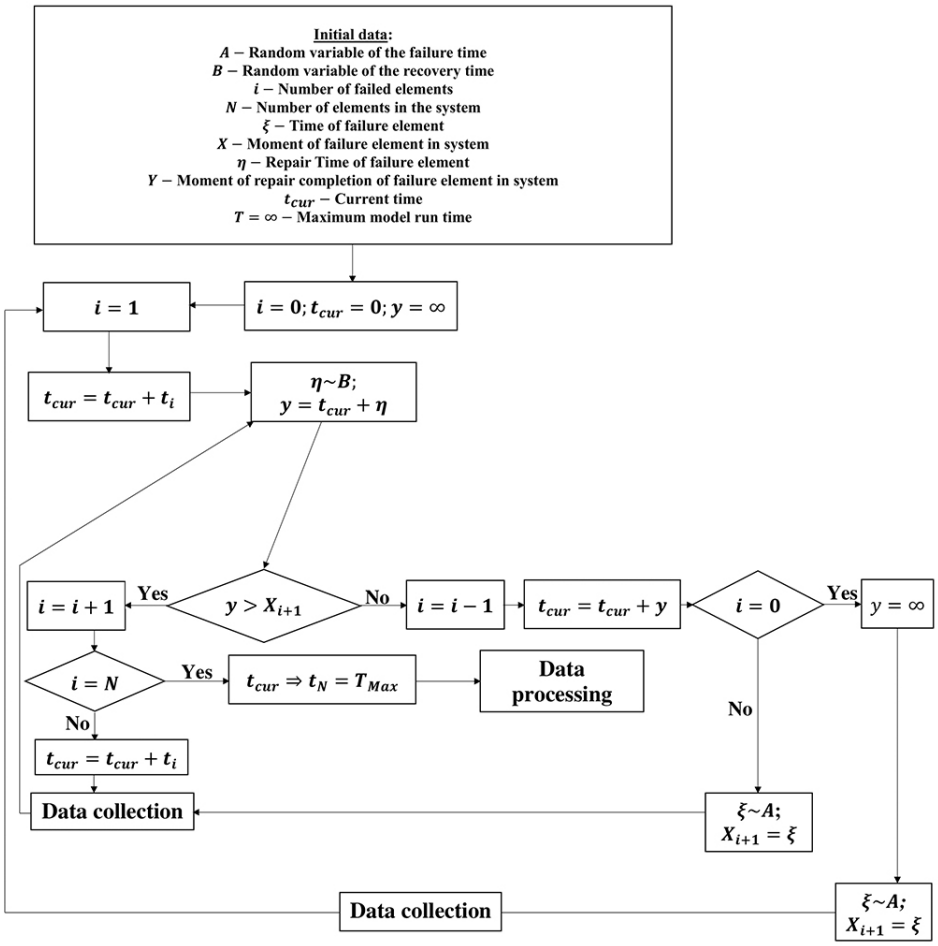


Fig. 3. Flowchart of the simulation model for evaluating system reliability.

Table 2. Values of the estimates of the mean time to failure of the $\langle GI_3/GI/1\rangle$ system.

T_1 \ T_2	$M(1/b_1)$	$WB(W)$	$LN(sig)$
$M(1/a_1)$	16530.34	19566.77	25.18033
$WB(W)$	28.57675	927.8087	564.099
$LN(sig)$	249458.5	71212.42	190780.8

Figure 4 presents graphs of the empirical distribution function $F^*(t)$ and the empirical reliability function $R^*(t)$.

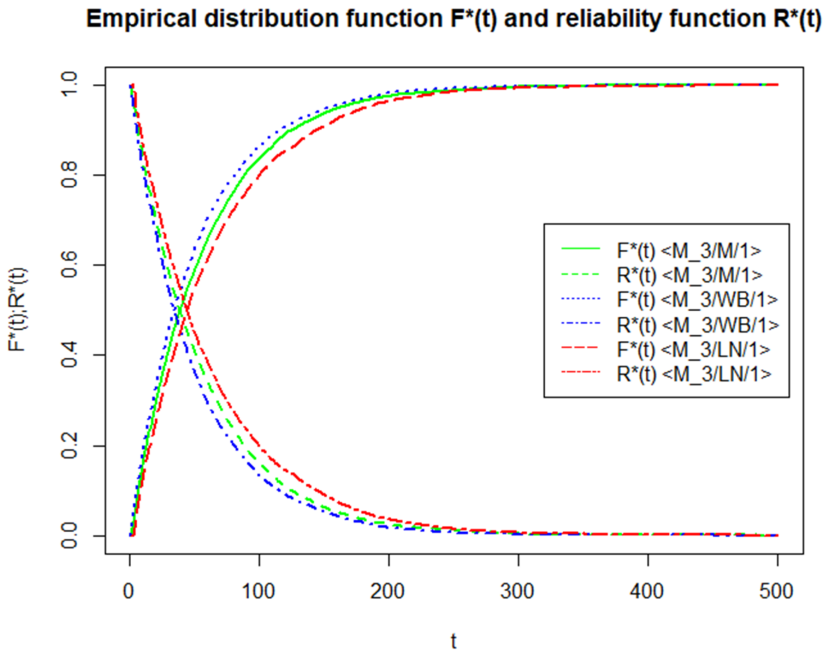


Fig. 4. Graphs of the empirical distribution function $F^*(t)$ and the empirical reliability function $R^*(t)$

The results also show the high asymptotic insensitivity of the empirical distribution function and the corresponding empirical reliability function of the system to the shapes of the uptime and repair time distributions of the system's elements.

3. Conclusion

We considered a repairable multiple cold standby system $\langle GI_N/GI/1\rangle$ with one repair device, with an arbitrary distribution function of uptime and an arbitrary distribution function of repair time of its elements. For the considered system we applied the discrete-event simulation approach to perform the assessment of the system-level reliability and obtained the values of the relative repair rate at which the given level of the system's stationary reliability is achieved. Graphic and numerical results show a high asymptotic insensitivity of the stationary system reliability to the input distributions. The differences between the curves under "fast" recovery become vanishingly small for all the studied special cases of distributions. It was shown that the most reliable case is the model with a lognormal distribution of uptime and an

exponential distribution of the repair time of a failed element. The graphic results also show a high asymptotic insensitivity of the empirical distribution function and the empirical reliability function of the system.

Acknowledgments

The publication has been prepared with the support of the "RUDN University Program 5-100" (recipient H.G.K. Houankpo, mathematical and simulation model development). The reported study was funded by RFBR according to the research projects No. 19-29-06043 (recipient Dmitry Kozyrev, formal analysis, validation) and No. 20-37-90137 (recipient H.G.K.Houankpo, methodology and numerical analysis).

REFERENCES

1. Houankpo H. G. K., Kozyrev D. V. (2017) Sensitivity Analysis of Steady State Reliability Characteristics of a Repairable Cold Standby Data Transmission System to the Shapes of Lifetime and Repair Time Distributions of Its Elements. In: CEUR Workshop Proceedings, vol. 1995, pp. 107–113 (2017). `http://ceur-ws.org/Vol-1995/`
2. Waqar Ahmed, Osman Hasan, Usman Pervez, Junaid zadir (2017) Reliability modeling and analysis of communication networks // J. Netw. Comput. Appl.,
3. Aleksandr Ometov, Dmitry Kozyrev, Vladimir Rykov, Sergey Andreev, Yuliya Gaidamaka, Yevgeni Koucheryavy. Reliability-Centric Analysis of Offloaded Computation in Cooperative Wearable Applications // Wireless Communications and Mobile Computing. Vol. 2017, Article ID 9625687, 15 pages, 2017. DOI:10.1155/2017/9625687.
4. Vladimir Rykov, Dmitry Kozyrev, Elvira Zaripova (2017) Modeling and Simulation of Reliability Function of a Homogeneous Hot Double Redundant Repairable System // Proceedings of the 31st European Conference on Modelling and Simulation, ECMS2017, pp. 701–705, 2017. DOI: 10.7148/2017-0701.
5. Efrosinin D., Rykov V. (2014) Sensitivity Analysis of Reliability Characteristics to the Shape of the Life and Repair Time Distributions // Communication in Computer and Information Science 487, (2014) 101-112.
6. Efrosinin D., Rykov V., Vishnevskiy V. (2014) Sensitivity of Reliability Models to the Shape of Life and Repair Time Distributions // 9th International Conference on Availability, Reliability and Security (ARES 2014) IEEE, 430-437, Published in CD: 978-I-4799-4223-7/14, DOI:10.1109/ARES.40.

7. Rykov V. V., Kozyrev D. V. Analysis of renewable reliability systems by Markovization method // Analytical and Computational Methods in Probability Theory (ACMPT 2017), Lecture Notes in Computer Science, Vol. 10684, 2017. Springer, Cham. Pp. 210-220. DOI: 10.1007/978-3-319-71504-9_19.
8. Rykov V. V., Kozyrev D. V. On Sensitivity of Steady-State Probabilities of a Cold Redundant System to the Shapes of Life and Repair Time Distributions of Its Elements, Statistics and Simulation // Springer Proceedings in Mathematics and Statistics vol. 231, Chapter 28, 2018. Springer, Cham. P. 391-402. DOI: 10.1007/978-3319-76035-3_28.
9. Lisnianski A, Laredo D., Hanoch Ben Haim (2016) Multi-State Markov Model for Reliability Analysis of a Combined Cycle Gas Turbine Power Plant // Published in: 2016 Second International Symposium on Stochastic Models in Reliability Engineering, Life Science and Operations Management (SMRLO), DOI: 10.1109/SMRLO.2016.31.
10. Wei Wang, Francesco Di Maio, Enrico Zio. Three-loop Monte Carlo simulation approach to Multi-State Physics Modeling for system reliability assessment // Reliability Engineering and System Safety, 167 (2017). 276–289.
11. Xiang-Yu Li, Hong-Zhong Huang, Yan-Feng Li, Enrico Zio. Reliability assessment of multi-state phased mission system with non-repairable multi-state components. Applied Mathematical Modelling, Elsevier, 2018, 61, pp.181-199.
12. Yan-Hui Lin, Yan-Fu Li, Enrico Zio. A comparison between Monte Carlo simulation and finite-volume scheme for reliability assessment of multi-state physics systems // Reliability Engineering and System Safety, 174 (2018). 1–11.
13. Bertz Tourgoutian, Yanushkevich, A., Riaan Marshall (2015) Reliability and availability model of offshore and onshore VSC-HVDC transmission systems // Published in: 11th IET International Conference on AC and DC Power Transmission, 13 July 2015, DOI: 10.1049/cp.2015.0101.
14. Junhai Cao, zinqin Wang, Ying Shen (2012) Research on Modeling Method of Complex System Mission Reliability Simulation // Published in: 2012 International Conference on quality, Reliability, Risk, Maintenance, and Safety Engineering, DOI: 10.1109/IC'R2MSE.2012.6246242.
15. Gu J., zhu C., Shang L., Dick R. (2008) Application-specific multiprocessor system-on-chip reliability optimization // Published in: IEEE Transactions on Very Large Scale Integration (VLSI) Systems. (Volume: 16, Issue: 5, May 2008), DOI: 10.1109/TVLSI.2008.917574.

UDC: 004.9

Creation and visualization of the subject area model

N.B. Bakanova[1], D.V. Volchkov[1], A.S. Bakanov[2]

[1]Keldysh Institute of Applied Mathematics of RAS, Miusskaya sq., 4, Moscow, Russia
[2]Institute of Psychology of the Russian Academy of Sciences, Yaroslavskaya st., 13, Moscow, Russia

Abstract

The article is devoted to the creation and visualization of a distributed information system model. To model and visualize the presentation of users and developers of the information system, to identify and integrate the functional tasks of the subject area, fuzzy cognitive maps are used.

Keywords: information systems, fuzzy cognitive maps, modeling, design of information systems

1. Introduction

Modern distributed information systems (IS) to support organizational management are quite complex software and hardware systems. This places high demands on the design, creation and support of IS. The list of requirements for the IS functionality is constantly growing, IS includes tasks and innovative directions related to the creation of integrated distributed information data warehouses, modes of analysis and forecasting, decision support modes. The main tool for developing complex distributed IS is modeling [1]. At all stages of the life cycle of an information system, it is necessary to have a visualized current structure of IS, which fully and comprehensively reflects the functioning of IS, presents the connections between functional modules, and shows interaction with the outside world. The development of methods and techniques to adequately represent and visualize the structure of large-scale information systems is an actual scientific direction. The evolution of information systems, the complication and development of functionalities implemented by information systems, determine the complexity of the used software systems, and, in turn, imposes high requirements on systems design methods. Most design methods for large-scale distributed information systems are based on the use of models. For quick and effective perception of information, it is advisable to use graphic images in

models as a means of presenting information. Most of the existing methodologies for designing information systems provide the ability to represent modeling objects in the form of various graphical notations for visualizing the model.

2. Model Development Methodologies

The feasibility of choosing a specific methodology in the development of a model is determined by an interconnected set of factors including the skills and experience of developers [2]. In the process of developing design methodologies, it was concluded that graphical languages are advantageous for modeling information systems, since descriptive languages do not have sufficient visibility and perception efficiency for modeling information processes of large-scale systems. As a justification for this conclusion, it was pointed out that descriptive languages do not provide a sufficient level of consistency and completeness of description, which is one of the mandatory requirements in the process of developing and designing information systems. The use of various types of visualization, graphic notations used in the design, due to the large number of classes of problems, the solution of which are oriented to certain models. Currently widely used approaches are "functional - modular" and "object-oriented". This article describes an approach to creating a visual representation of a distributed IS through the use of cognitive maps in order to identify and integrate the functional tasks of a management organization within a single information system.

3. Using cognitive maps for visualize IS

Regardless of the chosen design methodology, the process of developing an information system begins with a survey of the subject area. In the process of examining the subject area, a conceptual model of the subject area is created. Models created during the process of examining the subject area are developed in a static form. At the subsequent stages of development, models are created that can be represented both in a static and in a dynamic form. Using only static or only dynamic models in the development process can negatively affect the project being created and lead to unsuccessful implementation of the entire project as a whole. At the initial stage of designing a distributed multifunctional system, it is rather difficult to collect detailed requirements for functional tasks within the framework of the general task of informatization of the organizational structure, since the capabilities of the system are not yet clear for end users. Also, any requirements and wishes of users must be compared with the capabilities of the developers and with the general functional of the system, including multitasking, the work of specialists in other application areas, system restrictions and restrictions on the performance of maintenance tasks. A proven and well-proven approach for researching the subject area and developing information systems is the use of cognitive maps. Cognitive

maps allow you to reflect the views of various users on the system from the point of view of its functioning, since one of the distinctive elements characterizing a large-scale information system is its distributed. Users of distributed information systems can be located in different geographically remote points and interact with each other through the information system. The process of work is influenced by the characteristics of the information system, as well as the views of remote users on the functionality and characteristics of the system. To prepare a project for the development of the system, it is necessary to analyze the views of various users on the functionality of the system and the tasks of its development. In this case, cognitive maps are an effective tool, which is a graph model of a distributed IS. In the general case, cognitive maps are intended to reflect the subjective ideas of the subject or group of subjects about the spatial organization of the outside world. They are created and modified as a result of the active interaction of the subject with the outside world. Cognitive maps were originally proposed by R. Axelrod [3] in 1976 to visualize the "entity-relationship" representation. Later B. Kosco [4] was proposed to use elements of fuzzy logic in cognitive maps. Currently, cognitive maps provide ample opportunities for monitoring and modeling complex, large-scale, distributed systems. In the study, to study the representations of various user groups about a distributed information system, its relationships and functions, as well as the correspondence between different models, the approach described below was used. The user's view on the process of functioning of the system is reflected in the form of a cognitive map, which is represented by a graph whose vertices are the modules (objects) of the distributed system (functions, services and capabilities provided by these modules), and the arcs of the graph represent functional relationships between objects of the cognitive map. The following types of connections are indicated on the map: explicit connection, implicit (according to the subjective opinion of the user) connection. The weights of the arcs range from 0 to 1. Cognitive maps visualize the individual's mental representation [5] of a distributed system. It should be noted that there are a fairly large number of definitions of the term mental representation. The following definition is used in this article: mental representation - a subjective image of objective reality, a reflection of the inner and outer world in the human mind. Or in relation to this study, a subjective structured image of an distributed information system. In the research of O.I. Larichev and A.B. Petrovsky [6] was noted that in the course of interaction with the information environment, a specialist has to take into account a large number of different factors, as well as solve tasks of multicriteria choice. This leads to a load on the information processing system, forcing the individual to use different, sometimes very original heuristics to solve the tasks. A person's ability to receive and process information from the standpoint of cognitive psychology is described using various functional models of the user's

memory structure, mechanisms of the thinking process and other cognitive processes [7, 8]. The proposed approach is focused on modeling the activities of a group of applied users of a system of distributed large-scale information system. The task of the group of users of the system was to: develop technical specifications for the modernization of individual functional modules of the information system; to model the interaction between the various modules of a distributed information system; monitor the functioning of the developed modules. For this, a group of users of the system needed to have an idea of the functions and capabilities of the entire system as a whole, as well as the functions and capabilities of the individual modules of this system. After the distribution of tasks between the various modules of the system, the user needed to monitor the execution of tasks, as well as to analyze and evaluate the effectiveness of the tasks [9].

4. Conclusion

In the process of conducting research, a mental image of the distributed information system was revealed for different groups of users. The image had a hierarchical structure, which could be estimated quantitatively and qualitatively. The structure was in the form of a directed graph, the nodes of which corresponded to functional modules, and the directed arcs or edges corresponded to the connections that were identified by the user in the process of working with the information system. The number of hierarchy levels in the structure and the number of arcs converging to one node characterized both the structure of the mental representation of the group member and the nature (features) of his activity mediated by the information system. For example, if a large number of arcs converged to one node, it is obvious that the subject mainly uses the functionality of this particular module of the information system. Thus, by visualizing his perception of the information system, the subject represent own mental representation in the form of a graph and his function activities mediated by the information system. Article prepared by the assignment of the Ministry of Education and Science of the Russian Federation topic No. 0017-2019-0005 “Theoretical and applied problems of information technology, computer graphics, visual analytics and multidimensional data processing”, and topic No. 0159-2020-0001 “Psychological problems of professional mentality in the context of organizational and technological innovation”.

REFERENCES

1. Vishnevsky V., Semenova O.. “Polling Systems: Theory and Applications for Broadband Wireless Networks”. London: LAMBERT Academic Publishing, 2012, 317p.

2. Bakanov, A. S.; Zelenova, M. E. Cognitive Styles as Determinants of Success in Professional Activity // Social psychology and society 2015 Vol: 6 No.: 2 Pp.: 61-75.
3. Axelrod R. Structure of Decision: The Cognitive Maps of Political Elites. – Princeton, NJ : Princeton University Press, 1976.
4. Kosko B. Fuzzy Engineering. – Upper Saddle River, NJ : Prentice-Hall, 1997.
5. Kholodnaya M.A., Gelfman E.G.Development-focused educational texts as a basis for learners' intellectual development in studying mathematics (DET technology) / Psychology in Russia: State of the Art. Volume 9, Issue 3, 2016, pp. 24-37.
6. Petrovsky A.B. Multi-Attribute Sorting of Qualitative Objects in Multiset Spaces. // Multiple Criteria Decision Making in the New Millenium. Berlin, Springer-Verlag, 2001, p.124-131.
7. Bakanov A., Atanasova T. and Bakanova N., "Cognitive Approach to Modeling Human-Computer Interaction with a Distributed Intellectual Information Environment," 2019 Big Data, Knowledge and Control Systems Engineering (BdKCSE), Sofia, Bulgaria, 2019, pp. 1-4.
8. Sawaragi T., Iwai S., Katai O. An integration of qualitative causal knowledge for user –oriented decision support. Control theory and advanced technology. Vol. 2, No. 3, September 1986, pp. 451-483.
9. Volchkov D.V., Bakanov A.S., Tashev T.D. Resource approach to data mining based on network traffic. Proceedings of the Third International Conference "Natural Information Technologies" NIT 2012, October 02-05,2012, Polithechnika de Madrid, Madrid, Spain. – p.p.92-94.

UDC: 004.75

Faultless and timely multipath packets delivery probability in computer networks using UDP-based protocol

I.I. Noskov[1] and V.A. Bogatyrev[2]

[1,2]ITMO University, Kronverksky prospekt 49, Saint-Petersburg, Russia

noskovii@mail.ru, vladimir.bogatyrev@gmail.com

Abstract

Timely and faultless packets delivery problem in real-time systems is described in the paper. New method of packets transferring with high probability of faultless and timely delivery is presented. This approach is based on UDP protocol and using redundant transmissions via multipath reserve channels between a client and a server. For efficiency evaluation of our approach we use the multiplicative criteria based on faultless and timely packets delivery probability and average delivery time reserve relatively delivery time restriction defined in the real-time computer system. The efficiency of redundant multipath transmissions is analyzed and researched using obtained results from experiments with developed simulation models in OMNeT++ environment. This paper can be useful for network engineers who develop new transport or application layer protocols to provide reliable network transmissions in computer networks.

Keywords: multipath redundant transmissions; delivery probability; UDP; critical to delays packets; OMNeT++.

1. Introduction

Nowadays there are many research works and papers [1-8] described modern computer networks problems. Authors provide new ways and solutions that help to improve quality of interconnection between nodes. Fundamental problems of designing and developing information and communication systems are represented in [9, 10]. Security issues of computer networks are considered in [11, 12]. In papers [13, 14] authors provide researching connected with redundant multipath transmissions, but they didn't consider influence of real network protocols in developed models and didn't provide suggestions for improving or developing new reliable multipath network protocols. Real-time communication system reliability is associated not only with supporting the availability, fault tolerance and reliability of the system structure,

but also with timely delivery of critical to delays packets in real-time computer networks which provide computer communications in the client-server architecture [15]. Network engineers change physical network topology (add new links between routers/switches, add new network equipment etc.) or/and use different network protocols (FHRP, MPTCP, SCTP etc.) to achieve better reliability in computer networks. Approach with using new protocols is more suitable because it does not require new equipment and provides only software devices upgrading. It is important from economy point of view because we don't need to buy new equipment.

In the paper [16] authors provide survey of recent transport layer protocols. They describe new congestion control algorithms used in transport layer for better network performance, provide information about new transport layer protocols which control of packets delivery and use connection establishment between nodes before sending data. These protocols are developed based on TCP protocol architecture. Also they mentioned about multipath modification of TCP - MPTCP. But there are no papers about multipath redundant transmissions based on UDP protocol without delivery guarantee. But these protocols are widely used in real-time computer systems and delay sensitive applications.

UDP is transport layer network protocol which does not use handshaking for connection establishment and does not ensure of packets delivery. Time-sensitive applications often use UDP for real-time traffic sending scenario because new packets have bigger priority and loss out of date packets and receiving new packets is more preferable than waiting retransmissions for lost packets. In this cases packets are becoming outdated very quickly, and retransmissions of lost packets (like in TCP protocol) is not suitable for such systems. That is why it is important to research and improve UDP transport protocol for time sensitive systems and applications.

Developing and using new transport layer protocols needs to change kernel source code of operating system on communication nodes to provide opportunity to using new protocol. In most cases we cannot upgrade kernel without shutdown, also some operating systems are proprietary and we have no access to source code for its modification. That is why developing new reliable protocol over existing transport protocols on application layer is more suitable and scalable solution. New modification will be able to easy integrated to different systems and we shouldn't change kernel source code.

The main aim of this work is the developing simulation model of new multipath redundant transmissions protocol prototype based on UDP in the OMNeT++ environment and find out effective using areas of this protocol.

2. Developing model of UDP protocol multipath extension

Simulation environment OMNeT++ is modern specialized tool for simulating and researching computer network models. This environment contains a large library of real network protocols and equipment models [17]. There are many models developed in this environment [18-21]. It shows that it is flexible tool with implementations of different layers network protocols.

UDP is used as a transport protocol for our prototype. The OMNeT++ environment has implementations of various types of generator and sink applications which use UDP transport protocol. There are UDPBasicApp class and UDPSink class implementations in OMNeT++ environment. These classes provide functionality of client and server UDP-based applications. But these applications don't support redundant multipath transmissions. New classes of generator and sink applications have been developed using C++ programming language in order to provide redundant transmissions application layer UDP-based protocol.

New UDP application classes extend of the base classes and allow to specify several addresses in configuration file to provide redundant transmission via sending copy of datagrams to these addresses. Port number and packets length are set in configuration file. In Figure 1 you can see example of the traffic flow session between client and server using developed protocol based on UDP multipath transmissions.

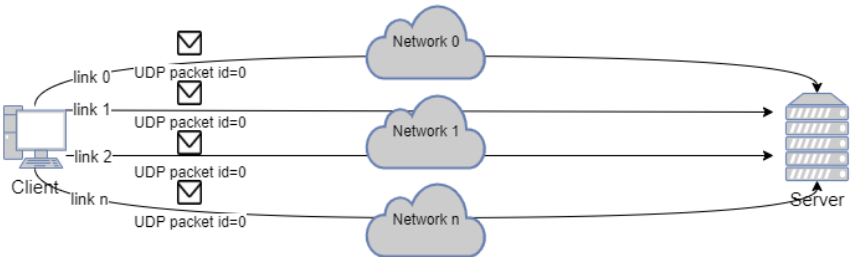


Fig. 1. Multipath transmissions using UDP-based protocol

User should set n interfaces which will be used for multipath data transmissions in new protocol. After that UDP sockets are created for n selected interfaces on the server and the client. UDP-based protocol is using all created sockets in one session for providing multipath data transmissions of user data flow. The protocol is creating copy of each datagram and sending copies via all sockets for current session. UDP datagrams are being transmitted from the client to the server via different physical channels simultaneously. Application layer header has ID to identify duplicates on the receiver side. The server side application recognizes copies in received data and drop extra packets. Fastest arrived datagram will be saved and transferred to application for further processing and providing data to user.

3. Developing computer network model using modified UDP protocol

The simulation model of system with server, five network switches and five clients was developed. This model is based on a EtherSwitch model which represents model of network switch and a StandardHost model which simulates client and server behaviour. Figure 2 shows this model in the OMNeT++ environment.

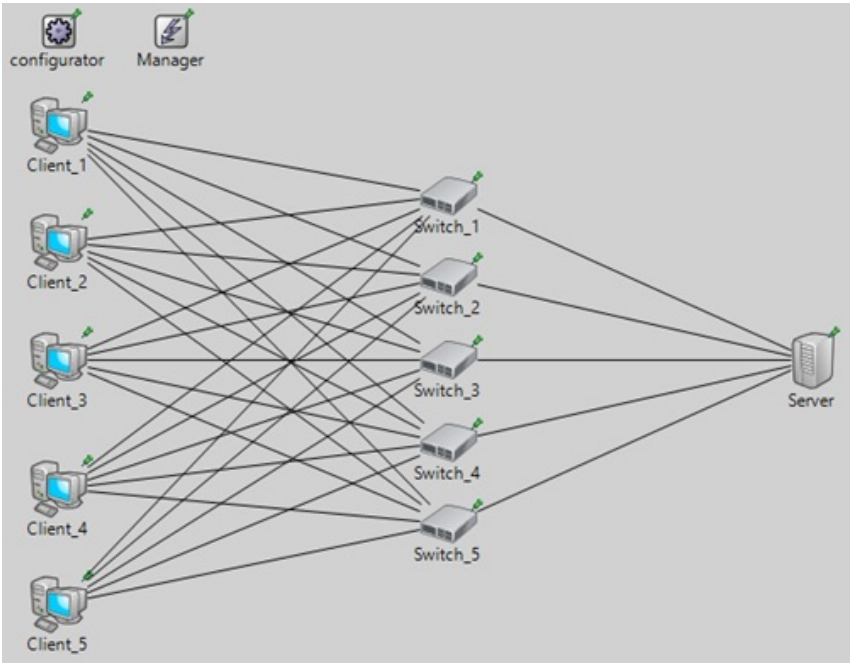


Fig. 2. Redundant computer network model

In this model network switches with connected to them clients and server present different network segments which can be used for redundant transmissions.

The criterion M has been used for efficiency evaluation of redundant transmissions.

$$M = P(t_0 - T) \tag{1}$$

This (1) represents the multiplicative criteria based on faultless and timely packets delivery probability and average delivery time reserve relatively delivery time restriction. P is a probability of timely (packet should be received before t_0 comes) and faultless packets delivery. Value of P is calculated after experiment. This parameter depends on bit error rate of channels (defined before start experiment) and network switches queues sizes which can increase delivery time for packets. t_0 is a

time limit for packets delivery in system (important parameter for real-time systems). T is a average delivery packets time during experiment. Considered criterion can be used for evaluating quality of transmissions in real-time systems where it is important to have a big value of faultless and timely delivery probability for transmissions and small value of packets delivery time.

In redundant transmissions via different network switches between nodes, each switch is used by more than one node. The coefficient K in our model shows how much redundant links between client and server we will use for multipath transmissions. We carried out simulation experiments with different values of the redundancy coefficient K.

These parameters have been used in simulation experiments with redundant transmissions: $L = 10$ Mbit/s - communication channels throughput. $B = 0.0001$ - bit error probability for channel, $\lambda = 1000$ 1/s - packet arrival rate, packets length for this simulation process is 100 B. Network delay time for 100 B packet for two links (client-switch and switch-server) has been calculated. This value includes data sending time according to throughput value and propagation delay in channel. Processing packet time on nodes (client, switch and server) was added to the total time. Obtained value was multiplied by two times and considered as a delivery time limit $t_0 = 0.0004$ s for our real-time system. This value can be different for other real-time systems and depends on their limitations for delivery packet time.

All described parameters are set in OMNeT++ environment in $*.ini$ (general simulation experiment parameters), $*.ned$ (contains network model topology description) and $*.xml$ (equipment parameters) configuration files.

4. Experiments with multipath redundant transmissions

Different experiments for researching multipath transmissions using new protocol were carried out. Plots from Figure 3 shows the value of criterion M at redundant transmissions coefficient K for different packet intensity (1000 1/s and 2000 1/s). From plots you can see that the selected criterion M takes large values at lower intensity, which indicates the small size of queues in the network switches. At greater intensity Switches queues are growing at greater intensity and this leads to increase of delivery packets time in the system and reduces the criterion M. According to these plots value of criterion M increases until $K = 3$ at 1000 1/s intensity and increases until $K = 2$ at 2000 1/s intensity. It says that we increase of faultless and timely delivery probability in developed real-time system simulation model using new protocol. In this network configuration we can conclude that packets transmissions with lower intensity is more efficient on all values of redundancy coefficient K. In the curve 2 you can see sharp decline at $K = 4$. It allows to make conclusion that switch buffers are overflowed and packets was being dropped.

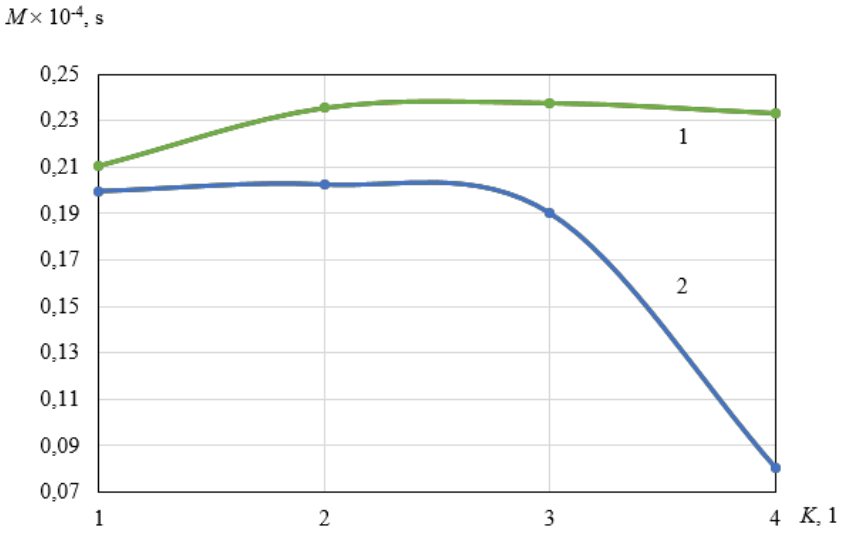


Fig. 3. The dependence of the efficiency criterion M on the redundancy coefficient K: at an intensity of 1000 1/s (curve 1); at an intensity of 2000 1/s (curve 2)

Increasing the number of redundant transmissions helps us to reduce packets lost probability. But the end to end packets delay in the system increases because queues in network switches are growing. Delays influence on probability of timely delivery packets and reduce value of criterion M. Efficiency criterion represents multiply probability of timely and faultless packet delivery and average delivery time reserve relatively delivery time restriction. This approach helps us to consider delivery time as a part of our criterion. It is needed for real-time systems where delivery time is very important characteristic of interaction because in such systems information is being out of date very fast. For systems where delivery time is not critical we can consider probability of faultless packet delivery and probability of faultless and timely delivery packets as a main efficiency criterion. In this case we don't consider delivery time reserve relatively delivery time restriction. It can be useful for system without strong packets delivery time limitations.

Figure 4 shows the criterion M at different packets arrival intensity for different values of redundant transmissions coefficient K (1 and 2). As you can see from plots increasing of redundancy coefficient is not effective for all area.

There is area in which transmissions with a high redundancy coefficient is more efficient on the same network configuration. After overcoming the intensity threshold of 2000 1/s, the redundant model is becoming less efficient than model without reservation. Increasing traffic in the network in redundant model leads to the growing

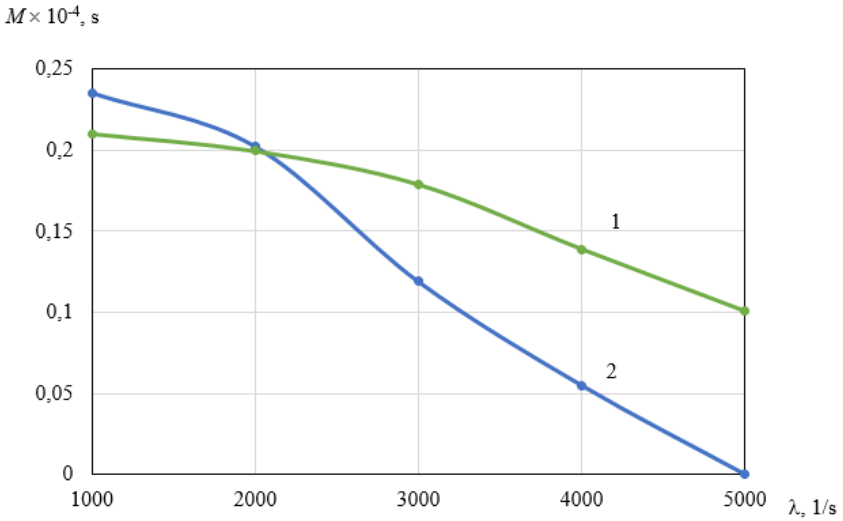


Fig. 4. The dependence of the efficiency criterion M on the intensity of packet arrival: with the redundancy coefficient $K = 1$ (curve 1); with the redundancy coefficient $K = 2$ (curve 2)

of queues in network switches. That is why delivery time of packets is growing. This case is unacceptable for real-time systems. But the probability of delivery packets increases using redundant transmissions. It is important for delivery time insensitive systems.

For delivery time insensitive systems faultless and timely delivery probability is more important criteria than M. Figure 5 shows faultless and timely delivery packet probability at packets arrival intensity for different value of redundancy coefficient K (1 and 2).

From plots above we can see that redundant transmissions help to achieve bigger packet delivery probability in most cases until $\lambda = 4000$ (after that packets are dropped because switches queues are overflowed). It helps us to transmit data with more faultless and timely delivery probability in insensitive computer networks using new developed protocol. This approach is useful in data centers which backup big amount of user data or servers where faultless data transmissions more important criteria than data delivery time.

5. Conclusions

The simulation model of computer network with the possibility of increasing the redundancy of packets transmissions has been developed in the OMNeT++ environment. Experiments to assess the effectiveness of packets transmissions with

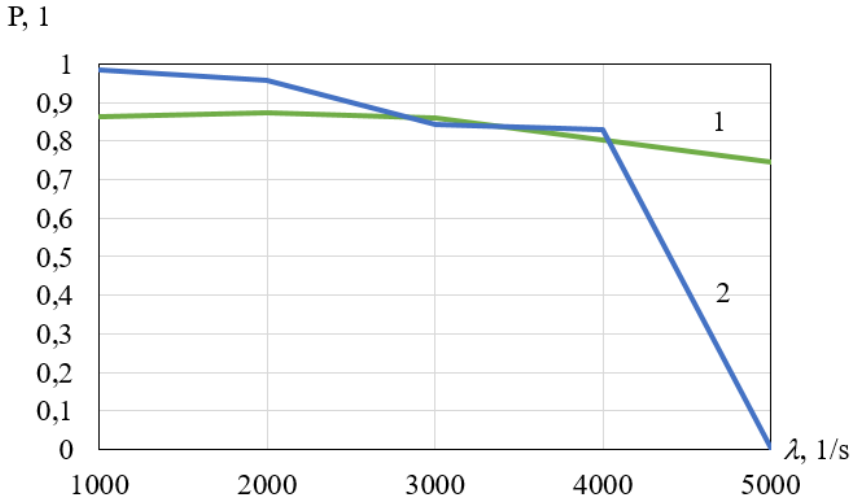


Fig. 5. The dependence of the faultless and timely delivery packet probability on the intensity of packet arrival: with the redundancy coefficient $K = 1$ (curve 1); with the redundancy coefficient $K = 2$ (curve 2)

different intensity and redundancy coefficient was carried out. Areas with effective using of redundant transmissions in computer networks with delivery time restrictions and delivery time insensitive systems have been described. Developed model allows to transmit packets via several physical channels and provides redundant data transfer. Modifications of UDP protocol on application layer are proposed. The presented results can be used in the design of high-reliable real-time computer network systems based on UDP protocol with strong restrictions to delivery time. This research can be used as a theoretical and practice base to develop new transport or application layer protocols based on UDP and provides more reliable and timely transmissions of important data in real-time computer systems.

REFERENCES

1. Birman K. P., Joseph T., Raeuchle T., Abbadi A. El. Implementing fault-tolerant distributed objects // IEEE Transactions on Software Engineering. — 1985. — 11(6). — Pp. 502–508.
2. Coulouris G., Dollimore J., Kindberg T., Blair G. Distributed Systems. Concepts and Design. Fifth edition. — Addison-Wesley, 2011. — 1080 p.
3. Defago X., Schiper A., Sergent N. Semi-Passive Replication // Proc. of the 17th IEEE Symposium on Reliable Distributed Systems (SRDS). — West Lafayette, IN, USA, Oct. 1998. — Pp. 43–50.

4. Gunnar A., Johansson M. Robust load balancing under traffic uncertainty-tractable models and efficient algorithms // Telecommun Systems. — 2011. — V. 48. — Iss. 1–2. — Pp. 93–107.
5. Kim Y., Righter R., Wolff R. Job replication on multiserver systems // Advances in Applied Probability. — 2009. — V. 41. — Pp. 546–575.
6. Kurose J. F., Ross J. F. Computer networking: a top-down approach. — 6th ed. — Boston Pearson, 2013. — 862 p.
7. Malichenko D. Optimization of Network Overhead for Transport Layer Coding // 9th Conference of Open Innovations Community FRUCT. — 2011. — P. 92–95.
8. Lee M. H., Dudin A. N., Klimenok V. I. The SM/M/N queueing system with broadcasting service // Math. Probl. Eng. — 2006. — Article ID 98171.
9. Sorin D. Fault Tolerant Computer Architecture. Morgan Claypool 2009, P. 103.
10. Kopetz H. Real-Time Systems: Design Principles for Distributed Embedded Applications. Springer, pp. 396, 2011.
11. Vishnevskii V.M. Teoreticheskie osnovy proektirovaniya (Theoretical Foundations of Design), Moscow: Tekhnosfera, 2003.
12. Aliev T.I. The synthesis of service discipline in systems with limits. Communications in Computer and Information Science, IET, vol. 601. 2016. pp. 151-156.
13. Bogatyrev V.A., Parshutina S.A. Redundant Distribution of Requests Through the Network by Transferring Them Over Multiple Paths // Communications in Computer and Information Science - 2016, Vol. 601, pp. 199-207.
14. Bogatyrev V.A., Parshutina S.A. Efficiency of Redundant Multipath Transmission of Requests Through the Network to Destination Servers // Communications in Computer and Information Science - 2016, Vol. 678, pp. 290-301.
15. Noskov I.I., Bogatyrev V.A. Interaction model of computer nodes based on transfer reservation at multipath routing // 2019 Wave Electronics and its Application in Information and Telecommunication Systems (WECONF) - 2019, pp. 8840607.
16. Polese M., Chiariotti F., Bonetto E., Rigotto F., Zanella A., Zorzi M. A survey on recent advances in transport layer protocols. CoRR, abs/1810.03884, 2018.
17. Varga A., Hornig R. An overview of the OMNeT++ simulation environment. In Simulation tools and techniques for communications, networks and systems workshops, Simutools '08, 2008.
18. Vesely V., Rek V., Rysavy O. Enhanced Interior Gateway Routing Protocol with IPv4 and IPv6 Support for OMNeT++ // Advances in Intelligent Systems and Computing. 2016, vol. 2015, no. 1, pp. 65-85. ISSN 2194-5357.
19. Vesely V., Sveda M. L2 protocols in OMNeT++ // IP Networking 1 - Theory and Practice. Zilina: Zilina University Publisher, 2012, pp. 37-40.

20. Vesely V., Rysavy O., Sveda M. Protocol Independent Multicast in OMNeT++ // The International Academy, Research and Industry Association, 2014, pp. 132-137.
21. Noskov I.I., Bogatyrev V.A. Simulating of fault-tolerant gateway based on VRRP protocol in OMNeT++ environment // CEUR Workshop Proceedings - 2019, Vol. 2522.

UDC: 519.246

Overbooking's problem for a case of a random environment existence

Alexander Andronov[1,2], Iakov Dalinger[2], Diana Santalova[3]

[1]Transport and Telecommunication Institute, Lomonosov Str., 1, 1019 Riga, Latvia
[2]Saint-Petersburg State University of Civil Aviation, Pilotov Str., 38, 196210 Saint-Petersburg, Russia
[3]Møreforsking AS, Britvegen 4, 6410 Molde, Norway

aleksander.andronov1@gmail.com, iakovdalinger@gmail.com, dsantalova@gmail.com

Abstract

An overbooking policy assumes that a booking of some product or service exceeds given possibilities. It takes into consideration that a part of the booking will be cancelled. This situation is considered following to examples of aviation ticket booking. It is supposed that an external random environment exists. The environment is described as a continuous-time finite irreducible Markov chain. A demand on the booking depends on the state of the random environment. We consider such indices as the average number of engaged seats, the probability that passenger with bought or booked ticket encounters a refusal etc. A numerical example is considered.

Keywords: continuous-time Markov chain, overbooking's problem

1. Introduction

An overbooking policy assumes that a sale and a booking of some product or service exceed given possibilities. It takes into consideration that a part of the sale or the booking will be canceled. The overbooking is used in different spheres of a transport, hotel's businesses etc. Numerous publications are devoted to this problem ([1]–[6]).

In this paper we consider the problem of the overbooking in the case of an external random environment existence. Airline overbooking will be considered for concreteness. Notably we use one word "to buy" both as for "to buy" and for "to book".

The following positing of the problem will be considered. We will follow to the text [6] as the later in below given bibliography.

It is considered one aircraft trip, having the capacity n^* passengers. The pre-departure time, when passengers buy tickets, is a random variable with the density

$f(x), x \geq 0$. The passengers buy tickets independently of each other. There is the probability q that the passenger with the ticket, doesn't come for the trip.

Then an external random environment exists, having k states with numbers $1, \ldots, k$. The environment is described as a continuous-time finite irreducible Markov chain $J(t)$ with the matrix $\lambda = (\lambda_{i,j})_{k \times k}$ of transition probabilities between states.

An average demand for a considered trip depends on the state of the random environment and equals d_i for the i-state, $i = 1, \ldots, k$. Therefore the intensity of customers' arrivals at time t till a departure, if the i-th state occurs, is calculated as follows:

$$\tilde{d}_i(t) = d_i f(t), t \geq 0.$$

At every moment of time t, the number of the sold tickets n and the state i of the environment J are known. A decision on overbooking is adopted with intervals Δ, at the instants $s\Delta$, $s = 1, 2, \ldots$, until departure. The i-th stage is called the time interval $(s\Delta, (s-1)\Delta)$. The average number of the passengers, whose buy tickets on the s-th stage, if $J((s^* - s)\Delta) = i$, is

$$\alpha_i(s) = \int_{t=\Delta(s-1)}^{\Delta s} \tilde{d}_i(t)dt, s = s^*, s^* - 1, \ldots, 1. \tag{1}$$

Additionally we are guided by the maximal value of the overbooking. Let it be $m_{n,i}(s)$: the maximal value of overbooking at instant $t = s\Delta$, if n place are busy and $J(t) = i$. Here $m_{n,i}(\Delta) \leq m^*$, where m^* is given.

2. Main results

We will consider the described process with the step $\Delta > 0$. Let τ be the time, when all seats are occupied, $s^* = \tau/\Delta$ be an integer, so the step number s belongs to the set $\{s^*, s^*–1, \ldots, 0\}$. The s-th step corresponds to the time interval $(\Delta s, \Delta(s–1))$.

The random environment is represented by the continuous-time irreducible finite Markov chain $J(t)$ with k states and matrix $\lambda = (\lambda_{i,j})$ of transition intensities between states. Let $P_{i,j}(t)$ be the probability that chain $J(t)$ will be in state j at instant t if the initial state is i, $P(t) = (P_{i,j}(t))$ be the corresponding matrix. This matrix is calculated as follows [7], [8].

Let $\begin{pmatrix} 1 & \ldots & 1 \end{pmatrix}^T_{k \times 1}$ be the column-vector from the units, $\Lambda = \lambda \begin{pmatrix} 1 & \ldots & 1 \end{pmatrix}^T_{k \times 1}$ be the column-vector, $diag(\Lambda)$ be the diagonal matrix with vector Λ on the main diagonal. The $k \times k$-matrix $A = \lambda - diag(\Lambda)$ is called generator of the Markov chain. We denote eigenvalues and eigenvectors of this matrix by $\chi_1, \chi_2, \ldots, \chi_k$ and $\beta_1, \beta_2, \ldots, \beta_k$ correspondingly. It is supposed that all values $\chi_1, \chi_2, \ldots, \chi_k$ are different.

Let $B = (\beta_1, \ldots, \beta_k)$ be the matrix, whose columns are eigenvectors $\beta_1, \ldots, \beta_k$ of the generator A, $\tilde{\beta}_1, \ldots, \tilde{\beta}_k$ be rows of the inverse matrix B^{-1}, so that $B^{-1} = (\tilde{\beta}_1^T, \ldots, \tilde{\beta}_k^T)^T$, $diag(\exp(t\chi))$ be the diagonal matrix with the vector $\exp(t\chi) = (\exp(t\chi_1), \ldots, \exp(t\chi_k))$ on the main diagonal. Then

$$P(t) = \sum\nolimits_{i=1}^{k} \exp(\chi_i t)\beta_i\tilde{\beta}_i = Bdiag(\exp(t\chi))B^{-1}, t \geq 0. \tag{2}$$

Now we can calculate the average value $ET_{i,\nu,j}(\Delta)$ of sojourn time in the state ν on interval $(0, \Delta)$ jointly with probability $P\{J(\Delta) = j\}$, if $J(0) = i$:

$$ET_{i,\nu,j}(\Delta) = \int_0^{\Delta} P_{i,\nu}(u)P_{\nu,j}(\Delta - u)du, \quad i, \nu, j \in \{1, \ldots, k\}. \tag{3}$$

Further let us calculating the probability $\tilde{P}r_{\eta,i,j}(s)$, that η new requests on tickets are received during stage s and the final state $J(\Delta(s-1))$ equals j, if the i–th state take place at instant Δs. With respect to the paper [6] we use the following formula:

$$\tilde{P}r_{\eta,i,j}(s) = \frac{1}{\eta!}\left(\sum\nolimits_{\nu=1}^{k} \alpha_\nu(s)ET_{i,\nu,j}(\Delta)\right)^{\eta} \exp\left(-\sum\nolimits_{\nu=1}^{k} \alpha_\nu(s)ET_{i,\nu,j}(\Delta)\right), \eta = 0, 1, \ldots. \tag{4}$$

Let $Pr_{n,i}(s)$ be the probability that at beginning of the s–th stage the following situation occurs: n claims are gotten, the state of MC $J(\Delta s)$ equals i. We will consider the following values of n: $n \in \{0, 1, \ldots, n^* + m^* + 1\}$. If values $n \leq n^*$ then $Pr_{n,i}(s)$ means the probability that n places are busy. If value n belong to interval $[n^* + 1, n^* + m^*]$ then $Pr_{n,i}(s)$ means the probability of corresponding overbooking. The probability $Pr_{n^*+m^*+1,i}(s)$ means the probability that the number of claims exceeds $n^* + m^*$.

We know values of $n = n^*$ and $i = i_0$ for the initial stage with number s^*, therefore

$$Pr_{n,i}(s^*) = \begin{cases} 1 & \text{if } n = n^*, i = i_0, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Further for $s = s^* - 1, s^* - 2, \ldots, 0$, $n \geq n^*$,

$$Pr_{n,j}(s) = \begin{cases} \sum\limits_{i=1}^{k} \sum\limits_{\eta=n^*}^{n} Pr_{\eta,i}(s+1)\tilde{P}r_{n-\eta,i,j}(s+1), \text{ if } n^* \leq n \leq n^* + m^*(s+1); \\ P_{i_0,j}((s^* - s)\Delta) - \sum\limits_{n=n^*}^{n^*+m^*(s+1)} Pr_{n,j}(s), \text{ if } n = n^* + m^*(s+1) + 1. \end{cases} \tag{6}$$

The part of the last formula, corresponding to the case $n = n^* + m^*(s+1) + 1$, follows from the equality

$$P_{i_0,j}((s^* - s)\Delta) = \sum\nolimits_{n=n^*}^{\infty} Pr_{n,j}(s).$$

The zero stage $s = 0$ corresponds to the instant of the trip beginning. Now n means the number of all booked or sold tickets. Each of the corresponding passengers can not arrive to the trip with probability q, independently on the other passengers. Therefore, the probability that n passengers are in front of the take off

$$PP_n = \sum\nolimits_{\eta=n}^{n^*+m^*} \frac{\eta!}{n!(\eta-n)!}(1-q)^n q^{\eta-n} \sum\nolimits_{j=1}^{k} Pr_{\eta,j}(0), n \le n^* + m^*. \qquad (7)$$

Finally the probability that n passengers have flown away is as follows:

$$PFA_n = \begin{cases} PP_n, & \text{if } n < n^*, \\ \sum\limits_{n=n^*}^{n^*+m^*} PP_n, & \text{if } n = n^*. \end{cases} \qquad (8)$$

The represented formulas allow to calculate various efficiency indices. Firstly, the average number of engaged seats:

$$Avr = \sum\nolimits_{n=1}^{n^*} n \times PFA_n. \qquad (9)$$

Secondly, the probability that n passengers with bought or booked tickets encounter a refusal equals $P_{n+n^*}, n = 1, \ldots, m^*$. Average number of those passengers AvrR is calculated as follows:

$$AvrR = \sum\nolimits_{n=1}^{m^*} nP_{n+n^*}. \qquad (10)$$

Finally, the probability that a customer encounters a refusal to purchase the ticket equals $\sum_{i=1}^{k} \sum_{s=1}^{s^*} Pr_{n^*+m^*+1,i}(s)$.

3. Numerical example

Our example has the following input data. The Markov chain has three states ($k = 3$) and the transition intensities matrix

$$\lambda = \begin{pmatrix} 0 & 0.3 & 0.4 \\ 0.5 & 0 & 0.4 \\ 0.5 & 0.6 & 0 \end{pmatrix}.$$

The initial state of Markov chain is known and fixed: $J(0) = i_0 = 1$. The capacity of the aircraft n^* equals 20. A time before departure, when passengers buy tickets, has Erlang distribution with parameters $\mu = 0.5$ and $\theta = 3$, and the density

$$f(x) = \frac{1}{4}(0.5x)^2 \exp(-0.5x), x \geq 0. \tag{11}$$

Further we assume, that the average demand for given trip depends on the state of the random environment only and equals $d_1 = 22.5$, $d_2 = 18$, $d_3 = 13.5$ for the first, second and third states. Now the arrivals intensity of passengers at time t until departure, if the i-th state occurs, is calculated by formula (1).

Let $\tau = 3$ be the time, when all 20 seats are occupied. We consider the selling and the booking process with the step $\Delta = 1$. Therefore we have $s^* = \tau/\Delta = 3$ stages. At last we assume that the probability q, that a passenger doesn't come to a trip, equals 0.15.

The results for these initial data are given below. Firstly, the case $m_{n,i}(s) = m^* = 3$ for all n and s is considered, when the maximal number of additional overbooking equals 3 and doesn't depend on the number s of the step and the number n of booked and sold tickets. Fig.1 contains graph of density (11).

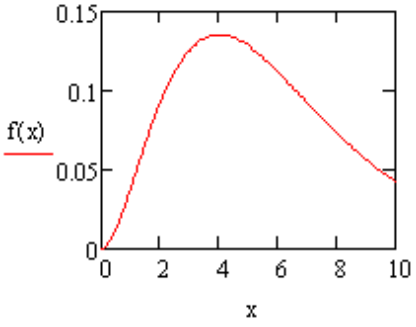


Fig. 1. Graph of density (11)

The expression (2) for the transition probabilities between states $J(t)$ has the following form:

$$P(t) = \begin{pmatrix} -0.704 & 0.577 & -0.323 \\ 0.503 & 0.577 & -0.323 \\ 0.503 & 0.577 & 0.889 \end{pmatrix} \begin{pmatrix} e^{-1.2t} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & e^{-1.5t} \end{pmatrix} \begin{pmatrix} -0.829 & 0.829 & 0 \\ 0.722 & 0.549 & 0.462 \\ 0 & -0.825 & 0.825 \end{pmatrix}.$$

Tables 1-3 contain the conditional average times $\tilde{ET}_{i,\nu,j}(\Delta) = ET_{i,\nu,j}(\Delta)/P_{i,j}(\Delta)$ of the sojourn of process $J(t)$ in the state ν on interval $(0, \Delta)$, if $J(0) = i, J(t) = j$.

The columns of the tables correspond to the initial states $i = 0, 1, 2$, the rows correspond to the intermediate states ν. Note that for each column the sum of all its elements equals $\Delta = 1$.

$\nu \setminus i$	1	2	3
1	0.944	0.466	0.500
2	0.026	0.440	0.056
3	0.030	0.094	0.444

Table 1. Conditional average sojourn times $\tilde{ET}_{1,\nu,j}(\Delta)$

$\nu \setminus i$	1	2	3
1	0.484	0.031	0.086
2	0.455	0.931	0.470
3	0.061	0.038	0.444

Table 2. Conditional average sojourn times $\tilde{ET}_{2,\nu,j}(\Delta)$

$\nu \setminus i$	1	2	3
1	0.484	0.055	0.042
2	0.086	0.487	0.044
3	0.430	0.458	0.914

Table 3. Conditional average sojourn times $\tilde{ET}_{3,\nu,j}(\Delta)$

Tables 4-7 show the calculation results according to formulas (5)–(6). The probabilities $\{Pr_{n,j}(s)\}$ are given for $s = 3, 2, 1, 0$, $j = 0, 1, 2$ and $n = 20, \ldots, 24$. Let us remind, that: 1) the value for $n = 24$ means the probability, that the number of claims exceeds $n^* + m^* = 23$; 2) the initial state $J(s^*) = i0 = 1$.

$j \setminus n$	20	21	22	23	24
1	1	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0

Table 4. Probabilities $Pr_{n,j}(3)$

The probabilities (7), that n passengers are in front of the take-off, are presented in Table 8.

$j \setminus n$	20	21	22	23	24
1	0.051	0.125	0.153	0.125	0.015
2	0.023	0.049	0.954	0.039	0.112
3	0.027	0.055	0.056	0.038	0.077

Table 5. Probabilities $Pr_{n,j}(2)$

$j \setminus n$	20	21	22	23	24
1	0.011	0.041	0.077	0.096	0.207
2	0.009	0.030	0.052	0.060	0.153
3	0.009	0.031	0.050	0.055	0.119

Table 6. Probabilities $Pr_{n,j}(1)$

$j \setminus n$	20	21	22	23	24
1	0.009	0.034	0.066	0.085	0.228
2	0.007	0.027	0.049	0.062	0.168
3	0.007	0.024	0.044	0.054	0.138

Table 7. Probabilities $Pr_{n,j}(0)$

n	12	13	14	15	16	17
PP_n	0.001	0.002	0.006	0.019	0.047	0.095
n	18	19	20	21	22	23
PP_n	0.158	0.209	0.213	0.158	0.075	0.017

Table 8. Probabilities PP_n

The probabilities (8) that n passengers have flown away are shown in Table 9.

n	13	14	15	16	17	18	19	20
PFA_n	0.002	0.006	0.019	0.047	0.095	0.158	0.209	0.464

Table 9. Probabilities PFA_n

These tables allow the calculating of various efficiency indices. The average number of engaged seats Avr, calculated by the formula (9), equals 18.853. The probability PP_{20+n} that n passengers with bought or booked tickets encounter a refusal equals 0.158, 0.075, and 0.017 for $n = 1, 2$, and 3. There the average number

of such passengers

$$1 \times 0.158 + 2 \times 0.075 + 3 \times 0.017 = 0.359.$$

It is noteworthy to compare these results with those whose will be without overbooking, when $m_{n,i}(s) = m^* = 0$. Instead of Table 9, Table 10 takes place.

n	11	12	13	14	15	16	17	18	19	20
PFA_n	0.001	0.005	0.016	0.045	0.103	0.182	0.243	0.229	0.137	0.039

Table 10. Probabilities PFA_n for the case $m_{n,i}(s) = m^* = 0$

The average number of engaged places Avr equals 17 instead of 18.853. We see that the difference is significant.

4. Conclusion

The considered model can be generalized in many ways. Firstly to discriminate between sold and booked tickets. Secondly it takes into account a possibility of a cancellation of booked tickets during a period of our consideration. Our future researches will be connected with the realization of these possibilities.

REFERENCES

1. Shlifer R., Vardi Y. An airline overbooking policy // Transportation Science. 1975. V.9(2), pp. 101 – 114.
2. Liberman V., Yechiali U. On the hotel overbooking problem – An inventory system with stochastic cancellations // Management Sci. 1978. V.24(11), pp. 1117–1126.
3. Chatwin R. E. Optimal control of continuous-time terminal-value birth-and-death processes and airline overbooking // Naval Res. Logist. 1996. V.43(2), pp. 159–168.
4. Chatwin R. E. Multi-period airline overbooking with a single fare class // Opns. Res. 1998. V.46(6), pp. 805–819.
5. Phillips R. Pricing and Revenue Optimization. Stanford University Press, Stanford, 2005.
6. Sulima N. Probabilistic model of overbooking for an airline // Automatic Control and Computer Sciences. 2012. V. 46(1), pp. 68 – 78.
7. Kijima M. Markov Processes for Stochastic Modeling. Chapman & Hall, London, 1997.
8. Pacheco A., Tang L. C. and Prabhu N. U. Markov-Modulated Processes & Semiregenerative Phenomena. World Scientific, Hoboken, New York, 2009.
9. Bellman R. Dynamic Programming. Princeton University Press, Princeton, N.J., 1957.
10. Bellman R. E., Dreyfus S. E. Applied Dynamic Programming. Princeton University Press, Princeton, N.J., 1962.

UDC: 519.872

Modeling D2D-enhanced IoT Connectivity

T.A. Milovanova[1], R.V. Razumchik[3], D.V. Kozyrev[1,2]

[1]Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation
[2]V.A.Trapeznikov Institute of Control Sciences of RAS, 65 Profsoyuznaya St, Moscow, 117997, Russian Federation
[3]Institute of Informatics Problems of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilova St, Moscow 119333, Russian Federation

milovanova-ta@rudn.ru, rrazumchik@ipiran.ru, kozyrev-dv@rudn.ru

Abstract

In the light of the proliferation of the Internet of Things (IoT), the device-to-device (D2D) communication is becoming a promising technology and a key enabler for enhancing the energy efficiency of the wireless network environment and reducing the traffic latency between user equipments (UEs) within their communication range. This papers considers an analytical framework for modeling a communication offloading scenario within the D2D communications underlaying cellular network, where a UE generates a session toward another UE in the same cell. The designed analytical model aims to improve the system capacity and energy-efficiency by offloading cellular traffic onto D2D communications when the source and destination UEs are proximate enough to satisfy their QoS requirements.

1. Introduction

The development of information and communication technologies (ICT) and increasing interest to their applications go hand in hand. The potential and capabilities of advanced ICT systems are still growing exponentially fueled by the progress in hardware, micro-systems, networking, data processing and human machine interfaces. This has led to an increased need for reliable connectivity of various electronic devices. This massive interconnection of proliferating heterogeneous physical objects together with services has formed a new ecosystem which is technically termed as the Internet of Things (IoT) [1]. However, this increase in connectivity creates many serious

The reported study was funded by RFBR, project number 20-07-00804 (recipient T.A. Milovanova and R.V. Razumchik, mathematical model development, numerical analysis) and project number 19-29-06043 (recipient D.V. Kozyrev, formal analysis, validation).

challenges. The rapid growth in the number of interconnected IoT devices creates a significant burden on existing and emerging wireless networks [2]. One of the ways to stem the tide of the number of newly deployed base stations is to take advantage of the device-to-device (D2D) connectivity. D2D communication is a promising technology that utilizes the proximity of communicating devices [3]. Although the idea of proximal communications is not new [4], the industrial standardization of D2D technology has only recently been started [5, 6]. D2D Communications underlaying cellular networks provide mobile broadband operators with supplementary transport and increase network capacity through spatial reuse of radio resources for cellular and D2D communications.

In this paper we propose a mathematical model of the wireless communication between mobile UEs inside an IoT cell. The area of interest has a base station providing infrastructure coverage and a constant number of UE devices. All UEs in the cell can initiate sessions that can be serviced by utilizing both the D2D links and the infrastructure links. We design an analytical framework for modeling the considered communications system as a closed multi-server queueing system. The established analytical model allows to analyze the system performance.

2. Mathematical model

Assume that the total number of UEs is fixed and equal to N. The total number of infrastructure links is also fixed and equal to c, whereas the maximum total number of D2D links is limited by the integer nearest to $N/2$. Each UE initiates a session* according to a Poisson flow and occupies either an infrastructure or a D2D link. The session initiation rate via an infrastructure link is equal to α and the one via a D2D link is equal to β. Duration of a UEs session through both the infrastructure link and the D2D link (does not depend on the UE type and cannot be interrupted) has an exponential distribution with parameters μ_I and μ_D respectively. Due to the adopted assumptions the pair† $(D(t), I(t))$ is the two-dimensional Markov chain in continuous time t with the discrete state space

$$\{(i,j):\ 0 \le i \le \lfloor N/2 \rfloor,\ 0 \le j \le c\}\,.$$

The Markov chain $\{(D(t), I(t)), t \ge 0\}$ is in fact the level-dependent QBD process. Following the terminology of [7] the level of this QBD process is the value of $D(t)$ and the phase is the value of $I(t)$. Denoting the infinitesimal generator of

*With the other UE, i.e. whenever an infrastructure or a D2D link is busy, it is busy simultaneously by 2 UEs.

†$D(t)$ is the total number of UE pairs, connected through the D2D links; $I(t)$ is the total number of UE pairs, connected through the infrastructure links.

$\{(D(t), I(t)), t \geq 0\}$ by A, the joint stationary distribution (which is unique and always exists whenever the μ_I and μ_D are not equal to zero simultaneously)

$$p_{ij} = \lim_{t\to\infty} \mathbf{P}\{D(t) = i, I(t) = j\},$$

can be found by solving (using one of the many methods in the literature) the system of linear algebraic equations $\vec{p}\mathsf{A} = \vec{0}$, $\vec{p}\vec{1} = 1$, where $\vec{p} = (p_{00}, \ldots, p_{0c}, p_{10}, \ldots, p_{1c}, \ldots$ $\ldots, p_{\lfloor N/2 \rfloor, c})$. Another way to obtain the joint distribution is build the uniformized two-dimensional discrete-time Markov chain $\{(D_t, I_t), t = 0, 1, 2, \ldots\}$ from $\{(D(t), I(t)), t \geq 0\}$. For the new, uniformized Markov chain the only possible transitions are

$$(D_t, I_t) = \begin{cases} (D_{t-1} - 1, I_{t-1}), & \text{w.p. } \frac{2\beta(N/2 - D_{t-1} - I_{t-1})}{\Delta_{t-1}}, \\ (D_{t-1}, I_{t-1} - 1), & \text{w.p. } \frac{2\alpha(N/2 - D_{t-1} - I_{t-1})}{\Delta_{t-1}}, \\ (D_{t-1} + 1, I_{t-1}), & \text{w.p. } \frac{2\mu_D D_{t-1}}{\Delta_{t-1}}, \\ (D_{t-1}, I_{t-1} + 1), & \text{w.p. } \frac{2\mu_I I_{t-1}}{\Delta_{t-1}}, \end{cases}$$

where $\Delta_{t-1} = 2\mu_D D_{t-1} + 2\mu_I I_{t-1} + 2(\alpha + \beta)(N/2 - D_{t-1} - I_{t-1})$.

From this relation it can be directly seen that within the considered model the session initiation rates and the session duration rates can be state-dependent (i.e. α, β, μ_D and μ_I can depend on the total number of idle UEs). This fact will be used in the next section to demonstrate the performance of the model under various assumptions on UEs behaviour.

3. Numerical example

Since under the adopted assumptions a requested UE session can always be established, the most natural QoS measure is the distribution of the number of busy D2D and infrastructure links. Below we present the numerical results for the two qualitatively different cases: (i) constant session initiation rate with $\alpha_n = \beta_n = 2$, $0 \leq n \leq N/2$ and (ii) pulsing session initiation rate with

$$\alpha_n = \beta_n = \left(\frac{N + 2 - 2n}{2}\right) \left|\sin\left(\frac{N + 2 - 2n}{2}\right)\right|. \tag{1}$$

Both in (i) and (ii) cases the mean duration of a UEs session was held fixed and equal to 1 i.e. $\mu_I = \mu_D = 1$. The distributions of the number of busy infrastructure links and busy D2D links, depending on the total number of users N and total number of infrastructure links c, are plotted in Fig. 1 and 2. The numbers above the graphs indicate the respective mean value of the distributions. From the figures it

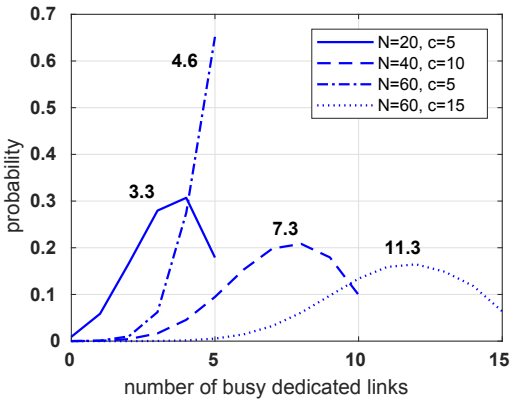


Fig. 1. Distribution of the number of busy infrastructure links for various values of N and c for case (i).

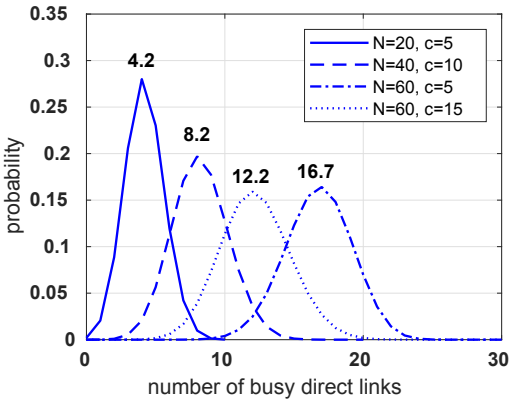


Fig. 2. Distribution of the number of busy D2D links for various values of N and c for case (i).

can be seen that the distributions behave quite regularly, showing tendency towards normal-shaped curves.

In case (ii), with quite a different, pulsing session initiation rates given by (1) (see Fig. 3), the QoS performance remains qualitatively the same as can be seen from Fig. 4 and 5.

Note that it is very appealing to conjecture that the distribution of busy D2D links are normal. But as can be seen from Fig. 5, such conjecture is too optimistic since it can happen that the distributions are skewed.

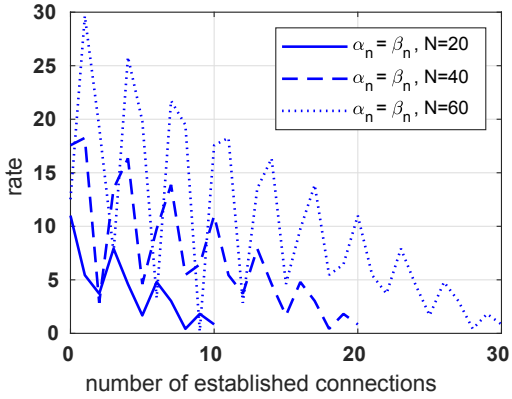


Fig. 3. Pulsing session initiation rates for various values of N.

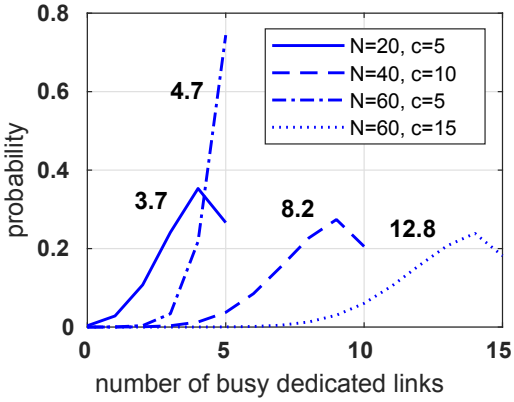


Fig. 4. Distribution of the number of busy infrastructure links for various values of N and c for case (ii).

If the model assumptions hold, then, as the numerical experiments show, the system never behaves irregularly in the sense that the joint distribution of the number of busy D2D and infrastructure links is visually smooth. For example, there never appear two (or more) spikes i.e. the distribution is always unimodal (see Fig. 6 for the joint distribution in the case $N = 60$, $c = 15$ and the arrival rates as given by (1)).

Therefore, within the model assumptions, the shape of the empirical joint distribution (estimated from the available data) may serve as the indication of the irregularity of the UEs' behaviour.

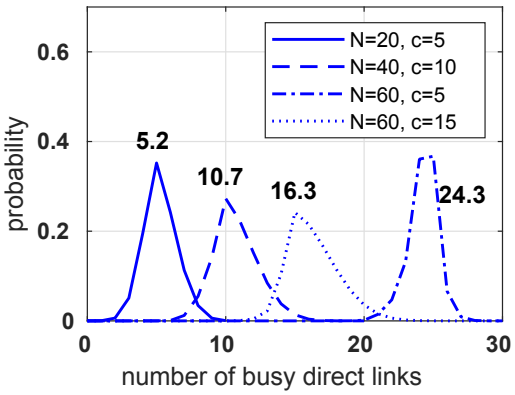


Fig. 5. Distribution of the number of busy D2D links for various values of N and c for case (ii).

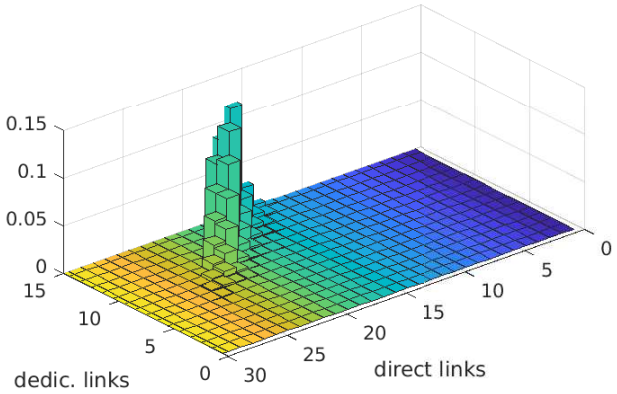


Fig. 6. Joint distribution of busy D2D and infrastructure links.

4. Conclusion

The performance evaluation study presented is for probably the simplest model, which remains tractable but ignores such features as the mobility of the UEs, possible impatience of the UEs, not 100% reliability of the D2D and infrastructure links etc. As our numerical experiments show such necessary additional features do have impact of the quality of service within the proposed model. Yet, as long as we assume memoryless distribution for the ongoing processes within the model, the general picture (as in Fig. 3 and 5) remains the same (even if all the rates depend on the current state of the process). Incorporation of other distributions in the model requires further study and effort to put into the mathematical formulation. With

this respect it is worth mentioning that (assuming memoryless distributions) there is a resemblance between the presented model and the 3-urn Ehrenfest model [8]. Indeed the evolution of the Markov chain $\{(D_t, I_t), t = 0, 1, 2, \dots\}$ can be imagined as the interchange of balls between the three labelled urns, under the restriction that the total number of balls remains fixed and equal to N. The only crucial difference is that in the classical multi-urn Ehrenfest model the rearrangement probabilities (although may depend on the total number of balls in some urns) are the same for any of the balls drawn. Yet in the presented model the rearrangement probabilities depend on the urn from which the ball is drawn. This fact makes the available analytical results for the multi-urn Ehrenfest model inapplicable.

REFERENCES

1. Kalla A., Prombage P., Liyanage M. Introduction to IoT // IoT Security (eds M. Liyanage, A. Braeken, P. Kumar and M. Ylianttila). 2020. doi:10.1002/9781119527978.ch1
2. Cisco. Global mobile data traffic forecast 2016–2021. White Paper. 2017.
3. Fodor G. et al.An Overview of Device-to-Device Communications Technology Components in METIS // IEEE Access. 2016. V. 4. P. 3288–3299.
4. Fitzek F. H. P. Cellular Controlled Short Range Communication for Cooperative P2P Networking // Wireless Personal Communications. 2009. V. 48. No. 1. P. 141–155.
5. 3GPP, TS 23.303, V15.1.0 (2018-06) Technical specification group services and system aspects; Proximity-based services (ProSe), Stage 2. Rel-15. 2018.
6. 3GPP, TS 36.746, V15.1.1 (2018-04) Study on further enhancements to LTE Device to Device (D2D), UE to network relays for Internet of Things (IoT) and wearables. Rel-15. 2018.
7. Neuts M.F. Matrix-geometric solutions in stochastic models: an algorithmic approach. Baltimore: The Johns Hopkins University Press, 1981.
8. Karlin S., McGregor J. Ehrenfest Urn Models. // Journal of Applied Probability. 1965. V. 2. No. 2. P. 352–376.
9. Kozyrev D., Ometov A., Moltchanov D., Rykov V., Efrosinin D., Milovanova T., Andreev S., Koucheryavy Y. Mobility-Centric Analysis of Communication Offloading for Heterogeneous Internet of Things Devices // Wireless Communications and Mobile Computing. 2018. doi: 10.1155/2018/3761075.

UDC: 519.872

OPTIMIZATION OF A SIGNAL PROCESSING STRATEGY IN SENSOR NODES WITH ENERGY HARVESTING AND CONSUMPTION FOR ADMISSION AND TRANSMISSION

A.N. Dudin[1,2], S.A. Dudin[1], O.S. Dudina[1]

[1]Belarusian State University, 4, Nezavisimosti Ave., Minsk, Belarus
[2]Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St., Moscow, 117198, Russia

Abstract

Operation of a sensor node of a wireless sensor network with energy harvesting is described by the single-server queue. Customers and energy units arrive according to the marked Markov arrival process. Service of a customer is possible only in presence of an energy unit. We assume that, besides the use of one energy unit for service of any customer, one more unit is expended at the moment of a customer arrival if the customer is accepted to the system. To optimize operation of the system, a parametric strategy of admission control is used. Under the fixed value of control parameter, the behavior of the system is described by the five-dimensional Markov chain. The generator of this Markov chain is obtained. Expressions for computation of the key performance indicators of the system are presented.

Keywords: Energy harvesting and consumption, admission control, marked Markov arrival process, impatience, phase-type distribution

1. Introduction

Wireless sensor networks have a huge number of applications including environmental applications (forest fire and flood detection, monitoring the pesticides level in the drinking water, the level of soil erosion, and the level of air pollution in real time), health applications (telemonitoring and tracking the human physiological data, location of doctors, patients and drugs inside a hospital), military applications, applications for home automation, control in office buildings, managing inventory control, vehicle tracking and detection, etc. The relevant surveys are given, e.g., in [1, 2].

Nodes of sensor networks have small batteries with limited power and storage space. When the battery of a node is exhausted, sometimes it is not possible to

replace it and the node dies. When a certain number of nodes will die, the network may not be able to perform its designated task. Recent advances in energy harvesting technology have resulted in the design of new types of sensor nodes which are able to extract energy from the surrounding environment. The major sources of energy harvesting include solar, wind, sound, vibration, thermal, and electromagnetic power. Since a signal detection occurs at the random instants and the process of energy harvesting can be also interpreted as the stochastic process, it is reasonable to apply the theory of queues for description of operation and optimization of sensor nodes with energy harvesting. As examples of papers where such application is made, we can refer to [3, 4, 5] and references therein. As more recent papers, we can mention [6] and [7] as well as several papers, which are published in 2019, cited in [7].

It is usually assumed in queueing models that energy units are spent only for service of a customer (transmission of a registered signal to the neighboring or gateway node). In our model, we account the fact that indeed the energy may be required not only for transmission of the signal but also for receiving and registering the signal from a sensor. Therefore, all harvested energy has to be split by the node between its activity related to the signal admission and buffering and the activity related to the signal transmission. Due to the possible deficit of energy, this sharing has to be organized in a proper way. From one hand, it does not make sense to accept too many signals (customers) if the chance to further provide service to them is small (due to the possible lack of energy or obsolescence of information delivered by these customers). From another hand, if the acceptance discipline is too strict, many signals are lost and the node badly implements its task. In addition, the risk of the future server starvation even in presence of energy may be high. To optimize the splitting of the harvested energy for customers admission and service in the system, in this paper, we introduce a parametric strategy of admission control and analyse influence of the parameter defining this strategy on the performance of the system.

2. Mathematical model

We consider a single-server queuing system with an infinite buffer for customers and a finite buffer of capacity R for energy. The structure of this system is presented in Figure 1.

Arrival of customers and energy units is defined by the marked Markov arrival process ($MMAP$). Such an arrival process is defined via the irreducible continuous-time Markov chain ν_t, $t \geq 0$, having a finite state space $\{0, 1, ..., V\}$. Let us denote the generator of this Markov chain by $D(1)$ and decompose it as follows: $D(1) = D_0 + D_1 + D_2$, where the matrices D_1 and D_2 are non-negative with some positive entries and the matrix D_0 has non-negative non-diagonal entries and negative diagonal entries. Arrivals may occur at the transition moments of this Markov chain. The

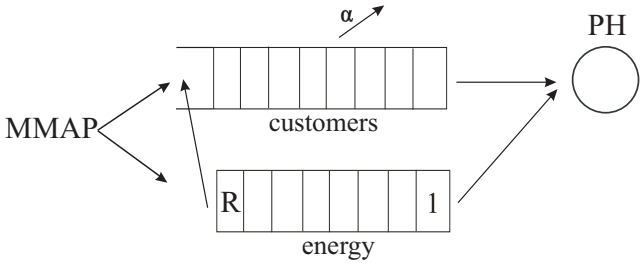


Fig. 1. System under study

entries of the matrix D_1 define the intensities of the transitions that are accompanied by the arrival of a customer. The entries of the matrix D_2 define the intensities of the transitions that are accompanied by the arrival of an energy unit. The non-diagonal entries of the matrix D_0 define the intensities of the transitions that are not accompanied by any arrival. The diagonal entries of the matrix D_0 define the intensity of departure of the Markov chain ν_t from the corresponding states.

The average intensity of customer arrival λ_c is defined by the formula $\lambda_c = \boldsymbol{\theta} D_1 \mathbf{e}$, where $\boldsymbol{\theta}$ is the row vector of the stationary probabilities of the Markov chain ν_t. This vector is the unique solution to the system $\boldsymbol{\theta} D(1) = \mathbf{0},\ \boldsymbol{\theta}\mathbf{e} = 1$. Here and throughout this paper, $\mathbf{e}$ is a column vector of appropriate size consisting of 1's, and $\mathbf{0}$ is a row vector of appropriate size consisting of zeroes. The average intensity of energy unit arrival λ_e is defined by the formula $\lambda_e = \boldsymbol{\theta} D_2 \mathbf{e}$.

An arriving unit of energy joins the buffer for energy if it is not full and is lost otherwise. We assume that the customers staying in the buffer are impatient. If any customer is not picked up for service during a period of time that is exponentially distributed with the parameter α, $0 \leq \alpha < \infty$, then the customer leaves the buffer and the system (is lost), independently of other customers.

We assume that energy units are spent both to receive customers and provide service to them. Namely, we assume that the number of energy units decreases by one during the epoch of a customer acceptance. Also, one energy unit disappears from the energy buffer during each service beginning epoch. If after the service completion epoch the energy is absent, new service is postponed until an arrival of an energy unit.

To optimize the energy consumption in the system, we account the following intuitive consideration. If during the arrival epoch of a customer the number of energy units in the system is small while the number of already accepted customers in the buffer is large, the arriving customer has a high chance to be lost due to the lack of energy when it will reach the server. Therefore, it is reasonable to reject it upon arrival. More formally, we assume that the parameter R_1, $R_1 \geq 0$, defines

the admissible virtual deficit of energy. A new customer is rejected if the number of available energy units is less or equal to $\max\{0, i - R_1\}$, where i is the number of customers in the buffer at the arrival epoch. The control parameter R_1 defines the tolerance of the admission strategy. If $R_1 = 0$, a new customer is rejected if availability of energy unit for its service is not guaranteed even if all accumulated energy will be spent for service of customers (without acceptance of future arrivals). If $R_1 \geq 1$, some temporal deficit of energy is admissible (in anticipation of arrival of new energy units during service of customers staying in the queue). It is clear that the system operation can be optimised via the proper choice of the threshold R_1.

If a customer arrives when the number of energy units in the energy buffer is more than $\max\{0, i - R_1\}$, then the customer is accepted for service. Two scenarios are possible:

1) If the server is idle and the number of energy units is not less than two, the customer immediately starts service and the number of energy units decreases by two. One unit of energy is spent on receiving this customer and one more unit is consumed for its service. If the server is idle and the number of energy units is equal to one, the unit of energy decreases to zero and the customer joins the buffer to wait until energy arrival (if it will not depart earlier due to impatience);

2) If the server is busy, the number of energy units decreases by one and the customer joins the buffer.

Customers are picked up for service from the buffer according to the First-In-First-Out discipline.

The service time of a customer by a server has the PH distribution with the irreducible representation $(\boldsymbol{\beta}, S)$. This service time can be interpreted as the time until the underlying Markov process $m_t, t \geq 0$, with a finite state space $\{1, \ldots, M, M+1\}$ reaches the single absorbing state $M+1$ conditional on the fact that the initial state of this process is selected among the states $\{1, \ldots, M\}$ according to the probabilistic row vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)$. The transition rates of the process m_t within the set $\{1, \ldots, M\}$ are defined by the sub-generator S, and the transition rates into the absorbing state (what leads to service completion) are given by the entries of the column vector $\mathbf{S}_0 = -S\mathbf{e}$. The mean service time is calculated as $b_1 = \boldsymbol{\beta}(-S)^{-1}\mathbf{e}$.

Our goal is to analyse the behavior of the described queueing model.

3. Process of system states and its stationary distribution

Let, during the epoch t, $t \geq 0$,

- i_t, $i_t = \overline{0, R + R_1}$, be the number of customers in the infinite buffer,
- n_t, be the state of the server: if $n_t = 0$, the server is idle, if $n_t = 1$, the server is busy,
- r_t, $r_t = \overline{0, R}$, be the number of available energy units,

- ν_t, $\nu_t = \overline{1,V}$, be the state of the underlying process of the $MMAP$ of customers,
- m_t, $m_t = \overline{1,M}$, be the state of PH service process.

The process $\xi_t = \{i_t, n_t, r_t, \nu_t, m_t\}$, $t \geq 0$, is the regular irreducible continuous-time Markov chain. It has the following state space:

$$\left(\{0,0,r,\nu\}\right)\bigcup\left(\{i,0,0,\nu\},\, i=\overline{1,R+R_1}\right)\bigcup\left(\{i,1,r,\nu,m\},\, i=\overline{0,R+R_1}\right),$$

$$r=\overline{0,R},\, \nu=\overline{1,V},\, m=\overline{1,M}.$$

Let us introduce the following notations:

I is the identity matrix and O is a zero matrix of an appropriate dimension. If it is necessary, dimension of the matrix is indicated by the suffix;

E^- is the square matrix of size $R+1$ with all zero entries except the entries $(E^-)_{i,i-1}$, $i=\overline{1,R}$, which are equal to 1;

E^+ is the square matrix of size $R+1$ with all zero entries except the entries $(E^+)_{i,i+1}$, $i=\overline{0,R-1}$, and $(E^+)_{R,R}$ which are equal to 1;

$\hat{I}_i$ is the square matrix of size $R+1$ defined as $\hat{I}_i = \mathrm{diag}\{\underbrace{1,1,\ldots,1}_{a_i},0,0,\ldots,0\}$ where $a_i = \min\{\max\{0,i-R_1\},R\}+1$, $i=\overline{0,R+R_1}$;

$\hat{\mathbf{e}}$ is the column vector of size $R+1$ defined as $\hat{\mathbf{e}}=(1,0,0,\ldots,0)$;

$\tilde{\mathbf{e}}$ is the column vector of size $R+1$ defined as $\tilde{\mathbf{e}}=(0,1,0,\ldots,0)$;

$\otimes$ and $\oplus$ are the symbols of the Kronecker product and sum of matrices.

Let us enumerate the states of the Markov chain ξ_t in the lexicographic order and refer to the set of states of the chain having value i of the first component of the Markov chain as *level* i, $i \geq 0$.

Let Q be the generator of the Markov chain ξ_t, $t \geq 0$.

Lemma 1. The generator Q has the following block-tridiagonal structure:

$$Q=\begin{pmatrix} Q_{0,0} & Q_{0,1} & O & O & \ldots & O & O & O \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & O & \ldots & O & O & O \\ O & Q_{2,1} & Q_{2,2} & Q_{2,3} & \ldots & O & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \\ O & O & O & O & \ldots & Q_{R+R_1-1,R+R_1-2} & Q_{R+R_1-1,R+R_1-1} & Q_{R+R_1-1,R+R_1} \\ O & O & O & O & \ldots & O & Q_{R+R_1,R+R_1-1} & Q_{R+R_1,R+R_1} \end{pmatrix}.$$

The non-zero blocks $Q_{i,j}$, $i,j \geq 0$, containing the intensities of the transitions from level i to level j have the following form:

$$Q_{i,j}=\begin{pmatrix} Q_{i,j}^{(0,0)} & Q_{i,j}^{(0,1)} \\ Q_{i,j}^{(1,0)} & Q_{i,j}^{(1,1)} \end{pmatrix},$$

where the sub-block $Q_{i,j}^{(n,n')}$ contains the intensities of transition from the states of level i with the value n of the component n_t of the Markov chain ξ_t to the states of level j with the value n' of the component n_t, $n, n' = 0, 1$.

The non-zero sub-blocks $Q_{i,j}^{(n,n')}$ are given by:

$$Q_{0,0}^{(0,0)} = I_{R+1} \otimes D_0 + \hat{I}_0 \otimes D_1 + E^+ \otimes I_V \otimes D_2,$$

$$Q_{0,0}^{(0,1)} = (I - \hat{I}_0)(E^-)^2 \otimes D_1 \otimes \boldsymbol{\beta},$$

$$Q_{0,0}^{(1,0)} = I_{(R+1)V} \otimes \boldsymbol{S}_0,$$

$$Q_{0,0}^{(1,1)} = I_{R+1} \otimes (D_0 \oplus S) + \hat{I}_0 \otimes D_1 \otimes I_M + E^+ \otimes D_2 \otimes I_M;$$

for $i = \overline{1, R + R_1}$:

$$Q_{i,i}^{(0,0)} = D_0 - i\alpha I_V + D_1,$$

$$Q_{i,i}^{(1,0)} = \hat{\mathbf{e}}^T \otimes I_V \otimes \boldsymbol{S}_0,$$

$$Q_{i,i}^{(1,1)} = I_{R+1} \otimes (D_0 \oplus S) + \hat{I}_i \otimes D_1 \otimes I_M + E^+ \otimes D_2 \otimes I_M - i\alpha I_{(R+1)VM};$$

$$Q_{0,1}^{(0,0)} = (I - \hat{I}_0)\tilde{\mathbf{e}}^T \otimes D_1,$$

$$Q_{0,1}^{(1,1)} = (I - \hat{I}_0)E^- \otimes D_1 \otimes I_M;$$

for $i = \overline{1, R + R_1 - 1}$:

$$Q_{i,i+1}^{(1,1)} = (I - \hat{I}_i)E^- \otimes D_1 \otimes I_M;$$

$$Q_{1,0}^{(0,0)} = \alpha\hat{\mathbf{e}} \otimes I_V,$$

$$Q_{1,0}^{(0,1)} = \hat{\mathbf{e}} \otimes D_2 \otimes \boldsymbol{\beta},$$

$$Q_{1,0}^{(1,1)} = \alpha I_{(R+1)VM} + E^- \otimes I_V \otimes \boldsymbol{S}_0\boldsymbol{\beta},$$

for $i = \overline{2, R + R_1}$:

$$Q_{i,i-1}^{(0,0)} = i\alpha I_V,$$

$$Q_{i,i-1}^{(0,1)} = \hat{\mathbf{e}} \otimes D_2 \otimes \boldsymbol{\beta},$$

$$Q_{i,i-1}^{(1,1)} = i\alpha I_{(R+1)VM} + E^- \otimes I_V \otimes \boldsymbol{S}_0\boldsymbol{\beta}.$$

Proof of the lemma is performed by means of analysis of the intensities of all possible transitions of the Markov chain ξ_t during the time interval having an infinitesimal length.

The Markov chain $\xi_t,\ t \geq 0$, is an irreducible and has a finite state space. Therefore, the stationary probabilities $\pi(i,1,r,\nu,m)$, $i = \overline{0, R+R_1}$, $r = \overline{0,R}$, $\nu = \overline{1,V}$, $m = \overline{1,M}$, and $\pi(0,0,r,\nu)$, $r = \overline{0,R}$, $\nu = \overline{1,V}$, and $\pi(i,0,0,\nu)$, $i = \overline{1,R+R_1}$, $\nu = \overline{1,V}$, of the system states exist. Let us form the row vectors $\boldsymbol{\pi}_i$ of these probabilities according to the lexicographic order.

It is well known that the probability vectors $\boldsymbol{\pi}_i$, $i = \overline{0,R+R_1}$, satisfy the following system of linear algebraic equations (equilibrium or Chapman-Kolmogorov equations):

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{R+R_1})Q = \mathbf{0}, \tag{1}$$

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{R+R_1})\mathbf{e} = 1$$

where Q is the infinitesimal generator of the Markov chain $\xi_t,\ t \geq 0$. System (1) is the finite one and there are several numerically stable methods for its solving that effectively use the sparse structure of the generator, see, e.g., [8].

4. Computation of performance measures

The average number of customers in the buffer is $N_c = \sum\limits_{i=1}^{R+R_1} i\boldsymbol{\pi}_i\mathbf{e}$.

The average number of energy units in the buffer is $N_e = \sum\limits_{r=1}^{R} r\boldsymbol{\pi}(0,0,r)\mathbf{e} + \sum\limits_{i=0}^{R+R_1}\sum\limits_{r=1}^{R} r\boldsymbol{\pi}(i,1,r)\mathbf{e}$.

The probability that at an arbitrary moment the server is busy is $P_{busy} = \sum\limits_{i=0}^{R+R_1} \boldsymbol{\pi}(i,1)\mathbf{e}$.

The probability that an arbitrary customer is lost due to impatience is $P_c^{imp-loss} = \frac{1}{\lambda_c}\sum\limits_{i=1}^{R+R_1} i\alpha\boldsymbol{\pi}_i\mathbf{e} = \frac{\alpha N_c}{\lambda_c}$.

The probability that an arbitrary customer is lost at the entrance to the system due to lack of energy is $P_c^{ent-loss} = \frac{1}{\lambda_c}\sum\limits_{i=0}^{R+R_1}\left[\boldsymbol{\pi}(i,0,0)D_1\mathbf{e} + \sum\limits_{r=0}^{\max\{0,i-R_1\}} \boldsymbol{\pi}(i,1,r)(D_1 \otimes I_M)\mathbf{e}\right]$.

The intensity of the output flow of successfully served customers from the system is $\lambda_{out} = \sum\limits_{i=0}^{R+R_1} \boldsymbol{\pi}(i,1)(\mathbf{e}_{(R+1)V} \otimes \mathbf{S}_0)\mathbf{e}$.

The loss probability of an arbitrary customer is $P_c^{loss} = 1 - \frac{\lambda_{out}}{\lambda_c}$.

The probability of the loss of an arbitrary unit of energy (due to the buffer overflow) is computed by $P_e^{loss} = \frac{1}{\lambda_e}\left[\boldsymbol{\pi}(0,0,R)D_2\mathbf{e} + \sum_{i=0}^{R+R_1}\boldsymbol{\pi}(i,1,R)(D_2 \otimes I_M)\mathbf{e}\right]$.

5. Conclusion

In this paper, we considered a novel queueing model describing operation of the node of a wireless sensor network. The problem of computation of the stationary distribution of the states of this model is solved what gives an opportunity to compute the variety of performance measures of the system and numerically analyse the impact of the control parameter.

6. Acknowledgments

This research has been partially supported by RUDN University Program 5-100.

REFERENCES

1. Yick J., Mukherjee B., Ghosal D. Wireless sensor network survey // Computer networks. 2008. V. 52(12), P. 2292–2330.
2. Demirkol I., Ersoy C., Alagoz F. MAC protocols for wireless sensor networks: a survey // IEEE Communications Magazine. 2006. V. 44(4). P. 115–121.
3. Gelenbe E. A sensor node with energy harvesting // ACM SIGMETRICS Performance Evaluation Review. 2014. V. 42(2). P. 37–39, .
4. Patil K., De Turck K., Fiems D. A two-queue model for optimising the value of information in energy-harvesting sensor networks // Performance Evaluation. 2018. V. 119. P. 27–42.
5. Dudin S.A., Lee M.H. Analysis of single-server queue with phase-type service and energy harvesting // Mathematical Problems in Engineering. 2016. V. 2016. ID592794. P. 1–16.
6. Kim Chesoong, Dudin S., Dudin A., Samouylov K. Multi-threshold control by a single-server queuing model with a service rate depending on the amount of harvested energy // Performance Evaluation. 2018. V. 127-128. P. 1–20.
7. Dudin A., Kim C.S., Dudin S. Optimal control by the queue with rate and quality of service depending on the amount of harvested energy as a model of the node of wireless sensor network // Lecture Notes in Computer Science. 2019. V. 11965. P. 165–178.
8. Baumann H., Sandmann W. Multi-server tandem queue with Markovian arrival process, phase-type service times, and finite buffers. // European Journal of Operational Research. 2017. V. 256. P. 187–195.

УДК 621.396

«Изучение сбоев при работе технологии MIMO»

V. M. Antonova[1,2] and A. M. Kuznetsova[1]

[1]Московский государственный технический университет им. Н.Э. Баумана,105005, Москва, 2-я Бауманская ул., д. 5, стр. 1

[2]Институт радиотехники и электроники им. В.А. Котельникова РАН,125009, Москва, ул. Моховая, 11-7

Аннотация

Для увеличения информативности, стабильности и точности в работе систем радиоканала беспроводных каналов связи по технологии MIMO требуется достаточно точное исследование радиолокационных характеристик, а также значимых информационных параметров, особенностей, отраженных от поверхностей и эхо сигналов. Значимое место в корректной работе подобных систем играет исследование и предотвращение сбоев каналов. В данной работе рассмотрен возможный способ упрощения реализации модели системы радиоканала беспроводных каналов связи по технологии MIMO, а также представлены данные зависимости пропускной способности радиоканала беспроводных каналов связи с переключением антенн на передаче и приеме.

Ключевые слова: MIMO, передача данных, разнесенный прием, пространственное мультиплексирование, антенны, пропускная способность радиоканала, комбинирование сигналов.

1. Введение

Под технологией MIMO (англ. Multiple Input Multiple Output) понимается метод пространственного кодирования сигнала, который позволяет увеличивать полосу пропускания канала, передача и прием данных в котором осуществляются системами, состоящими из нескольких антенн.

Передающие и приёмные антенны должны быть разнесены на такое расстояние, которое способно достичь слабой корреляции между соседними антеннами. Здесь имеет место допплеровские характеристики спектра.

Технология MIMO позволяет улучшить характеристики беспроводных систем связи двумя способами:

- разнесением, обеспечивающим повышение помехоустойчивости;

- пространственным мультиплексированием, позволяющим увеличить спектральную эффективность.

Методы разнесения могут и на приемной, и на передающей стороне. На приемной стороне разнесение в системах связи известно, как «разнесенный прием» [1]. Разнесение на передающей стороне реализуется методами пространственно-временного кодирования [2].

Пространственное мультиплексирование позволяет передавать параллельные независимые потоки информации и тем самым повышать спектральную эффективность системы в minMr,Mt раз, где: Mr – число приемных, а Mt – передающих антенн в системе MIMO [2]. Таким образом, спектральная эффективность системы связи MIMO растет линейно с увеличением числа антенн.

К сожалению, применению систем связи MIMO, особенно при большом числе антенн, препятствует высокая сложность их реализации. В этой связи актуальной становится разработка упрощенных алгоритмов формирования и приема сигналов [3]. На приемной стороне комбинирование сигналов осуществляется следующими тремя способами [1].

1) Автовыбор предусматривает выбор одного сигнала, обладающего наиболее высокими показателями отношения сигнал/шум (ОСШ) по сравнению со всеми принимаемыми копиями сигналов в ветвях разнесения.
2) Оптимальное линейное сложение предполагает сложение всех принимаемых копий с весами, которые зависят от ОСШ.
3) Простое сложение с равными весами предусматривает простое сложение принимаемых сигналов, не учитывая ОСШ.

В случае, когда на приеме число радиочастотных трактов реализовано в диапазоне от одного до числа Mr приемных антенн, предполагается использование подмножеств приемных антенн, у которых должны быть скомбинированы сигналы. Подобный подход известен, как гибридный (переключение) и обобщенный выбор антенн [4].

2. Сбои в работе технологии MIMO

Техника Wi-Fi MIMO использует неоднородность помещений и эффекты отражения, что позволяет сделать потоки данных независимыми. Таким образом, в чистом поле MIMO даст гораздо меньше эффекта, чем в офисе, и подобный подход с научной точки зрения следует считать очень конструктивным [5]. Что касается интерпретации данных, благодаря некоторым ухищрениям с модуляцией и более плотной математической обработке кодированных данных как на этапе

передачи, так и на этапе приема, становится возможным сохранение практически полной пропускной способности каждого из каналов, интерференция и взаимные помехи для которых решаются посредством все той же технологии Smart Antenna.

Сломанной называется технология MIMO, у которой не работает одна из нескольких антенн.

3. Разработка и описание модели радиоканала каналов связи в среде Matlab

На рисунке 1 проиллюстрирована структурная схема системы радиоканала беспроводных каналов связи по технологии MIMO.

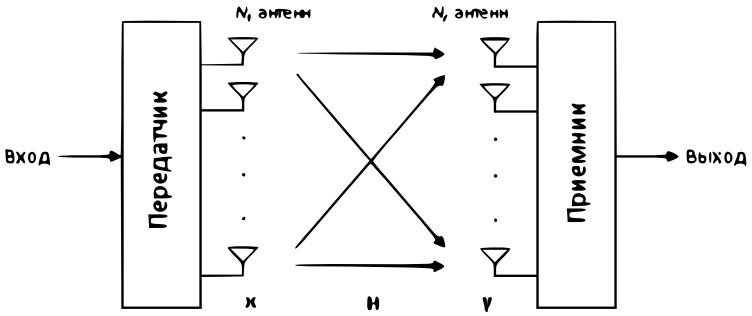


Рис. 1. Структурная схема системы

В ней предполагается реализация временного разделения каналов, а также реализация алгоритмов комбинирования и мультиплексирования антенн в режиме разнесенного приема.

На рисунке 2 представлен вариант реализации модели системы радиоканала беспроводных каналов связи по технологии MIMO в среде Matlab Simulink.

Модель состоит из передатчика, каналов связи, реализующих работу алгоритмов комбинирования и мультиплексирования антенн в режиме разнесенного приема для беспроводных каналов связи, приемника.

Модель реализации передатчика в среде Matlab Simulink представлена на рисунке 3. Передатчик включает в себя источники первичных цифровых сигналов, сумматора, выполняющего в данной реализации функции мультиплексора линий задержки, систем сжатия цифровых потоков и BPSK-модулятор, обеспечивающего функции переноса спектра общего потока на частоту радиоканала. С помощью блоков источников случайных сигналов с равномерным распределением реализуются входные потоки данных. К ним подсистемы сжатия цифрового потока. Данная подсистема при помощи блока Sign преобразует в биполярную

псевдослучайную последовательность случайный поток. В свою очередь последовательность имитирует входной поток данных.

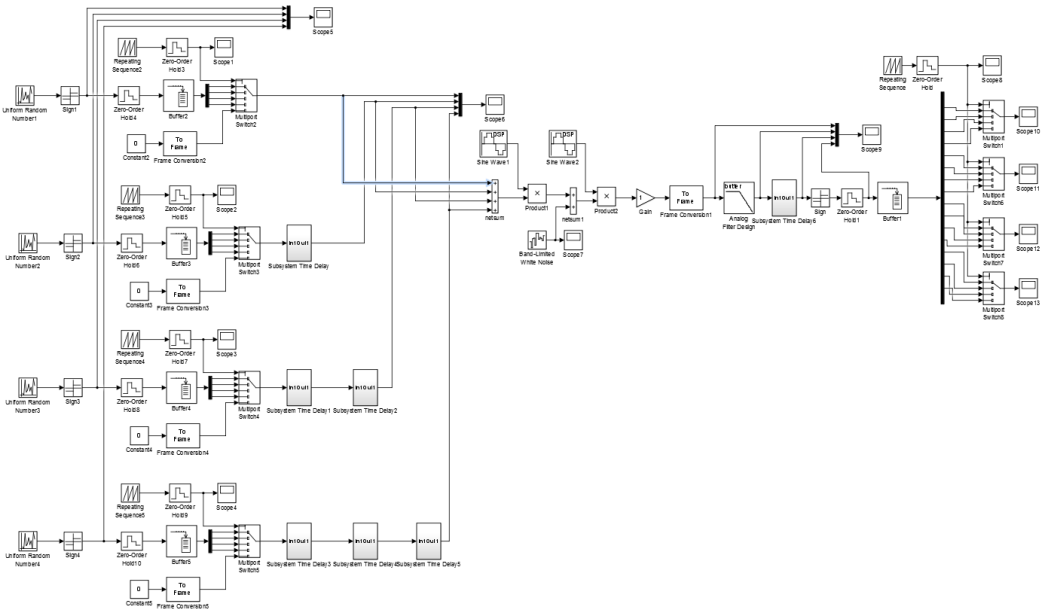


Рис. 2. Функциональная схема модели радиоканала беспроводных каналов связи по технологии MIMO

Данные, с 4-х источников передаются кадрами по 4-е временных слота, каждый слот по 4 бита с каждого источника. После этого данные сжимаются по времени при помощи мультипортового переключателя, мультиплексора и буфера, и далее подаются на сумматор с помощью блоков задержек (Subsystem TimeDelay). Мультиплексор общего потока, образованный подсхемой линии задержки и сумматором, поступает на модулятор, и после этого в канал связи. Модулятор представлен умножителем. На его первый вход поступает сформированный общий поток, на второй –подается гармоническое колебание несущей частоты с генератора.

Модель реализации приемника в среде Matlab Simulink представлена на рисунке 4.

Блоки приемника выполняют обратные операции, которые реализуются в передатчике. Вход приемника представлен демодулятором, возвращающим после фильтра нижних частот общий поток в область нижних частот. В данном случае необходимо соблюдать совпадение опорных частот приемника и генераторов

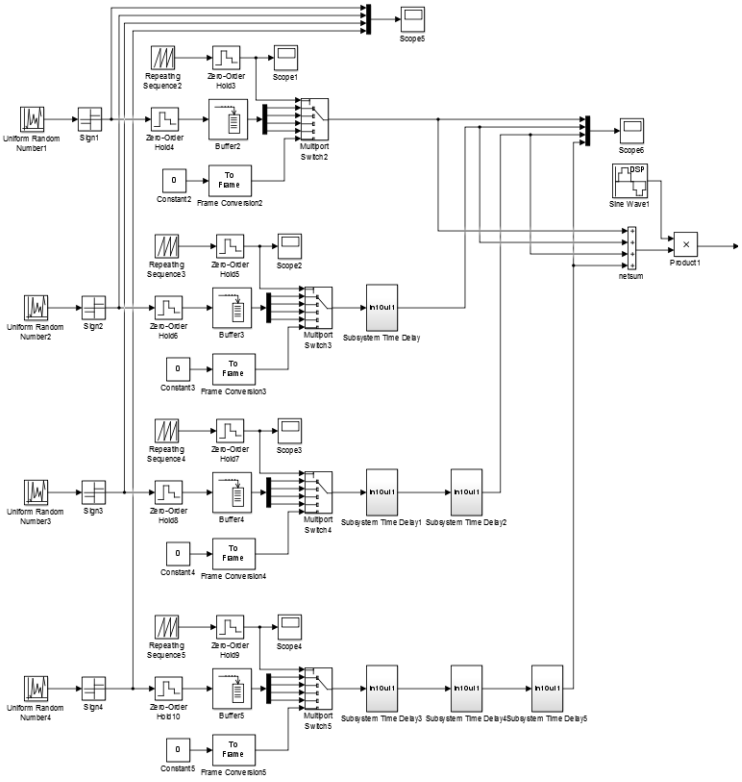


Рис. 3. Модель реализации передатчика в среде Matlab Simulink

передатчика. При помощи блока подсхемы линий задержки демодулированный поток задерживается на заданное число тактов с целью совмещения по времени начала периода генератора с началом кадра, который определяет длительность кадра.

После этого блоком Sign регенерируется демодулированный битовый поток и далее при помощи блоков демультеплексора, буфера и Zero-Order реализуется операция векторизации – параллельного представления кадра (по 4 бита в каждом временном слоте). По причине того, что 4 бита входят в один канальный интервал (временной слот), то выходы демультеплексора группируются по 4, а исходные потоки могут быть получены с выходов мультипортовых переключателей.

Блок осциллографов (Scope) служит в качестве условных получателей. С их помощью можно контролировать динамику процесса и правильность реализации систем по контрольным точкам.

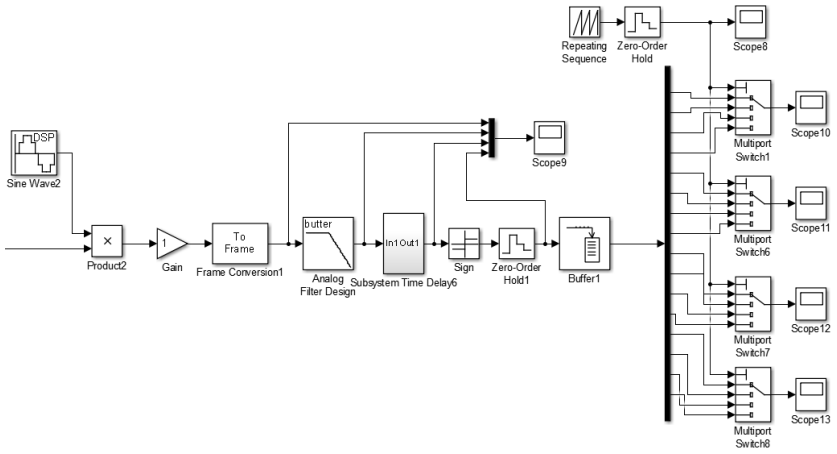


Рис. 4. Модель реализации приемника в среде Matlab Simulink

С помощью моделирования в среде MATLAB объекты каналов системы обеспечивают их компактную реализацию, которую можно конфигурировать и позволять задавать характеристики: допплеровские характеристики спектра, задержку распространения, K-фактор для каналов с замираниями Райса, максимальный сдвиг Доплера, среднее ослабление канала, моделирование ситуации сбоя работы, при обрыве одной из четырех антенн.

Для систем MIMO список параметров расширяется и включает принимающую и передающую корреляционную матрицу, а также число приемных и передающих антенн (до 8).

4. Результаты моделирования

На рисунке 5 приведены зависимости пропускной способности радиоканала беспроводных каналов связи с переключением антенн на передаче и приеме для следующих алгоритмов:

- оптимального по критерию максимума пропускной способности (3);
- оптимального по критерию максимума ОСШ (2);
- использующего на передающей стороне полную информацию о состоянии канала [2];
- не использующего на передающей стороне информацию о канале [2].

На рисунке 6 представлены K-факторы для каналов с замираниями Райса. В данном случае оно будет представлено сбоем первой из 4х антенн.

Полученные результаты формируют следующий вывод:

- только для сильных замираний (при K $<$ 0,5) методика работает;
- для значений значение замираний Райса K $>$0.5 значения теста на гауссовость не более $1,3 * 10^{-2}$, потому в данном случае данные замирания могут быть восприняты за гауссовский шум.

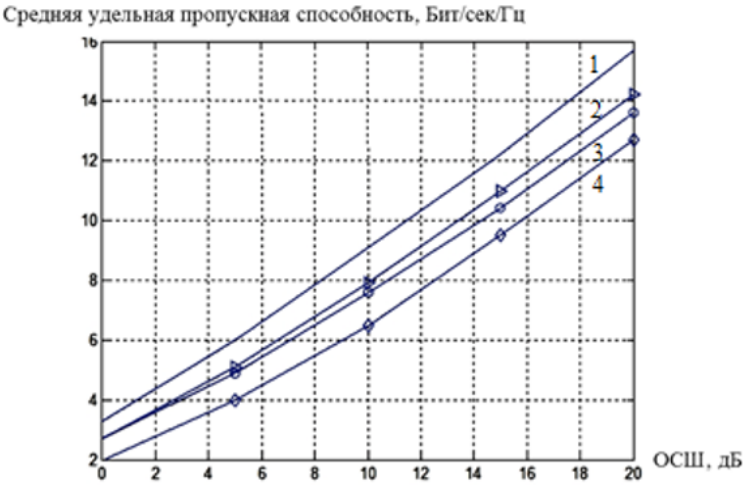


Рис. 5. Пропускная способность каналов связи: 1 - полная информация о состоянии радиоканала; 2 - критерий максимальной пропускной способности (4); 3 - критерий максимуму ОСШ (2); 4 - нет информации о состоянии канала

Рис. 6. K-факторы для каналов при сбое первой из 4х антенн: Красным – без шума; Розовым – ОСШ 3 дБ; Черным – ОСШ 5дБ; Синим – ОСШ 10 дБ

5. Заключение

Таким образом, одним из направлений развития можно выделить применение различных методов переключения антенн, характеризующееся выбором наименьшего числа передающих (приемных) антенн при числе передающих (приемных) трактов. Данное мероприятие упрощает реализацию всей системы, в том числе MIMO, и может быть достигнуто с помощью определенных энергетических потерь [1-4].

Несколько различных копий переданного сигнала получает приемник на приеме в системе связи с разнесением. С целью получения выигрыша в помехоустойчивости эти копии в приемнике комбинируются определенным образом. Представляется возможным комбинирование выбранных сигналов с помощью следующих методов: простым сложением или оптимальным линейным сложением.

Является очевидным, что более высокую помехоустойчивость позволяет получить оптимальное линейное сложение в сравнении с простым сложением. Одновременно нескольких приемных антенн (L ветвей разнесения), обладающих максимальным ОСШ, выбирают при обобщенном выборе антенн. После выбранные сигналы необходимо скомбинировать.

Работа выполнена при финансовой поддержке РФФИ No19-07-00525 А.

ЛИТЕРАТУРА

1. Molisch A., Win M. MIMO Systems with Antenna Selection // IEEE Microwave Mag. – March 2004. – Vol. 5. – P. 46–56.
2. Шувалов, Р. И. Место технологии MIMO в составе беспроводной сети / Р. И. Шувалов. - Текст: непосредственный, электронный // Молодой ученый. - 2019. - № 27 (265). - С. 36-39.
3. Крейнделин В.Б., Хазов М.Л. Проблемы применения технологии переключения антенн в многоантенных системах MIMO. Материалы X международной отраслевой научно-технической конференции “Технологии информационного общества”, 15-16 марта 2016 г, с.230.
4. Hampton J.R. Introduction to MIMO Communications. – UK, Cambridge University Press, 2014. – 288 p.
5. Интернет-источник: http://citforum.ru/nets/wireless/wifi_mimo/ (дата обращения 05.05.2020).

UDC: 681.533.3

Information-processing system for natural gas quality analysis

I.A. Brokarev[1] and S.V. Vaskovskii[2]

[1]National University of Oil and Gas «Gubkin University», 65 Leninsky Prospekt, Moscow, Russia

[2]V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, 65 Profsoyuznaya street, Moscow, Russia

brokarev.i@gubkin.ru, v63v@yandex.ru

Abstract

Task of development of natural gas quality analysis have been studied. The structure of the proposed system is discussed. The functionality and the most important features of component parts of the system are presented. The statistical model selection that is one of the most crucial stages of the proposed method is shown. The comparative analysis of statistical models has been conducted. Algorithms have been developed for correlation analysis of input and output variables. Opportunities for further development of the proposed system are considered.

Keywords: information-processing systems, neural network analysis, natural gas quality analysis, correlation analysis.

1. Introduction

The natural gas quality analysis is an important task for the gas industry. Slight fluctuations of natural gas composition and energy characteristics can lead to unexpected difficulties in calculating its cost indicators. Currently, a wide variety of different natural gas analysis systems are developed. Moreover, many alternative systems that are based on the correlation methods are under development [1]. The possibility to analyze gas quality in real time is the most significant benefit of this class of systems in comparison with systems based on the traditional gas chromatography methods. However, systems that are commonly used in gas industry have a number of drawbacks: expensive specialized equipment, significant amount of time of the analysis, the necessity of regular instrumentation calibration and checkout.

Various statistical models are used in correlation methods because of high complexity of solving the task with traditional computational methods. The choice of

statistical model for the gas quality determination is made by heuristic methods in most cases due to the lack of a general algorithm. That is why comparative analysis of statistical models for the discussed task is an urgent problem that should be solved for reaching the required goal.

This paper provides a structure and description of the main blocks of the proposed information-processing system. The system is based on the method of determination of the properties and the composition of natural gas by measuring of its physical parameters [2]. The conclusions are drawn about further development of the proposed system.

2. Development of the information-processing system for natural gas quality determination

The main structure of the proposed information-processing system is shown in fig. 1. The system consists of three blocks. We suggest using commercially available and relatively inexpensive sensors for natural gas physical parameters measurements to obtain necessary measurement data that are input data of the proposed system. The measurement data include following natural gas physical parameters: speed of sound, thermal conductivity and molar fraction of carbon dioxide. The aim of the system is to determine target natural gas quality parameters using input measurement data.

The first block (pseudogas composition determination) is the main block of the system that contains the majority of features of the proposed system. The task of this block are simplifying the studied object and minimizing amount of measured physical parameters and in its turn amount of applied sensors. This block we will describe below in more detail.

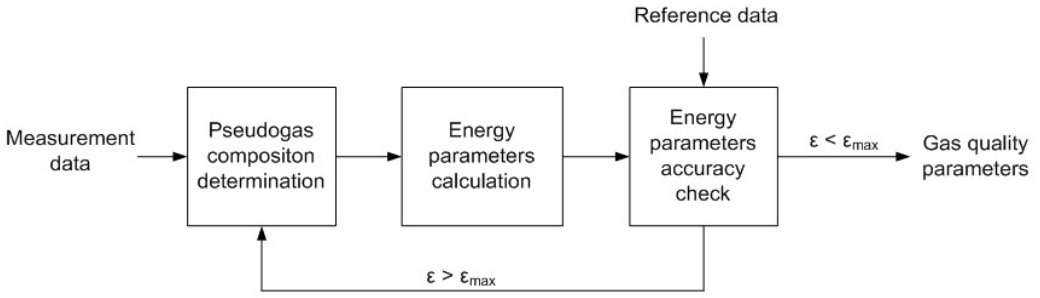


Fig. 1. Main structure of the proposed information-processing system

The obtained equivalent pseudogas composition is transmitted to the next block where energy parameters calculation occurs. To calculate the energy parameters of the gas under study, NIST REFPROP software is used [3]. The target energy

parameters for the discussed task are volumetric superior calorific value and Wobbe index. These parameters along with partial gas composition and relative density are considered to be final gas quality parameters that system should determine. To calculate the quality parameters, the GERG-2008 gas state equation was used at standard temperature and pressure conditions. The amount of output parameters can be decreased to simplify the calculations or increased by adding volumetric inferior calorific value in special cases. The next step involves energy parameters accuracy check that occurs in the corresponding block. The calculated in previous block gas quality parameters are compared with reference data. Any data obtained from traditional natural gas analyzers, e.g. gas chromatographs, can be used as the reference data. The final error parameter ε is calculated to receive deviation of system parameters from reference parameters. That parameter is based on a number of accuracy characteristics including maximum absolute error (MaxAE), mean absolute error (MAE), maximum absolute percentage error (MaxAPE) and mean absolute percentage error (MAPE). In case of final error parameter is less than maximum limiting value ε_{max} the system provides the target gas quality parameters. In the opposite case, the stage of pseudogas composition determination is repeated. That includes a number of procedures that will be carrying out until reaching the desired accuracy.

The first block includes many subblocks that should be described separately. It's structure is shown in fig. 2.

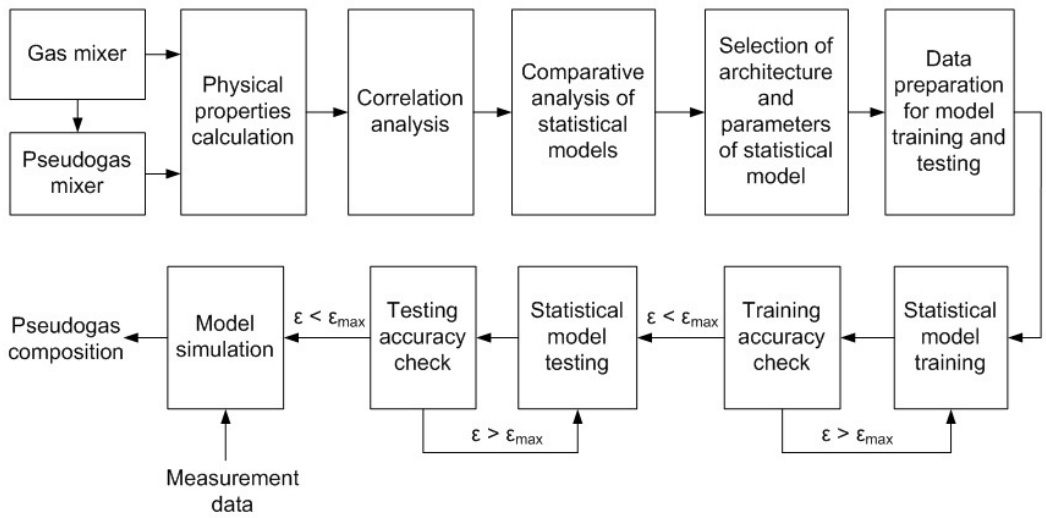


Fig. 2. Structure of pseudogas composition determination block

The gas mixer block forms a natural gas composition. For the data formation a sample of gas mixtures based on the typical natural gas is simulated taking into account the permissible ranges of the molar fractions of the components by sorting out all possible combinations of components. Then the simulated gas mixtures are reduced to equivalent fourcomponent pseudogas mixtures [4] in pseudogas mixer block. The physical properties calculation occurs in corresponding block. That process is similar to energy parameters calculation and includes calculation of theoretical values of parameters that will be used as measured. The correlation analysis is performed in corresponding block for selection of input parameters and elimination of their possible multicollinearity. Pearson correlation coefficients are calculated for each pair of the studied parameters. These coefficients can be used to determine a linear relationship between two parameters. The parameter list can be changed due to correlation analysis results. The next step is to get a number of statistical models usable to solve the task of the analysis for equivalent pseudogas composition and choose the most appropriate statistical model. This choice is based on an analysis of sources that address the problems arising when selecting statistical models for specific tasks of the gas industry, as well as the practical feasibility of implementing the selected statistical models. The following models were selected for comparative analysis based on the study results: multiparameter linear regression, ridge regression, Gaussian process regression and neural network model. All selected statistical models were trained and tested on the same data generated according to the previously described requirements. Then a number of criteria are used to estimate the model performance that are training time, accuracy on training data set and accuracy on test data set. Model with best performance is selected to the next step. Then architecture and parameters of selected statistical model are chosen. Two main architectures are available for neural networks: multilayer perceptron and simple recurrent neural network model. The main tuning parameters are amount of hidden layers (1 in default) and number of neurons in hidden layer (11 in default), activation function for a hidden layer (sigmoidal function in the form of a hyperbolic tangent in default) and activation function for the output layer (linear function in default).

The data preparation stage that occurs in corresponding block includes data division on training, validation and test sets. It should be noted that prior to training the model, the data are cross-validated and normalized in order to be able to be used uniformly and improve the determination results of the statistical model. Moreover, the amount of initial data can be reduced due to desired ranges of gas components. The statistical model training stage involves selected model training using the selectable learning algorithm (Levenberg-Marquardt algorithm in default) on prepared at the previous stage data.

The main tuning parameters of training are maximum number of training epochs (1000 in default), initial learning rate (0.001 in default), maximum validation failures (25 in default). The statistical model testing stage is the model simulation on the data that were not involved in training process. The accuracy check is provided both for training and testing stages. The procedure of accuracy check is similar to accuracy estimation in the energy parameters accuracy check block and involves calculation of error parameter ε. The final subblock of pseudogas composition determination block performs the function of model simulation on measurement data. The final parameter of the described subblock is the composition of equivalent pseudogas that will come to the next block of the proposed system.

3. Testing of the proposed information-processing system

The proposed information-processing system design and testing were carried out in the Matlab 2019b software [5] with NIST REFPROP plug-in. The main aim of testing is to verify system efficiency on the theoretical data. The initial data include 137214 gas mixtures that are based on typical natural gas. The ranges of its components are the following: 90-100% for methane, 0-3% for nitrogen and ethane, 0-1% for carbon dioxide and propane, 0-0.5% for butane and pentane, 0-0.2% for hexane. These mixtures were transformed to fourcomponent pseudogas mixtures. Then physical parameters of both types of mixtures were calculated. The number of parameters exceeds the number of statistical model input parameters to verify the previous results of correlation analysis. Additional physical parameters are dielectric permittivity, dynamic viscosity and isobaric heat capacity. The conducted correlation analysis proved the results of previous research. Speed of sound, thermal conductivity and molar fraction of carbon dioxide were selected as input parameters for the next stages.

The conducted comparative analysis and model tuning showed comparable results with previous papers [6]. The simple recurrent neural network with default architecture and parameters was chosen as working model. On the next step, the initial data was reduced to 111000 gas mixtures by eliminating gas mixtures with composition not close to the natural gas, e.g. pure methane. Then the data was divided on two sets for training and testing. The special data set was formed for simulation stage. It included 200 gas mixtures with calculated physical parameters. The selected recurrent neural network was trained, tested and simulated on the corresponding sets with accuracy characteristics shown in table 1. Each procedure was started only when the previous procedure (training in case of testing and testing in case of simulation) was successful. Carbon dioxide errors were set to zero, because the content of this component is input value and considered to be known.

Component	Characteristic	Stage		
		Training	Testing	Simulation
Methane	MaxAE, %	0.423	0.496	0.581
	MAE, %	0.007	0.008	0.012
	MaxAPE, %	0.531	0.625	0.751
	MAPE, %	0.008	0.010	0.014
Nitrogen	MaxAE, %	0.286	0.374	0.517
	MAE, %	0.011	0.012	0.018
	MaxAPE, %	0.301	0.372	0.521
	MAPE, %	0.013	0.015	0.022
Propane	MaxAE, %	0.251	0.333	0.491
	MAE, %	0.007	0.009	0.014
	MaxAPE, %	0.237	0.299	0.456
	MAPE, %	0.006	0.008	0.014

Table 1. Model accuracy characteristics at the training, testing and simulation stages

The calculated composition of simulation set was transmitted to energy parameters calculation block. Theoretical values of natural gas energy parameters was used as reference data. The volumetric superior calorific value and Wobbe index were calculated using determined pseudogas composition and compared with reference data. The accuracy of determination of target gas quality parameters (deviation between determined by system and reference values) is shown in fig.3 (for volumetric superior calorific value) and in fig. 4 (for Wobbe index). The maximum absolute error of gas quality parameters determination (0.0364 MJ/m3 for calorific value and 0.0914 MJ/m3 for Wobbe index) is less than the allowable error that is equal to 0.1 MJ/m3. The allowable error is permissible deviation of gas quality parameters determination for the first accuracy class according to current regulatory document [7].

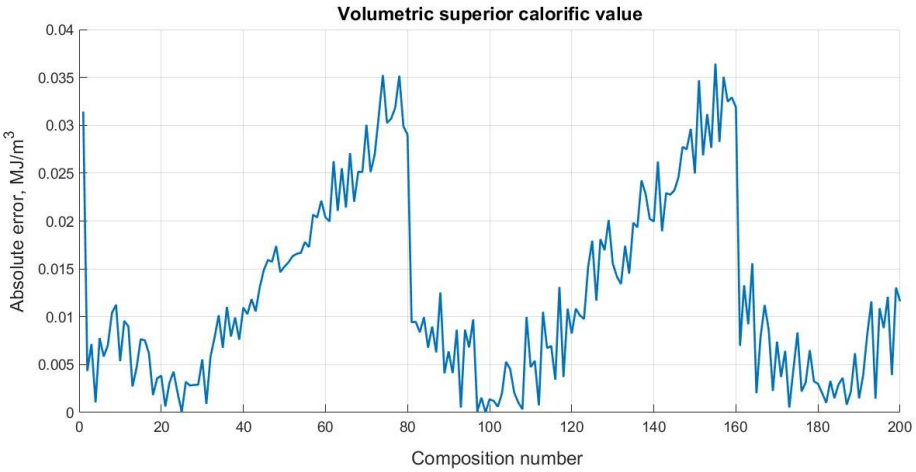


Fig. 3. The accuracy of determination of volumetric superior calorific value by proposed system

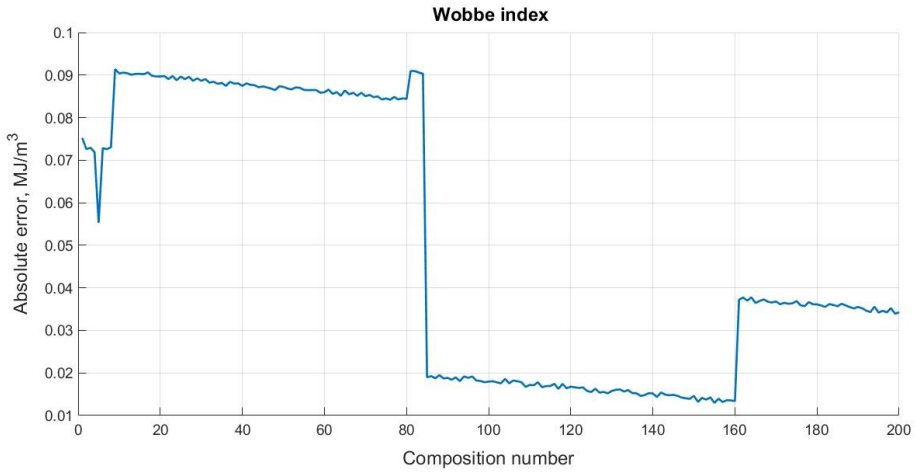


Fig. 4. The accuracy of determination of Wobbe index by proposed system

4. Conclusion

The information-processing system for determining the natural gas quality parameters was presented. The main advantages of the proposed system in comparison with the commonly used gas quality determination methods are the high adaptability and capability of operation in real time. The target gas quality properties, volumetric superior calorific value and Wobbe index determined by the system were compared

with reference data. The system showed acceptable performance on theoretical data. Further research is required in the field of testing the model on experimental data and adjusting system algorithms to solve the task of analyzing specific gas mixtures.

REFERENCES

1. Dörr H., Koturbash T., Kutcherov V. Review of impacts of gas qualities with regard to quality determination and energy metering of natural gas // Measurement Science and Technology. 2019. V. 30. №2. P. 1–20.
2. Koturbash T. T., Brokarev I. A. Method for determining the properties and composition of natural gas by measuring its physical parameters // Sensors and systems. 2018. №6. P. 43–50.
3. NIST REFPROP Software, `https://www.nist.gov/srd/refprop` (Accessed April 1, 2020).
4. Koturbash T. T., Brokarev I. A. Comparative analysis of the physical properties of natural gas and equivalent pseudogas mixtures // Sensors and systems. 2019. №3. P. 7–13.
5. Matlab 2019a Software, `https://www.mathworks.com` (Accessed April 1, 2020).
6. Vaskovskii S. V., Brokarev I. A. Comparative analysis of statistical models for the task of natural gas composition analysis // Information Technologies and Computing Systems. 2020. №1. P. 34–43.
7. ISO 15971:2008. Natural gas — Measurement of properties — Calorific value and Wobbe Index // International Organization for Standardization. 2008. 50 p.

UDC: 519.248

Channel Switching Strategies for multistep Markovian Controllable Queuing Systems (QS) Problems

A.S. Mandel, V.A. Laptin[1,2]

[1]V.A. Trapeznikov Institute of Control Sciences RAS, Profsoyuznaya 65, Moscow, Russia

[2]M.V. Lomonosov Moscow State University, Lenin hills 1, Moscow, Russia

almandel@yandex.ru, straqker@bk.ru

Abstract

This paper deals with a controllable queuing system in which the number of switching service channels monitor and modify at control time points spaced apart by a fixed time step. At transition from step to step, the intensity of the simplest incoming flow changes in accordance with a Markov's chain. The system is in a stationary mode between the steps. A cost function is the minimization of the total average cost of the system over a multi-step planning period. The problem is to find a channel switching strategy. The parametric structure of an optimal strategy significantly simplifies its construction.

Keywords: controllable queuing systems, Markovian incoming flow, switching strategies, strategy parameterization.

1. Introduction

The present authors consider a problem of the theory of controllable queuing systems (QS) which develops and generalizes problem statements, that have been discussed in works presented at conferences DCCN15 [1],DCCN17 [2] and DCCN18 [3]. In the investigated QS the number of service channels can change at the moments of control, standing apart from each other by the value of fixed time step. In this case, as in [1], it is believed that the QS receives the simplest incoming flow, the intensity of which at the moments of control undergoes sudden changes, taking a finite number of values λ_i, $i \in \overline{1,k}$, from a discrete set Λ. We use the assumption that the duration of the control step (which is what we choose for the unit of time) is sufficient to establish in the QS a stationary, in the probabilistic sense, mode of operation.

The aim is to form a strategy for switching service channels (disabling redundant service channels or introducing backup channels) in order to minimize the average cost of QS in a given N-step planning period.

2. Problem statement

We discuss the QS, in which the number of service channels is a controllable value and can be changed at control moments periodically located on the time axis (with step 1) of control of the QS state. A matrix of transition probabilities of the corresponding homogeneous Markov's chain $P = \|p_{ij}\|$, is given, where p_{ij} is the probability of transition (at the moment of control) from the intensity λ_i, $i \in \overline{1,k}$, at the previous step to the intensity λ_j, $j \in \overline{1,k}$ at the next step.

We use the assumption that the duration of the control step is sufficient to establish a stationary, in the probabilistic sense, operation mode in the QS in question at this step. If the intensity of the incoming flow at this step is equal to λ_i, while the intensity of service in one service channel is μ, then it is obvious [4, 5], that the number of service channels in the QS should be chosen to satisfy the following inequality:

$$u \geq u_{\text{critical}}(\lambda_i) = \underline{u_i} = \left[\frac{\lambda_i}{\mu}\right] + 1. \tag{1}$$

Let us introduce a control quality function which for given: (a) number of steps n, remaining till the end of the planning period ($n \leq N$), (b) intensity of the incoming flow λ_i, $i \in \overline{1,k}$, (c) actual number of service channels m and (d) decision on the switched number of service channels u is described by the value of average total cost $C_n(\lambda_i, m, u)$. The aim is to minimize $C_n(\lambda_i, m, u)$ by choosing the channel switching strategy. This cost is mathematical expectation of a sum of one-step costs at the remaining n steps, along the trajectory of the incoming flow whose intensity changes in the Markovian jump-wise manner.

Let us write an equation for one-step costs in the first step $C^{(1)}(\lambda_i, m, u)$:

$$C^{(1)}(\lambda_i, m, u) = C_{\text{oper}} + C_{\text{queue}} + C_{\text{switch}}. \tag{2}$$

Here, the operating costs of the active systems in the first step $C_{\text{oper}} = c_1 u$, the cost of queuing in the stationary mode is $C_{\text{queue}} = d\bar{l}_{\text{queue}}$, where $\bar{l}_{\text{queue}}$ is the average queue length, while the switching cost C_{switch} can be presented as:

$$C_{\text{switch}} = \begin{cases} A_1, & \text{if } u > m; \\ 0, & \text{if } u = m; \\ A_2 + c_2(m-u), & \text{if } u < m. \end{cases} \tag{3}$$

We will now use classical results [4, 5] to write down:

$$\bar{l}_{\text{queue}} = \left[\sum_{k=1}^{u-1} \frac{(u\rho_i)^k}{k!} + \frac{(u\rho_i)^u}{u!(1-\rho_i)}\right]^{-1} \frac{(u\rho_i)^u \rho_i}{u!(1-\rho_i)^2}, \tag{4}$$

where $\rho_i = \frac{\lambda_i}{u\mu}$.

Now, if under the assumptions made we denote through $C_n^*(\lambda_i, m)$ the minimum possible value of the average total cost in the last n steps of the control process, it is logical to write the following system of discrete dynamic programming equations:

$$C_1^*(\lambda_i, m) = \min_{u \geq \underline{u_i}} C^{(1)}(\lambda_i, m, u), \tag{5}$$

$$C_n^*(\lambda_i, m) = \min_{u \geq \underline{u_i}} (C^{(1)}(\lambda_i, m, u) + \alpha \sum_{j=1}^{k} p_{ij} C_{n-1}^*(\lambda_j, u)), \tag{6}$$

where $i \in \overline{1, k}$, α is the discount factor, $0 \leq \alpha \leq 1$, while $n \in \overline{2, N}$.

3. *A*-convexity in discrete problems of optimization

We consider function $g(i)$ of the variable i, taking values from the finite segment of the natural series: $i = 1, 2, \ldots, k$.

Definition 1. The function $g(i)$ is called convex (downward) if for all natural i and j:

$$g(i + j) - g(i) - \Delta^{(1)} g(i) \times j \geq 0, \tag{7}$$

where $\Delta^{(1)} g(i)$ is the first difference of $g(i)$ function in point i: $\Delta^{(1)} g(i) = g(i) - g(i - 1)$.

The necessary and sufficient condition for the convexity (downward) of the function is fulfilling the inequality $\Delta^{(2)} g(i) \geq 0$, where $\Delta^{(2)} g(i)$ is the second difference of $g(i)$ function in point i. In fact, let j from definition 1 be 1, then formula (7) is written as $\Delta^{(1)} g(i+1) \geq \Delta^{(1)} g(i)$, i.e. the first difference of $g(i)$ is rejected by the monotonically increasing function, i.e.$\Delta^{(2)} g(i) \geq 0$. Prove of the sufficiency is similar.

Definition 2. * Function $g(i)$ is called A-convex ($A \geq 0$), if for all natural i and j:

$$A + g(i + j) - g(i) - \Delta^{(1)} g(i) \times j \geq 0. \tag{8}$$

For functions of discrete variables, all the properties of A-convex functions noted by Herbert Scarf [7] are retained. Namely:

Property 1. If $g(i)$ is A-convex, then for any natural j the function $g(i + j)$ is also A-convex.

*The concept of A-convexity was proposed by Herbert Scarf [7] to analyze the properties of optimal decision making (control) strategies in inventory control problems in the presence of fixed supply costs that do not depend on the size of supply, adding to them the amounts determined by the size of the supply batch.

Property 2. If function $g_1(i)$ is A_1-convex, and function $g_2(i)$ is A_2-convex, then for any $\theta_1, \theta_2 > 0$, function $g(i) = \theta_1 g_1(i) + \theta_2 g_2(i)$ is also $(\theta_1 A_1 + \theta_2 A_2)$-convex.

Property 3. If function $g(i)$ is A_1-convex, then it is also A_2-convex for any $A_2 > A_1$.

Property 4. Let i be a random value with distribution $p_{i1}^k, p_i > 0, \sum_{i=1}^k p_i = 1$, and let function $g(i)$ be A-convex. Then, function $\alpha \sum_{i=1}^k p_i g(i)$, with $\alpha \geq 0$, is also A-convex.

4. Properties of optimal channel switching strategies

Let us revisit the problem of investigating properties of solutions to the system of dynamic programming equations (5)–(6). Some words about the plan of this section to present the following results:

- review of the qualitative features of "myopic" inventory control strategies;
- broadening of the concept of A-convexity for the problems of the examined type;
- transfer of results obtained for "myopic" strategies to multistep (dynamic) control problems.

4.1. "Myopic" strategies of control. This subsection briefly (with minor corrections) retells the content of the work [3] by the same authors.

In order to build an optimal "myopic" switching strategy, one has to find:

$$\min_{u \geq \underline{u_i}} C^{(1)}(\lambda_i, m, u) =$$

$$= \min_{u \geq \underline{u_i}} \{c_1 u + d\bar{l}_{\text{queue}} + \min \begin{cases} A_1 \mathbb{1}(u-m), \\ 0, \\ A_2 \mathbb{1}(m-u) + c_2(m-u), \end{cases} \}, \quad (9)$$

where $\mathbb{1}(u)$ denotes the function of a unit jump (Heaviside's function) which is equal to 1, if $u > 0$, or 0 in all other cases.

The main results obtained in [3] are that for each state $i = 1, 2, ..., k$, there are five critical parameters that fully determine the "myopic" inventory control strategy: $\underline{u_i}$ (defined by formula (1)), $r_{1,i}^{(1)}$, $R_{1,i}^{(1)}$, $r_{1,i}^{(2)}$ and $R_{1,i}^{(2)}$. Moreover, the last four parameters are arranged as follows: $r_{1,i}^{(1)} < R_{1,i}^{(1)} \leq R_{1,i}^{(2)} < r_{1,i}^{(2)}$, while a "myopic" channel switching strategy obeys the formula†:

†The application of formula (9) is associated with certain caveats, the essence of which is that in some cases listed in [3], the parameter $R_{1,i}^{(1)}$ is assigned equal to $\underline{u_i}$. There are other nuances as well.

$$u = \begin{cases} R_{1,i}^{(1)}, & \text{if } m \le r_{1,i}^{(1)} \text{ (switching on)}, \\ m, & \text{if } r_{1,i}^{(1)} < m \le R_{1,i}^{(1)}, \\ m, & \text{if } R_{1,i}^{(2)} \le m < r_{1,i}^{(2)}, \\ R_{1,i}^{(2)}, & \text{if } m \ge r_{1,i}^{(2)} \text{ (switching off)}. \end{cases} \tag{10}$$

It should be noted that in [3], as already noted, the discussion was conducted at a purely qualitative level and did not take into account, in particular, the discreteness of one of the QS state variables. Using the technique briefly outlined in Section 3 of this paper, it is possible to reformulate and strictly prove all the statements made in [3]. In contrast to [3], in this paper, the last four characteristic parameters of an optimal "myopic" strategy have one more sub-index, in this case equal to 1. In one of the subsequent paragraphs, this unit will transform into n the number of steps left until the end of the planning period for the multistep problem).

If this has not been the case, we would have to take care of "sliding" modes, when at first we would have to switch off part of the channels (bringing the number of active channels to $R_{1,i}^{(2)}$), but immediately after that, under certain circumstances, we would have to switch on part of the channels, returning the QS to another parameter – $R_{1,i}^{(1)}$), etc. It does not seem to be much of a trouble, but there is one more and probably more difficult task.

The above fact necessitates one more decomposition of $C_n(\lambda_i, m, u)$ criterion of the initial problem (5)–(6). In other words, two alternative managerial decision options are considered separately: (a) to enable additional channels or, conversely, (b) to disable some active channels. At the same time, the fee for enabling is described by the function $B_{\text{switch on}}(u) = c_1 u + d\bar{l}_{\text{queue}}$, while the fee for disabling is described by the function $B_{\text{switch off}}(u) = (c_1 - c_2)u + c_2 m + d\bar{l}_{\text{queue}}$, where $\bar{l}_{\text{queue}}$ is obtained from the formula (4) and all designations coincide with those introduced above. It is the type of functions that gives rise to the inequality $R_{1,i}^{(1)} \le R_{1,i}^{(2)}$. It is also important to note that $R_{1,i}^{(1)}$ is the point of the absolute minimum of the one-step cost function when channels are enabled, and $R_{1,i}^{(2)}$ is the point of absolute minimum of one-step cost function when channels are disabled.

4.2. Broadening of the notion of *A*-convexity. The proof of the facts listed in paragraph 4.1 based on the analogy between the problems of inventory control theory and those of queuing systems theory, which was important feature discussed in the present and previous works of the first author. An example of the original analysis of such analogies is the paper [8]. The problem of inventory control discussed

in [8] the authors called "fantastic"[‡] . Nonetheless, we cannot but admit that for the class of problems under consideration (both "fantastic" from the theory of inventory control and the problem of channel switching of QS theory), some unpreparedness and unsuitability of the existing theory of optimization and the theory of convexity for the solution of these problems was evident. The point is that in the mathematical description of the class of problems in question, not one but two different concepts of optimality (optimality criteria) are used. Speaking a language adequate to this class of problems, it is the *optimality of enabling channels* (when placing orders in inventory control theory), when we move along the axis of an integer variable u *left-to-right*, and the *optimality of disabling channels* (returning goods in inventory control theory), when we move along the axis of the variable *right-to-left*. We shall wrap this remark into new definitions.

Definition 3. Function $g(i)$ is referred to as A-convex from the left ($A \geq 0$), if for all natural i and j, which are to the left of the point of its absolute minimum,

$$A + g(i+j) - g(i) - \Delta^{(1)}g(i) \times j \geq 0.$$

Definition 4. Function $g(i)$ is referred to as A-convex from the right ($A \geq 0$), if for all natural i and j, which are to the right of the point of its absolute minimum,

$$A + g(i+j) - g(i) - \Delta^{(1)}g(i) \times j \geq 0.$$

We now can state that the enabling function in the one-step ("myopic") problem is A_1-convex from the left, whereas the disabling function for the same problem is A_2-convex from the right.

4.3. Multi-step dynamic problems of channel switching. Let us return to the solution of the multistep problem of channel switching control, which is described in equations (5)–(6). We repeat the notation of goal function in equation (6):

$$C_n^*(\lambda_i, m) = \min_{u \geq \underline{u_i}} (C^{(1)}(\lambda_i, m, u) + \alpha \sum_{j=1}^{k} p_{ij} C_{n-1}^*(\lambda_j, u)). \tag{11}$$

The expression in braces in the right hand side (r.h.s.) of this equation, regardless of which alternative is estimated (on or off), contains a non-modifying term $\alpha \sum_{j=1}^{l} p_{ij} C_{n-1}^*(\lambda_j, u)$. The one-step add-on $C^{(1)}(\lambda_i, m, u)$, by contrast, varies when the alternative is changed. From this we can conclude that there is every reason to believe that the expression in braces in the r.h.s. of formula (6) (also two-variant) is simultaneously A_1-convex from the left and A_2-convex from the right.

[‡]Fantastic in [8] was the assumption that a warehouse could not only submit replenishment orders, but also return the goods to their suppliers.

Assume mathematical induction that this statement is true for any n. As it follows from [3], for the case $n = 1$ it is so. Suppose now that this statement is true for the case $n - 1$. Further proof of the truth of this statement for the case n is quite tedious (in terms of mathematical computation), but, in fact, is carried out according to the classical scheme of proof of optimality, but for two-level strategies [6, 7].

As noted above, an important feature that helps avoiding the risk of "pitfalls", like sliding modes, is the ordering of the local minimum points of alternative functions in the form of inequalities $R^{(1)}_{n,i} \leq R^{(2)}_{n,i}$, $n \in \overline{1,N}, i \in \overline{1,k}$. It should be admitted here that the authors managed to prove this statement for strictly $n = 1$ case only. As for the arbitrary value of n, we can only argue that in the entire range of computational experiments performed there was not a single case of failure to fulfill these inequalities. The authors hope that in the near future they will manage to find a strong solution to this problem.

5. Conclusion

The paper concerns with channel switching strategies in a multi-line queuing system with a Markov description of the input flow intensity change process. The objective function in strategy development is to minimize the total average cost in a multistep planning period. Thus it is assumed, that the procedure of channel switching (enabling or disabling) leads to the expenses which consist of the fixed payments and costs which size depends on the number of enabled or disabled channels. As a result, the optimal channel switching strategies turn out to be parametric and the optimal solution at each step depends only on four characteristic parameters. The proof process required introducing some generalizations of the A-convexity concept, classic for the operations research.

REFERENCES

1. Mandel A. Econometric Models of Controllable Multiple Queuing Systems / Proceedings of the 18th International Conference, Distributed Computer and Communication Networks (DCCN 2015, Moscow, Russia). Geneva: Springer, 2016. – P. 296–304.
2. Mandel, A., Bakulin, K. Models of controllable multiple queuing systems for channel switching myopic strategies / Proceedings of the 20th International Conference, Distributed Computer and Communication Networks (DCCN 2017, Moscow, Russia). M.: Tecnhosphera, 2017. – P. 534–542. (In Russian)
3. Mandel A., Laptin V. Myopic Channel Switching Strategies for Stationary Mode: Threshold Calculation Algorithms / In: In: Vishnevskiy V., Kozyrev D. (eds),

Distributed Computer and Communication Networks. DCCN 2018. Communications in Computer and Information Science, vol. 919. Geneva: Springer, 2018. – DOI: 10.1007/978-3-319-99447-5_35
4. Gnedenko, B., Kovalenko, I. Introduction to the queuing systems theory. – M.: Nauka, 1966. (In Russian).
5. Vishnevskiy, V. Theoretical Principles of Computer Networks Design. M.: Tecnhosphera 2003. (In Russian)
6. Hadley, G, Whitin, T.M. Analysis of Inventory Control Systems. – Englwood Cliffs: Prentice Hall Inc., 1967.
7. Arrow, K., Karlin, S., and Scarf, H. Studies in the Mathematical Theory of Inventory and Production. – Stanford: Stanford University Press, 1958.
8. Mandel A., Granin S. Investigation of analogies between the problems of inventory control and the problems of the controllable queuing systems // In: Proceedings of 2018 Eleventh International Conference "Management of large-scale system development" (MLSD'2018). 2018. – IEEE, 2018. – P. 1–4. – DOI: 10.1109/MLSD.2018.8551852

УДК: 519.248

Стратегии переключения каналов в многошаговых марковских задачах управления СМО

А.С. Мандель, В.А. Лаптин[1,2]

[1]ИПУ РАН им. В.А. Трапезникова РАН, Профсоюзная ул., д. 65, Москва, Россия
[2]МГУ им. М.В. Ломоносова, Ленинские горы, д. 1, Москва, Россия

almandel@yandex.ru, straqker@bk.ru

Аннотация

Рассматривается управляемая система массового обслуживания, в которой число рабочих каналов обслуживания может изменяться в моменты контроля, отстоящие друг от друга на фиксированный временной шаг. При переходе от шага к шагу интенсивность простейшего входящего потока изменяется в соответствии с некоторой марковской цепью. Считается, что в интервале между шагами в системе устанавливается стационарный режим. В качестве целевой функции используется минимизация суммарных средних затрат на многошаговом периоде планирования. Строится стратегия переключения каналов. Выявлена параметрическая структура оптимальной стратегии, что существенно упрощает ее построение.

Ключевые слова: управляемые системы массового обслуживания, марковский входящий поток, стратегии переключения, параметризация стратегий.

1. Введение

Рассматривается задача теории управляемых систем массового обслуживания (СМО), которая развивает и обобщает постановки задач, что были рассмотрены в работах, представленных на конференциях DCCN15 [1], DCCN17 [2] и DCCN18 [3]. В исследуемой СМО число рабочих каналов обслуживания может изменяться в моменты контроля, отстоящие друг от друга на величину фиксированного временного шага. При этом, как и в работе [1], считается, что в СМО поступает простейший входящий поток, интенсивность которого λ в моменты контроля претерпевает скачкообразные изменения, принимая конечное число k значений λ_i, $i \in \overline{1,k}$, из дискретного множества Λ. Предполагается, что длительности шага контроля (его-то мы и выбираем за единицу измерения времени) достаточно для того, чтобы в СМО установился стационарный в вероятностном смысле режим

функционирования. Задача заключается в том, чтобы сформировать стратегию переключения рабочих каналов (отключения лишних работающих каналов или введение в действие резервных каналов), чтобы минимизировать средние затраты СМО на заданном N-шаговом периоде планирования.

2. Постановка задачи

Итак, рассматривается СМО, в которой число рабочих каналов обслуживания является управляемой величиной и может быть изменено в периодически (с шагом 1) расположенные на оси времени моменты контроля за состоянием СМО. Задана матрица вероятностей перехода соответствующей однородной марковской цепи $P = \|p_{ij}\|$, где p_{ij} – это вероятность перехода (в момент контроля) от интенсивности λ_i, $i \in \overline{1,k}$, на предыдущем шаге к интенсивности λ_j, $j \in \overline{1,k}$ на следующем шаге.

Считается, что длительность шага контроля достаточна для того, чтобы в рассматриваемой СМО на данном шаге установился стационарный в вероятностном смысле режим функционирования. Если на данном шаге интенсивность входящего потока равна λ_i, а интенсивность обслуживания на одном рабочем канале составляет μ, то, очевидно [4, 5], что число рабочих каналов в СМО должно выбираться удовлетворяющим следующему неравенству:

$$u \geq u_{\text{крит}}(\lambda_i) = \underline{u_i} = \left[\frac{\lambda_i}{\mu}\right] + 1. \tag{1}$$

Введем функционал качества управления, который при заданных: (а) числе шагов n, оставшихся до конца периода планирования ($n \leq N$), (б) интенсивности входящего потока λ_i, $i \in \overline{1,k}$, (в) фактическом числе рабочих устройств m и (г) принятии решения о включаемом числе рабочих устройств u – описывается величиной средних суммарных затрат $C_n(\lambda_i, m, u)$. $C_n(\lambda_i, m, u)$ предстоит минимизировать за счет выбора стратегии переключения каналов. Эти затраты представляют собой математическое ожидание от суммы одношаговых затрат на n оставшихся шагах, взятое по траектории входящего потока, интенсивность которого совершает марковские скачки.

Выпишем выражение для одношаговых затрат на первом шаге $C^{(1)}(\lambda_i, m, u)$:

$$C^{(1)}(\lambda_i, m, u) = C_{\text{экспл}} + C_{\text{очереди}} + C_{\text{переключ}}. \tag{2}$$

Здесь затраты на эксплуатацию рабочих устройств на первом шаге $C_{\text{экспл}} = c_1 u$, затраты на очередь в стационарном режиме равны $C_{\text{очереди}} = d\bar{l}_{\text{очереди}}$, где $\bar{l}_{\text{очереди}}$ – это средняя длина очереди, а затраты на переключения $C_{\text{переключ}}$ можно представить в следующем виде:

$$C_{\text{переключ}} = \begin{cases} A_1, & \text{если } u > m; \\ 0, & \text{если } u = m; \\ A_2 + c_2(m-u), & \text{если } u < m. \end{cases} \tag{3}$$

Воспользуемся классическими результатами [4, 5], чтобы записать:

$$\bar{l}_{\text{очереди}} = \left[\sum_{k=1}^{u-1} \frac{(u\rho_i)^k}{k!} + \frac{(u\rho_i)^u}{u!(1-\rho_i)}\right]^{-1} \frac{(u\rho_i)^u \rho_i}{u!(1-\rho_i)^2}, \tag{4}$$

где $\rho_i = \frac{\lambda_i}{u\mu}$.

Теперь, если в сделанных предположениях обозначить через $C_n^*(\lambda_i, m)$ минимально возможное значение средних суммарных затрат на последних n шагах процесса управления, то нетрудно записать следующую систему уравнений дискретного динамического программирования:

$$C_1^*(\lambda_i, m) = \min_{u \geq \underline{u_i}} C^{(1)}(\lambda_i, m, u), \tag{5}$$

$$C_n^*(\lambda_i, m) = \min_{u \geq \underline{u_i}} (C^{(1)}(\lambda_i, m, u) + \alpha \sum_{j=1}^{k} p_{ij} C_{n-1}^*(\lambda_j, u)), \tag{6}$$

где $i \in \overline{1,k}$, α – коэффициент дисконтирования, $0 \leq \alpha \leq 1$, а $n \in \overline{2,N}$.

3. *A*-выпуклость в дискретных задачах оптимизации

Рассматривается функция $g(i)$ от переменной i, принимающей значения из конечного отрезка натурального ряда: $i = 1, 2, \ldots, k$.

Определение 1. *Функция $g(i)$ называется выпуклой (вниз), если для всех натуральных i и j:*

$$g(i+j) - g(i) - \Delta^{(1)} g(i) \times j \geq 0, \tag{7}$$

где $\Delta^{(1)} g(i)$ – первая разность функции $g(i)$ в точке i : $\Delta^{(1)} g(i) = g(i) - g(i-1)$.

Необходимым и достаточным условием выпуклости (вниз) функции является выполнение неравенства $\Delta^{(2)} g(i) \geq 0$, где $\Delta^{(2)} g(i)$ – вторая разность функции $g(i)$ в точке i. В самом деле, пусть j из определения 1 равно 1, тогда формула (7) записывается как $\Delta^{(1)} g(i+1) \geq \Delta^{(1)} g(i)$, то есть первая разность $g(i)$ отказывается монотонно возрастающей функцией, то есть $\Delta^{(2)} g(i) \geq 0$. Аналогично доказывается и достаточность.

Определение 2. * *Функция $g(i)$ называется A-выпуклой ($A \geq 0$), если для всех натуральных i и j:*

$$A + g(i+j) - g(i) - \Delta^{(1)} g(i) \times j \geq 0. \tag{8}$$

Для функций от дискретных переменных сохраняются все свойства A-выпуклых функций, подмеченные Гербертом Скарфом [7]. А именно:

Свойство 1. *Если функция $g(i)$ A-выпукла, то при любом натуральном j функция $g(i+j)$ также A-выпукла.*

Свойство 2. *Если функция $g_1(i)$ A_1-выпукла, а функция $g_2(i)$ A_2-выпукла, то при любых $\theta_1, \theta_2 > 0$ функция $g(i) = \theta_1 g_1(i) + \theta_2 g_2(i)$ также является $(\theta_1 A_1 + \theta_2 A_2)$ - выпуклой.*

Свойство 3. *Если функция $g(i)$ A_1-выпукла, то она и A_2-выпукла для любого $A_2 > A_1$.*

Свойство 4. *Пусть i – случайная величина с распределением $p_i^k, p_i > 0, \sum_{i=1}^k p_i = 1$, и пусть функция $g(i)$ A-выпукла. Тогда функция $\alpha \sum_{i=1}^k p_i g(i)$, где $\alpha \geq 0$, также A-выпукла.*

4. Свойства оптимальных стратегий переключением каналов

Вернемся к задаче изучения свойств решений системы уравнений динамического программирования (5)–(6). Несколько слов о плане этого раздела, в котором будут представлены следующие результаты:

- анализ качественных особенностей «близоруких» стратегий управления запасами;
- расширение понятия А-выпуклости для задач рассматриваемого типа;
- перенос результатов, полученных для «близоруких» стратегий на многошаговые (динамические) задачи управления.

4.1. «Близорукие» стратегии управления запасами. В этом подразделе коротко (с небольшими коррекциями) пересказывается содержание работы [3] тех же авторов.

Итак, для того, чтобы построить оптимальную «близорукую» стратегию переключений необходимо найти:

*Понятие А-выпуклости было предложено Г. Скарфом [7] для анализа свойств оптимальных стратегий принятия решений (управления) в задачах логистики запасов при наличии фиксированных цен поставки, не зависящих от размера поставки, с добавлением к ним сумм, обусловленных размером партии поставки.

$$\min_{u\geq\underline{u_i}} C^{(1)}(\lambda_i, m, u) =$$

$$= \min_{u\geq\underline{u_i}}\{c_1 u + d\bar{l}_{\text{очереди}} + \min\begin{cases} A_1\mathbb{1}(u-m), \\ 0, \\ A_2\mathbb{1}(m-u) + c_2(m-u), \end{cases}\}, \quad (9)$$

где через $\mathbb{1}(u)$ обозначена функция единичного скачка, функция Хэвисайда, которая равна 1, если $u > 0$, и равна 0 в остальных случаях.

Основные результаты, полученные в работе [3], состоят в том, что для каждого состояния $i = 1, 2, \ldots, k$, существует 5 критических параметров, которые полностью определяют «близорукую» стратегию управления запасами: $\underline{u_i}$ (определен формулой (1)), $r_{1,i}^{(1)}$, $R_{1,i}^{(1)}$, $r_{1,i}^{(2)}$ и $R_{1,i}^{(2)}$. При этом последние 4 параметра упорядочены следующим образом: $r_{1,i}^{(1)} < R_{1,i}^{(1)} \leq R_{1,i}^{(2)} < r_{1,i}^{(2)}$, а «близорукая» стратегия переключения каналов реализуется по формуле[†]:

$$u = \begin{cases} R_{1,i}^{(1)}, & \text{если } m \leq r_{1,i}^{(1)} \text{ (включение)}, \\ m, & \text{если } r_{1,i}^{(1)} < m \leq R_{1,i}^{(1)}, \\ m, & \text{если } R_{1,i}^{(2)} \leq m < r_{1,i}^{(2)}, \\ R_{1,i}^{(2)}, & \text{если } m \geq r_{1,i}^{(2)} \text{ (отключение)}. \end{cases} \quad (10)$$

Отметим, что в [3], как уже отмечено, обсуждение велось на чисто качественном уровне и не учитывало, в частности, дискретности одной из переменных состояния СМО. Используя технику, кратко очерченную в разделе 3 настоящей работы, можно переформулировать и строго доказать все сделанные в [3] утверждения. В отличие от [3], в настоящей работе у четырех последних характеристических параметров оптимальной «близорукой» стратегии появился еще один нижний индекс, в данном случае равный 1. В одном из последующих пунктов эта единица превратится в n – число шагов, оставшихся до конца периода планирования, когда будет рассматриваться многошаговая задача.

Важной особенностью, характеристическим свойством «близорукой» задачи стало установление того факта, что между параметрами $R_{1,i}^{(1)}$ и $R_{1,i}^{(2)}$ существует упорядоченность, обусловленная выполнением неравенства $R_{1,i}^{(1)} \leq R_{1,i}^{(2)}$. Будь это

[†]Применение формулы (9) связано с некоторыми оговорками, суть которых заключается в том, что в некоторых случаях, перечисленных в работе [3], параметр $R_{1,i}^{(1)}$ назначается равным $\underline{u_i}$. Имеются и другие нюансы.

не так, пришлось бы думать, что делать со «скользящими» режимами, когда сначала пришлось бы отключать часть каналов (доводя число работающих каналов до $R^{(2)}_{1,i}$), но тут же, при некоторых обстоятельствах, часть каналов отключать, возвращая СМО к другому параметру – величине $R^{(1)}_{1,i}$, и т.д. Не такая уж, казалось бы, и беда, но еще одна и, быть может, более сложная задача.

Указанный выше факт требует еще одной декомпозиции критерия $C_n(\lambda_i, m, u)$ исходной задачи (5)–(6). А именно, отдельно рассматриваются два альтернативных варианта управленческого решения: (а) включить дополнительные каналы или, напротив, (б) отключить часть работающих каналов. При этом плата за включение описывается функцией $B_{\text{включения}}(u) = c_1 u + d\bar{l}_{\text{очереди}}$, а плата за отключение – функцией $B_{\text{отключения}}(u) = (c_1 - c_2)u + c_2 m + d\bar{l}_{\text{очереди}}$, где $\bar{l}_{\text{очереди}}$ задается формулой (4) и все обозначения совпадают с ранее введенными. Именно из вида этих функцией вытекает неравенство $R^{(1)}_{1,i} \le R^{(2)}_{1,i}$. Важно отметить и то обстоятельство, что $R^{(1)}_{1,i}$ является точкой абсолютного минимума функционала одношаговых затрат при включении каналов, а $R^{(2)}_{1,i}$ – точкой абсолютного минимума функционала одношаговых затрат при отключении каналов.

4.2. Расширение понятия *A*-выпуклости. Доказательство перечисленных в пункте 4.1 фактов опиралось на аналогию между задачами теории управления запасами и теми задачами теории массового обслуживания, которые рассматриваются в настоящей и предшествующих работах первого из авторов. В качестве примера первичного анализа таких аналогий можно назвать работу [8]. В этой работе была рассмотрена задача управления запасами, которую авторы назвали «фантазийной» . Тем не менее, нельзя не признать, что для рассматриваемого класса задач (как «фантазийной» из теории управления запасами, так и задачи о переключении каналов), была продемонстрирована некоторая неготовность, неприспособленность существующих теории оптимизации и теории выпуклости к решению этих задач. Дело в том, что в при математическом описании рассматриваемого класса задач используется не одно, а два разных видения оптимальности (критерия оптимальности). Говоря на языке, адекватном данному классу задач, это *оптимальность при включении* каналов (при подаче заказов в теории управления запасами), когда мы движемся по оси целочисленной переменной *и слева–направо*, и *оптимальность при отключении* каналов (возвращении товара в теории управления запасами), когда мы движемся по оси переменной *и справа-налево*. Облечем это замечание в новые определения.

Определение 3. *Функция $g(i)$ называется A-выпуклой слева ($A \ge 0$), если для всех натуральных i и j, которые находятся левее точки ее абсолютного*

минимума,

$$A + g(i+j) - g(i) - \Delta^{(1)}g(i) \times j \geq 0.$$

Определение 4. *Функция $g(i)$ называется A-выпуклой справа ($A \geq 0$), если для всех натуральных i и j, которые находятся правее точки ее абсолютного минимума,*

$$A + g(i+j) - g(i) - \Delta^{(1)}g(i) \times j \geq 0.$$

Теперь можно констатировать, что функционал включения в одношаговой («близорукой») задаче является A_1-выпуклым слева, а функционал отключения для той же задачи – A_2-выпуклым справа.

4.3. Многошаговые динамические задачи о переключении каналов. Вернемся к решению многошаговой задачи управления переключениями каналов, которая описывается уравнениями (5)–(6). Повторим запись целевого функционала в уравнении (6):

$$C_n^*(\lambda_i, m) = \min_{u \geq \underline{u}_i}(C^{(1)}(\lambda_i, m, u) + \alpha \sum_{j=1}^{k} p_{ij} C_{n-1}^*(\lambda_j, u)). \tag{11}$$

Выражение в фигурных скобках в правой части этого уравнения, независимо от того, какая альтернатива оценивается (включение или отключение), содержит не меняющийся при смене альтернативы член $\alpha \sum_{j=1}^{l} p_{ij} C_{n-1}^*(\lambda_j, u)$. Одношаговый добавок $C^{(1)}(\lambda_i, m, u)$, напротив, варьирует при смене альтернативы. Из этого можно сделать вывод о том, что есть все основания считать, что выражение в фигурных скобках в правой части формулы (6) (также двухвариантное) является одновременно A_1-выпуклым слева и A_2-выпуклым справа.

Выскажем предположение математической индукции о том, что это утверждение верно для любого номера n. Как следует из работы [3], для случая $n = 1$ так оно и есть. Предположим теперь, что это утверждение верно для случая $n - 1$. Дальнейшее доказательство истинности этого утверждения для случая n достаточно утомительно (по выкладкам), однако, по сути своей, выполняется по классической схеме доказательства оптимальности, но для двухуровневых стратегий [6, 7].

Как отмечалось выше, избежать скользящих режимов позволяет наличие упорядоченности точек локального минимума альтернативных функционалов в форме неравенств $R_{n,i}^{(1)} \leq R_{n,i}^{(2)}$, $n \in \overline{1, N}, i \in \overline{1, k}$. И тут следует признать, что авторам удалось доказать это утверждение строго только для случая $n = 1$. По поводу произвольного значения n можно только утверждать, что во всем объеме выполненных вычислительных экспериментов не было зарегистрировано ни единого случая нарушения выполнения этих неравенств.

5. Заключение

Исследованы стратегии переключения каналов в многолинейной системе массового обслуживания при марковском описании процесса изменения интенсивности входящего потока. Целевой функцией при построении стратегий является минимизация суммарных средних затрат на многошаговом периоде планирования. При этом предполагается, что процедура переключения каналов приводит к расходам, которые состоят из фиксированных платежей и затрат, величина которых зависит от числа включаемых или отключаемых каналов. В результате оптимальные стратегии переключения каналов оказываются параметрическими. Процесс доказательства потребовал введения некоторых обобщений классического в исследовании операций понятия A-выпуклости.

Литература

1. Mandel A. Econometric Models of Controllable Multiple Queuing Systems / Proceedings of the 18th International Conference, Distributed Computer and Communication Networks (DCCN 2015, Moscow, Russia). Geneva: Springer, 2016. – P. 296–304.
2. Мандель А.С., Бакулин К.Н. Models of controllable multiple queuing systems for channel switching myopic strategies / Proceedings of the 20th International Conference, Distributed Computer and Communication Networks (DCCN 2017, Moscow, Russia). М.: Техносфера, 2017. – С. 534–542.
3. Mandel A., Laptin V. Myopic Channel Switching Strategies for Stationary Mode: Threshold Calculation Algorithms / In: In: Vishnevskiy V., Kozyrev D. (eds), Distributed Computer and Communication Networks. DCCN 2018. Communications in Computer and Information Science, vol. 919. Geneva: Springer, 2018. – DOI: 10.1007/978-3-319-99447-5_35
4. Гнеденко Б.В., Коваленко И.Н. Введение в теорию массового обслуживания. – М.: Наука, 1966.
5. Вишневский В.М. Теоретические основы проектирования компьютерных сетей. М.: Техносфера. 2003.
6. Хедли Дж., Уайтин Т. Анализ систем управления запасами. – М.: Наука, 1969. 512 с.
7. Arrow K., Karlin S., and Scarf H. Studies in the Mathematical Theory of Inventory and Production. – Stanford University Press, 1958.
8. Mandel A., Granin S. Investigation of analogies between the problems of inventory control and the problems of the controllable queuing systems // In: Proceedings of 2018 Eleventh International Conference “Management of large-scale system development” (MLSD’2018). 2018. – IEEE, 2018. – P. 1–4. – DOI: 10.1109/MLSD.2018.8551852

UDC: 519.17

On improving the accuracy of the classification on imbalanced classes with machine learning

Eugene Yu. Shchetinin[1], Leonid A. Sevastianov[2,3], Dmitry S. Kulyabov[2,3], Edik A. Ayrjan[3,4]

[1]Financial University, Government of the Russian Federation, Moscow, Russian Federation
[2]Peoples' Friendship University of Russia (RUDN University), Moscow, Russian Federation
[3]Joint Institute for Nuclear Research, Dubna, Russian Federation
[4]Dubna State University, Dubna, Russian Federation

riviera-molto@mail.ru, sevastianov-la@rudn.ru, kulyabov-ds@rudn.ru, ayrjan@jinr.ru

Abstract

Imbalance of the classes, characterized by a disproportional ratio of observations in each class, is one of the significant problems in machine learning. Class imbalances can be detected in many areas, including medical diagnostics, spam filtering, and fraud detection. Most machine learning algorithms work optimally when the number of samples in each class is approximately the same. This is because most algorithms are designed to maximize accuracy and reduce error. However, under conditions of class imbalance, the model may be overfitted, which leads to incorrect estimates of object classification. Thus, in order to avoid this phenomenon and achieve better results, it is necessary to research methods for working with unbalanced data, as well as develop effective algorithms for classifying them. In this paper, we study machine learning methods to eliminate class imbalance in data in order to improve accuracy in multi-class classification problems. In this paper, to improve the accuracy of classification, it is proposed to use a combination of classification algorithms and feature selection methods RFE, Random Forest and Boruta with pre-balancing classes by random sampling, SMOTE and ADASYN. Using data on skin diseases as an example, computer experiments have shown that the use of sampling algorithms to eliminate the imbalance of classes, as well as the selection of the most informative features, significantly improves the accuracy of classification results. The Random Forest algorithm was the most effective in terms of classification accuracy when sampling data using the ADASYN algorithm.

Keywords: multiclass classification, imbalanced classes, machine learning, SMOTE, ADASYN, Random Forest.

The publication has been prepared with the support of the. RUDN University Program 5-100" and funded by Russian Foundation for Basic Research (RFBR) according to the research project No 19-01-00645.

1. Introduction

Classification problems are among the most popular in machine learning [1]. Supervised machine learning is most often used as the method for determining whether an object belongs to a particular class. The main idea of this approach is to inductively output a function based on labeled data for training. This means that the success of using a machine learning classification algorithm depends largely on the selection of objects that the algorithm "learns" from. Most of these algorithms require the researcher to include a comparable number of examples for each of the classes, but it is often not possible to make balanced data sets due to a number of factors. Often there are situations when the dataset number of examples of some of the minor class (this class will be called the minority, and the other, prevailing over first – majority class). The key ones are the specificity of the target area (balancing data can lower the indicator of its representativeness) and the different price of errors of the first and second types when classifying. Such trends are clearly visible, for example, in credit scoring, medicine and marketing [2,3].

This leads to the problem of training the model on imbalanced data (these are data whose distribution is skewed, and the mode and average values are not equal): according to the basic assumptions contained in most algorithms, the goal of training is to maximize the proportion of correct decisions relative to all decisions made, and the data for training and the general population are subject to the same distribution. However, taking into account these assumptions and unbalanced sampling results in the model being unable to classify data better than a trivial model that completely ignores a less represented class and marks all objects for classification as belonging to the majority class.

On the other hand, it is possible to build too much complex model that includes a large set of rules, but will cover a small number of objects. This classifier may be ineffective, which will lead the model to overfitting and incorrect estimates of the forecast. It should be noted that the consequences of erroneous classification may also differ. Moreover, an incorrect classification of examples of a minority class usually costs many times more than an erroneous classification of an object from a majority class. The correct selection of features may be more important than reducing data processing time or improving classification accuracy. for example, in medicine, finding the minimum set of features that is optimal for the classification task may be a prerequisite for making a diagnosis. Thus, to avoid this phenomenon and achieve a good result, it is necessary to research methods for working with imbalanced data.

In this paper, we investigate the methods for overcoming imbalanced classes problem in order to improve the quality of classification with a higher accuracy than when directly using classification algorithms for imbalanced classes. To improve the

accuracy of classification, we propose a scheme that consists of using a combination of classification algorithms and feature selection methods RFE, Random Forest and Boruta with the preliminary use of class balancing by random sampling, SMOTE and ADASYN.

2. Basic algorithms for balancing the classes

One approach to solving this problem is to use various sampling strategies, which can be divided into two groups: random and special [3]. In the first case, delete a certain number of examples of the majority class (undersampling), in the second – increase the number of examples of the minority class (oversampling).

2.1. The exclusion of examples of the majority class. Algorithm for random sampling of the majority class (random undersampling). To do this, we calculate the $K-$ number of majority examples that must be removed to achieve the required ratio of different classes. Then K majority examples are randomly selected and removed. In the case of the studied data, methods for increasing the minority class are natural. Let's move on to the consideration of such strategies.

2.2. The oversampling algorithm. Random naive sampling. The easiest way to increase the number of examples of a minority class is to randomly select observations from it and add them to the general dataset until a balance is reached between the majority and minority classes. Depending on what class ratio is needed, the number of random records to duplicate is selected. One of the problems with random naive sampling is that it simply duplicates existing data. The advantages of this approach include its simplicity, ease of implementation and the ability to change the balance in any desired direction. The disadvantages should be discussed separately according to which sampling strategy is used: although both of them change the overall size of the data in order to find a balance, their application has different consequences. In the case of undersampling, deleting data may cause the class to lose important information and, as a result, lower its presentation rate. In turn, the use of oversampling can lead to overfitting [4]. This approach to restoring balance is not always effective, so a special method was proposed to increase the number of examples of a minority class-the SMOTE algorithm (Synthetic Minority Oversampling Technique) [7]. The SMOTE algorithm is based on the idea of generating a certain number of artificial examples that are "similar" to those in the minority class, but do not duplicate them. To create a new record find the difference $d = X_b - X_a$, where X_a, X_b- feature vectors of "neighboring" examples a and b from the minority class. They are found using the nearest neighbors algorithm (KNN). In this case, it is necessary and sufficient for example b to get a set of k neighbors, from which the entry b will be selected later. Then from d by multiplying each of its elements by a random number in the interval $(0, 1)$ we get $\tilde{d}$. The feature

vector of the new example is calculated by adding X_a and $\widetilde{d}$. The SMOTE algorithm allows you to set the number of records to be artificially generated. The degree of similarity of examples a and b can be adjusted by changing the value of k (the number of nearest neighbors).

SMOTE solves many problems that are inherent to the random sampling method, and actually increases the initial data set in such a way that the model is trained much more efficiently. However, this algorithm has its drawbacks, the main of which is ignoring the majority class. This may result in a highly sparse distribution of objects of a minority class relative to a majority class, where data sets are "mixed", i.e. they are arranged in such a way that it is very difficult to separate objects of one class from another. An example of this phenomenon is when an object of a different class is located between an object and its neighbor, based on which a new instance is generated. As a result, the synthetically created object will be closer to the opposite class than to the class of its parents. In addition, the number of instances generated using SMOTE is set in advance, which reduces the ability to change the balance and flexibility of the method. It is important to note the significant limitations of SMOTE algorithm. Since it works by interpolating between rare examples, it can only generate examples inside the body of available examples – never outside. Formally, SMOTE can only fill in the convex hull of existing minority examples, but not create new external areas for them. The main advantage of SMOTE over traditional random naive over-sampling is that when creating synthetic observations instead of reusing existing observations, this classifier is less likely to be overfitted. At the same time, it is always necessary to make sure that the observations created by SMOTE are realistic.

2.3. Adaptive synthetic sampling algorithm and its generalizations. This method is based on synthetic sampling algorithms, the main ones being Borderline-SMOTE and Adaptive Synthetic Sampling (ADASYN) [9]. Borderline-SMOTE imposes restrictions on the selection of objects of the minority class that new instances are generated from. This happens as follows: for each object of a minority class, a set of k nearest neighbors is determined, then it is calculated how many instances of this set belong to the majority class (this number is taken as m). After this, we select those objects of the minority class for which the inequality $k/2 \leq m < k$ is true. The resulting set represents instances of the minority class located on the distribution boundary, and they are the ones that are more likely to be incorrectly classified than the others. It should be noted why the inequality that determines the selection of objects excludes cases in which all k neighbors belong to the majority class: this is due to the fact that such instances are located in the "mixing" zone of two classes, and only objects that distort the model learning process can be generated on their basis. In this regard, they are declared as noise and are

ignored by the algorithm. The ADASYN algorithm, in turn, is based on a systematic method that allows adaptive generation of different amounts of data in accordance with their distributions [6]. Input data for the algorithm – training data set: D_r with m samples with $\{x_i, y_i\}$, $i = 1, ..., m$, where x_i- is the $n-$ dimensional vector in the feature space, y_i- labels of corresponding class. Let's the m_r and m_x are the number of samples of minority and majority classes, respectively, such that $m_r \ll m_x$ and $m_x + m_r = m$. The algorithm's pseudocode looks like this:

1. Calculate the proportion of classes $d = \frac{m_r}{m_x}$; 2. If $d < d_x$ (where d_x is the specified threshold for the maximum allowable class imbalance):

a) Find the number of synthetically generated samples of the minor class $G = (m_x - m_r) \times \beta$, where β is the parameter used to determine the desired balance level ($\beta = 1$) indicates full class balance.

b) for each $x_i \in minorityclass$ find the K-nearest neighbors using the Euclidean distance and calculate $r_i = \triangle_i / K$;

c) normalize $r_x = \frac{r_i}{\sum_i r_i}$ so that r_x becomes the distibution density;

d) calculate $g_i = r_x \times G$ a synthetic sample formed for each image from the minority class, where $G-$ is the total number of examples of synthetic data;

e) for each example of data from a minority class x_i create the examples of synthetic g_i data in accordance with the following steps:

In a cycle from 1 to i :

(i) randomly select one example of minority data, x_u from K nearest neighbors for x_i data;

(ii) create an example of synthetic data: $g_i = x_i + (x_u - x_i) \times \lambda$, where $(x_u - x_i)$ is $n-$dimensional vector of Euclidean space, $\lambda-$ random number, $\lambda \in [0, 1]$.

The main difference between SMOTE and ADASYN is how to create synthetic sample samples for the minority class. ADASYN uses the r_x density function to determine the number of synthetic samples that will be created for a specific point, whereas SMOTE has a single weight for all minority points.

3. Researched dataset: description and characteristics

In this paper, a set of data on skin diseases was used for testing and comparative analysis of the methods described above to overcome the classes imbalance. Diagnosis of erythematous squamous cell diseases is a serious problem in dermatology, and modern principles of diagnosis and treatment are based on the earliest detection of the disease. All of them have common clinical features with very small differences. Another difficulty for diagnosis is that the disease may show signs of another disease at the initial stage and may have characteristic signs in subsequent stages.

The study data was created by Nielsen in 1998 and contains 366 observations forming 6 classes that can be characterized by 34 features [8]. The classes are:

psoriasis(class 1): – 112 cases; seborrheic dermatitis (class 2): – 72 cases; lichen planus (class 3): – 61 cases; pink lichen (class 4): – 49 cases; chronic dermatitis (class 5): – 52 cases; red hair lichen (class 6): – 20 cases. A full description of the data is given in [11].

4. Computer experiments

Data studies were performed using the following algorithm:

1) Data pre-processing: filling the gaps in the data and the coding of signs.
2) Balancing classes using the sampling algorithms described above.
3) Selecting attributes based on their importance.
4) Classification using logistic regression and the support vector method.
5) Assessment of classification quality.

In this paper, the selection of features based on their importance and informativeness was carried out by the following methods: a) recursive exclusion of RFE features [5]; b) Random Forest [10]; c) Boruta [4].

The Random Forest algorithm is an ensemble of numerous classification algorithms (decision trees). Each of these classifiers is built on a random subset of objects and a random subset of features. Let the training sample consist of N examples, the dimension of the feature space is equal to M, and an additional parameter m is set. All trees are built independently of each other using the following procedure:

1) Generate a random sub-sample with a repeat of size n from the training sample.
2) Let's build a decision tree that classifies the examples of this sub-sample, and during the creation of the next node of the tree, we will select the feature based on which the partition is made, not from all M features, but only from m randomly selected ones.
3) The tree is built until the subsample is completely exhausted and does not undergo the procedure of cutting off branches.

Object classification is carried out by voting: each tree of the ensemble refers the object to be classified to one of the classes, and the class that the largest number of trees voted for wins. To use Random Forest in the task of evaluating the importance of features, it is necessary to train the algorithm on the sample and calculate the out-of-bag error for each example of the training sample. Let X_n be a bootstrapped sample of the b_n tree. Bootstrapping is the selection of l objects from the selection with a return, as a result of which some objects are selected several times, and some – never. Placing multiple copies of the same object in a bootstrapped selection corresponds to setting the weight for this object, the corresponding term will be included in the functionality several times, and therefore the error penalty will be greater on it. Let $L(y, z)$ be the loss function, and y_i be the response on the

i-th object of the training sample, then the out-of-bag error is calculated using the following formula:

$$OOB = \sum_{a}^{b} L\left(y_i, \frac{\sum_{n=1}^{N}\left[x_i \ni X_n^l\right] b_n(x_i)}{sum_{n=1}^{N}\left[x_i \ni X_n^l\right]}\right).$$

Then, for each object, this error is averaged across the entire random forest. To evaluate the feature importance, its values are mixed for all objects in the training sample, and the out-of-bag error is counted again. The importance of the features is estimated by averaging the difference in out-of-bag errors across all trees before and after mixing the values. The values of such errors are normalized to the standard deviation.

Boruta is a heuristic algorithm for selecting significant features based on the use of Random Forest [4]. At each iteration, features that have a Z-measure less than the maximum Z-measure among the added features are removed. To get the Z-measure of a feature, you need to calculate the feature's importance obtained using the built-in algorithm in Random Forest, and divide it by the standard deviation of the feature importance. The added features are obtained as follows: the features that are present in the selection are copied, and then each new feature is filled in by shuffling its values. In order to get statistically significant results, this procedure is repeated several times, and variables are generated independently at each iteration.

Let's write down the Boruta algorithm step by step:

1) Add copies of all attributes to the data. In the future, copies will be called hidden signs.
2) Randomly shuffle each hidden attribute.
3) Run Random Forest and get the Z-measure of all attributes.
4) Find the maximum I-measure of all I-measures for hidden features.
5) Delete features that have a Z-measure smaller than the one found in the previous step.
6) Remove all hidden attributes.
7) Repeat all the steps until the Z-measure of all features is greater than the maximum z-measure of hidden features.

To solve the problem of multiclass classification on imbalanced data, next machine learning algorithms were chosen: logistic regression and the method of support vectors with a RBF core (RBF SVM). Three metrics were used to compare classification results: accuracy, precision and F1-measure. From the analysis of the results it can be seen that in all cases the use of sampling methods allowed to obtain a higher classification accuracy than on unbalanced data. Within the framework of the scheme described in this paper, the best classification accuracy was achieved by applying the

ADASYN class balancing algorithm and then selecting features using the random forest algorithm. For comparison, in the works of other researchers who conducted similar studies, for example, [9, 10], the classification accuracy reached only 93%.

5. Main results and conclusion

In this paper, we propose a scheme for improving the accuracy of classification on unbalanced data using algorithms for class balancing and feature selection, such as RFE, Boruta, Random Forest, and others. The results of computational experiments have shown the effectiveness of its application to solve this problem. In particular, the ADASYN algorithm has improved classification accuracy by up to 98% compared to other algorithms. In conclusion, it is worth noting that the problem discussed in this paper is still relevant , and existing methods can be improved. In recent time there are some new trends in data mining so called dee learning, developing the deep neural networks as a tool for solving various classification problems. So, we hope to apply them in our future researches of imbalanced classes classification.

REFERENCES

1. Patterson J., Gibson A. Deep Learning: A Practitioner's Approach. – O'Reilly Media, 2017. 532 p.
2. Murphy P. M., Aha D. W. UCI repository of machine learning databases. – Irvine: University of California, Department of Information and Computer Science, 1998. https://www.ics.uci.edu/mlearn/MLRepository.html.
3. He H., Garcia A. Learning from imbalanced data, IEEE Transactions on Knowledge and Data Engineering, 2009. Vol. 21. Iss. 9. P. 1263–1284. doi: 10.1109/TKDE.2008.239.
4. Japkowicz N., Stephen S. The class imbalance problem: A Systematic Study, Intelligent Data Analysis, 2002. Vol. 6. Iss. 5. P. 429–449. doi: 10.3233/IDA-2002-6504.
5. Lin X., Yang F., Zhou L. A support vector machine recursive feature elimination feature selection method based on artificial contrast variables and mutual information // Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences, 2012. Vol. 10. P. 149–155. doi: 10.1016/j.jchromb.2012.05.020.
6. Kursa M., Rudnicki W. Feature Selection with the Boruta Package. Journal of Statistical Software, 2010. Vol. 36. Iss. 11. P. 1–13. doi: 10.18637/jss.v036.i11.
7. Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P. SMOTE: synthetic minority over-sampling technique, Journal of Artificial Intelligence Research, 2002. Vol. 16. P. 321–357. doi: 10.1613/jair.953.

8. Murphy P. M., Aha D. W. UCI repository of machine learning databases. – Irvine: University of California, Department of Information and Computer Science, 1998. https://www.ics.uci.edu/mlearn/MLRepository.html.
9. He H., Bai Ya., Garcia A., Li Sh. ADASYN: Adaptive synthetic sampling approach for imbalanced learning, 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). – IEEE, 2008. P. 1322–1328.
10. Tuv E., Borisov A., Runger G., Torkkola K. Feature selection with ensembles, artificial variables, and redundancy elimination, The Journal of Machine Learning Research, 2009. Vol. 10. P. 1341–1366.

UDC: 004.046

Generation of metadata for information technology control

A.A. Grusho[1], N.A. Grusho[1], M.I. Zabezhailo[1], E.E. Timonina[1]

[1] Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Vavilova 44-2, 119333, Moscow, Russia

grusho@yandex.ru, info@itake.ru, zabezhailo@yandex.ru, eltimon@yandex.ru

Abstract

In this paper we propose a method of ensuring network information security by controlling network connections using metadata. The metadata contains information about admissible task interactions and application locations in a corporate network.

In this paper models of hierarchical decomposition of information technology in the form of directed acyclic graph have been built. Hierarchical decomposition makes it possible to optimally form blocks of information transformations for their distribution on hosts of distributed information and computing system. Those graphs are used for metadata construction.

Keywords: Information system; metadata; directed acyclic graph; information security

1. Introduction

To ensure the information security of network interactions [1] it has been proposed to manage network connections using metadata. The metadata contains information about permitted task interactions and locations of appliactions that are required to solve these tasks in a corporate network.

Information technology (IT) is reduced to information transformations, that is, solving tasks implemented by applications. A computer, as a node of a network, is called a host. Different tasks can be solved on different hosts of the network. The network then allows you to collect source data for tasks, and distribute the processing results.

IT security policy requires monitoring of host interactions at the corporate network. Control of host interactions across the network helps to reduce the threat of malicious code being introduced and distributed through network equipment and

The publication has been prepared with the partial support of Russian Foundation for Basic Research according to the research projects No.18-29-03081, 18-07-00274.

communication channels. Results of [1] show how the control of host interactions in the network can be implemented using metadata.

Prior to the development of metadata concept we have investigated speed characteristics of this class of architectures. Experiments were made on virtual networks. We got satisfactory results which were published in [2].

Let's assume that a mathematical model has been created for IT that defines all system actions that are required to perform the reqested computations or tasks. Complete information on performance of work can be represented by PERT (Project Evaluation and Review Technique) [3] diagrams, which are based on directed acyclic graphs (DAG).

2. Information Technology Description with DAG

IT models [4] presented as DAG are discussed. Capital Latin letters A, B,..., denote data (objects) serving as input or output data of information transformations in IT, in the figures this data will be represented by circles. The transformations will be called blocks, denoted with lowercase Latin letters and represented in the figures by boxes. The set of input data, transformations and output data will be indicated by lowercase Greek letters.

Each block corresponds to the transformation of information and corresponds to the solution of one or more tasks required for IT implementation. The DAG arcs correspond to the data transmission to blocks from previous blocks, that is, the arc exits the vertex corresponding to the output of the performed transformation and enters the vertex corresponding to the input data for the next transformation. Repeated transformations are valid if they have different inputs. IT can have directed cycles of transformations. Obviously, the input and output of each cycle must be different. Consequently, there are no directed cycles in the graph of IT. If the transformation output completely defines the input data of one or more of the following transformations, multiple arcs exit from the transform output. However, some outputs may be unused in the next block.

The simplest DAG describing the transformation of information is shown in Fig. 1, where A is the input of the transformation, B is the output of the transformation, f is the transformation itself. If the output data A of some transformation and the

Fig. 1. The simplest DAG

output data B of some other transformation are used for the transformation f, the input data for the transformation f is the vector (A, B) (see Fig. 2).

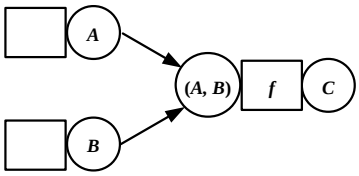


Fig. 2. Vector Inputs Data

Let's construct a hierarchical decomposition of DAG. The hierarchical decomposition is determined iteratively by applying two operations.

1. *Block division operation.* Let the data transformation f in the current DAG be as shown in Fig. 1. Suppose that f can be represented as a superposition of two transformations $B = f(A) = f_2(f_1(A))$ and $f_1(A) = C$. The result of block division operation on DAG fragment in Fig. 1 can be graphically represented as shown in Fig. 3. This transformation preserves the acyclicity of the source graph.

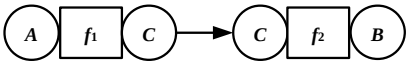


Fig. 3. Block division operation

2. *Block detailing operation.* Let the transformation f depend on the source data (A, B) and can be represented as $C = f(A, B)$. Suppose there are functions f_1 and f_2 such that $f(A, B) = f_2(f_1(A), B)$. The result of block detailing operation is then as shown in Fig. 4. Obviously, the block detail operation does not produce directed

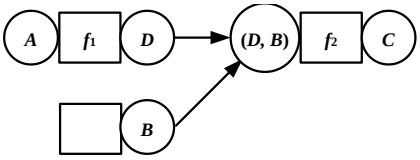


Fig. 4. Block detailing operation

cycles in the generated graph.

Definition 1. DAG $\mathcal{G}$ is called an adequate model of the given IT if it can be proved that any source data of IT transformed by blocks with the required functionality and transferred according to arcs of $\mathcal{G}$ produces the required result of IT.

In practice, the test on adequacy involves implementation of the IT model, repeated execution of IT and reproduction of the results obtained for a representative sample of source data.

Theorem 1. Each DAG $\mathcal{G}'$ obtained by applying a sequence of block detailing and block division operations on DAG $\mathcal{G}$, which is the adequate model of IT, produce a model of IT which is also adequate.

The inverse problem, i.e. the integration of DAG $\mathcal{G}$, which is the adequate model of this IT, has the following solution.

a) Let DAG $\mathcal{G}$ be the adequate model of IT. We highlight fragments of graph $\mathcal{G}$ corresponding to Fig. 3, i.e. fragments having transformations of the form $f(A) = f_2(f_1(A)) = B$. For this case, let's define the transform f as $f(A) = f_2(f_1(A))$. We replace the model of $\mathcal{G}$ containing two blocks of transformations $f_1(A) = C$ and $f_2(C) = B$ by the model with a single block. Clearly, the resulting graph $\mathcal{G}'$ is the adequate model of IT. We apply this operation repeatedly wherever possible and get a graph that is an adequate model of IT, but has fewer blocks.

b) Consider the fragment of graph $\mathcal{G}$ shown in Fig. 5.

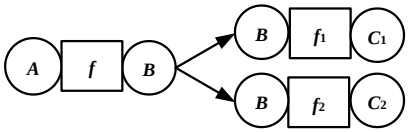


Fig. 5. Data transfer to different chains

This fragment can be transformed into DAG $\mathcal{G}'$ as follows, retaining the adequacy property of the model of IT (Fig. 6).

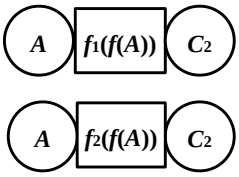


Fig. 6. Reduce the number of blocks by using parallel chains

Transforming the fragment in Fig. 5 to the fragment in Fig. 6 reduces the number of blocks in graph $\mathcal{G}'$ compared to graph $\mathcal{G}$.

c) The detailing fragment in Fig. 2 in the adequate model of IT can be reversed using vector data and vector transformations if vector (A, B) is the source vector

transformation data (f_1, f_2). Namely, the next piece of the adequate model of IT maintains the adequacy of the model of IT (Fig. 7).

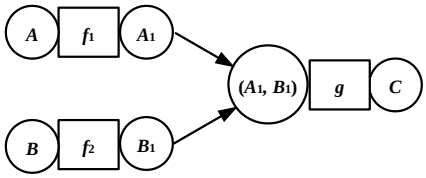


Fig. 7. Transformation using input from several other transformations

The fragment in Fig. 7 can be transformed into the fragment in Fig. 8 while maintaining the adequacy of the model of IT and reducing the number of blocks compared to the source graph.

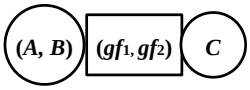


Fig. 8. Use of vector transformation to reduce the number of blocks

Thus, an iterative algorithm can be constructed that reduces the number of blocks in the next iteration and maintains the model of IT. This implies Theorem 2.

Theorem 2. Each adequate model of IT in the form of DAG can be derived from the adequate Fig. 1 model using a sequence of block detailing and block division operations. Conversely, each adequate model of IT in the form of DAG can be converted into the adequate model of IT presented in Fig. 1.

3. Transformation DAG of adequate models of IT to metadata

Usage of metadata for connection management in distributed information computing systems is described in detail in [1, 5]. However, in those papers, metadata that describes the order of solving the tasks is obtained from graphs of reduction [5] and are based on a complicated tree traversal algorithm.

In this paper, the source models of IT are DAG and the DAG hierarchy, which allows aggregation of tasks into blocks on separate computers. Therefore, it is necessary to construct the transformation of DAG into connection management metadata.

For the sake of simplicity, we build a rigid order of execution of IT task blocks represented by DAG [5]. This order must be supplemented by the scheme of input and output data exchange between blocks. All this information is called metadata $\mathcal{B}$

in IT. The network management algorithm uses the metadata to uniquely determine the order of accesses between blocks on the network.

For metadata $\mathcal{B}$, we define three additional tasks $\mathcal{M}$, $\mathcal{N}$, $\mathcal{R}$, wwhich control interactions on the network based on metadata $\mathcal{B}$. The task $\mathcal{M}$ distributes applications for execution of task blocks among hosts. For simplicity we will call it distribution of task blocks on hosts. The task $\mathcal{M}$ defines the binary relation $H(\alpha)$ meaning that the block α can be executed on host H. Note that the task $\mathcal{M}$ can use the hierarchical decomposition of DAG to form blocks and to distribute these blocks at network hosts while taking into account the information security requirements. In addition, task $\mathcal{M}$ allows for efficient block redundancy.

Results produced by the task $\mathcal{M}$ are used by the task $\mathcal{N}$. The task $\mathcal{N}$ keeps in contact with each host, and is responsible for permission checking and providing information to hosts on request for interaction of blocks at different hosts. Permission checking is based on metadata $\mathcal{B}$. The task $\mathcal{R}$ builds primary and backup routes for task $\mathcal{N}$.

Let the block α is legally running at host $H(\alpha)$. The block α_1 uses block output data α (denoted as (α, α_1)) and is located on another host. Each host H has an agent with cryptographic facilities and the key $k(H)$ for communication with the host $H(\mathcal{N})$. For each H, the connection to $H(\mathcal{N})$ does not produce an unacceptable delay.

In order to access block α_1 the block α contacts the task $\mathcal{N}$ through the agent of its host which checks whether (α, α_1) exists. Then, the information about whether it is necessary to connect to $H(\alpha)$, the connection protection key $k((\alpha, \alpha_1))$, the identifier, the port, and the time stamp are sent to the host $H(\alpha_1)$ through the agent of this host. Similar information is sent to the host $H(\alpha)$. After the data has been transmitted to the block α_1, the connection between hosts $H(\alpha)$ and $H(\alpha_1)$ is terminated.

Obviously, DAG defines the strict order [3] and the set of vertices of the graph $\mathcal{G}$ forms the partially ordered set. In order to construct metadata, first it is necessary to define the order of solving blocks of tasks $\{\alpha_1, \dots \alpha_m\}$ so that this order (the permutation of blocks $(\alpha_{i_1}, \dots \alpha_{i_m})$) has the property of consistency with the graph $\mathcal{G}$. That is, when metadata allows the transition from the block α_i to the block α_j, then the strict order of blocks defined by DAG $\mathcal{G}$ implies that α_i is less than α_j. In [3] it is proved that at least one such order exists, and algorithms for building permutations of task blocks under various additional restrictions are proposed. However, data exchange and the use of the built permutation are complicated.

The built permutation of blocks does not conflict with the sequence of task blocks execution in the DAG, meaning that if the block β needs data from the block α, the block α must wait for the block β to be executed in queue determined by the

permutation. If there are several blocks with data for the block β, they should all "remember" the data to be presented to the block β before that block appears and wait for their data transmission queue. Data collection is not provided in block execution order. Since there may be a lot of such cases and blocks may be repeated, each block in the permutation must remember which data it should transmit to which other block.

The second problem is that each particular block in the network needs to establish its own connection between hosts containing the respective blocks to transmit data. The host that is to transmit the data must initiate the connection, but does not know when to do it. In addition, queues to this particular application may appear due to transformation repetitions and other parallel IT.

It is possible to solve these problems by means of the matrix Γ, which is formed at the host $H(\mathcal{N})$ of the task $\mathcal{N}$. The square matrix $\Gamma = \|\gamma_{ij}\|$ of size $m \times m$, where m is the number of blocks, Γ is the matrix in which $\gamma_{ij} = 1$ if the block α_i is to transmit data to the block α_j, and $\gamma_{ij} = 0$ if the block α_i does not have to transmit data to the block α_j, or has already transmitted that data. When it's time for block α_j to execute, the j-th column of the matrix Γ defines blocks that wait for its queue to transmit their information or gain access to the block α_j.

The usage of the matrix Γ implies that the task $\mathcal{N}$ in turn establishes the connection to hosts containing blocks having nonzero values in the corresponding column of next block and indicates that data has to be transmitted to it.

4. Information security

Information security is governed by security policies [6]. For confidentiality these are traditional discretionary access control (DAC) security policies and modifications of this policy, such as Role Based Access Control (RBAC). Access control in the DAC is determined by the Access Control List (ACL) matrix of the permissions granted to the users and subjects on their behalf to access objects. The ACL must consider the value of the input data that is used or appears in IT during its execution.

In addition to these security policies, confidentiality protection uses the Multi-Level Security (MLS) policy, which prevents information flows from objects with valuable information to subjects and objects that are not allowed to access valuable information.

Obviously, information flow control supports the aforementioned classes of security policies. However, while providing information security in IT, the static picture of information flows as defined by DAG, does not take into account the possibilities of using valuable and not valuable source data simultaneously or generation of valuable information during IT execution. The valuable information may become exposed due to erroneous actions of users participating in IT. Thus, in many stages of IT, it

is necessary to consider the possibility of access rules alteration, i.e. analyze input and output data and make decisions on access rights.

When information flow control policy is used, an access denial means that IT is stopped, which is equivalent to the failure of that IT. Recall that $\mathcal{M}$ and $\mathcal{R}$ tasks provide backup of IT blocks and backup network routes. Stopping IT execution due to failure or security policy violation engages these backup capabilities. For this purpose, the task $\mathcal{N}$ must obtain the information on a possible failure or an appearance of valuable information, which is determined by the information classification. This classification is performed either at the input of the IT or in the executed block α_i according to the specified criterion.

In this case all non-zero elements in all columns of the i-th row of the matrix Γ are replaced with the symbol "v", which allows to engage the $\mathcal{N}$ task in order to create and use the specially protected IT continuation. Specifically, this requires network routes to be rebuilt in order to securely continue the IT and re-route non-valuable data. In fact, the system must be reconfigured and an additional secure IT perimeter must be created. That means that DAG and metadata must be changed, and the system itself becomes at least two-level (MLS).

5. Conclusion

The paper follows the idea to control interactions of hosts by metadata. Hierarchical decomposition models of IT in the form of DAG have been built. Hierarchical decomposition makes it possible to optimally form blocks of information transformations for their distribution across hosts of distributed information and computing system. Then DAG should be transformed to metadata.

Transformation of DAG to metadata consists of two tasks:

- ordering block execution in a way that does not contradict the DAG;
- distribution of data produced by executed blocks for use by other blocks of IT.

In order to distribute data, it was necessary to add data in the form of the matrix Γ to the task $\mathcal{N}$ field. The matrix Γ controls the queueing of already executed blocks to transmit its results to the next blocks. The relations between security policies and the model of IT in the form of DAG have been considered.

REFERENCES

1. Grusho A., Grusho N., Timonina E. Information Flow Control on the Basis of Meta Data // Lecture Notes in Computer Science. 2019. V. 11965. P. 548–562.
2. Grusho N. A., Senchilo V. V. Modeling of Secure Architecture of Distributed Information Systems on the Basis of Integrated Virtualization // J. Systems and Means of Informatics. 2018. V. 28. Is. 1. P. 110–122.

3. Tanayev V. S., Shkurba V. V. Introduction to the scheduling theory. Science, Moscow, 1975 (in Russian).
4. Samuylov K. E., Chukarin A. V., Yarkina N. V. Business processes and information technologies in management of the telecommunication companies. Alpina Publishers, Moscow, 2009.
5. Grusho A. A., Timonina E. E., Shorgin S. Ya. Hierarchical method of meta data generation for control of network connections // J. Inform. Primen. 2018. V. 12. Is. 2. P. 44–49.
6. TCSEC. Department of Defense Trusted Computer System Evaluation Criteria. DoD, 1985.

УДК: 004.716

Задача оптимального размещения базовых станций широкополосной сети для контроля линейной территории при ограничении на величину межконцевой задержки

В.В. Вишневский[1], А.А. Мухтаров[1], О.Ю. Першин[2]

[1]Институт проблем управления им. В.А. Трапезникова РАН, ул. Профсоюзная д.65, г. Москва, Россия

[2]РГУ нефти и газа (НИУ) им. И. М. Губкина, Ленинский проспект д.65, г. Москва, Россия.

vishn@inbox.ru, mukhtarov.amir.a@gmail.com, pershino@mail.ru

Аннотация

Статья посвящена проблеме проектирования беспроводной широкополосной сети связи. Сформулирована задача оптимального размещения базовых станций вдоль линейной территории, подлежащей контролю с ограничением на межконцевую задержку в сети. Целью решения задачи является максимизация области покрытия, попадающая под контроль множества размещенных станций, при выполнении технологических условий и бюджетного ограничения. В работе анализируются особенности технологической постановки и предлагается формулировка задачи в виде модели целочисленного линейного программирования (ЦЛП). При формирования математической модели задачи используется модель сети массового обслуживания с пуассоновским входящим потоком.

Ключевые слова: беспроводные сети, задача целлочисленного линейного программирования, пуассоновский поток

1. Введение

Беспроводные технологии нашли широкое распространение в различных сферах жизнедеятельности человека. Широкополосные сети связи используются для оперативного контроля и управления производственными или гражданскими объектами, технологическими установками, движущимися транспортными средствами и т.п. Применение беспроводных широкополосных технологий на базе семейства протоколов IEEE 802.11 для организации таких сетей имеет ряд

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований, проект №19-07-00919 от 29.12.2018.

преимуществ по отношению к проводным технологиям. К ним относятся быстрое развёртывание сетей связи, удобную модернизацию и масштабируемость архитектуры сети, снижение затрат на монтаж и обслуживание. В процессе проектирования такой современной инфраструктуры передачи информации стоит задача оптимального размещения контролирующего оборудования, в нашем случае базовых станций широкополосной сети связи, на избыточном множестве возможных точек размещения. Подобная проблема ставилась и обсуждалась в ряде работ [1] - [6].

Настоящая работа является продолжением исследований [8] -[9], где рассматривается частный случай задачи, когда подлежащая контролю территория представляет собой линейный участок, например, территория вдоль протяженных автомагистралей, линейную часть магистрального трубопровода, коридор промысловых коммуникаций и т. п. Была дана формулировка в виде модели целочисленного линейного программирования (ЦЛП). В статье [8] было представлено доказательство NP – полноты такой постановки задачи.

В отличии от задачи, рассмотренной в работе [9], где необходимо было разместить все заданное множество станций, сформированное на предыдущих этапах проектировании сети, в представленной работе рассмотрен более общий случай, где надо выбрать набор размещенных станция из заданного избыточного множества в результате решения оптимизационной задачи. Существенным обобщением исследуемой задачи в данной работе является наличие ограничения на межконцевую задержку сети.

2. Постановка задачи

Задача ставится следующим образом.

Для контроля заданной линейной территории необходимо разместить базовые приемо-передающие станции для контроля области, таким образом, чтобы получить максимальное покрытие данной территории при ограничениях на затраты и на время задержки передачи сигнала в тандемной сети, обеспечив при этом наличие связи между крайними шлюзами через систему размещенных станций.

Пусть задано множество станций $S = \{s_j\}, j = 1, 2, .., m$ с параметрами $\{r_j, R_j, \mu_j, c_j\}$: r_j - радиус покрытия станции, R_j - радиус радиорелейной связи, μ_j - интенсивность времени обслуживания и c_j - стоимость. Задан отрезок α длиной L с концами в точках a_0 и a_{n+1}. Внутри отрезка $\alpha = [a_0, a_{n+1}]$ задано множество точек размещения станций $A = \{a_i\}, i = 1, 2, ..., n$.

Точка a_0 имеет координату $l_0 = 0$, точка a_{n+1} имеет координату $l_{n+1} = L$. На концах отрезка, a_0 и a_{n+1} стоят станции специального вида s_0 и s_{m+1}, соответственно, для которых радиусы покрытия $r_0 = r_{m+1} = 0$, радиусами связи,

пропускной способностью и стоимостью в данной постановке задачи можно пренебречь $R_0 = R_{m+1} = \infty$, $\mu_0 = \mu_{m+1} = \infty$ и $c_0 = c_{m+1} = 0$.

Требуется разместить станции из множества S таким образом, чтобы максимизировать покрытие отрезка L контролирующими станциями при выполнении требования наличия связи между станциями s_0 и s_{m+1} (шлюзами) через систему размещенных базовых станций при выполнении ограничений на время межконцевой задержки T и суммарную стоимость размещенных станций C.

3. Система массового обслуживания для сети с линейной топологией

Для оценки характеристик производительности тандемной сети рассматривают модель многофазной сети массового обслуживания [10] - [11] . Рассмотрим нашу беспроводную широполосную сеть как сеть массового обслуживания с кросс-трафиком и с узлами $M/M/1$, то есть с простейшим входящим потоком и показательным распределением времени нахождения пакета в узле.

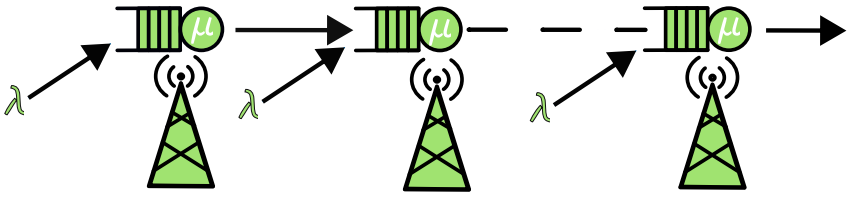


Рис. 1. Сеть массового обслуживания $M/M/1 \to ... \to \cdot/M/1$.

Интервал между поступлениями задается случайной величиной $A \sim f_A(t)$, $f_A(t) = \lambda e^{-\lambda t}$ и время обслуживания в такой системе задается непосредственно с помощью случайной величины $B \sim f_B(t)$, $f_B(t) = \mu e^{-\mu t}$. В такой системе выходящий поток с каждой станции также является простейшим с интенсивностью λ.

Среднее время между поступлениями пакетов известно и равно $\bar{t}$. Интенсивность входящих пакетов равно

$$\lambda = \frac{1}{\bar{t}}. \tag{1}$$

Для станции s_j коэффициент загрузки равен

$$\rho_j = \frac{\lambda}{\mu_j}. \tag{2}$$

Среднее число пакетов в такой системе

$$\overline{N} = \frac{\rho_j}{1-\rho_j}. \tag{3}$$

По теореме Литтла и с учетом (1), (2) и (3) средняя задержка по времени на каждой станции

$$\overline{T_j} = \frac{\overline{N}_j}{\lambda}. \tag{4}$$

4. Постановка задачи в виде модели ЦЛП.

Введем переменные:

y_i^+ и y_i^- для точек $a_i, i = 0, 1, ..., n, n+1$, т.е. для всех точек a_i. Данные переменные определяют размеры участков отрезка, покрываемые стоящими на них станциями (если на данной точке a_i станция не стоит, то y_i^+ и y_i^- равны нулю). Для шлюзов покрытия слева и справа $y_0^+, y_0^-, y_{n+1}^+, y_{n+1}^-$ равны нулю.

Целевая функция будет представлена как:

$$f = \sum_{i=1}^{n}(y_i^- + y_i^+) \to max. \tag{5}$$

Введем переменные x_{ij}, $i = 1, ..., n$; $j = 1, ..., m$.

$x_{ij} = 1$, если на месте a_i стоит станция s_j; $x_{ij} = 0$, в противном случае.

Также переменную $e_i, i = 1, ..., n$.

$e_i = 1$, если в точке a_i стоит станция; $e_i = 0$, в противном случае.

Для шлюзов определим $e_0 = 1$ и $e_{n+1} = 1$.

Запишем систему ограничений.

В каждой точке может стоять только одна станция, либо никакой

$$e_i = \sum_{j=1}^{m} x_{ij}, i = 1, ..., n. \tag{6}$$

Станции из множества S могут быть либо не размещены, либо размещены только один раз

$$\sum_{i=1}^{n} x_{ij} \le 1, j = 1, ..., m. \tag{7}$$

Величина покрытия не может быть больше радиуса покрытия станции, установленной в точке a_i, и равна 0, если станция в точке a_i не размещена:

$$y_i^+ \le \sum_{j=1}^{m} x_{ij} \cdot r_j, i = 1, ..., n; \tag{8}$$

$$y_i^- \le \sum_{j=1}^{m} x_{ij} \cdot r_j, i = 1, ..., n. \tag{9}$$

Суммарное покрытие между двумя любыми точками a_i и a_k , на которых стоят станции, не может быть больше расстояния между этими точками.

Для всех $i = 1, ..., n$

$$y_i^+ + y_k^- \le \frac{l_k - l_i}{2} \cdot (e_i + e_k) + (2 - e_i - e_k) \cdot L, k = i + 1, ..., n + 1; \tag{10}$$

$$y_i^- + y_k^+ \le \frac{l_i - l_k}{2} \cdot (e_i + e_k) + (2 - e_i - e_k) \cdot L, k = i - 1, ..., 0. \tag{11}$$

Данное условие исключает влияние эффекта перекрытия покрытий станций при подсчете суммарной величины покрытия всего отрезка.

По условиям задачи станция, стоящая в точке $a_i, i = 1, ..., n$, должна быть связана, по крайне мере, с одной станцией слева и с одной станцией справа, включая станции s_0 и s_{m+1}, стоящие в конечных точках a_0 и a_{m+1}.

Введём переменные $z_{ijk}, i = 1, 2, ..., n; j = 1, 2, ..., m; k = 1, 2, ..., n; k \ne i$.

$z_{ijk} = 1$, если в точке a_i стоит станция s_j и она связана со станцией, стоящей в точке a_k; $z_{ijk} = 0$, в противном случае.

$z_{ij0} = 1$, если в точке a_i стоит станция s_j, которая связана со станцией s_0 в точке a_{n+1}; $z_{ij0} = 0$, в противном случае.

$z_{ijn+1} = 1$, если в точке a_i стоит станция s_j, которая связана со станцией s_{n+1} в точке a_{n+1}; $z_{ijn+1} = 0$, в противном случае.

В обеих точках i и k должны стоять станции, чтобы можно было их связать.

Для всех $i = 1, ..., n; j = 1, ..., m; k = 1, .., n; k \ne i$

$$z_{ijk} \le e_i; \tag{12}$$

$$z_{ijk} \le e_k. \tag{13}$$

Необходимо, чтобы j-ая станция, стоящая на i-ом месте была связана, по крайне мере, с одной любой станцией, расположенной на k-ом месте, справа от $a_i (k > i)$.

Для всех $i = 1, ..., n; j = 1, ..., m$

$$\sum_{k=i+1}^{n+1} z_{ijk} \geq x_{ij}; \tag{14}$$

Также, по крайне мере, с одной любой станцией, расположенной на k-ом месте, находящейся слева от точки a_i $(k < i)$.

$$\sum_{k=0}^{i-1} z_{ijk} \geq x_{ij}, i = 1, ..., n; j = 1, ..., m; \tag{15}$$

$$\sum_{\substack{i=1 \\ i \neq k}}^{n} \sum_{j=1}^{m} z_{ijk} \geq e_k, k = 1, ..., n. \tag{16}$$

Неравенства (14) – (16) обеспечивают условие симметричности связи между станциями s_i и s_k для всех i, k.

Радиус связи станций s_j, стоящей в точке $a_i, i = 1, ..., n$, должен быть не меньше, чем расстояние до точки a_k, где стоит станция, с которой она связана.

Для всех $i = 1, ..., n$

$$z_{ijk} \cdot (R_j - (a_i - a_k)) \geq 0, k = i - 1, ..., 0; j = 1, ..., m; \tag{17}$$

$$z_{ijk} \cdot (R_j - (a_k - a_i)) \geq 0, k = i + 1, ..., n + 1; j = 1, ..., m. \tag{18}$$

Необходимо учитывать ограничение на время межконцевой задержки в сети T. Используя формулу (4) для расчета задержки на каждой станции запишем неравенство

$$\sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} \cdot \overline{T_j} \leq T. \tag{19}$$

И ограничение на стоимость

$$\sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} \cdot c_j \leq C. \tag{20}$$

5. Заключение

В работе рассмотрена оптимизационная задача выбора из заданного избыточного множества и расстановки базовых станций беспроводных широкополосных сетей связи на множестве возможных точек размещения с целью максимального охвата контролируемой линейной территории при выполнении технологических условий и ограничения на стоимость размещенных станций.

Для формулировки ограничения на время задержки в сети, рассмотрена тандемная сеть как сеть массового обслуживания с узлами $M/M/1$.

Исследуемая задача сформулированна в виде целочисленного линейного программирования (5) – (20). В дальнейших исследованиях планируется использование полученной модели в практических приложениях.

Литература

1. Amine, O., Khireddine, A.,: Base Station Placement Optimization Using Genetic Algortithm. International Journal of Computer Aided Engineering and Technology (IJCAET), Vol. 11, No. 6, 2019.
2. Brahim, M., Drira, W., Filali, F.: Roadside units placement within city scaled area in vehicular ad-hoc networks. 3rd International Conference on Connected Vehicles and Expo (ICCYE 2014), 2014.
3. Chattopadhyay, A., Blaszczyszyn, B., Keeler, H. P. (2018). Gibbsian On-Line Distributed Content Caching Strategy for Cellular Networks. IEEE Transactions on Wireless Communications, 17(2), 969–981.
4. H. E. Kiziloz, On Base Station Localization in Wireless Sensor Networks. Balkan Journal of Electrical and Computer Engineering - (BAJECE) Vol. 8, No. 1, January 2020, 57-61
5. Liu, H., Ding, S., Yang, L., Yang, T.: A connectivity-based strategy for roadside units placement in vehicular ad hoc networks. International Journal of Hybrid Information Technology 7, 91–108, 2014.
6. Reis, A., Sargento, S., Neves, F., Tonguz, O.: Deploying roadside units in sparse vehicular networks: What really works and what does not. IEEE Transactions on Vehicular Technology 63, 2794–2806, 2014.
7. Shen, C., Yun, M., Arora, A., Choi, H.-A. (2019). Efficient Mobile Base Station Placement for First Responders in Public Safety Networks. Advances in Biochemical Engineering/Biotechnology, 634–644.
8. R. Ivanov, O. Pershin, A. Larionov, V. Vishnevsky. On a Problem of Base Stations Optimal Placement in Wireless Networks with Linear Topology // Communications in Computer and Information Science. 2018. vol. 919. p. 505-513.

9. Ivanov R., Mukhtarov, A., Pershin, O.: Problem of Optimal Location of Given Set of Base Stations in Wireless Networks with Linear Topology / Proceedings of the 22nd International Conference on Distributed Computer and Communication Networks: Control, Computation, Communications (DCCN-2019, Moscow). Cham: Springer, 2019. p. 53-64.
10. Bendel, D., Haviv, M. Cooperation and sharing costs in a tandem queueing network. European Journal of Operational Research, 271(3), 926–933, 2018.
11. Wu, K., Shen, Y., Zhao, N. (2017). Analysis of tandem queues with finite buffer capacity. IISE Transactions, 49(11), 1001–1013.

UDC: 519.218

Modeling and Simulation of Reliability Function of a k-out-of-n:F System with Partial Repair

Ivanova N.M.[1,2]

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

[2]V.A.Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, 65 Profsoyuznaya street, Moscow, 117997, Russia

nm_ivanova@bk.ru

Abstract

A heterogeneous hot standby repairable k-out-of-n system is considered. In one of the previous papers reliability characteristics for such system for the case of $k = 2$ and $k = 3$ with exponential lifetime distribution have been found and sensitivity analysis of their reliability characteristics to the shape of the repair time distribution of their elements has been performed. In this paper the problem of asymptotic insensitivity of such system is studied with the help of a simulation approach for a special case of a 3-out-of-6 system when both life and repair time have general distributions.

Keywords: k-out-of-n system, reliability function, mathematical modeling and simulation, AnyLogic Environment, markovization method, sensitivity analysis

1. Introduction

Such systems as k-out-of-n are often found in various fields of human activity. They are used in areas such as telecommunication, transmission, transportation, manufacturing, and service applications. Therefore, reliability study of k-out-of-n systems are important not only from a theoretical point of view, but also from a practical one. Unfortunately, the calculation of some reliability characteristics by analytical methods becomes difficult even when considering relatively simple systems. This task becomes more complicated in case the repair time of the system's elements is considered non-exponential. The solution to this problem can be the usage of simulation methods for calculating the necessary characteristics.

A k-out-of-n system is a repairable system that consists of n elements. The study of the system is determined according to its definition. If we say that it remains operational when k out of n elements fail, it is a k-out-of-n:G system. In the other

way, it is a k-out-of-n:F system, which means that the system fails when k of n elements fail. In this paper we consider the second one [1, 2].

Such systems have been widely used in practical fields such as data transmission, redundant networks, production management, transportation, voting systems, radar systems, etc. It can have different configurations and dependencies, which greatly affects the reliability of the whole system. In paper [3], for example, a reliability analysis of a k-out-of-n system in the presence of two types of failure is performed. Definite system can be considered as a mathematical model in communication and engineering systems, in a nuclear power plant. Another important application is the reliability study of high-altitude unmanned rotor-craft platforms [4], in which the multi-rotor architecture of such platforms allows a platform with n rotary-wing engines to stay operational even after $k-1$ engines fail.

Due to the constant development and complication of such systems in practice, the implementation of their research is also complicated. Therefore, researchers resort to new methods, such as simulation. In paper [5] the study of the reliability function of a homogeneous hot double redundant repairable system is extended with the help of a discrete-event simulation model. In [6] a simulation method is used to calculate the steady-state probabilities of a heterogeneous double redundant hot standby repairable system.

This paper continues the study of a hot standby repairable system using simulation in the AnyLogic Environment. A k-out-of-n:F system is considered in a special case and its reliability function and the mean time to failure are studied.

2. Problem Setting and Notation

Consider a heterogeneous hot standby repairable system of the k-out-of-$n : F$ type. Such a system can be considered as a system with a parallel connection of elements, the failure of which will occur if less than $(n-k)$ elements out of n are operational. Fig. 1 shows the k-out-of-n system, which loses its working capability if any of its k elements fail. The dashed line means that k working elements are necessary for the system's functionally, and the allocation of elements $1, 2, ..., k$ is conditional. In reality all n elements are identical and any k of them may fail.

Suppose that

- the system works till it first enters state k, which is the state of the system failure;
- the failed elements of the system are repaired by a single repair facility;
- the elements fail according to a Poisson flow with intensity α;
- the random repair times of elements are independent and their common cumulative distribution function (c.d.f.) $B(t)$ is absolutely continuous with probability density function (p.d.f.) $b(t) = B'(t)$.

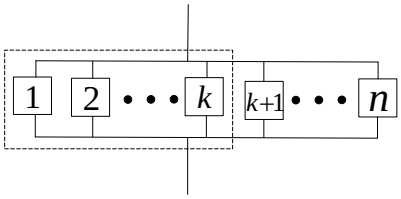


Fig. 1. A k-out-of-$n : F$ system.

The system state space can be represented as $\mathbf{E} = \{0, 1, ...k\}$, which means:

- 0 – all n elements operate;
- j – j elements out of n ($j = \overline{1, k-1}$) have failed, one of them is being repaired, and others $(n-k)$ operate;
- k – k elements have failed which means the system failure and its restoration, "DOWN" state.

Using the so-called *markovization method* [7] introduce as a supplementary variable $X(t)$ — the elapsed repair time of the element under repair, and consider a two-dimensional stochastic process

$$Z = \{Z(t) = (J(t), X(t)),\ \ t \geq 0\},$$

where the value $J(t)$ represents the number of failed elements at time t. 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

$$\pi_j(t; x) = \mathbf{P}\{J(t) = j,\ X(t) = x\}\ \ (j = \overline{1, k})$$

and corresponding macro-state probabilities for $t \geq 0$ by

$$\pi_j(t) = \mathbf{P}\{J(t) = j\} = \int_0^t \pi_j(t; x)dx.$$

The paper deals with the system's reliability function $R(t) = \mathbf{P}\{T > t\}$, where T is the system's lifetime, $T = \inf\{t :\ J(t) = k\}$.

3. Analytical Results

This paper deals with the special case of a k-out-of-n: F system, when $k = 3$ and $n = 6$. To construct the appropriate Markov process we introduce the following notations and present the transition graph of the 3-out-of-6 system (see fig. 2).

- j — the number of elements in the "DOWN" state,
- $\lambda_j = (n-j)\alpha$ — the system failure intensity in its j-th state,
- $\beta(x) = \frac{B'(x)}{1-B(x)}$ — conditional repair density of elements, given elapsed repair time is x,
- $\tilde{b}(s)$ — Laplace transform (LT) of the p.d.f. $b(t)$,
- $b = \int\limits_0^\infty (1-B(x))dx$ — mean repair time of a failed element.

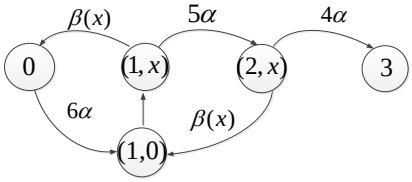


Fig. 2. Transition graph of the 3-out-of-6 system with absorption

3.1. Reliability Function. According to the fig. 2, the system of Kolmogorov forward partial differential equations for process Z in the scope $0 < x < t < \infty$ has the following form

$$\begin{aligned}
\frac{d}{dt}\pi_0(t) &= -\lambda_0\pi_0(t) + \int_0^t \pi_1(t,x)\beta(x)dx, \\
\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial x}\right)\pi_1(t;x) &= -(\lambda_1 + \beta(x))\pi_1(t;x), \\
\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial x}\right)\pi_2(t;x) &= -(\lambda_2 + \beta(x))\pi_2(t;x) + \lambda_1\pi_1(t;x) \\
\frac{d}{dt}\pi_3(t) &= \lambda_2 \int_0^t \pi_2(t;x)dx.
\end{aligned} \tag{1}$$

jointly with the initial

$$\pi_0(0) = 1, \tag{2}$$

and the boundary conditions

$$\begin{aligned}
\pi_1(t;0) &= \lambda_0\pi_0(t) + \int_0^t \pi_2(t;x)\beta(x)dx, \\
\pi_2(t;0) &= 0.
\end{aligned} \tag{3}$$

Remark 1. Note that the second boundary condition follows from the fact that the process never occurs into the state 2 with the elapsed time x equal to zero since

the process enters this state only as a result of failure of another element and the transition from the state $(1, x)$ with the same elapsed repair time.

Theorem 1. The Laplace Transforms (LT) $\tilde{R}(s)$ of the reliability function $R(t)$ of the 3-out-of-6: F system is

$$\tilde{R}(s) = \frac{C_3(s) \cdot s^2 + C_2(s) \cdot s + C_1(s) + C_0}{\Delta}, \tag{4}$$

where

$$\begin{aligned}
C_3(s) &= \lambda_1 \left(1 + \tilde{b}(s+\lambda_1) - \tilde{b}(s+\lambda_2)\right) - \lambda_2, \\
C_2(s) &= \lambda_1(1 - \tilde{b}(s+\lambda_2))(\lambda_0 + \lambda_1) - \\
&- \lambda_2(1 - \tilde{b}(s+\lambda_1))(\lambda_0 + \lambda_2) - \lambda_1\lambda_2(\tilde{b}(s+\lambda_2) - \tilde{b}(s+\lambda_1)), \\
C_1(s) &= \lambda_2\tilde{b}(s+\lambda_1)(\lambda_1^2 + \lambda_0\lambda_2) - \lambda_1^2\tilde{b}(s+\lambda_2)(\lambda_0 + \lambda_2), \\
C_0 &= (\lambda_1 - \lambda_2)\left(\lambda_0(\lambda_1 + \lambda_2) + \lambda_1\lambda_2\right), \\
\Delta &= (s+\lambda_1)(s+\lambda_2)\left((s+\lambda_0)(\lambda_1(1 - \tilde{b}(s+\lambda_2)) - \lambda_2) + \tilde{b}(s+\lambda_1)(s\lambda_1 + \lambda_0\lambda_2)\right)
\end{aligned}$$

Corollary 1. The expectation of the system lifetime has the following form:

$$\begin{aligned}
\mathbf{E}[T] = \frac{1}{\lambda_2} &+ \frac{(\lambda_2 - \lambda_1)(1 - \tilde{b}(\lambda_1))}{\lambda_1\left[\lambda_2(1 - \tilde{b}(\lambda_1)) - \lambda_1(1 - \tilde{b}(\lambda_2))\right]} + \\
&+ \frac{\lambda_2 - \lambda_1(1 + \tilde{b}(\lambda_1) - \tilde{b}(\lambda_2))}{\lambda_0\left[\lambda_2(1 - \tilde{b}(\lambda_1)) - \lambda_1(1 - \tilde{b}(\lambda_2))\right]}.
\end{aligned}$$

4. Simulation Results

In this section we present the results of simulation of the 3-out-of-6: F system with one repair unit and general distributions of both life and repair times of its elements.

To build a simulation model of the studied system, the AnyLogic simulation software developed by the Russian company "The AnyLogic Company" was chosen. Simulation modeling is conducted with the help of the discrete-event modeling method. It means that the system is modeled as a process, i.e. a sequence of operations being performed across entities.

All simulation experiments were conducted with the total simulation time $T = 10^4$.

In the first experiment, the Exponential distribution ($Exp(\alpha)$) is used for both life with a parameter α and repair times with a parameter β with increasing the

average repair time of systems elements. The average failure time α^{-1} equals 1. The comparison of both analytical and simulation results is shown in Fig. (3). Here we use the average repair time of system elements $b = 1, 4$. As it can be seen in Fig. (3) the analytical plots for reliability functions are higher than the corresponding simulation plots for all cases. However, they are very close to each other over the entire time period. For small values of t the reliability values differ by about $0.1 - 0.15$, while with an increase in t the difference between the curves decreases and equals about 0.01. Moreover it is seen that with the increasing value of b the difference between analytical and simulation results are decreasing. For $t = 1$ the reliability values can be considered as equal. Due to the fact that the difference between the obtained values for analytics and simulation does not exceed 2 %, the results of simulation modeling can be considered satisfactory.

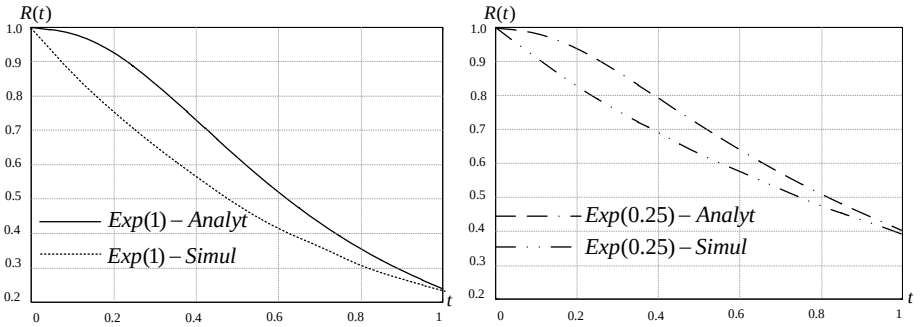


Fig. 3. Reliability function with exponential distribution for both life and repair times. Analytical (left) and simulation (right) results

The values of the average lifetime $\mathbf{E}[T]$ of the system for each case are shown in table 1. It is evident that they are very close to each other but simulation results are a little lower. These values of $\mathbf{E}[T]$ are confirmed by the behavior of the curves in the graphs above.

	$Exp(1)$, $b = 1$	$Exp(0.25)$, $b = 4$
analytical	0.708	1.083
simulation	0.685	1.033

Table 1. The mean time to failure $\mathbf{E}[T]$ of the system

In the second experiment, we can use only the simulation approach due to the impossibility of analytical calculations. Here we use Gamma distribution for the elements' lifetime and Gamma ($\Gamma(k, \theta)$), Pareto ($P(k, x_m)$) and Gnedenko-Weibull

($GW(k, \lambda)$) distributions for the repair times of the system's elements (see Fig. 4). The parameters of all distributions are selected in such a way as to fix the average repair time b as well as the value of its dispersion D. We consider the above mentioned repair time distributions with the fixed average time $b = 1$ and dispersion $D = 0.01$. For the lifetime distributions we compare two following cases:

1) the mean time to failure of the elements is $\alpha = 1$, while its dispersion $D = 1$. In this case the Gamma distribution turns to the Exponential distribution with parameter α^{-1} (curves above).
2) the mean time to failure of the elements is $\alpha = 0.5$, while its dispersion $D = 1$ (curves below).

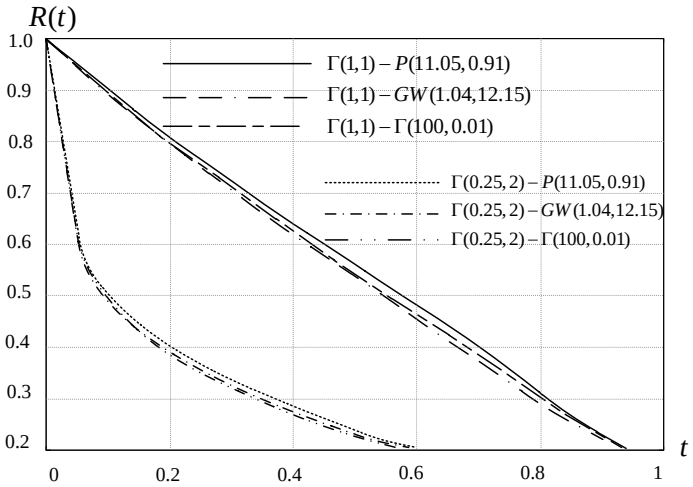


Fig. 4. Reliability function with gamma distribution for life time and general distributions for repair times

The graph shows how much the time to failure affects the behavior of the system. In the first case (when $\alpha = 1$), all the curves are higher relative to the second case ($\alpha = 0.5$). The average lifetime of the system will be higher in the case of less rare failures (see table 2). As we see, different cases of repair time distributions with the same parameters (average repair time of elements and dispersion) show very close results for the empirical reliability function and mean system lifetime. This confirms the asymptotic insensitivity to the form of the repair time distribution.

5. Conclusion

The problem of analytical calculation and simulation assessment of the reliability function for a k-out-of-n system has been considered. The analytical results of the

	$P(11.05, 0.91)$	$GW(1.04, 12.15)$	$\Gamma(100, 0.01)$
$\Gamma(1, 1)$	0.747	0.735	0.741
$\Gamma(0.25, 2)$	0.278	0.265	0.274

Table 2. The mean time to failure $\mathbf{E}[T]$ of the system

reliability function are presented in terms of the Laplace transform. The simulation approach allowed to demonstrate the asymptotic insensitivity of the considered system to the form of the repair time distribution of the system's elements. The analysis of the obtained results shows that the results of the exact analytical calculation and simulation have a close agreement.

REFERENCES

1. Trivedi K. S. Probability and Statistics with Reliability, Queuing and Computer Science Applications. John Wiley & Sons, New York, 2002.
2. Deborah K. Shepherd. k-out-of-n Systems. Encyclopedia of Statistics in Quality and Reliability, John Wiley & Sons, New York, 2008.
3. Medhat El-Damcese, Moustafa Salah Shama. Reliability Analysis of a New k-out-of-n: G Model // World Journal of Modelling and Simulation 16(1), 2020, pp. 3-17.
4. Perelomov V. N., Myrova L. O., Aminev D. A., Kozyrev D. V. Efficiency Enhancement of Tethered High Altitude Communication Platforms Based on Their Hardware-Software Unification. // In: Vishnevskiy V., Kozyrev D. (eds) Distributed Computer and Communication Networks. DCCN 2018. Communications in Computer and Information Science, Volume 919. Springer, Cham, pp. 184–200, 2018. DOI:10.1007/978-3-319-99447-5_16
5. Vladimir Rykov, Dmitry Kozyrev, Elvira Zaripova (2017) Modeling and Simulation of Reliability Function of a Homogeneous Hot Double Redundant Repairable System // Proceedings of the 31st European Conference on Modelling and Simulation, ECMS2017, pp. 701–705, 2017. DOI:10.7148/2017-0701.
6. Rykov V., Zaripova E., Ivanova N., Shorgin S. On Sensitivity Analysis of Steady State Probabilities of Double Redundant Renewable System with Marshal-Olkin Failure Model. // In: Vishnevskiy V., Kozyrev D. (eds) Distributed Computer and Communication Networks. DCCN 2018. Communications in Computer and Information Science, Volume 919. Springer Nature, Cham, pp. 234–245, 2018. doi:10.1007/978-3-319-99447-5_20
7. Cox D. The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables // Mathematical Proceedings of the Cambridge Philosophical Society, 51(3), 433-441. (1955). doi:10.1017/S0305004100030437

UDC: 621.391

The Modeling of Resource Sharing for Heterogeneous Data Streams over 3GPP LTE with NB-IoT Functionality

S.N. Stepanov[1], M.S. Stepanov[2], Umer Andrabi[3], Juvent Ndayikunda[4]

[1,2,4]Moscow Technical University of Communication and Informatics, Department of communication networks and commutation systems, 8A, Aviamotornaya str., Moscow, 111024, Russia

[3]Moscow Institute of Physics and Technology (State University), 9 Institutskiy per., Dolgoprudny, Moscow Region, 141701, Russia

stpnvsrg@gmail.com, mihstep@yandex.ru, umer.andrabi@rediffmail.com, juvndayi@mail.ru

Abstract

The model of resource allocation and sharing for conjoint servicing of real time video traffic of surveillance cameras and NB-IoT data traffic of smart meters and actuators over LTE cell facilities is constructed. In the model the access control is used to create the conditions for differentiated servicing of coming sessions. All random variables used in the model have exponential distribution with corresponding mean values but the obtained results are valid for models with arbitrary distribution of service times. Using the model the main performance measures of interest are given with help of values of probabilities of model's stationary states. The recursive algorithm of performance measures estimation is suggested. The model and derived algorithms can be used for study the scenarios of resource sharing between heterogeneous data streams over 3GPP LTE with NB-IoT functionality.

Keywords: NB-IoT technology, resource allocation and sharing, system of state equations, recursive algorithm

1. Introduction

The essential trend in the development of telecommunications is growth of the volumes and the diversity of Internet of Things (IoT) applications [1–4]. Together with usage of low-traffic smart meters we see the growing impact of multimedia traffic, for example collected by video surveillance systems deployed for security and safety reasons [5]. This trend has been recognized and supported by 3GPP with developing of NarrowBand IoT (NB-IoT) technology, which allows to use the same spectrum by 3GPP LTE high-end equipment and NB-IoT low-end devices [2, 3]. By providing the technical instruments that can be used for radio resources sharing between LTE and

NB-IoT technologies, 3GPP does not formulate the concrete solutions on how these resources should be shared. This problem can be solved by mathematical modeling with taking into account the features of traffic streams forming and accepting for servicing [5–14].

In this paper we address the above mentioned challenges by constructing an analytical framework for modeling the process of resource sharing for an operator planning to create and exploit surveillance system. The system consists of numerous video cameras to perform video monitoring and a large number of smart meters. Both network segments collecting and transfer heterogeneous data streams to analytical centers over existent infrastructure of LTE network (see, Fig. 1).

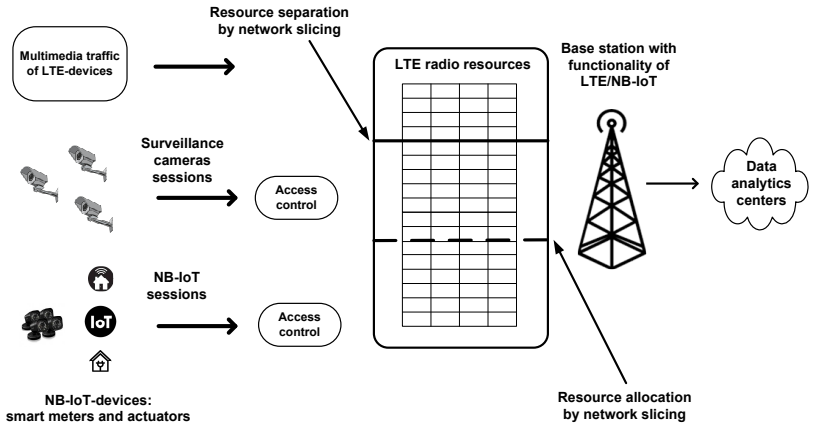


Fig. 1. The functional model of resource sharing between LTE surveillance cameras and NB-IoT sensors.

The proposed model generalizes the results of [5–8] by considering arbitrary number of traffic streams created by video cameras (LTE-devices) and one traffic stream originated from NB-IoT devices. In the model the access control [12] is used to create the conditions for differentiated servicing of coming sessions. All random variables used in the model have exponential distribution with corresponding mean values but the obtained results are valid for models with arbitrary distribution of service times. Three scenarios of resource sharing by coming traffic streams are considered: Slicing when resources are strictly divided among LTE and NB-IoT devices traffic streams; Fully shared, when resources are fully shared and Access controlled, when the access to resource is restricted depending on the amount of resource occupied by corresponding traffic stream.

2. Model Description

We consider an LTE cell with a base station placed in its center and formalize the process of resources sharing. The volume of available radio resources of LTE cell in uplink direction given by network slicing for serving traffic streams originated by surveillance cameras and NB-IoT sensors is measured in units of its smallest granularity. It is clear that the smallest requirement has NB-IoT device session so we can call it NB-IoT resource unit or simply resource unit. Let us suppose that total amount of given resource units is a function of the number of resource blocks (RB) and denote by v, the total number of resource units and by c denote the transmission speed provided by one unit.

Let us suppose that surveillance cameras are varying in quality. It means that corresponding traffic sessions produced by cameras are varying by volume. To take into account this property we consider n types of traffic sessions. Let us suppose that LTE devices traffic sessions of type k are coming after random time having exponential distribution with parameter λ_k, each session requires b_k resource units for servicing and occupies this resource for random time having exponential distribution with parameter μ_k, $k = 1, \ldots, n$. It is suggested that blocked LTE devices sessions are lost without resuming. Let us suppose that traffic sessions produced by NB-IoT devices are coming after random time having exponential distribution with parameter λ_d, each session requires b_d resource units for transmitting of files having exponential distribution with mean F. The service time of NB-IoT session has exponential distribution with mean value $\frac{F}{b_d}$ and parameter $\frac{b_d}{F}$. It is suggested that blocked LTE devices sessions are lost without resuming.

Let us formalize scenarios of resource sharing by coming traffic streams. The simplest scenario corresponds to the case when all v resource units are strictly divided among LTE devices sessions and NB-IoT devices sessions. Let us denote by v_ℓ the number of resource units that is given for exclusive usage to LTE devices sessions and by $v_b = v - v_\ell$ we denote the number of resource units given for exclusive usage to NB-IoT devices sessions correspondingly. By varying the values of v_ℓ and v_b we can give the priority in resource usage to the chosen traffic type but as we show later this way of resource sharing greatly decreases the usage of resource unit.

Next scenario is related with access control. Let us denote for k-th flow of LTE devices sessions by c_k the maximum allowed number of traffic sessions that can be on service at the same time. In a similar way let us denote for NB-IoT devices sessions by c_d maximum allowed number of traffic sessions that can be on service at the same time. For this type of resource usage the traffic session of k-th flow can be blocked for two reasons: (1) if $v_k = c_k b_k$ resource units have already been occupied by sessions from the k-th flow or (2) if total number of busy resource units is greater than $v - b_k$. The same is true for NB-IoT devices sessions. The coming session of this type can be

blocked for two reasons: (1) if $v_d = c_d b_d$ resource units have already been occupied by NB-IoT devices sessions or (2) if total number of busy resource units is greater than $v - b_d$. We show later that by using the access control (by choosing the values of v_k, $k = 1, \ldots, n$ and v_d) we can give the priority in resource usage to chosen traffic type and increase the usage of resource unit compare to static scenario.

The last scenario corresponds to the case when resources are fully shared without giving priority to some traffic streams. In this case we usually increase the usage of resource unit compare to formulated above scenarios but we are not able to reach the same level of sessions losses for all type of traffic streams considered in the model. All three formulated scenarios can be modeled by proper choosing of v and access boundaries v_k, $k = 1, \ldots, n$ and v_d so further we will study only model of resource sharing based on access control.

Let us denote by $i_k(t)$ the number of LTE devices sessions of k-th flow being on servicing at time t, and by $d(t)$ we denote the number of sessions of NB-IoT devices being on servicing at time t. The dynamic of a model states changing is described by Markov process $r(t) = (i_1(t), \ldots, i_n(t), d(t))$, defined on the finite set of model's states S. Let us denote by $(i_1, \ldots, i_n, d)$ the state of $r(t)$. The vector $(i_1, \ldots, i_n, d)$ belongs to S when i_k, $k = 1, \ldots, n$, d varies as follows

$$0 \le i_k \le c_k,\ k = 1, \ldots, n;\ \ 0 \le d \le c_d;\ \ i_1 b_1 + \ldots + i_n b_n + d b_d \le v. \qquad (1)$$

Let us denote by i for state $(i_1, \ldots, i_n, d) \in S$ the number of occupied resource units $i = i_1 b_1 + \ldots + i_n b_n + d b_d$.

Let us denote by $p(i_1, \ldots, i_n, d)$ the value of stationary probability of state $(i_1, \ldots, i_n, d) \in S$. It can be interpreted as portion of time the model stays in the state $(i_1, \ldots, i_n, d)$. This interpretation gives the possibility to use the values of $p(i_1, \ldots, i_n, d)$ for estimation of model's main performance measures. Let us define for k-th flow of LTE devices traffic by π_k the portion of lost sessions and by m_k the mean number of occupied resource units. Their formal definitions are as follows

$$\pi_k = \sum_{(i_1,\ldots,i_n,d) \in U_k} p(i_1, \ldots, i_n, d); \quad m_k = \sum_{(i_1,\ldots,i_n,d) \in S} p(i_1, \ldots, i_n, d) i_k b_k.$$

Here U_k is a subset of S having property $(i_1, \ldots, i_n, d) \in U_k$, if $i_k + 1 > c_k$ or $i + b_k > v$. In the same way we define the performance measures of NB-IoT devices traffic servicing: π_d the portion of lost sessions and m_d the mean number of occupied resource units

$$\pi_d = \sum_{(i_1,\ldots,i_n,d) \in U_d} p(i_1, \ldots, i_n, d); \quad m_d = \sum_{(i_1,\ldots,i_n,d) \in S} p(i_1, \ldots, i_n, d) d b_d.$$

Here U_d is a subset of S having property $(i_1, \ldots, i_n, d) \in U_d$, if $d+1 > c_d$ or $i+b_d > v$.

It can be proved that $r(t)$ is reversible Markov process. From relations of detailed balance follows that values of $P(i_1, \ldots, i_n, d)$ can be found as a unique solution of the system of state equation that has a product form [9–11]

$$p(i_1, \ldots, i_n, d) = \frac{1}{N} \frac{a_1^{i_1}}{i_1!} \cdots \frac{a_n^{i_n}}{i_n!} \frac{a_d^d}{d!}, \quad N = \sum_{(i_1,\ldots,i_n,d)\in S} \frac{a_1^{i_1}}{i_1!} \cdots \frac{a_n^{i_n}}{i_n!} \frac{a_d^d}{d!}. \tag{2}$$

Here $a_k = \lambda_k/\mu_k$ and $a_d = \lambda_d/\mu_d$ are offered traffic expressed in Erlangs and N is a normalizing constant. The values of introduced performance measures can be found by convolution algorithm [9, 12] based on the product form (2).

3. Numerical Assessment

By using the elaborated mathematical model and algorithms of it's performance measures estimation we can analyze the effectiveness of suggested scenarios of resource allocation. The level of traffic load can be characterized by ρ the offered load per one resource unit. To define ρ it is necessary to find the offered load of each traffic stream considered in the model. Let us denote by A_k the offered load expressed in resource units for k-th flow of LTE devices traffic $A_k = \frac{\lambda_k}{\mu_k} b_k = a_k b_k$. Let us denote by A_d the offered load expressed in resource units for flow of NB-IoT devices sessions $A_d = \frac{\lambda_d}{\mu_d} b_d = a_d b_d = \frac{\lambda_d F}{b_d}$. Than $\rho = \frac{A_1+\ldots+A_n+A_d}{v}$.

Let us consider the model with parameters: $v = 200$ resource units (r.u.); transmission rate that is provided by one resource unit is $c = 100$ kbit/c; $n = 1$; $b_1 = 10$ r.u.; $b_d = 1$ r.u.; $F = 100$ kbit; $1/\mu_1 = 10$ c; $1/\mu_d = 1$ c. We begin the model's numerical assessment with Fig 2 that presents the values of π_1 and π_d and Fig 3 with mean values of unit usage by LTE devices traffic — δ_1 and NB-IoT devices traffic — δ_d and the their sum $\delta = \delta_1 + \delta_d$ vs the value of ρ the offered load of one resource unit. The values of performance measures are obtained by recursive algorithm based on convolutions. Let us suppose that both traffic flows considered in the model generate the same offered load $A_1 = A_d = \frac{v\rho}{2}$. It allows to find the intensities λ_1, λ_d of sessions coming for each flow considered in the model from known values of ρ. The results presented in Fig. 2 and Fig. 3 show that despite equality of offered traffic NB-IoT-devices sessions obtain priority in occupying the transmission resource that is clearly seen in overload conditions, when $\rho > 1$, (see Fig. 3).

The only way to overcome mentioned difficulties is to create the conditions for differentiated servicing of coming sessions. Three scenarios of resource sharing are compared: Slicing when resources are strictly divided among LTE devices and NB-IoT devices traffic streams, Fully shared, when resources are fully shared and Access controlled, when the access to resource is restricted depending on the amount of

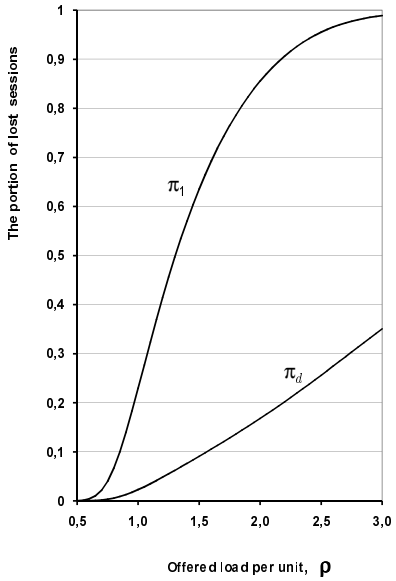


Fig. 2. The portions of lost sessions for LTE and NB-IoT devices

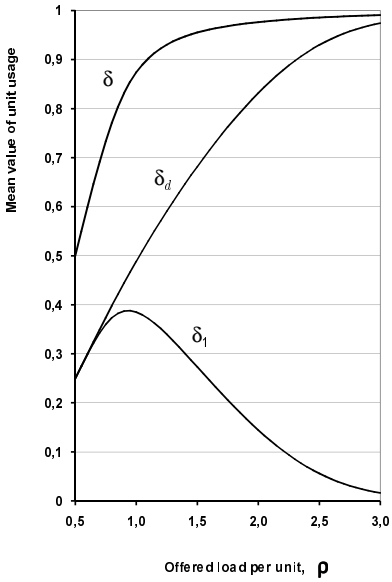


Fig. 3. The values of unit usage by LTE and NB-IoT devices

resource occupied by corresponding traffic stream. We compare the properties of the discussed resource allocation procedures with Fig 4 that presents the portion of the lost LTE devices sessions vs intensity of offered NB-IoT devices sessions and Fig 5 that presents the mean value of resource unit usage vs intensity of offered NB-IoT devices sessions.

The model input parameters are the same as was used in Fig 2 and Fig 3 except $a_1 = 10$ Erl. For Slicing scenario $v_\ell = v_b = 100$ r.u. For Access controlled scenario $v_1 = 200$ r.u., $v_d = 100$ r.u. The presented results can be summarized as follows.

1) The simplest for usage Slicing scenario when resources are strictly divided among LTE devices and NB-IoT devices traffic streams can be used for achievement of prescribed values of performance indicators but have two drawbacks. The first is the high degree of sensitivity of characteristics to the value of offered load that requires a priory knowledge of the traffic intensity. The second is the lower values of resource unit usage compare to the Access controlled and Fully shared scenarios.
2) Fully shared scenario have the best values of resource unit usage but allows the degradation of losses for heavy traffic especially in situation of overload (see, Fig 3).

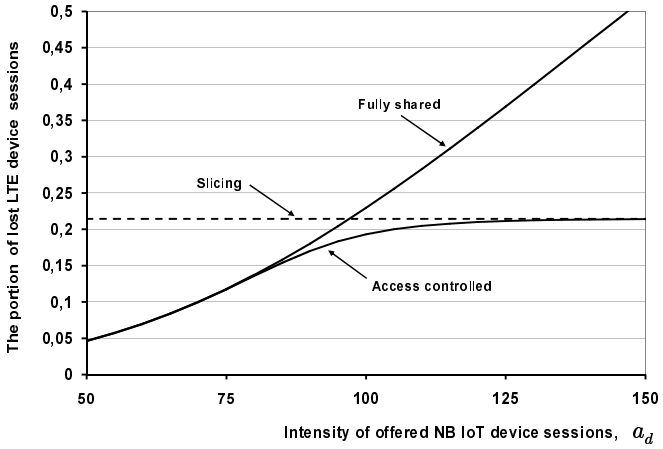


Fig. 4. The portion of the lost LTE devices sessions vs intensity of offered NB-IoT devices sessions.

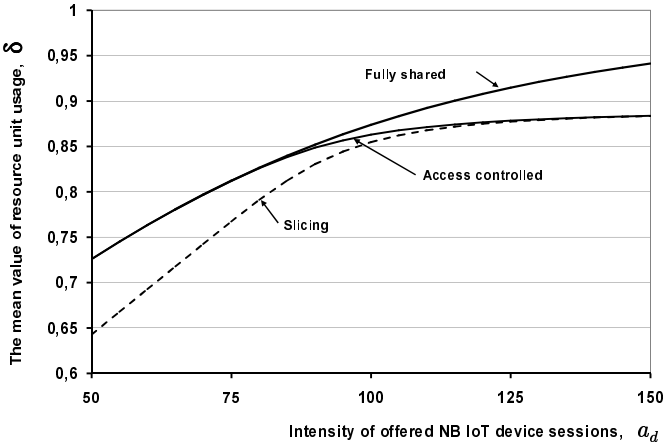


Fig. 5. The mean value of resource unit usage vs intensity of offered NB-IoT devices sessions.

3) Access controlled scenario outperform Slicing scenario and is free from negative features of Fully shared scenario. The suggested procedure of resource allocation is recommended for implementation over 5G mobile networks.

4. Conclusion

The model of resource allocation and sharing for conjoint servicing of real time video traffic of surveillance cameras and NB-IoT data traffic of smart meters and

actuators over LTE cell facilities is constructed. In the model the access control is used to create the conditions for differentiated servicing of coming sessions. All random variables used in the model have exponential distribution with corresponding mean values but the obtained results are valid for models with arbitrary distribution of service times. Using the model the main performance measures of interest are given with help of values of probabilities of model's stationary states. The recursive algorithm of performance measures estimation is suggested.

The constructed analytical framework additionally offers the possibility to find the volume of resource units and access control parameters required for serving incoming traffic with given values of performance indicators. Proposed model can be further developed to include the possibility of reservation and using the processor sharing discipline for serving NB-IoT sessions traffic [14].

REFERENCES

1. Mehmood,Y., Ahmad, F., Yaqoob,I., Adnane, A., Imran, M., Guizani, S.: Internet-of- Things-Based smart cities: recent advances and challenges. IEEE Commun. Mag. **55**(9), 16-24 (2017)
2. Rico-Alvarino, A., Vajapeyam, M., Xu, H., Wang, X., Blankenship, Y., Bern, J., Tirronen, T., Yavuz, E.: An overview of 3GPP enhancements on machine to machine communications. IEEE Commun. Mag. **54**(6), 14–21 (2016)
3. 3GPP. Standardization of NB-IOT completed. `http://www.3gpp.org/news-events/ 3gpp-news/1785-nb_iot_complete, June 2016.`
4. Nokia. Dynamic end-to-end network slicing for 5G. White Paper. (2017)
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. Gudkova, I., Samouylov, K., Buturlin, I., Borodakiy, V., Gerasimenko, M., Galinina, O., Andreev, S.: Analyzing Impacts of Coexistence between M2M and H2H Communication on 3GPP LTE System. Lecture Notes in Comput. Scie., Springer, Cham. **8458**. 162-174 (2014)
7. Stepanov, S., Stepanov, M., Tsogbadrakh, A., Ndayikunda, J., Andrabi, U.: Resource Allocation and Sharing for Transmission of Batched NB-IoT Traffic over 3GPP LTE. The Proc of the 24th Conference of Open Innovations Association (FRUCT). Moscow Technical University of Communications and Informatics. Moscow, Russia. 422-429. (2019)
8. Stepanov, S.N. and Stepanov, M.S.: Efficient Algorithm for Evaluating the Required Volume of Resource in Wireless Communication Systems under Joint

Servicing of Heterogeneous Traffic for the Internet of Things. Automation and Remote Control. **80**(8). 1970-1985 (2019)
9. Iversen, V. B.: Teletraffic Engineering and Network Planning. Technical University of Denmark. (2010)
10. Ross, K.W.: Multiservice Loss Models for Broadband Telecommunications Networks. Springer. (1995)
11. Kelly, F.P.: Blocking probabilities in large circuit-switched networks. Adv. Appl. Prob. **18**. 473–505 (1986)
12. Iversen, V. B.: The exact evaluation of multi-service loss system with access control. Teleteknik. **31**(2). 56–61 (1987)
13. Iversen, V.B., Stepanov, S.N.: The Usage of Convolution Algorithm with Truncation for Estimation of Individual Blocking Probabilities in Circuit-Switched Telecommunication Networks. In: V.Ramaswami and P.E.Wirth (editors). Proceedings ITC 15. Elservier, Amsterdam. 1327-1336 (1997)
14. Stepanov, S.N., Stepanov, M.S.: Planning Transmission Resource at Joint Servicing of the Multiservice Real Time and Elastic Data Traffics. Automation and Remote Control. **78**(11). 2004-2015. (2017)

UDC: 621.391

Estimation of Performance Measures of Emergency Services for Overload of Calls

S.N. Stepanov[1], M.S. Stepanov[2], M.O. Shishkin[3]

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

stpnvsrg@gmail.com, mihstep@yandex.ru, mackschischkin1@yandex.ru

Abstract

The mathematical model of public-safety answering points (PSAP) functioning is constructed and analyzed. In the model the usage of interactive voice response (IVR) and the possibility of call repetition are taken into account. Algorithm of characteristics estimation based on truncation of used infinite space of states and solving the system of state equations is suggested. Relative error of characteristics calculation caused by truncation is found. Approximate algorithm of performance measures estimation is constructed. The usage of the model for elimination of negative effects of PSAP overload is considered.

Keywords: Public-safety answering points, system of state equations, approximate algorithms, repeated attempts

1. Introduction

Technically the possibility to reach primary emergency services such as police, ambulance etc is organized through public-safety answering points (PSAP) [1, 2]. PSAP handles requests from the citizens and dispatches an intervention resources if necessary. In some countries, one number is used for all the emergency services (e.g. 112 in continental Europe including Russia). The functional model of the emergency service in network is presented on Fig 1. After entering to the emergency service call should be distributed to one of available call-taker. In the case of an overload in PSAP1, the call can be routed to the PSAP2 responsible for another geographical region (see Fig 1).

Numerous references describing mathematical [3–9] and engineering [10–13] backgrounds of information services modeling can be found in the literature. Special attention is paid to the modeling of overload when customer with some probability repeats the unsuccessful request [1–3, 8–10, 14, 15]. The detailed description of the process of call formation and servicing complicates the estimation of characteristics. Often it can be done only by solving the system of state equations by standard

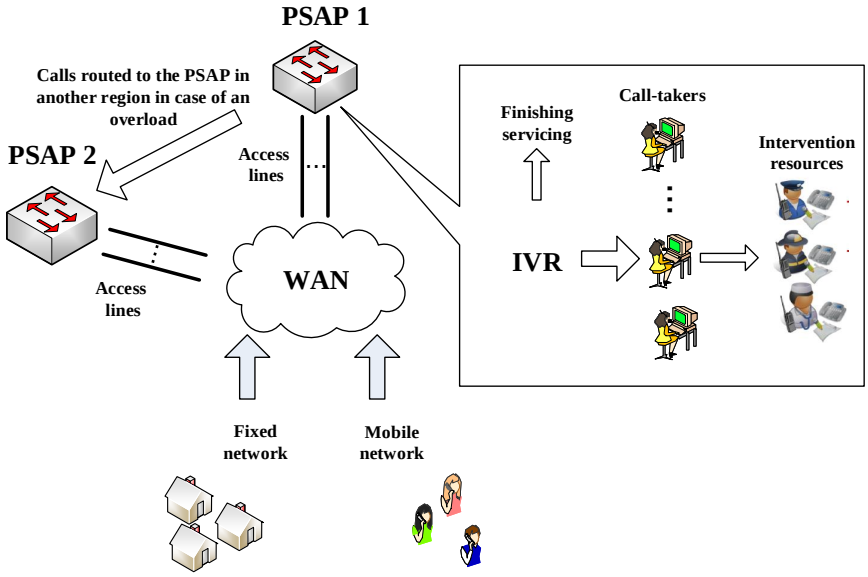


Fig. 1. The functional model of the PSAP place in the network.

algorithms of linear algebra. In doing so it is necessary to truncate the model's state space. The correct usage of this approach needs in determination of the error caused by truncation. Another actual problem is elaborating of simple approximations for main performance measures that can be used for calculation of characteristics in case of overload. In the paper formulated tasks were solved for the model of PSAP where usage of interactive voice response (IVR) and the possibility of call repetition in case of blocking or unsuccessful waiting time are taken into account.

2. Model Description

Calls for getting emergency service are entering the PSAP through telephone access lines. After occupying an access line a call can be served by IVR and if it is necessarily by PSAP call-takers. Let us denote by v the overall number of call-takers and by $w + v$ we denote the overall number of access lines. The PSAP functioning is considered in case of overload. It means that apart from primary calls that arrive for servicing according to the Poisson model with intensity λ the emergency center serves the flow of repeated calls caused by insufficient number of free call-takers and access lines or by unsuccessful finishing the time of waiting the beginning of service. In both situations, a calling citizen with probability H repeats the request for servicing after

random time having exponential distribution with parameter ν and with additional probability $1 - H$ a citizen stops his attempts to find free call-taker and leaves the system unserved. It is supposed that maximum allowed time of waiting the beginning of servicing at PSAP has exponential distribution with parameter σ.

The process of call servicing at PSAP includes two stages. The first stage consists in getting a recorded message from the IVR and second consists in exchanging of information with call-taker. It is supposed that duration of call-taker's service has exponential distribution with parameter μ. The transition to call-taker's servicing is depending on the type of the call: primary or repeated. With probability q_p for primary call and with probability q_r for repeated attempt after getting service from IVR a citizen is trying to get servicing from PSAP call-taker. With additional probabilities $1 - q_p$ and $1 - q_r$ correspondingly a citizen leaves the system satisfying by the servicing at IVR.

Let us denote the state of the model by $(j, i) \in S$ where j is the number of citizens repeating a call, i is the number of occupied call-takers and access lines $j = 0, 1, \ldots, ; \quad i = 0, 1, \ldots, v + w$ and S is the model space of states. Let us denote by $j(t)$ the number of citizens repeating a call at time t and by $i(t)$ we denote the number of busy call-takers and occupied waiting places at time t. The model functioning is described by Markov process $r(t) = (j(t), i(t))$, defined on the infinite space of states S.

Let us denote by $p(j, i)$ the probability of stationary state $(j, i) \in S$ of the considered PSAP model and define main performance measures of the process of calls serving. The first group of characteristics are mean values of components of the model state. Let us denote by M_r, M_i, M_w mean numbers of, respectively, citizens repeating a call, occupied call-takers and occupied waiting positions. Next group are intensities of coming and blocking calls. We denote by I_b, I_o and I_t correspondingly the intensity of calls lost in attempt to get service from call-takers, arrived to get service from call-takers and arrived to get service at PSAP. Key performance measures of PSAP functioning are defined as follows. Let us denote by π_t the portion of time when all call-takers and waiting positions are occupied, by π_c we denote the ratio of lost calls arrived to get service at PSAP, by T_w we denote the mean time for call to be on waiting or servicing, by M we denote the mean number of retrials per one primary call, by τ we denote the portion of repeated calls in the total flow of calls. The definitions of introduced performance measures are looking as follows

$$M_r = \sum_{j=0}^{\infty} \sum_{i=0}^{v+w} p(j, i) j; \quad M_i = \sum_{j=0}^{\infty} \left(\sum_{i=0}^{v} p(j, i) i + v \sum_{i=v+1}^{v+w} p(j, i) \right); \tag{1}$$

$$M_w = \sum_{j=0}^{\infty} \sum_{i=v+1}^{v+w} p(j,i)(i-w); \quad I_b = \sum_{j=0}^{\infty} p(j,v+w)(\lambda q_p + j\nu q_r);$$

$$I_o = \lambda q_p + M_r \nu q_r; \quad I_t = \lambda + M_r \nu; \quad \pi_t = \sum_{j=0}^{\infty} p(j,v+w);$$

$$\pi_c = \frac{I_b + M_w \sigma}{I_t}; \quad T_w = \frac{M_i + M_w}{I_o - I_b}; \quad M = \frac{M_r \nu}{\lambda}; \quad \tau = \frac{M_r \nu}{I_t}.$$

3. Estimation of performance measures

The introduced performance measures are expressed through values of $p(j,i)$. To find them it is necessary to compose and solve the system of state equations. It can be done by standard procedures used in theory of network modeling. By algebraic transform of the system of state equations it is possible to derive equations that relates (1)

$$M_r \nu = (I_b + M_w \sigma) H; \tag{2}$$

$$I_t = \lambda(1 - q_p) + M_r \nu (1 - q_r) + I_b + M_w \sigma + M_i \mu. \tag{3}$$

To solve the system of state equations it is necessarily to truncate the number of repeating citizens by applying inequality $j \le j_m$, where j_m is some integer number, and find the values of $p(j,i)$ by ordinary algorithms of linear algebra. Let us denote performance measures of truncated model by the same symbols that used for initial model only with superscript $*$ and find the error caused by truncation. The analog of (2) for truncated model is looking as follows

$$M_r^* \nu = (I_b^* + M_w^* \sigma) H - \gamma, \tag{4}$$

where γ defined as $\gamma = p^*(j_m, v+w)\lambda H q_p + \sum_{i=v+1}^{v+w} p^*(j_m, i)(i-v)\sigma H$.

Let us denote by Δ the difference between exact value of characteristic and their estimate obtained with help of truncated model, for example, $\Delta M_r = M_r - M_r^*$. By using the basic property of exponentially distributed variables and ideas used in [14] it can be proved that the following inequalities are true

$$\Delta M_r \ge 0; \quad \Delta I_b + \Delta M_w \sigma \ge 0; \quad \Delta M_i \ge 0. \tag{5}$$

For main performance measures from (2)–(5) follows upper estimates of error caused by truncation as function of γ

$$\gamma \le \Delta M_r \nu \le \frac{\gamma}{1 - q_r H}; \quad 0 \le \Delta I_b + \Delta M_w \sigma \le \frac{\gamma q_r}{1 - q_r H}; \quad 0 \le \Delta M_i \mu \le \gamma q_r. \tag{6}$$

For other model's characteristics that can be expressed as function of M_r, M_i, M_w, I_t and model's input parameters the estimation of relative error can be obtained with help of (6). For example, for π_c the following inequality is true

$$\delta\pi_c = \left|\frac{\Delta\pi_c}{\pi_c}\right| \leq \frac{\gamma}{1 - q_r H}\left(\frac{q_r}{I_b + M_w\sigma} + \frac{1}{\lambda + M_r\nu}\right). \tag{7}$$

From (6) follows that error of estimation of M_r is proportional to γ. Let us denote by $\Delta^b M_r$ and by $\Delta^a M_r$ correspondingly the lower and upper estimates of ΔM_r presented at (6) and consider a numerical example that illustrates their accuracy. Model input parameters are as follows: $\lambda = 30$; $q_p = 0{,}5$; $q_r = 0{,}9$; $H = 0{,}9$; $\nu = 5$; $\mu = 1$; $\sigma = 0{,}5$; $v = 10$; $w = 5$. The value of j_m varies from 2 to 40. The values of characteristics found for $j_m = 40$ are considered as found for unlimited interval of varying j. As a time unit was chosen the mean time of servicing a request by call-taker. In the Table 1 are presented the values of j_m and depending on j_m the values of π_c, M_r, ΔM_r, the lower $\Delta^b M_r$ and upper $\Delta^a M_r$ estimation of ΔM_r found from (6) and value of γ. From the content of the table it is seen that upper estimate of ΔM_r has very good accuracy. As result after making calculation of performance measures with help of truncated model we can find the error caused by truncation in terms of characteristics of truncated model. More details about using the concept of truncation can be found in [14].

j_m	π_c	M_r	$\Delta^b M_r$	ΔM_r	$\Delta^a M_r$	γ
2	0,280257	1,07070753	$7{,}13 \cdot 10^{-1}$	$3{,}70 \cdot 10^{0}$	$3{,}75 \cdot 10^{0}$	$3{,}56 \cdot 10^{0}$
4	0,360144	2,12947699	$5{,}06 \cdot 10^{-1}$	$2{,}64 \cdot 10^{0}$	$2{,}66 \cdot 10^{0}$	$2{,}53 \cdot 10^{0}$
8	0,451382	3,79243499	$1{,}86 \cdot 10^{-1}$	$9{,}73 \cdot 10^{-1}$	$9{,}77 \cdot 10^{-1}$	$9{,}28 \cdot 10^{-1}$
10	0,472605	4,28008825	$9{,}25 \cdot 10^{-2}$	$4{,}85 \cdot 10^{-1}$	$4{,}87 \cdot 10^{-1}$	$4{,}63 \cdot 10^{-1}$
20	0,491770	4,76361833	$2{,}80 \cdot 10^{-4}$	$1{,}47 \cdot 10^{-3}$	$1{,}48 \cdot 10^{-3}$	$1{,}40 \cdot 10^{-3}$
30	0,491825	4,76509130	$3{,}02 \cdot 10^{-8}$	$1{,}59 \cdot 10^{-7}$	$1{,}59 \cdot 10^{-7}$	$1{,}51 \cdot 10^{-7}$
40	0,491825	4,76509146				

Table 1. The dependence of error caused by truncation on j_m

4. Approximate calculation of performance measures

Let us derive the approximate algorithm of performance measures calculation. In doing this we construct simplifying model of PSAP functioning by supposing that the flow of retrials in the considered model is poissonian with some intensity $x - \lambda$, where x is unknown intensity of total poissonian flow of primary and repeated calls. Let us indicate the obtained estimates by the same symbols that was used

for corresponding characteristics of the initial model only with superscript $*$ and suppose that for obtained in this way estimates the relation (2) is true. It gives

$$x = (I_b^* + M_w^* \sigma)H + \lambda, \tag{8}$$

where characteristics

$$I_b^*(x) = \lambda q_p \pi_t^*(x) + (x - \lambda)\pi_t^*(x) q_r; \quad \pi_t^*(x) = p(v + w); \quad M_w^*(x) = \sum_{i=v+1}^{v+w} p(i)(i - v)$$

are functions of x. Values of $p(i)$, $i = 0, \ldots, v + w$ are calculated from relations of detailed balance $p(i)\Lambda = p(i + 1)((i + 1)\mu I(i < v) + (v\mu + (i + 1 - v)I(v \geq v))$; $i = 0, \ldots, v + w - 1$, where $\Lambda = \lambda q_p + (x - \lambda)q_r$.

From (8) we obtain equation for determination of x

$$x = \frac{\lambda(1 + \pi_t^*(x)H(q_p - q_r)) + M_w^*(x)\sigma H}{1 - \pi_t^*(x) q_r H}. \tag{9}$$

It is easy to prove that (9) has solution, this solution is unique and can be obtained by successive substitutions. By implementing the approach used in [15] it is possible to prove that obtained estimates are asymptotically correct when $\lambda \to \infty$.

5. The usage of the model for elimination of PSAP overload

The process of normal functioning of PSAP can be disturbed by increasing the intensity of coming requests. The overload can be caused by many reasons that are discussed in Section 1. To decrease the negative consequences of input flow fluctuations we can use the procedures of the filtering the input flows of primary calls or repeated attempts. Another possibility is to redirect part of the input flow of primary calls to other PSAPs (see, Fig 1). The consequences of usage these and other procedures aimed to elimination of overload and estimation of necessary volumes of access lines and call-takers can be studied with help of constructed model. Because shortage of place let us consider only one example.

In case of overload part of primary flow can be redirected to other PSAP with similar service facilities. The exact proportion can be found with help of constructed model. We illustrate the procedure of redirecting by numerical example. Model's input parameters are are as follows: $\lambda = 24$; $v = 10$; $w = 5$; $H = 0{,}9$; $\nu = 5$; $j_m = 50$; $\mu = 1$; $\sigma = 0{,}1$. As a time unit was chosen the mean time of servicing a request by call-taker. The portion r of redirected primary calls defined as $r = \frac{24 - \lambda_c}{24}$ and varies from 0 to 0,4. Here λ_c is current value of intensity of primary calls that in the considered case consequently decreases from 24. The required level of service should

satisfy the inequality $\pi_c < 0{,}05$. The Fig 2 shows the dependence of π_c on r. The results of calculations show that by redirecting primary calls we can decrease the value of losses in efficient way. The constructed model allows to study the process of forming and servicing requests in case of overload and choose the right values of parameters that can be used for control the PSAP functioning in case of overload.

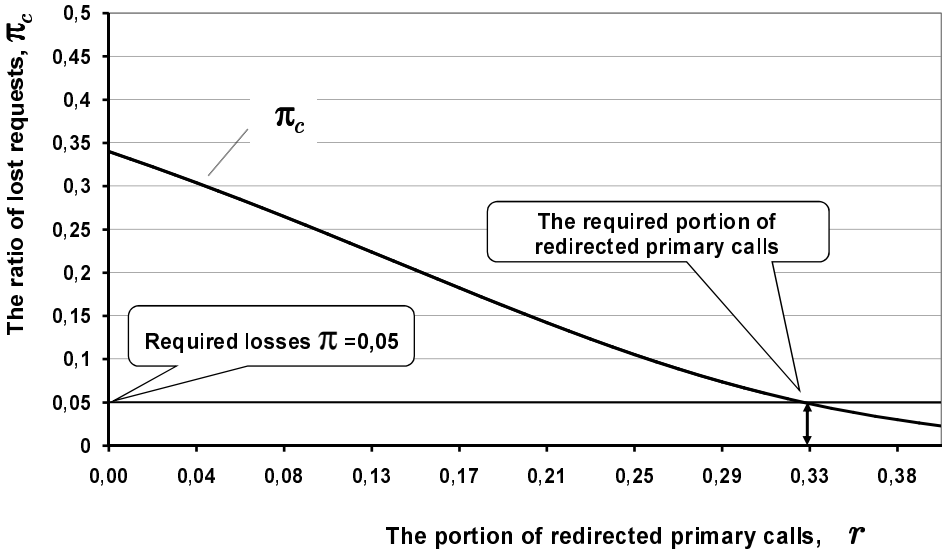


Fig. 2. The dependence of π_c on the portion of primary calls redirected to other PSAP.

6. Conclusion

In this paper the mathematical model of PSAP is constructed and analyzed. In the model the usage of interactive voice response and the possibility of call repetition in case of blocking or unsuccessful waiting time are taken into account. Algorithm of characteristics estimation is suggested based on truncation of the used infinite state space and solving the system of state equations. Relative error of characteristics calculation caused by truncation is found. Approximate algorithm of performance measures estimation is constructed. It is shown that obtained estimates are asymptotically correct in case of overload. The usage of the model for elimination of negative effects of PSAP overload is considered.

REFERENCES

1. The European Emergency Number Association website. Overload of calls. https://eena.org/document/overload-of-calls/
2. Technion website. A Routing Policy for Call Centers Designed to Respond to Unexpected Overloads. http://iew.technion.ac.il/msom2010//msom.technion.ac.il/confprogram/papers/MC/4/38.pdf
3. Stepanov S.N., Stepanov M.S., Zhurko H.: The Modeling of Call Center Functioning in Case of Overload. In: Vishnevskiy V., Samouylov K. (eds) DCCN 2019. Lecture Notes in Computer Science (LNCS). **11965**, 391–406 (2019)
4. Stolletz, R., Helber, S.: Perfomance Analysis of an Inbound Call-Center with Skills-Based Routing. Springer-Vellag, Hannover (2004)
5. Mandelbaum, A., Zeltyn, S.: Staffing many-server queues with impatient customers: constraint satisfaction in call centers. Operations Research. **57**(5), 1189–1205 (2009)
6. Bhulai, S., Koole, G.: A queueing model for call blending in call centers. IEEE Transactions on Automatic Control. **48**(8), 1434–1438 (2003)
7. Colin, M.: Call center service level: A customer experience model from benchmarking and multivariate analysis. Esic Market Economics and Business Journal, **51**(3), 467– 496 (2020)
8. Stepanov, S. N., Stepanov, M. S.: Construction and Analysis of a Generalized Contact Center Model. Automation and Remote Control. **75**(11), 1936–1947 (2014)
9. Stepanov, S. N., Stepanov, M. S.:Algorithms for Estimating Throughput Characteristics in a Generalized Call Center Model. Automation and Remote Control. **77**(7), 1195–1207 (2016)
10. Aguir, S., Karaesmen, F., Aksin, O.Z., Chauvet F.: The impact of retrials on call center performance. OR Spectrum. **26**(3), 353–376 (2004)
11. Aksin, Z., Armony, M., Mehrotra, A.: The modern call center: a multi- disciplinary perspective on operations management research. Production and Operations Management. **16**(6), 665–688 (2007)
12. Whitt, W.: Engineering solution of a basic call-center model. Management Science. **51**(2), 221–235 (2005)
13. Whitt, W.: Staffing a call center with uncertain arrival rate and absenteeism. Production and Operations Management. **15**(1), 88–102 (2006)
14. Stepanov, S. N.: Markov Models with Retrials: The Calculation of Stationary Performance Measures Based on the Concept of Truncation. Mathematical and Computer Modelling. **30**, 207–228 (1999)
15. Stepanov, S. N.: Generalized model with retrials in case of extreme load. Queueing Systems. **27**, 131–151 (1998)

UDC: 001.891

Performance of MATLAB clustering algorithms

A. Ivanov[1], N. Ziazina[1,2], V. Antonova[1,3]

[1]BMSTU, ul. Baumanskaya 2-ya, 5/1, Moscow, Russia
[2]ISP RAS, Alexander Solzhenitsyn st., 25, Moscow, Russia
[3]IRE RAS, Mokhovaya 11-7, Moscow, Russia

iam18u032@student.bmstu.ru, nataliacs@yandex.ru, ant_veronika@bmstu.ru

Abstract

Clustering is one of machine learning's tasks when given objects must be split into specific groups based on distance between them. Its applications include different fields such as pattern matching, data compression and image analysis. Many programing languages allow to create clustering algorithms, though using already implemented ones is much easier. MATLAB includes a few of them. Knowing the performance of MATLAB's cluster analysis algorithms may help choose the more optimal hardware for a given problem.

1. Clustering analysis

In mathematical notation, clustering problem is such: given is a set X: x_1, x_2, x_3, x_4, ... x_m, their labels Y: $y(x_1)$, $y(x_2)$, $y(x_3)$, $y(x_4)$, ..., $y(x_m)$. On the set X a metric $\rho(x, x')$ is given. It is necessary to group the sample into subsets (clusters), assign the label $y_i \in Y$ to each object $x_i \in X$, so that the objects inside each cluster are close relative to the metric ρ, and objects from different clusters are significantly farther. The clustering algorithm is a function F: $X \to Y$, which associates the cluster identifier $y \in Y$ with any object $x \in X$. It is postulated that:

1. The clustering algorithm a is scale invariant.
2. The set of clustering results of algorithm a, depending on the change in the distance function ρ, must coincide with the set of all possible partitions of the set of objects X.
3. The clustering algorithm is consistent.

Clustering results vary between different algorithms. Also, some may require a predefined number of clusters, while others don't.

The work is partially supported by the Russian Foundation for Basic Research (project No. 19-07-00525 A – Developing flow-based models of routing problems in telecommunications networks).

2. Clustering algorithms in MATLAB.

An overview of some of MATLAB's clustering algorithms (built-in):

1. Hierarchical Clustering – does not require number clusters, cannot detect anomalies. It creates a dendrogram, consisting of multiple levels of clusters. Resembles a tree of clusters [3, 4].
2. k-Means – input should specify the number of groups. The shape of the cluster is spheroidal. Not useful for outlier detection. It is assumed that every object belongs to one of k classes, which are defined by a central vector. Classes are formed so that the square of distance from an object to the centroid is minimal [1, 2, 3].
3. Density-Based Spatial Clustering of Algorithms with Noise (DBSCAN) – does not need the number of clusters, can detect oddity in data. It takes the density of objects into account. The shape of clusters is arbitrary [3, 6].
4. Nearest Neighbors – can work without specified number of groups. As they name states, the algorithm is distance-based, and because of that the shape of clusters is arbitrary [3, 5].

Because of the clustering analysis' nature, there is no right or wrong algorithm. It all depends on the data that should be split up. The performance of the algorithms is also dependent on the data.

3. Benchmark details

MATLAB has built-in tools to measure time. It is either *tic*/*toc* start timer/end timer or *timeit* function. This research will use *tic*/*toc* [11]. Time intervals of 500 clustering function calls will be recorded, their average and sample standard deviation will be calculated. In order to minimize distortion, the MATLAB process' priority will be set to Realtime before the benchmark is run. The following hardware will be used during tests:

1. AMD Ryzen 3900X, 12 cores @ 3.8GHz base on x570 chipset and 32GB of DDR4 3200MHz RAM.
2. The processor above with Nvidia RTX 2080ti, 4352 CUDA cores & 11 GB of VRAM.
3. Intel core i7-3770k, 4 cores @ 3.5GHz base, overclocked to ... with 16GB of DDR3 ... MHz RAM.
4. The processor above with Nvidia GTX 660, 960 CUDA cores & 2 GB of VRAM.
5. Intel core i7-8750h, 6 cores @ 2.2GHz base and 16GB of DDR4 ... MHz RAM.
6. The processor above with Nvidia GTX 1070 (mobile), 2048 CUDA cores and 8 GB of VRAM.
7. AMD Ryzen 3700U, 4 cores @ 2.3GHz base and 6 (effectively) GB of RAM.

For GPUs to be utilized, the Parallel Toolbox in MATLAB is installed. The dataset is NIPS Conference Papers 1987-2015 from UCI Machine Learning Repository [7]. Papers are referenced by "Vision", "Neural" and "Learning" values. All clustering algorithms use squared Euclidean distance as a metric:

$$|a-b|_2^2 = \sum_i (a_i - b_i)^2 \tag{1}$$

4. CPU benchmark

The benchmark reads the dataset (transposed and optimized, available at [12]). k-Means, DBSCAN, hierarchy and nearest-neighbors clustering algorithms' implementations are run. The mean and sample standard deviation is calculated:

$$\bar{t} = \frac{1}{N}\sum_{i=1}^{N} t_i \tag{2}$$

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(t_i - \bar{t}\right)^2} \tag{3}$$

Algorithm	3900x			3700u			8750h			3770k		
	average	deviation	first run	average	deviation	first run	average	deviation	first run	average	deviation	first run
k_means	0.1485647	0.0294064	0.1733067	0.2369192	0.3332865	7.6030114	0.1607835	0.0688289	0.1874	0.197224	0.10638	0.302635
hierarchy	0.6195843	0.1145432	0.7813761	1.005445	0.351111	3.366557	0.6148698	0.0583684	0.9526	0.986554	0.069673	1.412833
neighbors	0.0240679	0.1134024	0.1685378	0.0428162	0.0239203	0.5702964	0.0268133	0.0119936	0.1322	0.031593	0.01869	0.23805
my_dbscan	0.2443005	0.0141887	0.3879294	0.1941701	0.0228209	0.5046445	0.1049712	0.0118161	0.1969	0.135808	0.015145	0.455906

Table 1. CPU-only benchmark

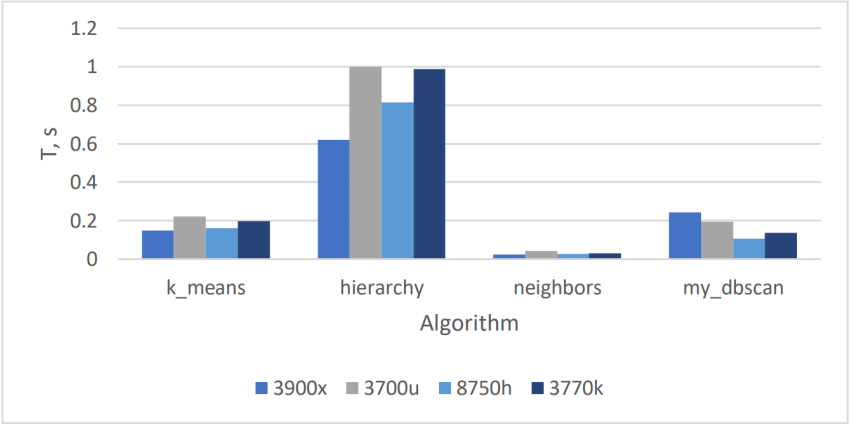


Fig. 1. CPU-only benchmark (average), lower – better

Since, probably, clustering will be performed once, first run results have been recorded as well.

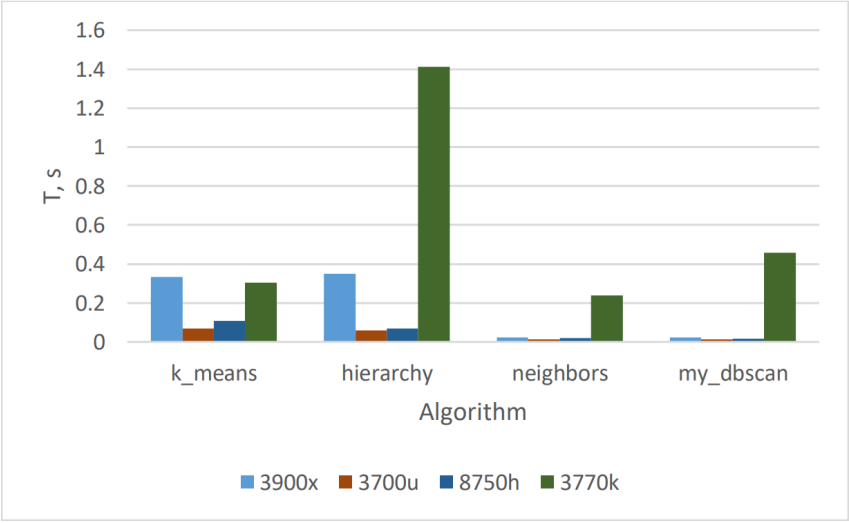


Fig. 2. CPU-only benchmark (first run), lower – better

MATLAB can plot the results of clustering, so we can view the results in a more comprehensible format.

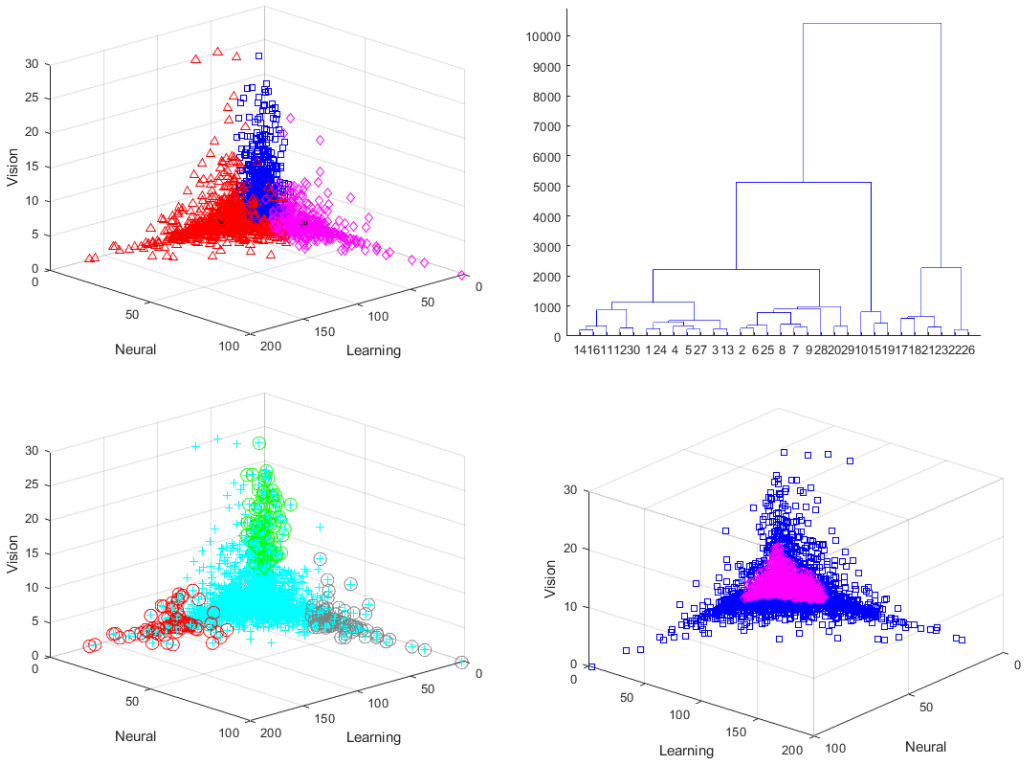


Fig. 3. k-means, hierarchy, nearest-neighbor, DBSCAN clustering

5. GPU benchmark

This benchmark uses the same dataset, as the one above. It only differs from the previous by utilizing GPUs. Parallel computing option is specified for k-means, so that the GPU is used for computing [8]. Also, 'IncludeTies', 'NSMethod', and 'SortIndices' name-value pair arguments are not used for knnsearch function [9], 'squaredeuclidean' Distance argument is supplied to pdist function for the same reason [10].

Algorithm	2080ti			1070 mobile			660			3770k		
	average	deviation	first run	average	deviation	first run	average	deviation	first run	average	deviation	first run
k_means	0.101600341	0.022818093	0.3723852	0.193905003	0.559473012	12.6737349	0.263266735	1.013714704	22.88024	0.197224	0.10638	0.302635
hierarchy	0.304797988	0.012092614	0.471679	0.539515566	0.044483149	1.4714957	0.598769451	0.089768183	2.596127	0.986554	0.069673	1.412833
neighbors	0.029876952	0.010741517	0.2659459	0.038634962	0.014030987	0.3483474	0.043220559	0.021544788	0.521507	0.031593	0.01869	0.23805
my_dbscan	0.090224443	0.008407704	0.2682983	0.137256744	0.008500473	0.3158782	0.165918617	0.012611378	0.34114	0.135808	0.015145	0.455906

Table 2. GPU benchmark

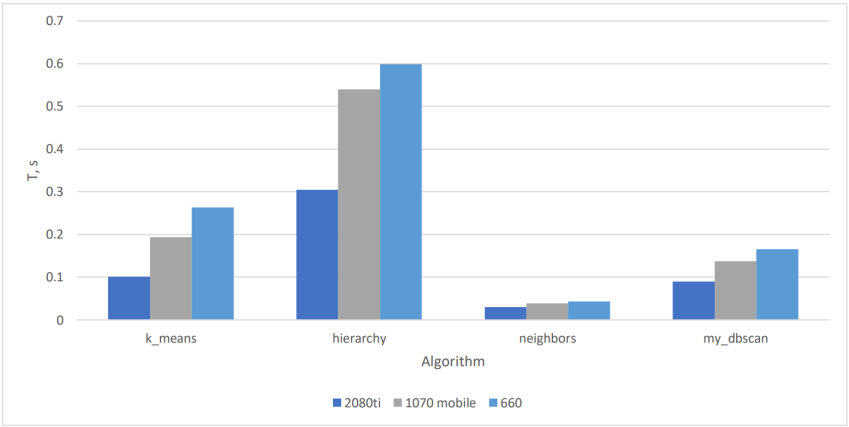


Fig. 4. GPU performance (average), lower – better

GPUs perform much better than CPUs in such tasks because of their architecture: they can perform the same step on different data in one tick (known as single instruction, multiple data, SIMD).

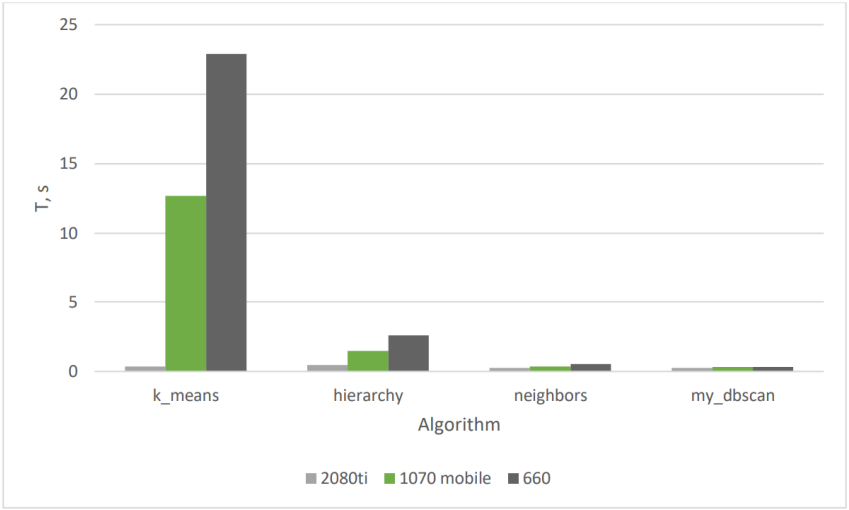


Fig. 5. GPU benchmark (first run), lower – better

Nevertheless, data that is to be analyzed, must get from RAM to VRAM and that takes some time. That's why first run results are rather poor (especially noticeable for k-means algorithm).

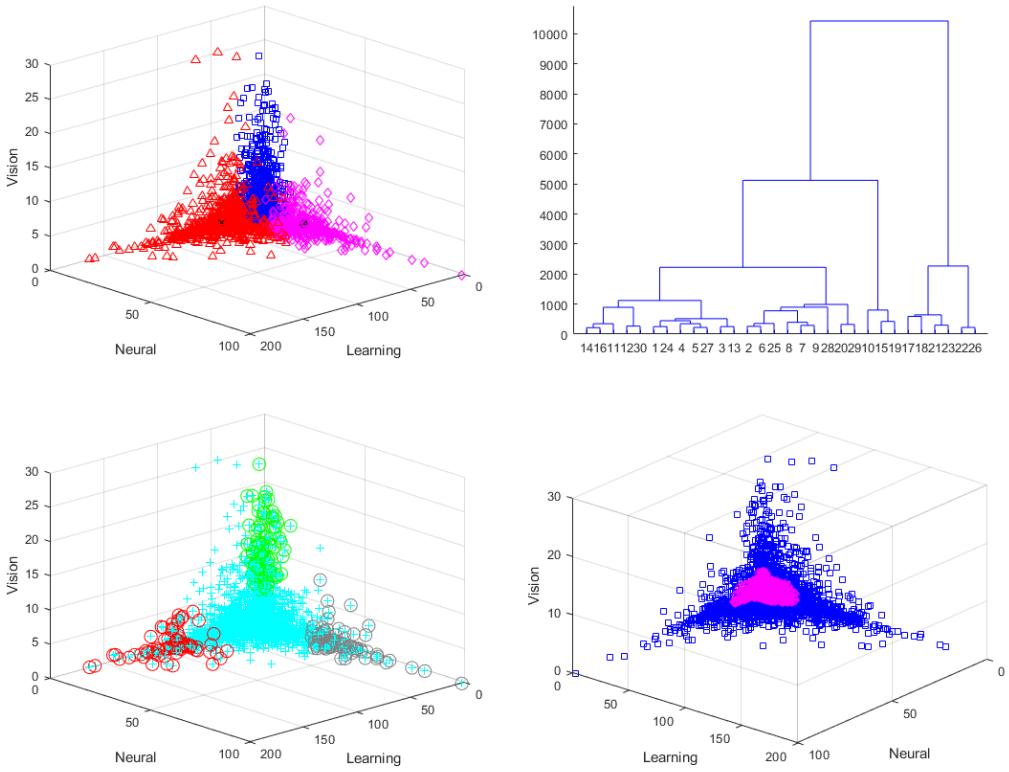


Fig. 6. k-means clustering

6. Conclusion

GPUs are usually considered much better in mathematical calculations because of their core count – they can perform way many more operations than most CPUs. Nevertheless, there is always a time lag for data synchronization – it must get from cache/RAM into VRAM and GPU's cache in order to be processed. The most powerful piece of hardware, the Nvidia RTX 2080ti, outperformed almost all other pieces of equipment, as expected, although the AMD Ryzen 9 3900X can run 7 more iterations of the nearest neighbor clustering method that the graphics card. The best performance per ruble is delivered by Intel core i7-3770k, which is still a quite capable processor, despite its age. CPU-only test has shown that the algorithms are more subject to single core performance, rather than core count. The parallel toolbox can unlock the processor's and the graphics card' potential by employing more threads on more cores.

REFERENCES

1. Babichev S., Lytvynenko V., & Taif M. A. (2016). Estimation of the inductive model of objects clustering stability based on the k-means algorithm for different levels of data noise. Radio electronics, computer science, management,(4 (39)), 54-60. doi:10.15588/1607-3274-2016-4-7
2. MacKay, David (2003). "Chapter 20. An Example Inference Task: Clustering". Information Theory, Inference and Learning Algorithms. Cambridge University Press. pp. 284–292. ISBN 978-0-521-64298-9. MR 2012999
3. Choose Cluster Analysis Method. (n.d.) Retrieved 5/13/2020 from MATLAB & Simulink: https://www.mathworks.com/help/stats/choose-cluster-analysis-method.html
4. Frank Nielsen (2016). "Chapter 8: Hierarchical Clustering". Introduction to HPC with MPI for Data Science. Springer.
5. Cover, Thomas M.; Hart, Peter E. (1967). "Nearest neighbor pattern classification" (PDF). IEEE Transactions on Information Theory. 13 (1): 21–27. CiteSeerX 10.1.1.68.2616. doi:10.1109/TIT.1967.1053964.
6. Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M. (eds.). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231. CiteSeerX 10.1.1.121.9220. ISBN 1-57735-004-9.
7. Valerio Perrone and Paul A. Jenkins and Dario Spano and Yee Whye Teh (2016). Poisson Random Fields for Dynamic Feature Models. arXiv 1611.07460
8. k-means clustering - MATLAB kmeans (n.d.) Retrieved 5/13/2020 from MathWorks Help Center: `https://www.mathworks.com/help/stats/kmeans.html` (13.05.2020)
9. Find k-nearest neighbors using input data - MATLAB knnsearch (n.d.) Retrieved 5/13/2020 from MathWorks Help Center: `https://www.mathworks.com/help/stats/knnsearch.html` (13.05.2020)
10. Pairwise distance between pairs of observations – MATLAB pdist (n.d.) Retrieved 5/13/2020 from MathWorks Help Center: `https://www.mathworks.com/help/stats/pdist.html` (13.05.2020)
11. Measure the Performance of Your Code - MATLAB & Simulink (n.d.) Retrieved 5/13/2020 from MathWorks Help Center: `https://www.mathworks.com/help/matlab/matlab_prog/measure-performance-of-your-program.html` (13.05.2020)
12. `https://github.com/berkut126/MatlabPerformance/blob/master/NIPS.csv` (13.05.2020)

UDC: 519.872

Gaussian asymptotics for a multiclass M/M/1/1 retrial queueing system

A.A.Nazarov[1], T.Phung-Duc[2], Y.E. Izmailova[1]

[1]Institute of Applied Mathematics and Computer Science, National Research Tomsk State University, 36 Lenina ave., 634050, Tomsk, Russian Federation

[2]Faculty of Engineering Information and Systems, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan

nazarov.tsu@gmail.com, tuan@sk.tsukuba.ac.jp, evgenevna.92@mail.ru

Abstract

In this paper, we consider the multiclass M/M/1/1 retrial queueing system where customers of each type have type-dependent arrival process, service time distribution and retrial rate. In this system, a blocked customer enters the orbit of its own type and retries for service after an exponentially distributed time with the parameter depending on its type. Under the condition that the retrial rates are small, i.e., long retrial time, we obtain the asymptotic joint stationary distribution of the numbers of customers in the orbits, which turns out to be Gaussian with explicit means and covariance matrix.

Keywords: retrial queueing system, a multiclass system, asymptotic analysis

1. Introduction

Retrial queues have become popular in the queueing researches due to the challengings in analysis as well as the needs of modelling retrial phenomenon in real world systems, e. g. telecommunication and service systems [1, 2, 6, 13, 15]. Analytical solutions for single-class pure Markovian models (single or multiple servers) are obtained for some special cases [2, 4, 5, 10, 11, 12].

For multiclass retrial queues, the analysis is even more difficult and analytical solution for the joint stationary distribution has not been obtained even for single server case. To the best of our knowledge, only the stability conditions [7, 9, 14] and moments of the numbers of customers in the orbits have been obtained [2, 8]. Series expansion is obtained for the joint stationary queue length of a multiclass single server model with finite orbits under light traffic, i.e., small arrival rate [3]. However, analytic result of the joint queue length distribution has not been obtained for the model with infinite orbit sizes.

The reported study was funded by RFBR according to the research project No.18-01-00277.

The difficulty is that the joint generating functions of the numbers of customers in the orbits are the solution of a system of partial differential equations. In this paper, we consider the system under an asymptotic regime of slow retrials. First, we obtain the first order asymptotic result that the scaled numbers of customers in the orbits converge to the constants having a clear physical meaning. Next, we obtain the second order asymptotic result which states that the joint distribution of the centered numbers of customers in the orbits converges to a multidimensional Gaussian distribution with explicit mean and covariance matrix.

The rest of our paper is organized as follows. Section 2 presents the model and problem formulation. Section 3 shows the detailed analysis of the first and second order asymptotics. Section 4 presents some concluding remarks.

2. Model Description and Problem Statement

We consider a multiclass single server retrial queueing system. Let N be the number classes of incoming customers. Customers of class n arrive from outside the system according to a Poisson process with a rate $\lambda_n, n = \overline{1,N}$. If an arriving customer finds the server free, the customer occupies the server and gets a service. The service times for customers of class n are assumed to be exponentially distributed with service rate $\mu_n, n = \overline{1,N}$. If the server is busy incoming customers of class n join the orbit for class n and make a delay for an exponentially distributed time with rate $\sigma_n, n = \overline{1,N}$ then repeat their request for service.

Let $i_n(t), n = \overline{1,N}$ be the random processes of the numbers of customers in the orbits. We denote in vector notation as $\mathbf{i}(t) = [i_1(t) \dots i_N(t)]$. The aim of the current research is to derive the stationary probability distribution of this vector process. Let $k(t)$ be the random process that defines the server states: 0 if the server is free, n if the server is busy serving an incoming call of type n, $n = \overline{1,N}$.

The process $\mathbf{i}(t)$ is not Markovian, therefore we consider the $(N+1)$-dimensional continuous time Markov chain $\{k(t), \mathbf{i}(t)\}$.

Denoting $P_k(\mathbf{i},t) = P\{k(t) = k, i_1(t) = i_1, \dots, i_N(t) = i_N\}, k = \overline{0,N}$ it is possible to write down the following equalities

$$P_0(\mathbf{i}, t+\Delta t) = P_0(\mathbf{i},t) \prod_{m=1}^{N} (1-\lambda_m \Delta t)(1 - i_m \sigma_m \Delta t) + \sum_{m=1}^{N} P_m(\mathbf{i},t)\mu_m \Delta t + o(\Delta t),$$

$$P_n(\mathbf{i}, t+\Delta t) = P_n(\mathbf{i},t)(1-\mu_n \Delta t) \prod_{m=1}^{N} (1-\lambda_m \Delta t) + P_0(\mathbf{i},t)\lambda_n \Delta t +$$

$$+ P_0(\mathbf{i}+\mathbf{e}_n, t)(i_n+1)\sigma_n \Delta t + \sum_{\nu=1}^{N} P_n(\mathbf{i}-\mathbf{e}_\nu, t)\lambda_\nu \Delta t + o(\Delta t), n = \overline{1,N}.$$

Here $\mathbf{e}_n$ is the vector whose n-th component is equal to unity, and the rest are zero.

We will consider the system in a steady state regime under the condition (See [9, 14]):

$$\sum_{m=1}^{N} \frac{\lambda_m}{\mu_m} < 1.$$

We denote $P_k(\mathbf{i}) = \lim_{t\to\infty} P_k(\mathbf{i}, t)$ the stationary probability distribution of the system states $\{k(t), \mathbf{i}(t)\}$.

Let us write the system of equations for the probability distribution $\{P_0(\mathbf{i}), P_1(\mathbf{i}), \ldots, P_N(\mathbf{i})\}$, $\mathbf{i} \geq 0$, using equalities the above:

$$\begin{gathered} P_0(\mathbf{i}) \sum_{m=1}^{N} (-\lambda_m - i_m\sigma_m) + \sum_{m=1}^{N} P_m(\mathbf{i})\mu_m = 0, \\ -P_n(\mathbf{i}) \left(\mu_n + \sum_{m=1}^{N} \lambda_m \right) + P_0(\mathbf{i})\lambda_n + +P_0(\mathbf{i} + \mathbf{e}_n)(i_n + 1)\sigma_n + \\ + \sum_{\nu=1}^{N} P_n(\mathbf{i} - \mathbf{e}_\nu)\lambda_\nu = 0, n = \overline{1, N}. \end{gathered} \tag{1}$$

Here it is assumed that $P_k(\mathbf{i}) = 0, k = \overline{0, N}$, if at least one component of the vector $\mathbf{i}$ is negative.

Lets introduce the multidimensional partial characteristic functions

$$H_k(\mathbf{u}) = \sum_{i_1=0}^{\infty} \ldots \sum_{i_N=0}^{\infty} P_k(i_1, \ldots, i_N) \exp\left\{ j \sum_{m=1}^{N} u_m i_m \right\} = \sum_{\mathbf{i}=0}^{\infty} e^{j\mathbf{u}^T\mathbf{i}} P_k(\mathbf{i}), k = \overline{0, N}, \tag{2}$$

where $j = \sqrt{-1}$ is an imaginary unit and $\mathbf{u}$ is vector with components $u_n, n = \overline{1, N}$.

Using functions (2) and transform (1), the following system of equations is obtained

$$\begin{gathered} -H_0(\mathbf{u}) \sum_{m=1}^{N} \lambda_m + j \sum_{m=1}^{N} \frac{\partial H_0(\mathbf{u})}{\partial u_m} \sigma_m + \sum_{m=1}^{N} H_m(\mathbf{u})\mu_m = 0, \\ -H_n(\mathbf{u}) \left(\mu_n + \sum_{m=1}^{N} \lambda_m \right) + H_0(\mathbf{u})\lambda_n - j\sigma_n e^{-ju_n} \frac{\partial H_0(\mathbf{u})}{\partial u_n} + \\ + \sum_{m=1}^{N} H_n(\mathbf{u})\lambda_m e^{ju_m} = 0, n = \overline{1, N}. \end{gathered} \tag{3}$$

3. Asymptotic analysis under the long delay condition

Denote

$$\sigma_n = \sigma\gamma_n, n = \overline{1,N}.$$

The main idea of this paper is to find the solution of system (3) by using an asymptotic analysis method under the limit condition of the long delay of customers in the orbits, i.e., when $\sigma \to 0$.

3.1. Asymptotic of the first-order. We make the following substitutions in the system (3):

$$\sigma = \epsilon, \mathbf{u} = \epsilon\mathbf{w}, H_k(\mathbf{u}) = F_k(\mathbf{w},\epsilon), k = \overline{0,N}.$$

As the result, we get the following equations

$$\begin{aligned}
&-F_0(\mathbf{w},\epsilon)\sum_{m=1}^{N}\lambda_m + j\sum_{m=1}^{N}\frac{\partial F_0(\mathbf{w},\epsilon)}{\partial w_m}\gamma_m + \sum_{m=1}^{N}F_m(\mathbf{w},\epsilon)\mu_m = 0,\\
&-F_n(\mathbf{w},\epsilon)\left(\mu_n + \sum_{m=1}^{N}\lambda_m\right) + F_0(\mathbf{w},\epsilon)\lambda_n - j\gamma_n e^{-j\epsilon w_n}\frac{\partial F_0(\mathbf{w},\epsilon)}{\partial w_n} + \\
&\qquad + \sum_{m=1}^{N}F_n(\mathbf{w},\epsilon)\lambda_m e^{j\epsilon w_m} = 0, n = \overline{1,N}.
\end{aligned} \tag{4}$$

Denoting the asymptotic solution of the system of equations (4) in the form $F_k(\mathbf{w}) = \lim\limits_{\epsilon\to 0} F_k(\mathbf{w},\epsilon), k = \overline{0,N}$, we obtain solution named as "first-order asymptotic". The following statement is true.

Theorem 1. The first-order asymptotic characteristic function of the probability distribution of the numbers of customers in the orbits has the form:

$$F_k(\mathbf{w}) = R_k \exp\left\{\sum_{m=1}^{N} j w_m x_m\right\}, k = \overline{0,N},$$

where parameter

$$R_n = \frac{\lambda_n}{\mu_n}, n = \overline{1,N}, R_0 = 1 - \sum_{m=1}^{N}\frac{\lambda_m}{\mu_m} \tag{5}$$

is the stationary probability distribution of the state of the server ($\mathbf{R} = \{R_k\}, k = \overline{0,N}$ in matrix form),

$$x_n = \frac{\lambda_n}{\gamma_n}\frac{1-R_0}{R_0}, n = \overline{1,N}. \tag{6}$$

The value x_n has the meaning of the average value of the number of customers in the orbit of type n, normalized by the value σ. The numerator $\lambda_n(1-R_0)$ represents the arrival rate to the orbit of type n while the denominator expresses the departure rate of customers from orbit n that successfully occupy the server upon arrival.

3.2. Asymptotic of the second-order. In the system (3) let us denote

$$H_k(\mathbf{u}) = H_k^{(2)}(\mathbf{u}) \exp\left\{\sum_{m=1}^{N} j\frac{u_m}{\sigma_m}\gamma_m x_m\right\}, k = \overline{0,N}. \tag{7}$$

The functions $H_k^{(2)}(\mathbf{u})$ are partial characteristic functions of the centered random processes $i_m(t) - \dfrac{x_m}{\sigma}$. Substituting

$$\sigma_n = \sigma\gamma_n, \sigma = \epsilon^2, \mathbf{u} = \epsilon\mathbf{w}, H_k^{(2)}(\mathbf{u}) = F_k^{(2)}(\mathbf{w},\epsilon), k = \overline{0,N}. \tag{8}$$

and expression (7) into the system (3) we get:

$$\begin{gathered}
-F_0^{(2)}(\mathbf{w},\epsilon)\sum_{m=1}^{N}(\lambda_m+\gamma_m x_m) + j\epsilon\sum_{m=1}^{N}\frac{\partial F_0^{(2)}(\mathbf{w},\epsilon)}{\partial w_m}\gamma_m + \sum_{m=1}^{N}F_m^{(2)}(\mathbf{w},\epsilon)\mu_m = 0,\\
-F_n^{(2)}(\mathbf{w},\epsilon)\left(\mu_n + \sum_{m=1}^{N}\lambda_m(1-e^{j\epsilon w_m})\right) + F_0^{(2)}(\mathbf{w},\epsilon)(\lambda_n+\gamma_n x_n e^{-j\epsilon w_n}) -\\
-j\epsilon\gamma_n e^{-j\epsilon w_n}\frac{\partial F_0^{(2)}(\mathbf{w},\epsilon)}{\partial w_n} = 0, n = \overline{1,N}.
\end{gathered} \tag{9}$$

Denoting the asymptotic solution of the system of equations (9) in the form $F_k^{(2)}(\mathbf{w}) = \lim\limits_{\epsilon\to 0} F_k^{(2)}(\mathbf{w},\epsilon), k = \overline{0,N}$, we obtain this solution, named as "second-order asymptotic". The following statement is true.

Theorem 2. The second-order asymptotic characteristic function of the probability distribution of the numbers of customers in the orbits has the form:

$$F_k^{(2)}(\mathbf{w}) = R_k \exp\left\{-\frac{1}{2}\sum_{\nu=1}^{N}\sum_{m=1}^{N} w_\nu K_{\nu m} w_m\right\}, k = \overline{0,N}, \tag{10}$$

where parameters $K_{\nu m}$ are the solution of the following system:

$$\gamma_m R_0 K_{mm} - \lambda_m R_0 \sum_{l=1}^{N} \frac{\gamma_l}{\mu_l} K_{lm} = \lambda_m (1-R_0)(1-R_m) + \lambda_m^2 \sum_{l=1}^{N} \frac{R_l}{\mu_l}, \nu = m,$$

$$\gamma_m R_0 K_{m\nu} + \gamma_\nu R_0 K_{\nu m} - \lambda_m R_0 \sum_{l=1}^{N} \frac{\gamma_l}{\mu_l} K_{l\nu} - \lambda_\nu R_0 \sum_{l=1}^{N} \frac{\gamma_l}{\mu_l} K_{lm} = \tag{11}$$

$$= 2\lambda_m \lambda_\nu \sum_{l=1}^{N} \frac{R_l}{\mu_l} - (R_m \lambda_\nu + R_\nu \lambda_m)(1-R_0), \nu \neq m.$$

Substituting (8) and (7) to (10), we can write approximation expressions for the partial characteristic functions at small values σ:

$$H_k(\mathbf{u}) \approx R_k \exp\left\{ j \sum_{m=1}^{N} \frac{u_m}{\sigma} x_m - \frac{1}{2} \sum_{m=1}^{N} \sum_{\nu=1}^{N} \frac{u_m}{\sqrt{\sigma}} K_{m\nu} \frac{u_\nu}{\sqrt{\sigma}} \right\}, k = \overline{0, N}.$$

Summing up all values $k = \overline{0, N}$, we obtain approximation of the characteristic function of probability distribution of number customers in the orbits

$$H(\mathbf{u}) \approx \exp\left\{ j \sum_{m=1}^{N} \frac{u_m}{\sigma} x_m - \frac{1}{2} \sum_{m=1}^{N} \sum_{\nu=1}^{N} \frac{u_m}{\sqrt{\sigma}} K_{m\nu} \frac{u_\nu}{\sqrt{\sigma}} \right\}.$$

Thus, the distribution of the numbers of customers in the orbits in the multiclass retrial queueing system is asymptotically Gaussian.

4. Example

We consider a particular case with $N = 3$.

Table 1. The parameters of the model.

The rate of arrival flow,λ_n	The service rate, μ_n	The rate of a delay in the orbit, $\sigma_n = \sigma\gamma_n$
$\lambda_1 = 0.7$	$\mu_1 = 2$	$\sigma_1 = 0.01$
$\lambda_2 = 0.6$	$\mu_2 = 3$	$\sigma_2 = 0.02$
$\lambda_3 = 0.5$	$\mu_3 = 4$	$\sigma_3 = 0.03$

Figure 1 shows the asymptotic probability distribution $P(i)$ for the total number of customers in the orbits.

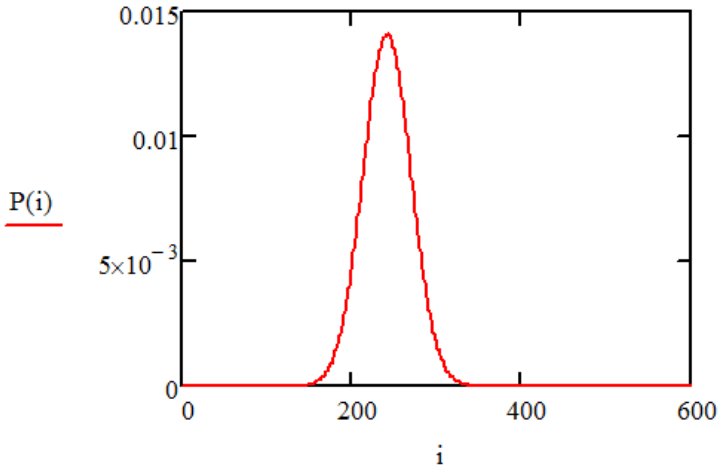


Fig. 1. Graph of distribution $P(i)$ for $\sigma = 0.01$

5. Conclusion

In this paper, we have considered a multiclass single server retrial queueing system. Equations for characteristic functions of the multi-dimensional probability distribution of the numbers of customers in the orbits are obtained. We then used the method of asymptotic analysis under condition a long delay of customers in the orbits to find the limiting probability distribution of the number of the customers in the orbits. This probability distribution turned out to be Gaussian. We as well derived the expressions for the means of the Gaussian distribution and the stationary probability distribution of the state of the server. A system of linear equations is obtained for finding the elements of the covariance matrix.

REFERENCES

1. Artalejo, J. R., and Gómez-Corral, A. (2008). Retrial Queueing Systems. Springer.
2. Falin, G., and Templeton, J. G. (1997). Retrial queues. CRC Press.
3. Fiems, D., and Phung-Duc, T. (2019). Light-traffic analysis of random access systems without collisions. Annals of Operations Research, 277(2), 311-327.
4. Hanschke, T. (1987). Explicit formulas for the characteristics of the M/M/2/2 queue with repeated attempts. Journal of Applied Probability, 24, 486–494.
5. Hanschke, T. (1999). A matrix continued fraction algorithm for the multiserver repeated order queue. Mathematical and Computer Modelling, 30, 159–170.

6. Hu, K., Allon, G., and Bassamboo, A. (2016). Understanding customers retrial in call centers: Preferences for service quality and service speed. Available at SSRN 2838998.
7. Kim, B., and Kim, J. (2020). Stability of a multi-class multi-server retrial queueing system with service times depending on classes and servers. Queueing Systems, 94(1-2), 129-146.
8. Kulkarni, V. G. (1983). On queueing systems by retrials. Journal of Applied Probability, 20(2), 380-389.
9. Morozov, E., and Phung-Duc, T. (2017). Stability analysis of a multiclass retrial system with classical retrial policy. Performance Evaluation, 112, 15-26.
10. Phung-Duc, T., Masuyama, H., Kasahara, S., and Takahashi, Y. (2009). M/M/3/3 and M/M/4/4 retrial queues. Journal of Industrial and Management Optimization, 5(3), 431.
11. Phung-Duc, T., Masuyama, H., Kasahara, S., and Takahashi, Y. (2010). State-dependent M/M/c/c+ r retrial queues with Bernoulli abandonment. Journal of Industrial and Management Optimization, 6(3), 517-540.
12. Phung-Duc, T., Masuyama, H., Kasahara, S., and Takahashi, Y. (2013). A matrix continued fraction approach to multiserver retrial queues. Annals of Operations Research, 202(1), 161-183.
13. Phung-Duc, T. (2017), Retrial Queueing Models: A Survey on Theory and Applications, to appear in Stochastic Operations Research in Business and Industry (eds. by Tadashi Dohi, Katsunori Ano and Shoji Kasahara), World Scientific Publisher.
14. Shin, Y. W. and Moon, D. H. (2014). M/M/c retrial queue with multiclass of customers. Methodology and Computing in Applied Probability, 16(4), 931-949.
15. Tran-Gia, P., and Mandjes, M. (1997). Modeling of customer retrial phenomenon in cellular mobile networks. IEEE Journal on selected areas in communications, 15(8), 1406-1414.

UDC: 004.77

How to build a hyper-local Internet

Dmitry Namiot [1] and Manfred Sneps-Sneppe [2]

[1]Lomonosov Moscow State University, GSP-1, 1-52, Leninskiye Gory, Moscow, Russia
[2]Ventspils University of Applied Sciences, Inzenieru 101a, LV-3601, Venspils, Latvia

dnamiot@gmail.com, manfreds.sneps@gmail.com

Abstract

The paper describes the model of hyper-local Internet. This refers to a set of Internet resources that are, to one degree or another,relevant (useful) for users located in a certain limited area. For example,these resources discuss the functioning of a housing complex, an educational institution, contain information about local services, etc. The paper proposes both a model for organizing the markup of such areas based on the use of wireless technologies and a scheme for describing (presenting) resources. Collections of this kind can be dynamically created and maintained by any users. The result of the work is the presentation of a working model of spatial marking of the Internet, which allows you to combine existing resources together in the spatial community of their content. As a technological basis for such services, a new model of using Wi-Fi Direct is advocated.

Keywords: network proximity; Wi-Fi Direct; physical web

1. Introduction

This work is a continuation of a series of articles on information services based on the concept of proximity [1, 2]. We are talking about services for mobile users (that is, about mobile services), when access to any information is provided depending on the proximity of the mobile device (and, accordingly, the mobile user) to a certain selected point. As such a point acts as a node in wireless networks. It can be some fixed element of the network infrastructure (for example, a Wi-Fi access point), or it can be some node that is specially created (often dynamically) just to act as a reference node to represent such services.

In other words, it is the spatial proximity. But only instead of calculating the distances and evaluating whether to consider such a distance as close (small) or not (which, of course, depends on the service), the fact of physically limited signal

propagation of wireless networks is used here Here is the distance over which such a signal extends and is considered close. This allows you to determine the proximity directly, without any work with geo-coordinates.

It is a complete rejection of the calculation of coordinates that allows us to evaluate such proximity for arbitrary devices, including those created specifically for this type of task. For example, the position of a mobile device can also be estimated by the signals of wireless networks. But in all such cases, there is some previously known (prepared) marking of the terrain with affixed nominal RSSI signal strength values [3]. And the essence of navigation is to, comparing the measured value of the signal strength with the reference values, to determine the deviation from the known coordinates of the wireless node [4]. Metrics that are used to determine deviations, methods of organizing and constructing such markups may vary, but the essence of the process remains the same - it is still working with geo-coordinates [5].

Why does the idea of not working with geo-coordinates come up? Here we can specify several reasons. For mobile services, working with coordinates is GPS systems. All the rest is just GPS refinement and adjustment. Accordingly, the refusal to work with geo-coordinates is explained precisely by the shortcomings (problems) in using GPS. This is for example:

- Indoor services
- Ability to block signal (GPS spoofing)
- Cold start
- Measurement accuracy. GPS exists in two versions - military and commercial. In public services, a commercial version is used, and its accuracy can be significantly exceeded by other means
- Moving objects (coordinates are constantly changing)

Accordingly, for modern navigation systems using information about wireless networks, two points can be noted. The need for preliminary markup excludes public (third-party) services from consideration, since for them, in most cases, markups on third-party sites will not be available. Such markups needs to be updated, which, of course, affects the economy of services. With this approach, navigation, of course, can only be tied to fixed wireless nodes with known coordinates.

If you refuse to use geo-coordinates, then arbitrary nodes of wireless networks can be used as reference nodes (their coordinates are unknown and will never be used). Instead of some computational model, proximity will be described by a set of rules, such as:

If Node1 AND Node2 are available then ...

Moreover, in the conditions can be used any measured characteristics, and not just the signal strength. The most suitable models here are fuzzy logic systems [6].

Another consequence of this approach is the ability to use advertising information for wireless nodes. From a software point of view, the visibility (accessibility) of a wireless node means the ability to receive some information that this node sends out (distributes). For Wi-Fi Direct, there is an advertisement for services where a wireless node can advertise (represent) a certain service. The service description is distributed (advertised) in this case, which is represented as an abstract set of pairs

<property name, property value >

The point is that all such "advertising" of wireless nodes can be customized. Accordingly, in this way it is possible to transmit some information of services. It turns out some useful dualism in practice - obtaining this kind of information is fixing the fact of proximity (accessibility / visibility of the wireless node) and, at the same time, obtaining some useful (as part of the service) information. This allows, in many cases, to refuse the use of server (cloud components) in services. What, for example, looks like a classic service using location information:

- Mobile device receives location information
- The received data is used as a key when accessing a cloud service that will search for data

In the case of network spatial proximity, this can be reduced simply to searching for the nearest nodes, when the necessary data will be transmitted through the advertising presentation of these nodes, simultaneously with the search.

The remainder of the article is structured as follows. In section 2, we consider hyper-local Internet. In section 3, we discuss the existing prototypes. In section 4, we discuss the technical details, and section 5 provides the conclusion.

2. On hyper-local Internet

In this section, we would like to dwell on the model of services that are considered in this article. As shown in the previous section, the network proximity model (the term spatial proximity is still used) allows you to mark (outline, limit) a certain spatial area. Mobile users (mobile applications or even mobile web applications) can determine the presence (visibility) of network nodes and, thereby, determine (fix) their affiliation at a particular point in time to a given site (spatial area). Moreover, such a fixation of belonging to the spatial domain (fixing the fact of being nearby a

wireless network node) is accompanied (may be accompanied) by the receipt of some information (data set) from this node [7].

The idea of our service is to use a similar approach in marking up Internet resources. We want to describe in similar way resources that are relevant in some local context. It is known that services using location information in most cases are used precisely for searching for local information, information that relates to a certain area near the requestor. However, there is no reliable way to describe the resources of the Internet related specifically to a certain local area. What is meant here is a description of the resources, and not the issuance of any geo-coded information upon request. For example, all sorts of wiki sites and discussion forums for residents of a community are very popular. It can also be not only traditional sites, but also specially created groups (communities) in social networks. You can also mention, for example, the increased popularity of channels in Telegram. Widely used. For example, dedicated Twitter accounts for publishing any data (including from some sensors / measuring devices). The question is how can new users of such resources find them?

Traditional models would consist of organizing some centralized catalog that would contain links with corresponding geo-coordinates. The client application would determine the coordinates of the user and refer to this directory for a list of resources. This is a completely working model (both theoretically and practically), but there is one blocking point that explains why this did not happen (why many attempts to create such directories did not work). The very decentralized nature of Internet services suggests that authors create content (services) without any communication (verification) with some "authorizing" authority. Accordingly, the creators of the service have no incentive to register their resources somewhere. A centralized collection of such information is not possible because collectors themselves cannot find out about local resources.

Based on this, our idea is that the creators (authors) of such content (local services) themselves would advertise it, and local subscribers would have the opportunity to receive such advertising. This means that we want to create a wireless network node that will "advertise" some existing Internet service (content). Such advertising (in fact - a description of the Internet service) will be available to mobile subscribers (applications on mobile devices) located near this site. Such a node can be created (opened), including directly on the phone of the author of this content (service) [7].

At the same time, we will use standard Wi-Fi Direct mechanisms for advertising services, and for the presentation (description) of services - a system with open

code Hypercat [8]. This means that there can be many programs for scanning (viewing) such advertisements. This is not only tied to our application, which is just one example. There is a complete analogy with web browsers. Our proposal defines the layout format (conceptually plays the same role as HTML). The browser implementation can be any.

We also note that obtaining resource descriptions in the proposed scheme is carried out without organizing a connection between devices - that is, in safe mode.

The term hyper-local is used in Internet services in the sense of indicating short distances [9]. The proposed scheme can be called a model of hyper-local Internet.

3. On prototypes and existing works

Firstly, as our prototypes and previous works, we can name our previous works on network proximity models. For example, when a node name modification (SSID) or customization of an advertising presentation was used to send information about a user's profile on a social network, this is also a link to a web resource. This web resource was relevant in this local context, since the user sending the link was here.

In general, we considered services based on the network proximity model as context-sensitive services. The visibility of a particular wireless node (s) is the replenishment of context information. The attributes of each such node found are also context information. Accordingly, the host name (SSID), host address, signal strength (RSSI) - all this is context information. Any service available on a particular device is also context. As types of possible actions (operations) with context, we can indicate the following:

- Entering a device into the accessibility (visibility) zone of specified devices / services or leaving this zone causes a change in status (state) in the application
- Entering a device into the accessibility (visibility) zone of specified devices / services or exiting from such a zone causes an information request (some kind of access to the data store) for subsequent processing
- Staying in the accessibility (visibility) zone of the specified devices / services causes a change in status or request for information upon the occurrence of some other conditions (for example, if the time spent is exceeded)
- Recording of events (entry / exit from the availability (visibility) zone of the specified devices / services and stay in such an area) for use in subsequent processing

Application Examples:

- Notification of the intersection (at the entrance or exit) of a certain virtual perimeter (analog of a geo-fence)
- Sending notification with a coupon / special offer in case of repeated presence in a certain area
- Turn off the call on your mobile phone when you fall into a certain area
- Notification when changing the set of received (available) codes, etc.

In this case, we are talking about a model of a geo-information system. Instead of working with geo-coordinates, a network proximity model is used. As a result, we want to get lists of Internet resources, which (lists) are tied to the area in which some mobile device is located (Fig.1). Available (visible) wireless nodes contain information

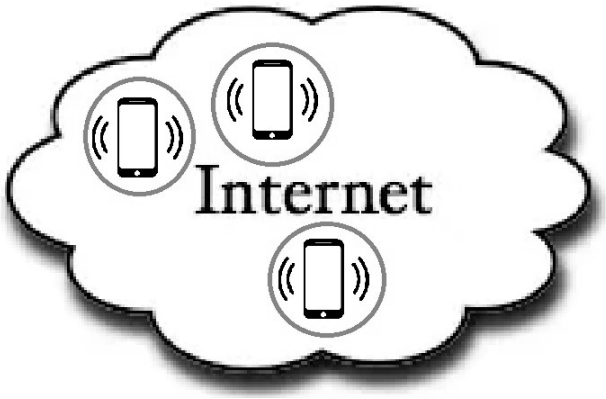


Fig. 1. On Internet proximity markup [7].

about Internet resources. And getting information about available (visible) network nodes will be equivalent to getting information about network resources described with their help.

Other models that can be mentioned in this regard are floating content [10] and partially ICN [11].

4. On Wi-Fi Direct usage

As a means of markup, we will use Wi-Fi Direct services. This is a technology that involves the direct interaction (in the sense - the connection) of Wi-Fi devices. Here, in fact, there are two technologies: Wi-Fi Direct and Wi-Fi Aware [12]. The latter is based on the Neighbor Awareness Networking Specification - the definition of services that are provided by local (nearby) Wi-Fi devices. The Wi-Fi alliance

talks about technology similarities, the difference is that Wi-Fi Direct requires some kind of coordinator to make connections, and Wi-Fi Aware creates decentralized, dynamic peer-to-peer connections. At the moment, phones with Wi-Fi Aware are not yet widespread, so all the considerations below relate specifically to Wi-Fi Direct.

Wi-Fi Direct supports the ability to define services before forming groups and connections [13]. It is this property that can be used to organize models based on network proximity. A service in such a model is simply a dataset associated with a particular device. Search (disclosure) of a service is, in fact, simply a determination of the characteristics of a wireless node [14].

<key, value>

Here is an illustrative fragment from the Android SDK manual: three keys with their values.

```
// Create a string map containing information about your service.
Map record = new HashMap ();
record.put ("listenport", String.valueOf (SERVER_PORT));
record.put ("Name", "Links");
record.put ("Description", "test service description");
record.put ("URL", "https://some-server.org/catalogue.json");
// Service information. Pass it an instance name, service type
// _protocol._transportlayer, and the map containing
// information other devices will want once they connect to this one.
WifiP2pDnsSdServiceInfo serviceInfo =
WifiP2pDnsSdServiceInfo.newInstance ("_ test", "_presence._tcp", record);
```

Accordingly, advertising a service in Wi-Fi Direct is, in fact, broadcasting a hash table over the network (in this example, record). That allows you to implement all of the above schemes for the implementation of information services without contacting the server (cloud) for processing or intermediate data storage. Confirmation of the fact that you are in the vicinity of a device will mean simultaneously receiving some information from it without establishing a connection. This form of presentation makes Wi-Fi Direct the most convenient for implementing models based on network proximity.

Our idea is to describe on a Wi-Fi Direct device a service that contains links to web resources. The presentation scheme will be as follows. On a device, each service defines three characteristics:

Name - The default value of “Links”
Description - search string
URL - Hypercat directory link

The name of the service is used for searching, the description is used for possible refinement of the search (filtering), and the web resources themselves are described as the Hypercat directory.

Hypercat is an open source project that solves the problem of finding (addressing) services in projects related to the Internet of Things. This is a fairly actively developing project. Its results form the basis of standards for the Internet of Things. The British Standardization Institute (BSI) even claims to be the first standard in the world for the Internet of Things. Obviously, of course, this is more of a marketing statement, but, nevertheless, the importance and usefulness of this product are obvious. The corresponding BSI developments were translated into Russian and distributed by the working group, which is engaged in the domestic standards of Smart City and the Internet of Things. The Hypercat specification is designed to provide IoT application clients with search and discovery (disclosure) of information about available services on the Internet.

The specification is based on the concept of a directory that describes an unsorted collection of links. So, as a result, we get a scheme where a mobile device (mobile phone) can determine a link to a collection of arbitrary Internet resources, and this collection will be available to other mobile subscribers in the vicinity of this device. And this model will work both indoors (Fig.2) and outdoors. The model will support

Fig. 2. Indoor Internet proximity markup [7].

both static determining devices and devices that are in motion. In the latter case, the scope of the resources will “follow” the determining device.

5. Conclusion

This article describes a model for using Wi-Fi Direct services to advertise Internet resources. In fact, this proposal can be described as marking up space in terms of linking Internet services. The paper proposes both a markup scheme and a method for describing services. Together, this leads to a new scheme for representing web resources (more precisely, arbitrary resources that can be represented using a URI). Such a scheme is a hyper-local Internet. It is not proposed to use any new resources or a new programming scheme. The proposed model is focused on the description (markup) of existing resources.

REFERENCES

1. Namiot, Dmitry, and Manfred Sneps-Sneppe. "On Content Models for Proximity Services." 2019 24th Conference of Open Innovations Association (FRUCT). IEEE, 2019.
2. Namiot, Dmitry, Manfred Sneps-Sneppe, and Romass Pauliks. "On Mobile Applications Based on Proximity." 2019 7th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW). IEEE, 2019.
3. Kasantikul, Kittipong, et al. "An enhanced technique for indoor navigation system based on WIFI-RSSI." 2015 Seventh International Conference on Ubiquitous and Future Networks. IEEE, 2015.
4. Varshney, Vibhu, Rajat Kant Goel, and Mohammed Abdul Qadeer. "Indoor positioning system using wi-fi & bluetooth low energy technology." 2016 Thirteenth International Conference on Wireless and Optical Communications Networks (WOCN). IEEE, 2016.
5. Zhu, Julie Yixuan, et al. "Spatio-temporal (ST) similarity model for constructing WIFI-based RSSI fingerprinting map for indoor localization." 2014 international conference on Indoor positioning and indoor navigation (IPIN). IEEE, 2014.
6. Bai, Ying, and Dali Wang. "Fundamentals of fuzzy logic control—fuzzy sets, fuzzy rules and defuzzifications." Advanced Fuzzy Logic Technologies in Industrial Applications. Springer, London, 2006. 17-36.
7. Namiot, Dmitry. "Wi-Fi Direct as a technological basis for hyper-local Internet." International Journal of Open Information Technologies 8.6 (2020)
8. Namiot, Dmitry, and Manfred Sneps-Sneppe. "On Search Services for Internet of Things." International Conference on Distributed Computer and Communication Networks. Springer, Cham, 2017.
9. Venetis, Petros, et al. "Hyper-local, directions-based ranking of places." Proceedings of the VLDB Endowment 4.5 (2011): 290-301.

10. Ott, Jörg, et al. "Floating content: Information sharing in urban areas." 2011 IEEE International Conference on Pervasive Computing and Communications (PerCom). IEEE, 2011.
11. Ahlgren, Bengt, et al. "A survey of information-centric networking." IEEE Communications Magazine 50.7 (2012): 26-36.
12. What is the relationship between Wi-Fi Aware and Wi-Fi Direct? https://www.wi-fi.org/knowledge-center/faq/what-is-the-relationship-between-wi-fi-aware-and-wi-fi-direct Retrieved: May, 2020
13. Camps-Mur, Daniel, Andres Garcia-Saavedra, and Pablo Serrano. "Device-to-device communications with Wi-Fi Direct: overview and experimentation." IEEE wireless communications 20.3 (2013): 96-104.
14. Use Wi-Fi Direct (P2P) for service discovery https://developer.android.com/training/connect-devices-wirelessly/nsd-wifi-direct Retrieved: May, 2020

UDC: 654.1

Forecasting the incoming load of a contact center using chaos theory methods

Boris S. Goldstein[1] and Sergey V. Kislyakov[2]

[1]The Bonch-Bruevich Saint-Petersburg State University of Telecommunications (SPbSUT) and RTC ARGUS, Saint-Petersburg, Russia

[2]The Bonch-Bruevich Saint-Petersburg State University of Telecommunications (SPbSUT), Saint-Petersburg, Russia

bgold@niits.ru, s.v.kislyakov@gmail.com

Abstract

The work is devoted to the search for optimal methods of contact center management, in particular, methods for forecasting the load for further calculation of the required number of employees. If the number of staff will always be more than is required, then the owners of the contact center will incur financial losses on salaries. If there are too few employees, the quality of service will drop. To forecast the load, forecasting models based on simple nonlinear forecasting, local linear forecasting and global polynomial approximation are used. The article presents the results of the application of chaos theory methods to predict the incoming contact center load.

Keywords: Contact center, workforce management, chaos theory,load forecasting, OSS/BSS

1. Introduction

Contact centers are a powerful interaction tool with a large number of customers. Contact centers provide customers with information services through voice channels, as well as using chats, email, IP-telephony. To ensure the required quality of customer service with a minimum number of operators, it is necessary to know (or predict with the greatest possible accuracy) the amount of incoming load.

2. Service Quality Metrics

One of the main parameters characterizing the processed incoming calls is the waiting time for the client in the queue. This parameter greatly affects the overall impression of using the services of the contact center. It is believed that the optimal value will be the 80/20 formula, that is, 80 percents of calls expect processing less

than 20s.Another metric is the average call processing time by the agent. Calls that are too long may indicate unprofessional work for operators, and too short can indicate that they do not really provide services to consumers. If the call cannot be processed by the operator, then the service will not be provided to the client. The optimal value of this indicator is considered 4 - 8per cents. Customer satisfaction assessment is the most important metric and is usually determined during post-call surveys, although other indicators, such as the Net Promoter Score [1], can be included in the assessment. All of these indicators are affected in one way or another by the organization of work of the operators themselves - contact center employees - their schedule and their number per shift. If there is less than optimal number of operators, the queue will increase and the level of service will drop. If there are too many of them, then downtime will increase and losses on wages will increase. Therefore, it is extremely important to optimize the schedule of operators, which directly depends on the incoming load. The load is a variable value and depends on a number of factors.

3. Work Force Management for contact center

In the documents of the TeleManagement Forum organization, a set of such tasks is defined as Workforce Management (WFM) - the common name for a set of planning processes, the result of which is a schedule for employees for some future period. The analysis is based on data on incoming traffic for previous periods and operator productivity. The result of the process is a schedule for each contact center operator. WFM tasks for contact centers are:

- Prediction of the load at certain time intervals (usually 30 min);
- Determining the number of operators, and, if necessary, operators with certain skills, who must be at a certain time interval at his workplaces;
- Building a work schedule for each contact center employee; It is quite difficult to achieve a minimum error for forecasting, because of many factors that can affect the flow of incoming calls must be considered. For different business profiles, the factors are:
- Leaps of the number of calls as a result of marketing promotion;
- Changes in demand - for example, the acquisition of a new company or the appearance of new products;
- Weather factors - snow, floods and very hot weather can have a big impact on the number of calls received at the contact center;
- Special events - events such as the World Cup can cause a large drop in calls, but do not occur every year.
- Equipment failures - power failure, broken telephone lines, etc.

The influence of these factors must be minimized to obtain the most accurate forecasting results.

4. Input data

We use data on the calls received for 1999 of an anonymous Israeli bank.The data are freely available for various studies. The contact center provides the following services:

- Information on transactions and banking operations for customers;
- Interactive voice response services (using VRU modules - Voice Response Unit);
- Providing information to prospective customers;
- Support for users of the bank website.

The data is organized into 12 text files, each of which contains data on calls per month (from 20,000 to 30,000 calls per month). Each call is characterized by 17 parameters.

5. Random process of the contact center load

To continue the work, it is necessary to verify the randomness of the process under study. To do this, we use the following criteria of chaos [2]:

- Non-negative Lyapunov Index, which indicates chaotic dynamics;
- Fractal structure of the trajectory in the phase space (state space), indicating the presence of a strange attractor.

We use the first criterion. Using the Lyapunov Index, the sensitivity of the system to variations in the initial conditions is checked. The calculations were carried out using the TISEAN 3.0.1 and MATLAB software. Maximum of the Lyapunov Index was calculated using the $lyap_k$ utility of the TISEAN software.

We have determined the optimal step of the time series Δt . As the initial data, we have selected incoming calls for January. We formed the time series and determine the Lyapunov Index for different Δt . The obtained values are presented in Table 1.
Table 1.

Δt	0,1	0,5	1	2
$\lambda 1$	0,022935	0,051363	0,010193	0,010186

Table 1. Time steps and Lyapunov Index

For small Δt , the duration of the conversation theoretically can exceed the step of the time series, leading to big leaps of the input load [3]. This is due to the fact that in the formation of the time series we attribute this or that conversation to the

corresponding time interval, based on its beginning. That is, the larger the step of the time series, the less the effect. From table 1 it is obvious that the condition for the presence of chaos is satisfied, because the obtained Lyapunov Index is greater than zero [4]. Further measurements were performed at $\Delta t = 0.5h$, with a maximum value of the Lyapunov Index.

6. Forecasting Methods

We have chosen one week For the forecast interval in the experiment. First, we will form a time series for forecasting according to the algorithm that has already been discussed above. The first part of the series will be input data for the forecasting unit, the second part will be used to verify the calculated forecast. Thus, comparing the predicted and real values, we calculate the forecasting error, which we will use later as a criterion for comparing the methods.

7. Simple nonlinear prediction

The first method we used here is Simple Nonlinear Forecasting method [5] . Its essence lies in the local approximation of the nearest neighbors in the phase space. This algorithm is implemented in the lzo-run utility of the TISEAN 3.0 software. The first step is to determine the minimum fractal dimension of the phase space m. A method for determining the minimum sufficient dimension of an embedding m was proposed by Kernell [5,6]. For each point x_i in the m-dimensional phase space, the nearest neighbor x_j is sought. If R_i exceeds some predetermined threshold, then this point is marked as having a false neighbor. Such an algorithm is implemented in the False-nearest utility of the TISEAN software, the result of which will be the dependence of the proportion of false neighbors depending on the dimension of the attachment for this system. The criterion that the fractal dimension is chosen large enough is the zero or small fraction of false neighbors. In our case, we choose m = 23, going from the 10% threshold. To calculate the delay d, the lzo-test utility was used. The utility makes a zero-order ansatz and estimates errors of one-step forecasting of the model on a multidimensional time series. The forecast errors presented are normalized to the standard deviations of each component. Using this utility, a pair m, d was selected at which the forecast error turned out to be the smallest. The fractal dimension m = 23 was determined at the previous step of the study. Based on the calculations, the delay is d = 2, because it gives the best result in the initial steps of forecasting, while the error is still acceptable. Thus, the selected pair of values m = 23, d = 2 . The last value to be determined is radius of the neighborhood in which the search for the nearest neighbors will occur in the forecasting process. As practice shows, it does not greatly affect the quality of the forecast; nevertheless, in this paper we will not neglect it and obtain a more accurate value. It is necessary to

obtain the normalized forecast error as a function of the radius of the neighborhood. Such an algorithm is implemented by the lzo-gm utility. For calculations, we will use a pair of values that were determined at the previous step, m = 23, d = 2 . The calculated size of the neighborhood is r = 3.438 for m = 23, d = 2. We proceed directly to the forecasting process. Recall that the values of fractal dimension, delay, and neighborhood size were calculated m = 23, d = 2, r = 3.438. As an elementary step in the time series, $\Delta t = 0.5h$ was chosen. Prediction interval is 1 week.

As the initial data, we used the time series, formed earlier on the basis of data on the incoming calls of the contact center for May. The forecast will be based on the first 3 weeks, the last week will be compared with the forecast obtained and used as a reference for calculating the forecast error. Prediction is performed using the simple nonlinear forecasting method implemented by the lzo-run utility.

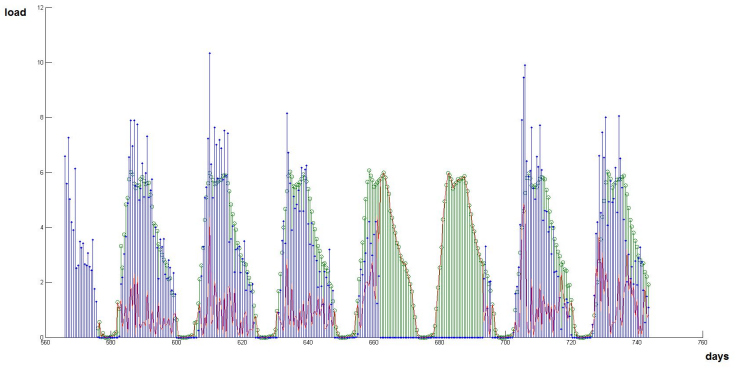


Fig. 1. Forecast using the method of simple nonlinear forecasting

The graph with a dot marker shows a portion of the original time series - the real values of the last week of May. A graph with a circle marker shows the forecast obtained. The last curve is the absolute forecast error, which is the absolute value of the difference between the predicted values and the real values throughout the entire weekly forecasting interval.

8. Local linear prediction

The difference between the Local Linear Forecasting method from the previous one is the use of local linear approximation to obtain a forecast. Using the same algorithm, we choose the values of dimension, delay m = 30, d = 2. We have chosen the radius of the neighborhood r = 6, based on the calculations. For forecasting, we use the following values m = 30, d = 2, r = 6 . As in the previous experiment, the

number of forecasting steps is L = 336 (with a time series step of t = 0.5h and a forecasting interval of 1 week). The forecast is based on the data of the first three weeks of May (as in the previous experiment). The last week we use as real values to calculate the error.

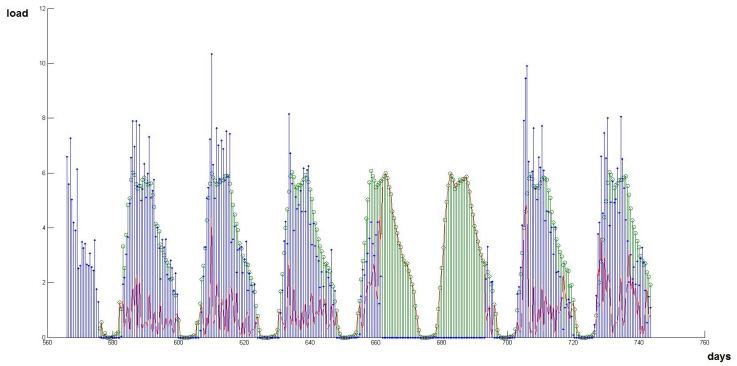


Fig. 2. Forecast using the local linear forecasting method

The graph shows a portion of the original time series (the graph with a dot marker) - the real values of the last week of May. A graph with a circle marker - forecast obtained. The last curve shows the absolute forecast error.

9. Global polynomial approximation method

We have chosen the values m = 16, d = 3, selecting them in order to obtain the smallest values of the forecast error. It should be noted that despite the fact that the parameters for the forecasting methods were selected based on the same time series, they vary depending on the methods. Changes in this parameter are caused by obtaining the smallest forecast error. Now we get the forecast based on the data m = 16, d = 3, L = 336. We have chosen the order of the polynomial p = 2 (we selected the value based on experimental data).

Similarly to previous experiments, a blue graph with a dot marker - real values, a green graph with a circle marker - forecast and a red graph - absolute forecast error, calculated as the modulus of the difference between real values and forecast.

10. Comparison of the Results

In order to compare the considered forecasting methods, we calculate the normalized forecasting error for each method [8]. An analysis of the results showed that the best method for simple nonlinear forecasting did the trick, showing the best results

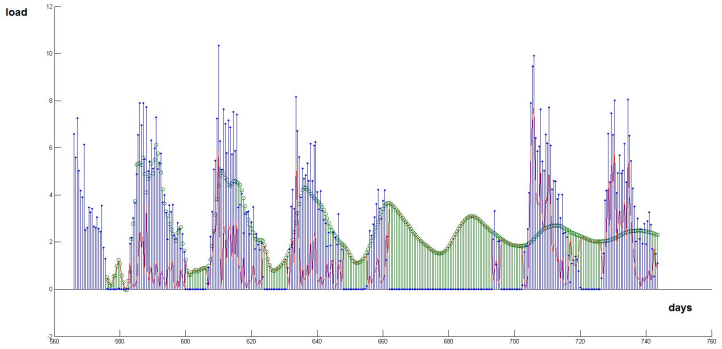


Fig. 3. Forecast using the global polynomial approximation method

both for the short-term forecast (about a day) and the medium-term (week). Other methods gave an acceptable result at approximately the daily interval. Moreover, the method of local linear forecasting was expected to be more accurate than polynomial approximation.

11. Conclusion

Using one of the criteria for randomness (non-negative Lyapunov Index), using the example of real data on calls, the chaotic nature of the process of the load on the contact center was proved. Thus, with confidence we can talk about the possibility of using chaos theory to predict this process. According to the results of the experiment, the method of simple nonlinear forecasting is best suited for this purpose. The method of local linear forecasting gave good results only for short-term prediction (about a day). The global approximation method allows you to get a satisfactory result only with a short-term forecast. Thus, to predict the load of contact centers, it is advisable to choose a method of simple nonlinear forecasting.

REFERENCES

1. Bobab C., Gatej C., Ciobanu O. Developing a Scale to Measure Customer Loyalty. // Procedia Economics and Finance. – December 2012. – 3. – pp. 623–628.
2. George Contopoulos Highlights of chaos research// the Nonlinear Sciences archive, 2018. https://arxiv.org/pdf/1807.09492.pdf
3. Christian Oestreicher, PhD, A history of chaos theory // Dialogues Clin Neurosci. 2007 9, 3 . 279–289

4. Moon F.C. Chaos and fractal dynamics: An Introduction for Applied Scientists and Engineers. New York: Wiley, 1992
5. J. Doyne Farmer John J. Sidorowich Predicting chaotic time series //Physical review letters, 1987, 8. c.845-848
6. Moon, F.C. Chaotic and Fractal Dynamics. New York: Wiley, 1992.
7. Anderson C. Business Intelligence. // Data Science in Practice. – September 2019.
8. Bracht O. Five ways to handle Big Data in R. – November 2013.

UDC: 519.87

On ergodicity of some stochastic networks and its applications

Elmira Yu. Kalimulina[1]

[1]Institute of Control Sciences of Russian Academy of Sciences, Russia, Moscow

Abstract

This paper is a continuation of previous research in ergodicity of some models for unreliable networks. The set of random graphs and the sequence of matrixes describing the failure and recovery process has been used instead of the fixed graph for network structure. The main results about an ergodicity and bounds for rate of convergence to stationary distribution are formulated under more general assumptions on intensity rates.

Keywords: ergodicity, stochastic networks, convergence rate, spectral gap

1. Model description

The standard queueing network with following parameters is considered (see Fig.1) [2, 9]:

- the network consists of m nodes, $M = \{1, 2, \ldots, m\}$;
- each node is a multi-server system with an infinite waiting room;
- the algorithm of service is FCFS (First Come First Served);
- all customers are supposed to be indistinguishable;
- there is an external Poisson arrival flow with intensity Λ (so the open queueing network is considered);
- denote the routing matrix as $R = (r_{ij}), i, j = 0, 1, \ldots, m$; without loss of generality R is supposed to be regular;
- denote the traffic vector as $\lambda = (\lambda_0, \lambda_1, \ldots, \lambda_m)$;
- denote service rates as $\mu = (\mu_1(n_1), \ldots, \mu_m(n_m))$;
- the number of customers in the system is denoted as $\mathbf{n} = (n_1, \ldots, n_m)$

The following modification of the standard model (unreliable network) is considered [3, 4]:

- each node may break down and repair with intensities α_i, β_i, $i = 1, ..., m$;

The publication has been prepared with the support of the Russian Foundation for Basic Research according to the research project No.20-01-00575 A.

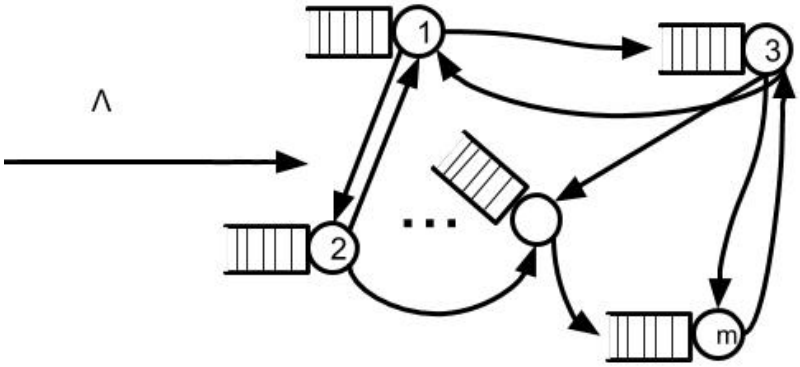


Fig. 1. Standard queueing network.

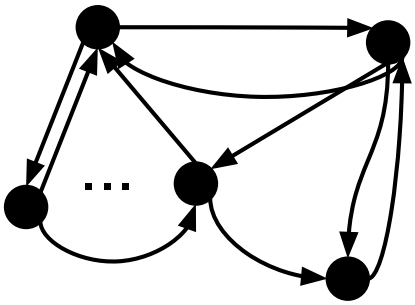


Fig. 2. Initial structure of the network.

- A dynamic routing has been applied as a failure management mechanisms. The principle of dynamic routing is in selecting the alternative node if the target node is under failure. The alternative node is selected from the nearest to the failed one. This modification make this model different from another similar ones [7, 8].

This failure management mechanisms results adding some component to standard state space of the process. The Fig.2 shows the initial structure of the network. The standard approach implies this graph to be fixed. Recoveries and failures form a way to transform this graph to another in the suggested model.

The number of nodes is fixed, but nodes can be blocked (by deleting/adding edges to it). This way of transformation is shown on Fig.3. We denote the state space of the graph transformation process as the set G. It is formed in the following way: the node i is "removed" with some intensity α_i (failure rate for this node) or it can be restored with some intensity β_i.

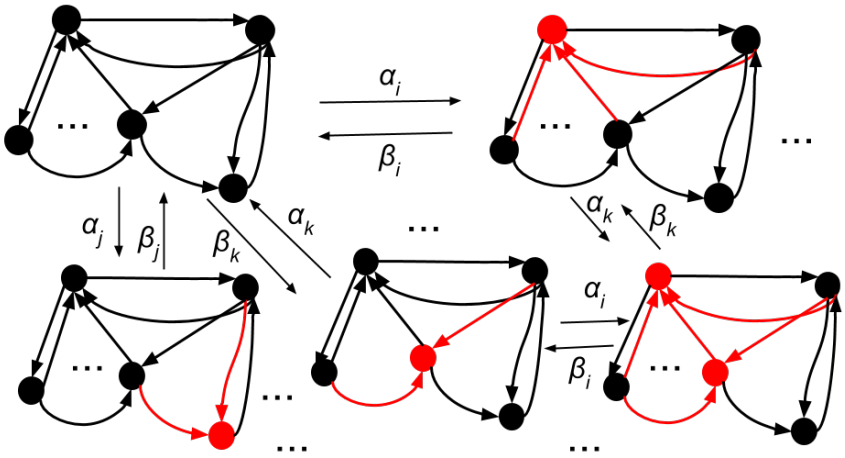


Fig. 3. The network graph evolution process.

So, the state space for our network process is the following, it is extended by adding the component G:

$$\tilde{\mathbf{n}} = (G, n_1, n_2, ..., n_m) \in |G| \times \mathbb{Z}_+^m =: \mathbb{E},$$

where G is a component describing the graph (or transition matrix) transformation. We can find a degree distribution for the process from state space G. The average number of vertices of degree k at time t: $\{M(k,t)\} = M\ P(k,t)$ can be described by the equation:

$$\begin{aligned}\{M(k,t+1)\} = \{M(k,t)\} - \frac{\alpha_k}{M\sum_k P(k)\alpha_k}\{M(k,t)\} + \\ + \frac{\alpha_{k-1}}{M\sum_{k-1} P(k-1)\alpha_{k-1}}\{M(k-1,t)\} + \\ + \frac{\alpha_{k+1}}{M\sum_{k+1} P(k+1)\alpha_{k+1}}\{M(k+1,t)\}.\end{aligned}$$

It describes the evolution of graph of our network structure in time and for the continuous time takes the form:

$$\begin{aligned}M\frac{\partial P(k,t)}{\partial t} = -\alpha_k P(k,t) + \alpha_{k-1}P(k-1,t) + \\ P(k+1,t) + \alpha_{k+1}P(k+1,t).\end{aligned} \tag{1}$$

Is easy to see for this equation that

Lemma 1. (1) is linear homogeneous equation (under assumption of constant failure and recovery rates) and has a stationary solution: $P(k) = \lim_{t\to\infty} P(k,t)$.

The network process state is described by the following vector

$$\vec{n} = ((n_1, s_1), (n_2, s_2), ..., (n_m, s_m)),$$

where n_i – the number of customers at the i-th node and

$$s_i = \begin{cases} 0, & \text{if the } i\text{-th node works,} \\ 1, & \text{otherwise.} \end{cases}$$

2. Main results

The behaviour of $\vec{n}$ is a Markov chain in continuous time. It includes an embedded homogeneous Markov chain with positive probabilities for transitions:

$$\begin{aligned} s_i &\longrightarrow (1 - s_i), \\ n_i &\longrightarrow (n_i \pm 1). \end{aligned} \tag{2}$$

Exponential convergence of reliability process $\vec{S} = (s_1, \ldots, s_m)$ converges to stationary distribution with exponential rate.

Let's consider the reliability process $X_S(t)$ of our model separately.

$$\{X_{S_i}(t+1) = X_{S_i}(t)\} = \frac{\sum\limits_{j=1}^{m} \gamma_j - \gamma_i}{\sum\limits_{j=1}^{m} \gamma_j},$$

$$\{X_{S_i}(t+1) = 1 - X_{S_i}(t)\} = \frac{\gamma_i}{\sum\limits_{j=1}^{m} \gamma_j},$$

where

$$\gamma_i = \alpha_i \mathbf{1}\{s_i = 0\} + \beta_i \mathbf{1}\{s_i = 1\}.$$

Convergence of process $X_R(t)$. The behaviour of the process $X_R(t)$ is defined by the process $X_S(t)$ with the same transition probabilities. It takes values from the finite set ($R = \|r_{ij}(t)\|$), so $X_R(t)$ has the stationary distribution and converges to

it exponentially. The sequence of $R = \|r_{ij}(t)\|$ has a limit $\tilde{R} = \|\tilde{r}_{ij}\|$, where $\tilde{r}_{ij}$ are dependent random variables.

Processes $X_S(t)$ and $X_R(t)$ describe only reliaibility of our network. At this moment we still haven't took into consideration the service process and input flow, that are our main interest of studying.

But they are ergodic and don't depend on the input flow and service process (in further we will apply these facts).

Now we are ready to define the network process.
Process definition. A new state space

$$\tilde{\mathbf{n}} = (G, n_1, n_2, \dots, n_m) \in G \times \mathbb{Z}_+^m =: \mathbb{E}$$

The following transitions in a network are possible:

$$\begin{aligned}
T_{ij}\tilde{\mathbf{n}} &:= (G, n_1, \dots, n_i - 1, \dots, n_j + 1 \dots, n_m),\\
T_{0j}\tilde{\mathbf{n}} &:= (G, n_1, \dots, n_j + 1, \dots, n_m),\\
T_{i0}\tilde{\mathbf{n}} &:= (G, n_1, \dots, n_i - 1, \dots, n_m),\\
T_f\tilde{\mathbf{n}} &:= (G^+, n_1, \dots, n_m),\\
T_r\tilde{\mathbf{n}} &:= (G^-, n_1, \dots, n_m).
\end{aligned}$$

Model description. The process $\mathbf{X} = (X(t), t \geq 0)$ defined by

$$\begin{aligned}
\mathbf{Q}f(\mathbf{n}) = {} & \sum_{i=1}^{m}\sum_{j=1}^{m}(f(T_{0j}\mathbf{n}) - f(\mathbf{n}))\lambda_i r_{ij} + \sum_{i=1}^{m}\sum_{j=1}^{m}(f(T_{ij}\mathbf{n}) - f(\mathbf{n}))\mu_i(n_i)r_{ij} + \\
& \sum_{k\in G^+}(f(T_k\mathbf{n}) - f(\mathbf{n}))\alpha_k + \sum_{k\in G\setminus G^+}(f(T_k\mathbf{n}) - f(\mathbf{n}))\beta_k + \\
& \sum_{i=1}^{m}(f(T_{i0}\mathbf{n}) - f(\mathbf{n}))\mu_i(n_i)r_{i0},
\end{aligned} \tag{3}$$

where $\lambda = (\lambda_0, \lambda_1, \dots, \lambda_m)$ satisfies the balance equations.

Assumptions for (3)

1) $\inf\limits_{\mathbf{n},i} \sum\limits_{i=1}^{m} \dfrac{\alpha_i\mu_i(\mathbf{n})}{\alpha_i + \beta_i} > \Lambda$;

2) $\tilde{R} = \|\tilde{r}_{ij}\|$ is irreducible, so the expectation of steps visited by one customer within the network is finite;

The second condition may be checked for $R(t)$ under large t. The convergence rate of $R(t)$ may be estimated from the Markov-Doeblin condition (see, e.g. Doeblin, 1938 [12]).

The second condition guarantees the existence on non-zero values for the traffic vector $\lambda = (\lambda_0, \lambda_1, \ldots, \lambda_m)$. It leads every customer to leave the system with non-zero probability. So the number of nodes each customer visited within the network is less than some geometrically distributed random variable and has a finite expectation.

Notations for network process. $\mathbf{X} = (X_t, t \geq 0)$ – Markov process;
$Q = [q(\mathbf{e}, \mathbf{e}')]_{e,e' \in \mathbb{E}}$ – transition intensities;
π – stationary distribution;
infinitesimal generator:

$$\mathbf{Q}f(\mathbf{e}) = \sum_{\mathbf{e}' \in \mathbb{E}} (f(\mathbf{e}') - f(\mathbf{e}))q(\mathbf{e}, \mathbf{e}');$$

scalar product on $L_2(\mathbb{E}, \pi)$: $\langle f, g \rangle_{pi} = \sum\limits_{\mathbf{e} \in \mathbb{E}} f(\mathbf{e})g(\mathbf{e})\pi(\mathbf{e})$.
Spectral gap for $\mathbf{X}$ [1, 6]:

$$Gap(\mathbf{Q}) = \inf\{-\langle f, \mathbf{Q}f \rangle_\pi : \|f\|_2 = 1, \langle f, \mathbf{1} \rangle_\pi = 0\}$$

Theorem 1. *If* $\mathbf{X}$ *- the process, with* $\mathbf{Q}$ *- infinitesimal generator (suppose bounded), minimal service intensity* $\mu > 0$*, and assumptions satisfy the conditions (1-2), then*
$Gap(\mathbf{Q}) > 0$ *iff for each* $i = 1, \ldots, m$*, the birth and death process with* λ_i *and* $\mu_i(n_i)$ *have* $Gap_i(\mathbf{Q}_i) > 0$*.*

Theorem 2. *If* $\mathbf{X}$ *- the process with* $\mathbf{Q}$ *- infinitesimal generator (suppose bounded), minimal service intensity* $\mu > 0$*,* $X(t)$ *satisfies the condition (1–2),*

then $Gap(\mathbf{Q}) > 0$ *iff for each* $i = 1, \ldots, m$*, distribution* $\pi = (\pi_i), i \geq 0$ *is strongly light-tailed, i.e.* $\inf\limits_k \dfrac{\pi_i(k)}{\sum\limits_{j>k} \pi_i(j)} > 0$.

Theorem 3. *Let* $\mathbf{X}$ *be unreliable queueing network with generator* $\mathbf{Q}$ *, given above, and the corresponding transition semigroup* P_t*. Suppose that* G *satisfies the condition (1).*
If π_i *is strongly light-tailed, for each* $i = 1, \cdots, m$*, then equivalently*

- *for all* $f \in L_2(\mathbb{E}, \pi)$

$$\|P_t f - \pi(f)\|_2 \leq e^{-Gap(\mathbf{Q})t} \|f - \pi(f)\|_2, t > 0,$$

- *for each* $\mathbf{e} \in \mathbb{E}$ *there exists* $C(\mathbf{e}) > 0$ *such that*

$$\|\delta_{\mathbf{e}} - \pi(f)\|_{TV} \leq C(\mathbf{e})^{-Gap(\mathbf{Q})t}, t > 0.$$

The proofs of these results are based on the standard techniques developed by T.Ligget and extended for queueing systems by other researchers [6, 10, 11].

REFERENCES

1. Liggett, T. H. Interacting Particle Systems, 1999.
2. E. van Doorn. Many papers on the rate of convergence of birth-death processes.
3. Elmira Yu Kalimulina *Analysis of Unreliable Open Queueing Network with Dynamic Routing* DCCN, Springer, 2015.
4. Elmira Yu Kalimulina *Rate of convergence to stationary distribution for unreliable Jackson-type queueing network with dynamic routing.* DCCN, Springer, 2016.
5. Chen, M.-F. *Eigenvalues, Inequalities, and Ergodic Theory.* Springer, 2005.
6. Liggett, T. H. Exponential l_2 convergence of attractive reversible nearest particle systems *Ann. Prob*, 1989(17), 403–432.
7. Sauer, C. and Daduna, H. Availability formulae and performance measures for separable degradable networks (2003). Economic Quality Control 18(2), 165-–194. *Economic Quality Control* , 18(2), 165—194.
8. Lawler, G. F. and Sokal, A. D. Bounds on the L_2 Spectrum for Markov Chains and Markov Processes: A Generalization of Cheeger's Inequality. Trans. Amer. Math. Soc., 1988, 309(2), 557-–580.
9. E. van Doorn. Representations for the rate of convergence of birth-death processes. Theory Probab. Math. Statist, 2002, 65, 37—43.
10. Lorek P., Szekli R. Computable bounds on the spectral gap for unreliable Jackson networks. Adv. in Appl. Probab, Volume 47, Number 2 (2015), 402-424.
11. Lorek P. The exact asymptotic for the stationary distribution of some unreliable systems. arXiv:1102.4707 [math.PR].
12. Doeblin W. Mathématique de l'Union Interbalkanique, 1938

UDC: 004.7

Quantifying the round-trip delay in Cloud-RAN

E.S. Sopin[1,2], A.V. Darmolad[1], D.N. Bixalina[1]

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

[2]Institute of Informatics Problems, FRC CSC RAS, 44-2 Vavilov Str., Moscow 119333, Russia

{sopin-es, 1032162870, 1032163087}@rudn.ru

Abstract

Cloud-based Radio Access Network (C-RAN) is a centralized cloud computing architecture for radio access networks (RANs) that provides large-scale deployment, joint support for radio technologies, and real-time virtualization capabilities. By moving signal processing functions to a data center, C-RAN significantly reduces power consumption and deployment cost. The architecture of the cloud radio access network consists of three main components: a pool of base-band units (BBU pool), remote radio heads (RRHs), and a transport network. In C-RAN, base stations are replaced by remote radio heads: data blocks are digitized, transmitted through the fiber-optical infrastructure, and remotely processed in BBU pool. One of the main issues is to control the round-trip delay between the remote radio heads and the BBU pool. In the paper, we describe a C-RAN in terms of queuing network and accurately evaluate all delay components. Besides, we analyze the required computational resources of the BBU pool required to satisfy the strict round-trip delay budget in C-RAN.

Keywords: Cloud-RAN, queuing network, round-trip delay

1. Introduction

Cloud-based Radio Access Network (C-RAN) is a new architectural concept for mobile communication networks designed to support high data rates with lower costs and are expected to provide low latency, high flexibility, and low power consumption to meet 5G requirements. In the traditional RAN architecture, the baseband processing and radio functions are located inside the base station (BS), while in the C-RAN, the functionality of the base station is separated from the cellular node and distributed between the Remote Radio Head (RRH) and the BaseBand Unit pool (BBU pool),

The publication has been prepared with the support of RFBR according to the research projects No.18-07-00576, 20-07-01052.

which located far from each other [1]. Processing functions in the main frequency band are virtualized and moved to the BBU pool in the central cloud. RRH is located in the BS and contains low power antennas and performs all the radio frequency functions necessary to emit a signal in a cell. They perform amplification, analog-to-digital conversion of radio signals, and send the digitized radio signals to the central BBU pool, where the received signals are processed, and cloud resources are dynamically allocated on demand [2]. Flexible distribution of computing resources across all RRHs and central processing of radio signals in the BBU pool improves the statistical multiplexing coefficient and simplifies the maintenance of cellular networks. In addition, the C-RAN architecture allows deploying a large number of RRH's at low interference using coordinated multi-point(CoMP) techniques such as coordinated transmission and reception, which reduces the number of base stations required on cellular nodes, resulting in reduced operational and capital expenditures.

However, the described architecture may suffer from increased delays due to signal processing in the BBU pool. In this paper, we carefully indicate all delay components of the round trip delay, formalize the process in terms of queuing theory and provide formulas for the mean response time and amount of computational resources required to satisfy the delay budget.

2. System model

The components of the round-trip delay are shown in the figure 1.

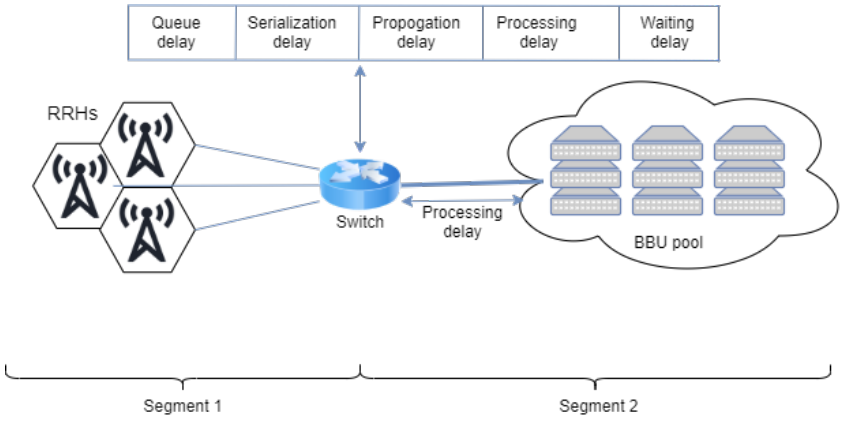


Fig. 1. Round-trip delay components in a C-RAN

User Equipment (UE), sends its signal to the RRH, which is grouped into data blocks, and then transmitted for further processing to the BBU pool. Data blocks pass through several network segments between the RRH and the BBU pool. Delays

may occur in each network segment, which include a propagation delay w_p and a serialization delay w_t for sending a block of data over the network. The transmission delay w_t at each intermediate node between the BBU pool and RRH can be expressed as the ratio of the number of sent bits to the bandwidth of the communication line. A number of articles consider the ideal case where each RRH is connected by a separate fiber-optic channel directly to the BBU pool, but this is not a realistic scenario. We assume the presence of a router that combines several RRH signal streams, and then sends it in one channel. Hence, there are additional queuing delay w_q on this router. Also, processing delay of w_s occurs in the BBU pool. Forward Error Correction (FEC) is a signal coding / decoding technique with the ability to detect errors and correct information by the forward method. Thus, the receiving equipment can detect and correct errors that occur in the transmission channel. FEC dramatically reduces the number of bit errors (BER), which allows you to increase the transmission distance of the signal without regeneration. The largest component of processing delay is due to direct error correction [3]. Also, when there is a shortage of resources in the BBU pool, a waiting delay w_b occurs. After that, the BBU sends the response back to the UE. Consequently, the round-trip delay includes transmission delays twice. The total delay must meet the strict latency requirements. In the case of LTE, the delay budget is about 3 ms [4]. Formula (1) shows all the components of the round-trip delay:

$$\xi_T = 2\left(w_{p_1} + w_{t_1} + w_q + w_{p_2} + w_{t_2}\right) + w_s + w_b \tag{1}$$

3. Delay components evaluation

The propagation delay w_p for the fiber-optic channel from RRH to the router is $\frac{d_1}{c_0}$ (segment 1), where d_1 is the distance from the RRH to the switch and c_o is the speed of light in fiber-optic cable. Accordingly, the delay in segment 2 from the router to the BBU pool is $\frac{d_2}{c_0}$, where d_2 is the distance between them. The serialization delay w_{t1} in the segment 1 is $\frac{b}{r_1}$, in the second section $w_{t2} = \frac{b}{r_2}$, where b is a fixed length of code block, and r_1, r_2 are the bandwidth of the segments.

To analyze queuing delay at the router, we employ the G/G/1-type queuing system. Analytical review [5, 6] showed that one of the most successful approximations for calculating the average waiting time w_q in this type of queuing system the following formula:

$$w_q = \frac{\rho_1 \frac{b}{r_2}(v_a^2 + v_b^2)}{2(1-\rho_1)} f(v_a) \tag{2}$$

where $\rho_1 = \frac{\lambda b}{r_2} < 1$ is the offered load on the system, λ is the average flow rate of customers, v_a is the variation coefficient of the interarrival times and v_b is the variation coefficient of the service times. Finally, $f(v_a)$ is a correction function, which

depends on the value of the variation coefficient v_a:

$$f(v_a) = \begin{cases} exp[-\frac{2(1-\rho_1)}{3\rho_1}\frac{(1-v_a^2)^2}{v_a^2+v_b^2}], & v_a < 1; \\ exp[-(1-\rho_1)\frac{v_a^2-1}{v_a^2+4v_b^2}, & v_a \geq 1. \end{cases} \quad (3)$$

The processing delay in the BBU pool is the time to process the radio signal, for example, demodulation and coding. The decoding calculation has its own performance directly related to the number of cycles performed by the FEC, and the average processing delay w_s can be expressed as

$$w_s = \frac{kbF}{pO} + J. \quad (4)$$

Formula (4) can be obtained using the following considerations. The BBU pool executes k cycles of the FEC algorithm for each code block. Parameter b, as before, is the length of the code block in bits, which may vary, depending on the technology in use, the coding rate, and the puncturing rate adjustment algorithm [7]. Each bit of the code block is usually processed by decoder with complexity F, expressed in bitwise operations. The processor clock speed allocated for BBU is denoted as p (in Hz) and O is the processor efficiency in operations per clock cycle, which is defined by the number of processor cores. In addition, we denote J the time required to process other wireless functions [3].

To calculate the waiting time in the BBU pool w_b we model it in terms of $G/G/m$ queuing system. Since the amount of computational resources in the BBU pool is optimized to serve the offered load, we use the asymptotic approximation for the$G/G/m$ queuing system [8] :

$$w_b = \frac{\lambda(\sigma_T^2 + \frac{\sigma_X^2}{m})}{2(1-\rho_2)} \quad (5)$$

Where m is the number of BBUs, $\rho_2 = \frac{\lambda \bar{X}}{m}$ is the offered load, $\bar{X} = \frac{kbF}{pO} + J$ is the average service time, σ_T is the variance of the interarrival times and σ_X is the variance of the service times. Combining all the delay components discussed above, the total delay between the BBU pool and the RRH can therefore be expressed as:

$$\mathbb{M}\xi_T = 2\left(\frac{d_1}{c_0} + \frac{b}{r_1} + \frac{\rho_1 \frac{b}{r_2}(v_a^2+v_b^2)}{2(1-\rho_1)} f(v_a) + \frac{d_2}{c_0} + \frac{b}{r_2}\right) + \frac{kbF}{pO} + J + \lambda\frac{[\sigma_T^2 + \frac{\sigma_X^2}{m}]}{2[1-\rho_2]} \quad (6)$$

b	6144 bits
c_0	$2 \cdot 10^8$ m/s
$r_1 = r_2$	10 Gbit/s
k	7 cycles
F	200 operations per bit
p	3,47GHz
O	2 operations per cycle
J	0,3 ms
λ	1000000 applications per second
ρ_2	0,9
v_a	0,3
v_b	0,4

Table 1. Input parameters

4. Numerical analysis

In this section we show the results of the numerical analysis. We consider the scenario from [3], where the BBU pool must decode the code block from RRH and, setting ξ_T less than or equal to the delay budget, we find the maximum distance $d = d_1 + d_2$ between the BBU pool and RRH. Take $d_1 = \frac{1}{3}d$ and $d_2 = \frac{2}{3}d$. The input parameters for the evaluation are summarized in table 1. We calculate the variance of σ_T^2 and σ_X^2 in terms of the variation coefficients. In the case of the interval between receipts:

$$v_a = \frac{\sqrt{\sigma_T^2}}{\frac{1}{\lambda}} \Rightarrow \sigma_T^2 = (\frac{v_a}{\lambda})^2$$

In case of service:

$$v_a = \frac{\sqrt{\sigma_X^2}}{\bar{X}} \Rightarrow \sigma_X^2 = (v_b \bar{X})^2$$

Finally, we accept the delay budget = 3 ms[4].

Figures 2 and 3 show the calculated average round-trip delay with total distance $d = 20$km in cases $\lambda = 10^6$ and $1.5 \cdot 10^6$ respectively, as a function of number of BBUs m in BBU pool. One can note that the dependence is nearly linear. In case of $\lambda = 10^6$, the minimum required number of BBUs to satisfy the delay budget is 1688, while in case $\lambda = 1.5 \cdot 10^6$ it is 2526.

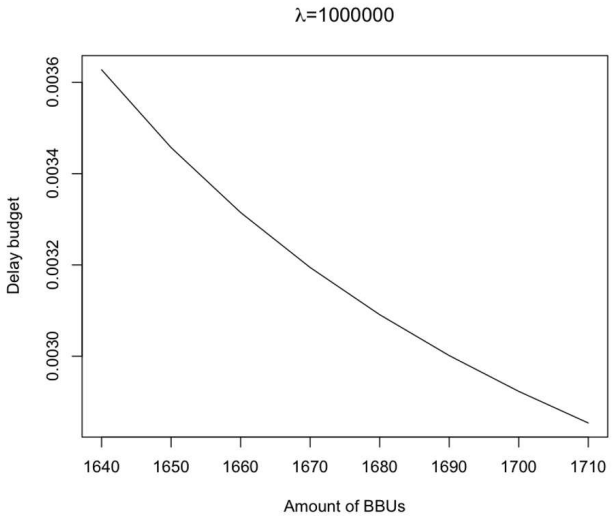


Fig. 2. The round-trip delay for data block arrival intensity $\lambda = 10^6$

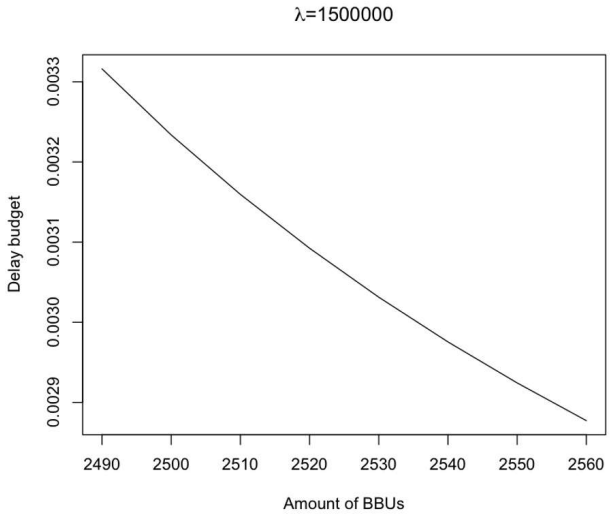


Fig. 3. The round-trip delay for data block arrival intensity $\lambda = 1.5 \cdot 10^6$

5. Conclusion

In the paper, we developed a mathematical model of the Cloud RAN that carefully takes into account all delay components. To quantify the average round-trip delay

we used well-known methods of queuing theory. The developed model was used to evaluate the required computational resources in the BBU pool to satisfy the strict delay budget in C-RAN.

REFERENCES

1. C. Kuilin, D. Ran, C-ran the road towards green ran, China Mobile Research Institute, White Paper.
2. F. Musumeci, C. Bellanzon, N. Carapellese, M. Tornatore, A. Pattavina, S. Gosselin, Optimal bbu placement for 5g c-ran deployment over wdm aggregation networks, Journal of Lightwave Technology 34 (8) (2015) 1963–1970.
3. S. Bhaumik, S. P. Chandrabose, M. K. Jataprolu, G. Kumar, A. Muralidhar, P. Polakos, V. Srinivasan, T. Woo, Cloudiq: A framework for processing base stations in a data center, in: Proceedings of the 18th annual international conference on Mobile computing and networking, 2012, pp. 125–136.
4. I. Alyafawi, E. Schiller, T. Braun, D. Dimitrova, A. Gomes, N. Nikaein, Critical issues of centralized and cloudified lte-fdd radio access networks, in: 2015 IEEE International Conference on Communications (ICC), IEEE, 2015, pp. 5523–5528.
5. T. Aliev, Osnovy modelirovaniya diskretnyh sistem.
6. G. Basharin, P. Bocharov, Y. Kogan, Analiz ocheredey v vychislitelnyh setyah. Teoriya i metody rascheta, Nauka, 1989.
7. M. F. Brejza, L. Li, R. G. Maunder, B. M. Al-Hashimi, C. Berrou, L. Hanzo, 20 years of turbo coding and energy-aware design guidelines for energy-constrained wireless applications, IEEE Communications Surveys & Tutorials 18 (1) (2015) 8–28.
8. S. K. Bose, Simulation techniques for queues and queueing networks, in: An Introduction to Queueing Systems, Springer, 2002, pp. 257–281.
9. E. U. T. R. Access, Multiplexing and channel coding,(3gpp ts 36.2. 212 version 12.6. 0 release 12),", Technical Specification.
10. H. Holma, A. Toskala, LTE for UMTS: OFDMA and SC-FDMA based radio access, John Wiley & Sons, 2009.

UDC: 519.872

Resource Queuing System with Preemptive Priority for URLLC and eMBB Coexistence in 5G NR

E. Sopin[1,2], V. Begishev[1], D. Moltchanov[3], A. Samuylov[3]

[1]Peoples' Friendship University (RUDN University), 117198, 6 Miklukho-Maklaya St, Moscow, Russian Federation
[2]Institute of Informatics Problems, FRC CSC RAS, 119333, Vavilova 44-2, Moscow, Russian Federation
[3]Tampere University, Address, Kuntokatu 3 FI-33520, Tampere, Finland

{sopin-es, begishev-vo}@rudn.ru, {dmitri.moltchanov, andrey.samuylov}@tuni.fi

Abstract

One of the ways to enable smooth coexistence of ultra reliable low latency communication (URRLC) and enhances mobile broadband (eMBB) services at the air interface of perspective 5G New Radio (NR) technology is to utilize preemptive priority service. In this paper, we provide approximate analysis of the queuing system with random resource requirements, two types of customers and preemptive priority service procedure. The distinctive feature of the systems – the random resource requirements – allows to capture the essentials of 5G NR radio interface but inherently increases the complexity of analysis. We present the main performance metrics of interest including session drop probability and system resource utilization as well as assess their accuracy by comparing with computer simulations.

Keywords: resource queuing system, preemptive priority, blocking probability, interruption probability, URLLC traffic, network slicing

1. Introduction

In recent years, queuing systems with random resource requirements, where customers require not only a server but also a random volume of resources, have drawn significant attention for their ability to capture specifics of session serving process in prospective cellular systems including 5G New Radio (NR) technology [3, 1]. However, despite many research activities in the field, resource queuing systems with priorities have not been addressed yet.

The model presented in this paper has a large scope of applications in 5G systems. Of particular interest is provisioning of ultra reliable low latency service (URLLC) at

The publication has been prepared with the support of RFBR according to the research projects No.19-07-00933, 20-07-01052.

New Radio (NR) base stations in industrial applications in context of network slicing [4, 7]. Recall that URLLC service requires extremely small delays and loss guarantees at the air interface. To ensure it when mixed with conventional enhanced mobile broadband (eMBB, [6]) service at a single NR BS, several approaches ranging from the use of intentional overlapping by using non-orthogonal multiple access (NOMA, [5]) to explicit static bandwidth reservations have been proposed in the past. In this context, explicit prioritization may provide an alternative approach to maintain extreme service characteristics of URLLC traffic.

The specified model allows to account for random resource requirements at the air interface of both eMBB and URLLC service induced by random locations of user equipment in the coverage area of NR BS [3, 2]. Preemptive priority discipline simultaneously accounts for efficient use of resources at the NR BS and ensures that URLLC traffic receives absolute priority over conventional eMBB traffic reaching the prescribed loss guarantees. Supplementing the model with a certain deployment of NR BSs and UEs in the considered area one may characterize the required density of NR BSs needed to maintain the prescribed performance provided to both URLLC and eMBB traffic types.

The behavior of preemptive priority customers is equivalent to the behavior of these customers in the same queuing system without non-priority customers. Thus, we focus on the metrics of interest associated with non-priority customers.

2. Model description and analysis

We consider a multiserver queuing system with N servers and resource volume R. Two types of customers are served in the queuing system: first type are the preemptive priority customers and the second type are non-priority customers. Customers arrive according to Poisson process with intensities λ_1 and λ_2 correspondingly. An arriving customer of type l requires discrete random volume of resources according to probability distribution $\{p_{l,r}\}$, $l = 1, 2$, $r = 1, 2, ...$, where $\sum_{r=1}^{\infty} p_{l,r} = 1$. The service times are exponentially distributed with intensities μ_1 and μ_2, correspondingly.

Assume that there are n_1 customers of the first type totally occupying r_1 resources and n_2 customers of the second type occupying r_2 resources. Upon arrival of a second type customer that requires j resources, if there is free server ($n_1 + n_2 < N$) and total volume of unoccupied resources is greater than the required volume of the customer ($R - r_1 - r_2 \geq j$), then the customer is accepted and the required resource volume is allocated to the customer. If arriving customer is the priority customer, then it is still accepted if $n_1 < N$ and $R - r_1 \geq j$, but the service of one or more customers of second type is interrupted. In this case, the system randomly choose and terminates

second type customers one by one, until the required server and resources can be allocated to the customer $R - r_1 - r_2^* \geq j$.

To decrease the complexity of the stochastic process that describes the behavior of the system, we employ the technique originally proposed in [8]. According to it, instead of keeping track of resources allocated to all the customers, we follow only a total amount of occupied resources for each type of customers. Then, the behavior of the system can be described by a simplified process, $X(t) = (\xi_1(t), \gamma_1(t), \xi_2(t), \gamma_2(t))$, where $\xi_l(t)$ is the number of l-type customers at time t and $\gamma_l(t)$ is total resource volume occupied by l-type customers.

Let $q(n_1, r_1, n_2, r_2)$ be the stationary distribution of $X(t)$, $Q(n, r)$ and $P(n, r)$ – marginal stationary distributions of first and second type customers, respectively. Since preemptive priority customers are not affected by non-priority customers, then, according to [8] we have

$$Q(n, r) = Q(0, 0)\frac{\rho_1^n}{n!}p_{1,r}^{(n)}, \tag{1}$$

where $\rho_l = \frac{\lambda_l}{\mu_l}$, $p_{1,r}^{(n)}$ is the probability that n first type customers totally occupy r resources and $Q(0, 0)$ is calculated using the normalizing condition. Note that probabilities $p_{1,r}^{(n)}$ are evaluated from distribution $\{p_{1,r}\}$ by n-fold convolution.

To derive equations for approximation of stationary probabilities $P(n, r)$, we introduce additional notation. Let $\Pi(n_1, r_1|n_2, r_2)$ be the probability that there are n_1 first type customers occupying r_1 resources conditional to n_2 second type customers are in the system with r_2 resources occupied, i.e.,

$$\Pi(n_1, r_1|n_2, r_2) = \frac{Q(n_1, r_1)}{\sum\limits_{n=0}^{N-n_2}\sum\limits_{r=0}^{R-r_2} Q(n, r)}. \tag{2}$$

Further, let $\beta_2(m, j|n, r)$ be the probability that m customers of second type totally occupy j resources under condition that n second type customers occupy r resources, $n \geq m, r \geq j$,

$$\beta_2(m, j|n, r) = \frac{p_{2,j}^{(m)} p_{2,r-j}^{(n-m)}}{p_{2,r}^{(n)}}. \tag{3}$$

Finally, let $\varphi(k, j|n_2, r_2)$ be the probability that exactly k non-priority customers occupying j resources are terminated upon arrival of a priority customer under

condition that there are n_2 customers of second type with r_2 resources occupied, i.e.,

$$\varphi(k,j|n_2,r_2) = \sum_{n_1=0}^{N-n_2}\sum_{r_1=0}^{R-r_2} \Pi(n_1,r_1|n_2,r_2)\beta_2(k,j|n_2,r_2)\times$$
$$\times \sum_{s=R-r_1-r_2+1}^{R-r_1} p_{1,s} \prod_{m=1}^{k-1}\sum_{l=m}^{r_1+r_2+s-R} \beta_2(m,l|n_2,r_2). \quad (4)$$

The product may be interpreted as the probability that dropping of $1,2,...,k-1$ second type customers is not enough to allocate the required resources to the priority customer. Then, multiplication of $\beta_2(k,j|n_2,r_2)$ and sum over s gives the probability that dropping k customers is enough and the last two sums reflect all the possible values of number of first type type customers and resources occupied by them.

Utilizing the the introduced probabilities, (2), (3), and (4), we can derive the system of equilibrium equations as follows:

$$P(0,0)\lambda_2 \sum_{j=0}^{R} p_{2,j}\hat{Q}(R-j) = \mu\sum_{j=1}^{R} P(1,j) + \lambda_1 \sum_{k=1}^{N}\sum_{j=k}^{R} P(k,j)\varphi(k,j|k,j); \quad (5)$$

$$P(n,r)\left(n\mu + \lambda_2\sum_{j=0}^{R-r} p_{2,j}\hat{Q}(R-r-j) + \lambda_1\sum_{k=1}^{n}\sum_{j=k}^{r-n+k}\varphi(k,j|n,r)\right) =$$
$$= \lambda_2\hat{Q}(R-r)\sum_{i=1}^{r-n+1} P(n-1,r-i)p_{2,i} + (n+1)\mu\sum_{j=1}^{R-r} P(n+1,r+j)\beta_2(1,j|n+1,r+j) +$$
$$+ \lambda_1\sum_{k=1}^{N-n}\sum_{j=k}^{R-r} P(n+k,r+j)\varphi(k,j|n+k,r+j), \qquad n \le r \le R, \quad n < N; \quad (6)$$

$$P(N,r)\left(N\mu + \lambda_1\sum_{k=1}^{N}\sum_{j=k}^{r-N+k}\varphi(k,j|N,r)\right) = \quad (7)$$
$$= \lambda_2\hat{Q}(R-r)\sum_{i=1}^{r-N+1} P(N-1,r-i)p_{2,i}, \qquad N \le r \le R.$$

where $\hat{Q}(r) = \sum_{n=0}^{N-1}\sum_{j=0}^{r} Q(n,j)$. By solving the system (5) - (7) we obtain the marginal stationary probabilities $P(n,r)$.

Now, we proceed with performance metrics of interest. First, we analyze the blocking probabilities upon arrival of both types of customers, namely π_{b1} and π_{b2}. For priority customers, the blocking probability is obtained in [8], while π_{b2} is deduced using similar logic

$$\pi_{b1} = 1 - Q(0,0) \sum_{n=0}^{N-1} \frac{\rho_1^n}{n!} \sum_{r=0}^{R} p_{1,r}^{(n+1)}; \tag{8}$$

$$\pi_{b2} = 1 - \sum_{n_1+n_2 \le N-1} \sum_{r_1+r_2 \le R-1} P(n_2, r_2) \Pi(n_1, r_1 | n_2, r_2) \sum_{j=1}^{R-r_1-r_2} p_{2,j}. \tag{9}$$

Finally, we proceed with the probability π_i that a non-priority customer is interrupted. The intensity of customer interruption is obtained as $\lambda_2(1-\pi_{b,2}) - \bar{N}_2\mu_2$ from the equality of arrival intensity and intensity of leaving the system. Here $\bar{N}_2$ is the average number of 2-type customers in the system. Then, the interruption probability is given by the ratio of interruption and arrival intensities, that is,

$$\pi_i = 1 - \frac{\bar{N}_2 \mu}{\lambda_2(1-\pi_{b,2})}. \tag{10}$$

3. Numerical results

To analyze the accuracy of the proposed analysis, we have developed a simulation tool that models the considered resources queuing system with two types of customers and preemptive priority service discipline. Here, we assume that resource requirements of both priority and non-priority customers have geometric distribution with parameter 0.5. Further, we assume that $N = 11$, $R = 18$, the arrival intensity of priority customers is $\lambda_1 = 4$, and service intensities are $\mu_1 = \mu_2 = 1$. The arrival intensity of non-priority customers λ_2 varies from 4 to 8.

Figure 1 shows the comparison results. One can note that the blocking probability of the priority customers shows almost perfect match with the simulations, while other two probabilistic metrics have bigger relative error. Notably, the relative error of all probabilities decreases with the increase of the non-priority customers intensity.

4. Conclusion

Motivated by coexistence of high-priority URLLC and low priority eMBB traffic at 5G NR air interface, in the paper, we have analyzed the limited resources queuing system with two types of customers and preemptive priority service discipline. We have developed an approach for approximate analysis performance measures for both type of customers. Although the complexity of the proposed approach is lower compared to the direct solution of equilibrium equations, it is still high requiring

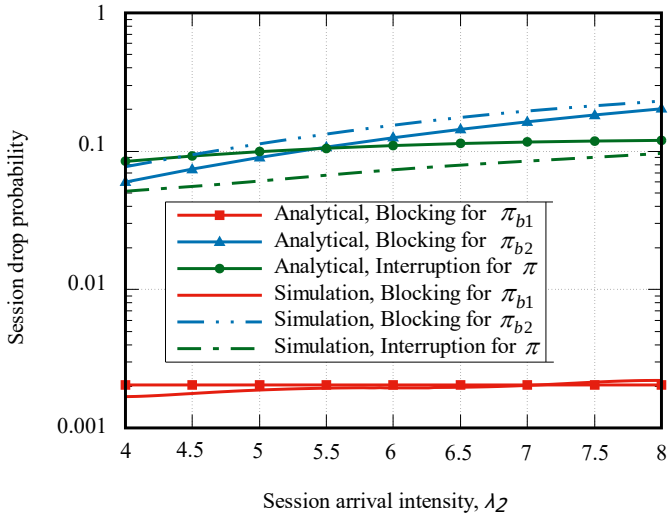


Fig. 1. Comparison of analytical and simulation results

significant computational efforts. In our future work, we will work on improving approximation accuracy.

REFERENCES

1. Lu X., Sopin E, Petrov V., Galinina O., Moltchanov D., Ageev K., Andreev S., Koucheryavy Y., Samouylov K., Dohler M. Integrated Use of Licensed- and Unlicensed-Band mmWave Radio Technology in 5G and Beyond // IEEE Access. 2019. V. 7. P. 24376–24391.
2. Kovalchukov R., Moltchanov D. Gaidamaka Yu., Bobrikova E. An Accurate Approximation of Resource Request Distributions in Millimeter Wave 3GPP New Radio Systems // Internet of Things, Smart Spaces, and Next Generation Networks and Systems. NEW2AN 2019. Lecture Notes in Computer Science. V. 11660. P. 572–585.
3. Begishev V., Moltchanov D., Sopin E.,Samuylov A., Andreev S., Koucheryavy Y., Samouylov K. Quantifying the impact of guard capacity on session continuity in 3GPP new radio systems // IEEE Transactions on Vehicular Technology. 2019. V. 68(12). P. 12345–12359.
4. Orsino A., Kovalchukov R., Samuylov A., Moltchanov D., Andreev S., Koucheryavy Y., Valkama M. Caching-aided collaborative D2D operation for predictive

data dissemination in industrial IoT // IEEE Wireless Communications. 2018. V. 25(3). P. 50–57.
5. Kassab R., Simeone O., Popovski P. Coexistence of URLLC and eMBB services in the C-RAN uplink: an information-theoretic study // 2018 IEEE Global Communications Conference (GLOBECOM). P. 1–6.
6. Chen Y., Cheng L., Wang L. Prioritized resource reservation for reducing random access delay in 5G URLLC // 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC). P. 1–5.
7. Foukas X., Patounas G., Elmokashfi A., Marina M. Network slicing in 5G: Survey and challenges // IEEE Communications Magazine. 2017. V. 55(5). P. 94–100.
8. Naumov V.A., Samuilov K.E., Samuilov A.K. On the total amount of resources occupied by serviced customers // Automation and Remote Control. 2016. V. 77(8). P. 1419–1427.

UDC: 519.234.6

Leader Election in Communities for Information Spreading

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

The paper is devoted to the information spreading in random complex networks. Our objective is to elect leader nodes or communities of the network, which may spread the content among all nodes faster. We consider a well-known SPREAD algorithm by Mosk-Aoyama and Shah (2006), which provides the spreading and the growth of the node set possessing the information. Assuming that all nodes have asynchronous clocks, the next node is chosen uniformly among nodes of the network by the global clock tick according to a Poisson process. The extremal index measures the clustering tendency of high threshold exceedances. The node extremal index shows the ability to attract highly ranked nodes in its orbit. Considering a closeness centrality as a measure of a node's leadership, we find the relation between its extremal index and the minimal spreading time.

Keywords: Complex network, random graph, community, extremal index, information spreading

1. Introduction

Analyze and modeling of a fast content spreading is an important issue in distributed computing [1], [2], and social networks. Proposed solutions for this issue do not only help to observe the information diffusion but also serve as a valuable resource to predict the characteristics of the network. The spreading time of infection within a human contact network [3] can dramatically reflect on the life of humanity.

In [4], [5], a statistical clustering of the random network by the node extremal index (EI) is proposed. The EI plays a key role in the extreme value analysis since it allows to obtain a limit distribution of maximum when observations are dependent.

Definition 1. A stationary sequence $\{Y_n\}_{n\geq 1}$ with distribution function (df) $F(x)$ and $M_n = \max_{1\leq j\leq n} Y_j$ is said to have EI $\theta \in [0,1]$ if for each $0 < \tau < \infty$ there is a sequence of real numbers $u_n = u_n(\tau)$ such that

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

hold ([6], p.53).

For independent values, the EI is equal to one. The converse is incorrect. As closer θ to zero, as stronger the dependence. The EI measures the clustering tendency of high threshold exceedances. Its reciprocal $1/\theta$ approximates the mean number of exceedances per cluster (the mean cluster size). In classical settings, the cluster is defined as a block of data with at least one exceedance over threshold u. In [7] the cluster is a set of consecutive exceedances between two consecutive non-exceedances. We follow this definition and modify it with respect to graphs. There are several problems here. First, the real network may be non-stationary with regard to any characteristic of the nodes. The node degree (i.e. the number of its links), PageRank, and a centrality index may show the node leading in the network. The network may be partition into communities by interests, that are sets of nodes with a large number of internal links. One can expect that the communities can be rather homogeneous and even stationary distributed. Communities can be selected by applying such measures like the conductance, clustering coefficient and modularity [8], [9]. Next, nodes in the graph are not numerated as in a random sequence, but clusters require an order. Thus, in [4], [5], generations of followers of a node are used as blocks and potential clusters. The hypothesis in [4], [5] states that the node EI shows the ability to attract highly ranked nodes in its orbit and to spread information faster. This means that a coupled tree-like graph of the node taken as root may contain influential nodes. Since the EI relates to the stationary sequence, we use a community to which the node belongs instead of the sequence.

We aim to study the dependence between the EI of a closeness centrality [10] used as a node's leadership measure and minimum time needed to spread the information to all nodes of a undirected graph starting from a node. To this end, we use a well known SPREAD algorithm by [2] which provides the spreading and the growth of the set possessing the node information. We partition the graph into communities to find leaders. The question arises, what EI the node or its community must have to be a leader with regard to the minimum spreading time?

The paper is organized as follows. In Section 2, related works regarding the community and leading nodes identification as well as the information spreading algorithm, the tail index (TI) estimation and stationarity tests are recalled. In Section 3, our main result concerning the EI estimation in random graphs based on the intervals estimator by [7] is presented. Simulation study is given in Section 4. The exposition is finalized by conclusions in Section 5.

2. Related works

2.1. Graph community characteristics. For graph $G = (V, E), |V| = n$, the conductance measures the minimum relative connection strength between "isolated"

subsets $\{S\}$ and the rest of the network [1, 2, 11]. The conductance is defined by

$$\Phi(G) = \min_{S\subseteq V, |S|\le n/2} \phi(S,V), \qquad \phi(S,V) = \frac{\sum_{i\in S, j\in V\setminus S} P_{i,j}}{|S|},$$

where P is the stochastic probability matrix associated with the communication of nodes [2]. The conductance satisfies $0 < \Phi(G) \le 1$. In [1] the node i chooses a neighbor j with probability $P_{i,j} = 1/D_i$, where D_i is the degree of node i. In [2] it is proposed to use $P_{i,j} = 1/D_{max}$ if $(i,j) \in E$ and $P_{i,j} = 1 - D_i/D_{max}$ if $i = j$, but it requires to know the maximum degree in the graph $D_{max} = \max_{i\in V} D_i$.

The modularity Q is a measure to partition the network into communities [8]. It shows how many edges exist within communities and between them:

$$Q = \frac{1}{2m}\sum_{vw}\left[A_{vw} - \frac{D_w D_v}{2m}\right]\delta(c_w, c_v),$$

$m = \frac{1}{2}\sum_{vw} A_{vw} = |E|/2$ is a number of edges in the undirected graph, A is an adjacency matrix, $\delta(c_w, c_v)$ is equal to 1 when nodes w and v belong to the same community. A Greedy Modularity Maximization Algorithm (GMMA) [8] is used to detect the community structure fast.

2.2. Information spreading time. The information spreading time is determined in [2] for global broadcast problem, i.e. when one aims to disseminate a content to all nodes in the network. Suppose node $i \in V$ has a message m_i. $S_i(t)$ denotes the set of nodes that have the message m_i at time t. For $\delta \in (0,1)$ the δ-information-spreading time of the algorithm P is determined as

$$T_P^{spr}(\delta) = inf\{t \ge 1 : Pr(\cup_{i=1}^n \{S_i(t) \ne V\}) \le \delta\},$$

In [2] it is derived that there exists an information dissemination algorithm P such that, for any $\delta \in (0,1)$, $T_P^{spr}(\delta) = O((log(n) + log(\delta^{-1}))/\Phi(G))$, where n is the number of nodes in the network. Spreading Algorithm 1 has been proposed in [2].

Algorithm 1 Algorithm SPREAD(P)

When a node i initiates communication at round t:

1: Node i chooses a node u at random, and contacts u. The choice of the communication partner u is made independently of all other random choices, and the probability that node i chooses any node j is $P_{i,j}$.
2: Node u sends all of the messages it has to node i, so that $M_i(t+1) = M_i(t) \cup M_u(t)$, where $M_i(t)$ is a set of received messages on round t.

2.3. Leader node identification. Let $G = (V, E)$ be undirected connected graph of order n. A standard graph index used for the leader election is the closeness centrality C_x [10]:

$$C_x = \frac{n-1}{\sum_{y,y\neq x} d(x,y)}, \qquad 0 < C_x \leq 1, \tag{1}$$

where d(x, y) is the shortest path $(x, \ldots, y)$ between nodes x and y. The closer a node to other nodes, the closer its value C_x to 1. The node degree D_x may also be used as a measure of the leadership.

2.4. Tail index identification. Let $X_1, \ldots, X_n$ be a stationary sequence of i.i.d r.v.s. A TI $\alpha = 1/\gamma$ is reciprocal of the extreme value index γ. γ may be estimated by the moment estimator $\widehat{\gamma}_M(k)$ [12] and the Hill's estimator $\widehat{\gamma}_H(k)$:

$$\widehat{\gamma}_M(k) = \widehat{\gamma}_H(k) + 1 - 0.5\left(1 - \frac{\widehat{\gamma}_H^2(k)}{S_{n,k}}\right)^{-1}, \qquad \widehat{\gamma}_H(k) = \frac{1}{k}\sum_{i=1}^{k} log(\frac{X_{(n-i+1)}}{X_{(n-k)}}),$$

by order statistics $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$, where $S_{n,k} = \frac{1}{k}\sum_{i=1}^{k}\left(log(\frac{X_{(n-i+1)}}{X_{(n-k)}})\right)^2$. k is a number of the largest order statistics. Its optimal value is further chosen by the bootstrap method [13]. Bootstrap confidence intervals are calculated by [14].

2.5. Stationarity tests. We use the stationarity test statistic

$$V/S = V_n/\widehat{s}^2_{n,q}, \quad V_n = \frac{1}{n^2}\left[\sum_{k=1}^{n}(S_k^*)^2 - \frac{1}{n}\left(\sum_{k=1}^{n} S_k^*\right)^2\right], \quad \widehat{s}^2_{n,q} = q^{-1}\sum_{i,j=1}^{q}\widehat{\gamma}_{i-j}, \tag{2}$$

$S_k^* = \sum_{j=1}^{k}(X_j - \overline{X}_n)$, $\widehat{\gamma}_j = n^{-1}\sum_{i=1}^{n-j}(X_i - \overline{X}_n)(X_{i+j} - \overline{X}_n)$, $0 \leq j < n$, proposed in [15]. The null hypothesis of stationarity is rejected, if $V/S > c_\alpha$, c_α is a quantile of the asymptotic df of the Kolmogorov statistic $F_K(\pi\sqrt{x})$. $c_\alpha \in \{0.190, 0.153, 0.1, 0.069\}$ holds for significant level $\alpha \in \{5, 10, 30, 50\}\%$, respectively.

3. Extremal index estimation

Let $G = (V, E)$ be undirected graph of the order n and $(X_1, \ldots, X_n)$ be a sample of node characteristics with a marginal df $F(u)$. For stochastic sequence, a cluster is determined as consecutive exceedances over a predefined threshold u between two consecutive non-exceedances. An inter-cluster size is [7]

$$T(u) = \min\{t \geq 1 : X_{j+t} > u\} \quad \text{given} \quad X_j > u. \tag{3}$$

The intervals estimator of the EI is [7]

$$\widehat{\theta}(u) = \min(1, \theta^*), \ \theta^* = \begin{cases} \frac{2(\sum_{i=1}^{N-1} T(u)_i - 1)^2}{(N-1)\sum_{i=1}^{N-1}(T(u)_i - 1)(T(u)_i - 2)}, & max\{T(u)_i\} > 2 \\ \frac{2(\sum_{i=1}^{N-1} T(u)_i)^2}{(N-1)\sum_{i=1}^{N-1}(T^2(u)_i)}, & \text{otherwise,} \end{cases} \tag{4}$$

where $N = N(u) = \sum_{i=1}^{n} \mathbb{1}(X_i > u)$. To introduce the intervals estimator for graphs, we propose to define $T(u)$ as the number of edges of the shortest path between nodes which characteristics are larger than u.

3.1. Extremal index of the community. Let us choose a node x with characteristic X_x and a community of nodes to which it may belong. To determine the EI of the node x, we take a high quintile of $F(x)$ as the threshold u^*. Let a community S be a strict-sense stationary set with EI θ.

Algorithm 2 Node EI estimation

1: Set sequences $X_{xy} = \{X_x, X_{i_1}, X_{i_2}, \ldots, X_{i_m}, X_y\}$, $m \geq 1$ corresponding to shortest paths $(x, \ldots, y)$ from a fixed node x to each node y of the community S.
2: Define $\{T(u^*)_i\}$, $i = 1, 2, ...N$ by (3) for all sequences, where N is a total number of inter-cluster times over all possible shortest paths X_{xy}.
3: Estimate the EI $\widehat{\theta} = \widehat{\theta}_x(u^*)$ of node x by (4).

To estimate the EI of the community $\theta(S)$ one has to determine $\{T(u^*)_i\}$ for all possible pairs $(x, y) \in S$.

4. Community and leader election by simulation study

For the simulated undirected connected graph $G = (V, E)$, the spreading time T^{spr} is modeled as the time needed to send the message m_x from node x to other nodes by Algorithm 1. The clock ticks are modeled as Poisson process with the rate $n = |V|$. We partition the graph into communities by algorithm GMMA [8]. Using the closeness centrality C_x as a node characteristic, the TIs and EIs of communities are estimated as in Sections 2.4 and 3.1. To compare the node and community leadership we analyze relations between D_x, C_x, T^{spr}, the TI and the EI.

We simulate a geometric graph with the number of nodes $n = 200$ and the radius $r = 0.11$ [16]. This is the undirected graph constructed by n nodes uniformly placing in a unit square. An edge connects two nodes, if the distance between them is less than r. We partition the graph into communities Fig. 1 (left) and state the stationarity of all communities by test (2) Fig. 1 (right). Nodes with high C_x spread information faster, but the tendency is weaker preserved with regard to node degrees, Fig. 2 (left). In Fig.2 (right), we compare communities of the graph in Fig. 1 regarding the spreading time, the TI, the EI and the conductance. S_0 denotes the entire graph. The EI and the TI of S_4 are close to ones of S_0. In this respect, S_4 is a leading community. Its T^{spr} has the smallest variation and it contains a leader node with the smallest spreading time. S_2 has the smallest EI value and the highest conductance, but its minimum T^{spr} is not the best. S_7 is the worse spreader despite

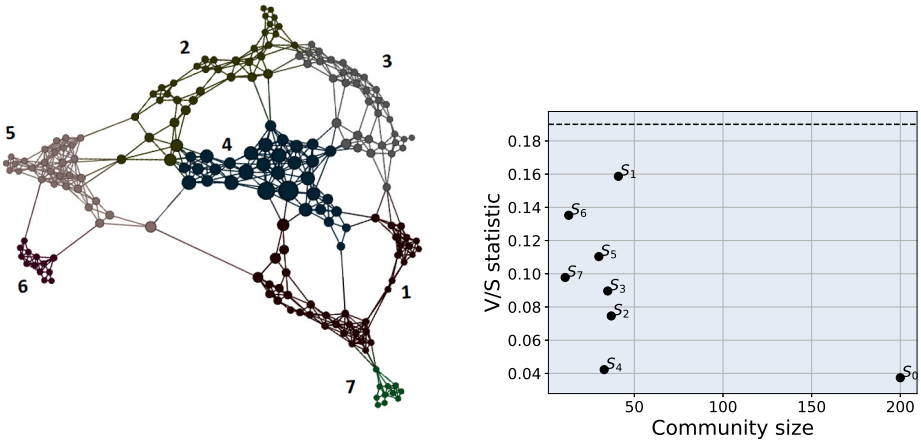


Figure 1. Geometric graph with communities $\{S_i\}_{i=1}^7$ and circle sizes of nodes marked proportional to C_x (left); the V/S-statistics of communities against the community sizes and with the 5% critical value 0.19 (dotted line) (right).

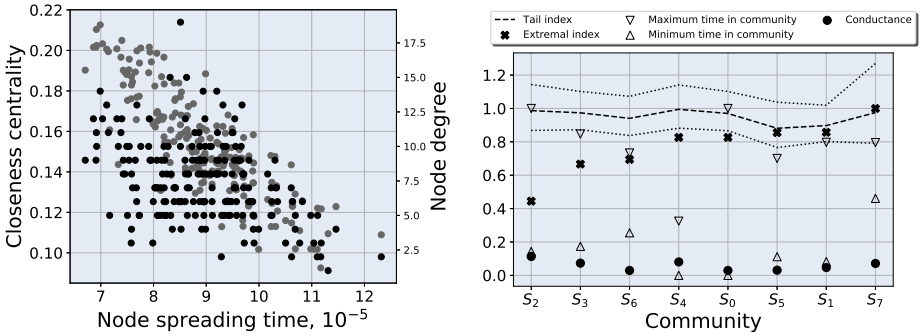


Figure 2. C_x (black) and D_x (grey) against the minimum T^{spr} (left); Minimum and maximum T^{spr}s over nodes in communities $\{S_i\}$ normalized to $[0, 1]$, the EIs $\theta(S_i)$, the conductances $\phi(S_i, V)$, the TIs with 95% bootstrap confidence intervals (dotted lines) (right).

its conductance is middle since it contains weak connected independent nodes, and its EI is equal to 1. The TIs of all communities are close, that indicates their similar heaviness of tails.

Now, we analyze communities over 100 simulated geometric graphs. We choose leading communities by the minimum EI, TI among all other communities in the

graph or those which have the same EI as an entire graph. Fig. 3 shows that the leading communities determine the best spreading time T_{min}^{spr} in the whole graph if they have the same EI. The linear dependence is much weaker when the leading community is chosen by the minimum EI or TI.

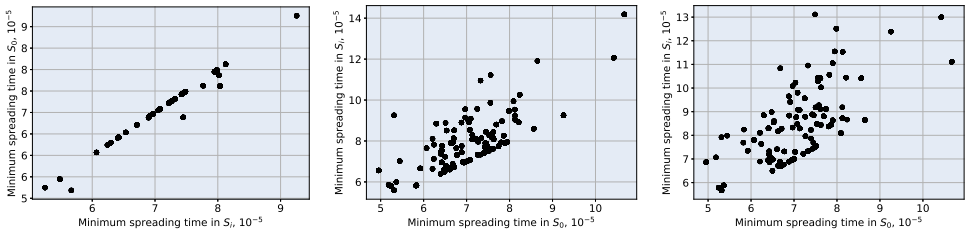


Figure 3. T_{min}^{spr}s in leading communities and entire graphs when they have the same EIs (left), when the communities have the minimum TI (middle) and the minimum EI (right).

5. Conclusion

We study the leadership of nodes and communities regarding their minimum spreading time. The modification of the intervals estimator of the EI for graphs is proposed. Taking the closeness centrality as a node characteristic and estimating its EI, we found that a community with the same EI as its graph identifies the best spreading time of the entire graph. Our future research will concern to analytical relations between the EI and the minimum spreading time as well as the fast algorithm of the intervals estimator of the EI on graphs.

Acknowledgments

The authors were partly supported by Russian Foundation for Basic Research (grant 19-01-00090).

REFERENCES

1. Censor-Hillel K., Shachnai H. Partial Information Spreading with Application to Distributed Maximum Coverage // In Proceedings of the 29th ACM SIGACT-SIGOPS symposium on Principles of distributed computing (PODC '10), ACM, New York, USA. 2010. P. 161-170.
2. Mosk-Aoyama D., Shah D. Computing separable functions via gossip // In Proceedings of the twenty-fifth annual ACM symposium on Principles of distributed computing (PODC '06), ACM, New York, USA. 2006. P. 113-122.

3. Holme P., Litvak N. Cost-efficient vaccination protocols for network epidemiology // PLoS Comput Biol. 2017. V. 13(9): e1005696.
4. Markovich N.M., Ryzhov M.S., Krieger U.R. Nonparametric Analysis of Extremes on Web Graphs: PageRank versus Max-Linear Model // CCIS. 2017. V. 700. P. 13–26.
5. Markovich N.M., Ryzhov M.S., Krieger U.R. Statistical Clustering of a Random Network by Extremal Properties// CCIS. 2018. V. 919. P.71-82.
6. Leadbetter M. R. Extremes and local dependence in stationary sequences// Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete. 1983. P. 291-306.
7. Ferro C., Segers J. Inference for clusters of extreme values // J. R. Statist. Soc. B. 2003. V. 65, Part 2, P. 545–556.
8. Clauset A., Newman M. E., Moore C. Finding community structure in very large networks // Physical Review E. 2004. V. 70(6). P. 066111.
9. Newman M. E. J. Networks: An Introduction // Oxford University Press, Second edition, 2018.
10. Stephenson K., Zelen M. Rethinking centrality: Methods and examples // Social Networks. 1989. V. 11(1). P. 1-37.
11. Leskovec J., Lang K.J., Dasgupta A., Mahoney M.W. Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters // eprint arXiv:0810.1355. 2008.
12. Dekkers A. L. M., Einmahl J. H. J., Haan, L. De. A Moment Estimator for the Index of an Extreme-Value Distribution // Ann. Statist. 1989. V. 17(4). P. 1833-1855.
13. Hall, P. Using the Bootstrap to Estimate Mean Squared Error and Select Smoothing Parameter in Nonparametric Problems // Journal of Multivariatte Analysis. 1990. V. 32. P. 177-203.
14. Caers J., Beirlant J., Vynckier P. Bootstrap confidence intervals for tail indices // Computational Statistics and Data Analysis. 1998. V. 26(3). P. 259-277.
15. Giraitis L., Leipus R., Philippe A. A test for stationarity versus trend and unit root for a wide class of dependent errors // Econometric Theory. 2006. V. 22. P. 989-1029.
16. Penrose M. Random Geometric Graphs // Oxford Studies in Probability. 2003.

УДК: 519.872

Система массового обслуживания $MMAP/PH_{1,2}/N/0$ с неоднородными запросами и приоритетами

В.И.Клименок[1], А.Н. Дудин[1], В.М. Вишневский[2]

[1]Факультет прикладной математики и информатики, Белорусский государственный университет, проспект Независимости, 4, Минск, Беларусь

[2]Институт проблем управления Российской академии наук„ ул. Профсоюзная, 65, Москва, Россия

klimenok@bsu.by, dudin@bsu.by, vishn@inbox.ru

Аннотация

В статье рассматривается многолинейная система массового обслуживания с неоднородными запросами, которые поступают в систему в маркированном марковском потоке. Запросы двух типов отличаются приоритетами и параметрами фазового распределения процесса обслуживания. Рассматриваемая система может использоваться для моделирования узла телекоммуникационной сети, в который поступает коррелированный поток запросов двух типов из соседних узлов. Функционирование системы описывается многомерной цепью Маркова. Вычисляются стационарное распределение вероятностей состояний и характеристики производительности системы, включая вероятности потерь запросов разных типов.

Ключевые слова: многолинейная система массового обслуживания, маркированный марковский поток, приоритеты, фазовое распределение, вероятности потерь

1. Введение

Одним из существенных разделов теории массового обслуживания является теория приоритетных систем, в которых запросам разных классов присваиваются разные категории важности и обслуживание производится в соответствии с приоритетной схемой, при которой более важные запросы имеют разного рода приоритеты по сравнению с менее важными. Модели систем с приоритетами возникают во многих приложениях. В частности, в телекоммуникационной сети приоритет клиента может определяться ее владельцем посредством соглашения об уровне обслуживания, в соответствии с которым клиент готов платить больше, чтобы получить высокоприоритетный доступ к популярному ресурсу.

Работа выполнена при финансовой поддержке РФФИ, проект №19-29-06043.

Различные приоритетные схемы используются в медицине в отделениях скорой помощи при сортировке поступающих пациентов по степени тяжести заболевания. В случае ненадежных систем приоритетный запрос может рассматриваться как поломка оборудования. Приоритет может устанавливаться, чтобы максимизировать прибыль компании. Например, интернет-магазин может установить высокий приоритет для запросов крупных потребителей с целью предотвращения их ухода на другие интернет-ресурсы.

Приоритетные системы широко представлены в литературе. Обзор ранних работ в этой области можно найти в известных монографиях [1, 2, 3, 4, 5]. Большинство из этих работ посвящено системам со стационарными пуассоновскими потоками. Вместе с там, потоки в современных телекоммуникационных сетях являются, как правило, коррелированными и неоднородными. В случае однородных запросов хорошей математической моделью таких потоков является марковский поток (Markovian arrival process – MAP), см., например, [6]. Системы массового обслуживания с MAP-потоками и приоритетами рассмотрены в статьях [7, 8, 9, 10]. Однако особый интерес для приложений представляют системы с коррелированными потоками разнородных запросов. Такие потоки хорошо моделируются маркированным марковским потоком (Marked Markovian arrival process – $MMAP$), см. [11]. По нашим сведениям, на сегодняшний день существует немного работ, посвященных приоритетным системам с $MMAP$-потоками. Мы можем сослаться на статью [12], в которой исследована система $MMAP/MAP/1$ с абсолютными приоритетами. Статья посвящена анализу моментов длины очереди. В статье [13] рассмотрена многолинейная система с $MMAP$-потоком и абсолютными приоритетами, которые могут меняться в течение времени ожидания начала обслуживания. Для этой сложной системы авторам удалось найти условие существования стационарного режима и границы для длин очередей.

В данной статье мы рассматриваем многолинейную систему без буфера, в которую поступает коррелированный поток запросов двух типов. Запросы одного из типов имеют абсолютный приоритет. Времена обслуживания запросов обоих типов распределены по фазовому закону (известная аббревиатура PH - Phase type distribution) с разными параметрами. Рассматриваемая система может служить адекватной моделью узла телекоммуникационных сетей различного назначения. Построена многомерная цепь Маркова, описывающая функционирование системы, вычислено ее стационарное распределение и важнейшие характеристики производительности.

2. Описание модели

Рассматривается N-линейная система массового обслуживания без буфера. Запросы разных типов поступают в $MMAP$ -потоке под управлением неприводимой цепи Маркова с непрерывным временем ν_t, $t \geq 0$, которая принимает значения в множестве $\{0, 1, 2, \ldots, W\}$. В случае двух типов запросов $MMAP$ полностью определяется пространством состояний управляющего процесса ν_t, $t \geq 0$, и $(W+1) \times (W+1)$ матрицами D_k, $k = 0, 1, 2$, или их производящей функцией $D(z) = D_0 + D_1 z + D_2 z^2$. Элементами матриц D_k, $k = 1, 2$, являются интенсивности переходов процесса ν_t, сопровождающиеся генерацией заявки k-го типа. Аналогичный смысл имеют недиагональные элементы матрицы D_0, а диагональные элементы этой матрицы есть взятые с противоположным знаком интенсивности выхода процесса ν_t из соответствующих состояний. Матрица $D(1)$ является инфинитезимальным генератором управляющего процесса $\nu_t, t \geq 0$. Стационарное распределение этого процесса, представленное в виде вектор-строки $\boldsymbol{\theta}$, определяется как решение системы линейных алгебраических уравнений: $\boldsymbol{\theta} D(1) = \mathbf{0}$, $\boldsymbol{\theta}\mathbf{e} = 1$. Здесь и далее $\mathbf{e}$ – вектор-столбец, состоящий их единиц, $\mathbf{0}$ – нулевая вектор-строка. Интенсивность поступления запросов k-го типа задается формулой: $\lambda_k = \boldsymbol{\theta} D_k \mathbf{e}$, а суммарная интенсивность поступления запросов равна $\lambda = \lambda_1 + \lambda_2$. Дисперсия v_k длин интервалов между моментами поступления запросов k-го типа вычисляется по формуле $v^{(k)} = \frac{2\boldsymbol{\theta}(-D_0 - D_{\bar{k}})^{-1}\mathbf{e}}{\lambda_k} - \left(\frac{1}{\lambda_k}\right)^2$, $\bar{k} \neq k$, $k, \bar{k} = 1, 2$. Коэффициент корреляции $c_{cor}^{(k)}$ длин двух соседних интервалов между моментами поступления групп запросов k-го типа вычисляется по формуле $c_{cor}^{(k)} = \left[\frac{\boldsymbol{\theta}(D_0 + D_{\bar{k}})^{-1}}{\lambda_k} D_k (D_0 + D_{\bar{k}})^{-1}\mathbf{e} - \left(\frac{1}{\lambda_k}\right)^2\right](v^{(k)})^{-1}$, $\bar{k} \neq k$, $k, \bar{k} = 1, 2$. Более подробное описание $MMAP$ можно найти, например, в [11]

Время обслуживания любым прибором заявки k-го, $k = 1, 2$, типа имеет фазовое распределение (PH– Phase type distribution), которое задается парой $(\boldsymbol{\beta}_k, S_k)$. Здесь $\boldsymbol{\beta}_k$–вектор-строка порядка M_k, а S_k – квадратная матрица порядка M_k. Время обслуживания интерпретируется как время, за которое некоторая управляющая цепь Маркова $m_t^{(k)}$, $t \geq 0$, с пространством состояний $\{1, \ldots, M_k, M_k + 1\}$ достигнет единственного поглощающего состояния $M_k + 1$. Переходы цепи $m_t^{(k)}$, $t \geq 0$, в пространстве несущественных состояний $\{1, \ldots, M_k\}$ задаются субгенератором S_k, а интенсивности переходов в поглощающее состояние задаются вектором $\boldsymbol{S}_0^{(k)} = -S_k \mathbf{e}$. В момент начала обслуживания состояние процесса $m_t^{(k)}$, $t \geq 0$, выбирается из пространства состояний $\{1, \ldots, M_k\}$ на основании вероятностного вектора-строки $\boldsymbol{\beta}_k$. Интенсивности об-

служивания задаются как $\mu_k = -(\boldsymbol{\beta}_k S_k^{-1}\mathbf{e})^{-1}$. Более подробное описание PH-распределения можно найти в [14].

Предполагаем, что запросы первого типа обладают абсолютным приоритетом. Если неприоритетный запрос, поступающий в систему, застает все приборы занятыми, то он покидает систему навсегда. В подобной ситуации, если все приборы заняты приоритетными запросами, то поступающий приоритетный запрос также теряется. Если же хотя бы один прибор занят неприоритетным запросом, то поступающий приоритетный запрос вытесняет этот неприоритетный запрос (который теряется) и занимает его место на приборе.

Нашей целью является расчет стационарного распределения системы и ее характеристик производительности.

3. Цепь Маркова, описывающая функционирование системы

Пусть в момент времени t,

- n_t - число занятых приборов , $n_t = \overline{0, N}$;
- r_t - число приборов, занятых обслуживанием запросов 1-го типа, $r_t = \overline{0, n_t}$;
- ν_t - состояние управляющего процесса MAP, $\nu_t = \overline{0, W}$;
- $m_t^{(j,k)}$ - состояние управляющего процесса обслуживания на j-ом приборе, обслуживающем запрос k-го типа, $m_t^{(j,1)} = \overline{1, M_1}, j = \overline{1, r_t}$, $m_t^{(j,2)} = \overline{1, M_2}$, $j = \overline{1, n_t - r_t}$. Полагаем, что приборы, обслуживающие запросы 2-го типа, расположены после приборов, обслуживающих запросы 1-го типа. Кроме того, предполагаем, что приборы, обслуживающие запросы k-го типа, нумеруются в порядке их занятия, т.е. прибор, который начинает обслуживание, нумеруется максимальным числом среди всех приборов, занятых обслуживанием запросов этого типа. Когда прибор заканчивает работу, происходит перенумерация).

Тогда функционирование системы описывается цепью Маркова

$$\xi_t = \{n_t, r_t, \nu_t, m_t^{(1,1)}, m_t^{(2,1)}, \dots, m_t^{(r_t,1)}, m_t^{(1,2)}, m_t^{(2,2)}, \dots, m_t^{(n_t-r_t,2)}\}$$

с пространством состояний

$$\Omega = \{(n, r, \nu, m^{(j,1)}, m^{(l,2)}),\ n = \overline{0, N}, r = \overline{0, n}, m^{(j,1)} = \overline{1, M_1},$$
$$j = \overline{1, r}, m^{(l,1)} = \overline{1, M_2}, l = \overline{1, n - r}\}.$$

Упорядочим состояния цепи в лексикографическом порядке ее компонент. Обозначим через $Q_{i,l}$ матрицу интенсивностей переходов цепи из состояний, соответствующих значению i первой компоненты, в состояния, соответствующие значению l этой компоненты, $i, l = \overline{0, N}$.

Теорема 1. *Инфинитезимальный генератор Q цепи Маркова ξ_t, $t \geq 0$, имеет следующую блочную структуру:*

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & O & \dots & O & O \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & \dots & O & O \\ O & Q_{2,1} & Q_{2,2} & \dots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \dots & Q_{N-1,N-1} & Q_{N-1,N} \\ O & O & O & \dots & Q_{N,N-1} & Q_{N,N} \end{pmatrix},$$

где

$$Q_{n,n} = diag\{D_0 \oplus S_1^{\oplus r} \oplus S_2^{\oplus n-r},\ r = \overline{0,n}\},\ n = \overline{0,N-1},$$

$$Q_{N,N} = diag\{(D_0 + D_2)) \oplus S_1^{\oplus r} \oplus S_2^{\oplus N-r},\ r = \overline{0,N-1},\ D(1) \oplus S_1^{\oplus N}\} +$$
$$+ diag^{+}\{D_2 \otimes I_{M_1^r} \otimes \mathbf{e}_{M_2}\boldsymbol{\beta}_1 \otimes I_{M_2^{N-r-1}}, r = \overline{0,N-1}\},$$

$$Q_{n,n-1} = \left(\begin{array}{c} diag\{I_{\bar{W}} \otimes I_{M_1^r} \otimes (\boldsymbol{S}_0^{(2)})^{\oplus n-r},\ r = \overline{0,n-1}\} \\ \hline O_{\bar{W}M_1^n \times \bar{W}\sum_{r=0}^{n-2} M_1^r M_2^{n-r-1}} \mid I_{\bar{W}} \otimes (\boldsymbol{S}_0^{(1)})^{\oplus n} \end{array} \right) +$$
$$+ diag^{-}\{I_{\bar{W}} \otimes (\boldsymbol{S}_0^{(1)})^{\oplus r} \otimes I_{M_2^{n-r}}\ r = \overline{1,n-1}\}, n = \overline{1,N},$$

$$Q_{n,n+1} = \left(diag\{D_2 \otimes I_{M_1^r} \otimes I_{M_2^{n-r}}\boldsymbol{\beta}_2, r = \overline{0,n}\} \mid O_{\bar{W}\sum_{r=0}^{n} M_1^r M_2^{n-r} \times \bar{W} M_1^{n+1}} \right) +$$
$$+ \left(\begin{array}{c} O_{\bar{W}\sum_{r=0}^{n-1} M_1^r M_2^{n-r} \times \bar{W}\sum_{r=0}^{n+1} M_1^r M_2^{n-r+1}} \\ \hline O_{\bar{W}M_1^n \times \bar{W}\sum_{r=0}^{n} M_1^r M_2^{n-r+1}} \mid D_1 \otimes I_{M_1^n} \otimes \boldsymbol{\beta}_1 \end{array} \right) +$$
$$+ \left(diag^{+}\{D_1 \otimes I_{M_1^r} \otimes \boldsymbol{\beta}_1 \otimes I_{M_2^{n-r}},\ r = \overline{0,n-1}\} \mid O_{\bar{W}\sum_{r=0}^{n} M_1^r M_2^{n-r} \times \bar{W} M_1^{n+1}} \right),$$
$$n = \overline{1,N-1}.$$

Здесь использованы следующие обозначения: $\bar{W} = W + 1$; $A^{\oplus l} = \sum_{m=0}^{l-1} I_{n^m} \otimes A \otimes I_{n^{l-m-1}}$, $l \geq 1$, *для матрицы (или вектор-столбца) A, имеющей (имеющего) n строк, $\otimes(\oplus)$ – символ кронекерова произведения (суммы) матриц, см., например, [15].*

Доказательство леммы выполняется путем анализа вероятностей всех возможных переходов цепи Маркова $\xi_t, t \geq 0$, в течение интервала времени, имеющего бесконечно малую длину.

4. Стационарное распределение. Характеристики производительности

Пусть $\mathbf{p}$ является вектором-строкой стационарного распределения вероятностей состояний цепи. Этот вектор определяется как единственное решение системы линейных алгебраических уравнений

$$\mathbf{p}Q = \mathbf{0}, \ \mathbf{p}\mathbf{e} = 1.$$

В случае большой размерности данной системы для ее решения целесообразно использовать специальный алгоритм, предложенный в [16] и основанный на идее сенсорных цепей Маркова. В результате получим вектор $\mathbf{p} = (\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_N)$, где $\mathbf{p}_n$ – вектор- строка стационарных вероятностей, соответствующих значению n первой компоненты. Вычислив векторы $\mathbf{p}_n$, можно найти ряд важных стационарных характеристик производительности системы. Приведем некоторые из них.

- Распределение числа занятых приборов в системе $p_n = \mathbf{p}_n\mathbf{e}$, $n = \overline{0, N}$.
- Среднее число занятых приборов $N_{busy} = \sum\limits_{n=1}^{N} np_n$.
- Распределение числа приборов, занятых обслуживанием запросов 1 типа

$$q_r^{(1)} = \delta_{r,0}p_0 + \sum_{n=1}^{N} \mathbf{p}_n \begin{pmatrix} \mathbf{0}^T \\ \bar{W}\sum\limits_{l=0}^{r-1} M_1^l M_2^{n-l} \\ \mathbf{e}_{\bar{W}M_1^r M_2^{n-r}} \\ \mathbf{0}^T \\ \bar{W}\sum\limits_{l=r+1}^{n} M_1^l M_2^{n-l} \end{pmatrix}, \ r = \overline{0, N},$$

  где $\delta_{r,0}$ -символ Кронекера.
- Среднее число приборов, занятых обслуживанием запросов 1 типа $N_{busy}^{(1)} = \sum\limits_{r=1}^{N} rq_r^{(1)}$.
- Распределение числа приборов, занятых обслуживанием запросов 2 типа

$$q_m^{(2)} = \delta_{r,0}p_0 + \sum_{n=1}^{N} \mathbf{p}_n \begin{pmatrix} \mathbf{0}^T \\ \bar{W}\sum\limits_{l=0}^{n-m-1} M_1^l M_2^{n-l} \\ \mathbf{e}_{\bar{W}M_1^{n-m} M_2^m} \\ \mathbf{0}^T \\ \bar{W}\sum\limits_{l=n-m+1}^{n} M_1^l M_2^{n-l} \end{pmatrix}, \ m = \overline{0, N}.$$

- Среднее число приборов, занятых обслуживанием запросов 1 типа $N_{busy}^{(2)} = \sum_{m=1}^{N} m q_m^{(2)}$.
- Вероятность того, что приоритетный запрос будет потерян

$$P_{loss,1} = \frac{1}{\lambda_1}\mathbf{p}_N \begin{pmatrix} O_{\bar{W}\sum_{r=0}^{N-1} M_1^r M_2^{N-r}\times\bar{W}} \\ I_{\bar{W}} \otimes \mathbf{e}_{M_1^N} \end{pmatrix} D_1\mathbf{e}.$$

- Вероятность того, что неприоритетный запрос будет потерян вследствие занятости буфера

$$P_{loss,2}^{input} = \frac{1}{\lambda_2}\mathbf{p}_N \begin{pmatrix} I_{\bar{W}} \otimes \mathbf{e}_{M_1^0 M_2^N} \\ I_{\bar{W}} \otimes \mathbf{e}_{M_1^1 M_2^{N-1}} \\ \vdots \\ I_{\bar{W}} \otimes \mathbf{e}_{M_1^N M_2^0} \end{pmatrix} D_2\mathbf{e}.$$

- Вероятность того, что поступающий приоритетный запрос вытеснит с обслуживания неприоритетный запрос

$$P_{loss,2}^{serv} = \frac{1}{\lambda_1}\mathbf{p}_N \begin{pmatrix} I_{\bar{W}} \otimes \mathbf{e}_{M_1^0 M_2^N} \\ I_{\bar{W}} \otimes \mathbf{e}_{M_1^1 M_2^{N-1}} \\ \vdots \\ I_{\bar{W}} \otimes \mathbf{e}_{M_1^{N-1} M_2^1} \\ I_{\bar{W}} \otimes \mathbf{0}_{M_1^N M_2^0} \end{pmatrix} D_1\mathbf{e}.$$

5. Заключение

В статье исследована приоритетная многолинейная система массового обслуживания с коррелированным потоком запросов двух типов. Система анализируется на основе довольно общих предположений о процессах поступления и обслуживания. Предложены алгоритмы вычисления стационарного распределения состояний системы и основных характеристик производительности, включая вероятности потерь приоритетных и неприоритетных запросов.

ЛИТЕРАТУРА

1. Гнеденко Б. В. Даниелян Э. А., Димитров Б. Н., Климов Г. П., Матвеев В.Ф. Приоритетные системы обслуживания. МГУ, Москва, 1973.

2. Матвеев В. Ф., Ушаков В. Г. Системы массового обслуживанияю МГУ, Москва, 1984.
3. Miller R. Priority queues// Annals of Mathematical Statistics. 1960. V.31. P. 86–103.
4. Kleinrock L. Queueing systems. Volume II: Computer applications. John Wiley and Sons, New York, 1976.
5. Takagi H. Queueing analysis: a foundation of performance evaluation, volume 1:vacation and priority systems, part 1. North-Holland, 1991.
6. Lucantoni D. New results on the single server queue with a batch Markovian arrival process // Communication in Statistics-Stochastic Models. 1991. V.7. P. 1–46.
7. Bocharov P. P., D'Apice C., Pechinkin A. V., Salerno S. Queueing Theory. VSP, Utrecht, Boston, 2004.
8. Choi B. D., Hwang G. U. The $MAP, M/G_1, G_2/1$ queue with preemptive priority // Journal of Applied Mathematics and Stochastic Analysis. 1997. V. 10 (4). P. 407—421.
9. Machihara F. A bridge between preemptive and nonpreemptive queueing models// Performance Evaluation. 1995. V. 23(2). P. 93-–106.
10. Takine T., Sengupta B. A single server queue with service interruptions// Queueing Systems. 1997. V. 26 (3-4). P. 285-–300.
11. He Q. M. Queues with marked customers// Advances in Applied Probability. 1996. V. 28. P. 567-587.
12. Horvath G. Efficient analysis of the queue length moments of the MMAP/MAP/1 preemptive priority queue// Performance Evaluation. 2012. V. 69. P. 684–700.
13. He Q.-M., Xie J., Zhao X. Priority Queue with Customer Upgrades // Naval Research Logistics. 2012. V. 59. P. 362–375.
14. Neuts M. Matrix-geometric solutions in stochastic models. The Johns Hopkins University Press, Baltimore, 1981.
15. Graham A. Kronecker Products and Matrix Calculus with Applications. Ellis Horwood, Cichester, 1981.
16. Klimenok V.I., Kim C.S., Orlovsky D.S., Dudin A.N. Lack of invariant property of Erlang $BMAP/PH/N/0$ model // Queueing Systems. 2005. V. 49. P. 187–213.

UDC: 519.872

Research of demographic processes by methods of queuing theory

M.G. Nosova[1]

[1]Tomsk State University of Control Systems and Radioelectronics, 634050, pr. Lenina 40, Tomsk, Russia

nosovamgm@gmail.com

Abstract

The paper considers an autonomous non-Markov queuing system with two types of applications, the research of which is performed by the method of asymptotic analysis of the stochastic density of the numbers of applications. Its asymptotic normality is shown. The main characteristics determining this distribution are found. Such a queuing system sufficiently and adequately simulates the process of changing the age structure of the population and can be used to analyze demographic situations.

Keywords: queuing system, asymptotic analysis, mathematical model, population projections, population growth

1. Introduction

To develop an effective demographic policy in the modern economy, the analysis and forecasting of the processes of reproduction of the size and structure of the population are of particular relevance. Population projections are used in various fields of state and regional management, in marketing research and insurance.

In the modeling of demographic processes, deterministic models (discrete and continuous) and stochastic discrete are most common. However, demographic processes proceed in continuous time and are stochastic. Research methods for such processes in mathematical demography are not sufficiently developed [1, 2, 3]. That is why the urgent task is to significantly expand the mathematical models of the process of changing the demographic situation, as well as the development of research methods.

The article proposes to apply models and methods of queuing theory to analyze the processes of changing the demographic situation.

2. Mathematical model

In this paper, we consider a mathematical model of human population growth as an autonomous non-Markov queuing system with an unlimited number of servers and two type of applications (applications of the first type and applications of the second type) [1] . Define the process of servicing applications. Each application at the time of its receipt occupies a free device and is on it for the entire service time, the duration of which is random. Durations of servicing various requirements are stochastically independent, have the same distribution determined by the function $S_i(x)$=1-$B_i(x)$, where $B_1(x)$ and $B_2(x)$ are the distribution functions of the service time of applications of the first and second type, respectively. After completing the service, the application leaves the system.

For applications in the system, we define the age $x \geq 0$ as the length of the interval from the time t-x of the beginning of its service (the moment of entry into the system) to the current time t. Each application of the first type of age x at time t with intensity $b(x,t)$ generates a new requirement, that is, the probability that the application of the first type of age x from time t for an infinitely small time interval of duration Δt will generate a new requirement is $b(x,t)\Delta t + o(\Delta t)$, and the probability of generating two or more requirements is an infinitely small quantity of a higher order than Δt. A new application (of the first or second type) at the time of its appearance takes a free device and begins the process of its maintenance, generating the requirements of a new generation.

In terms of demography, the served application is interpreted as a female or male person, the application service time is the person's life expectancy, $S_1(x)$ and $S_2(x)$ are the survival function for women and men, respectively, the application age x is the person's age at the considered time t, function $b(x,t)$ is the birth rate of women of age x in year t (fertility function). We assume that with a probability r a girl is born and with a probability of $(1\text{-}r)$ a boy. The incoming flow of applications is the process of the birth of children, that is, the sequence of moments of the birth of children from the entire population of women.

To determine the stochastic density, we denote $N_1(x_1,x_2,t)$ and $N_2(x_1,x_2,t)$ – the number of applications of the first and second type with the age $x \in [x_1, x_2)$, served in this system at time t. Here x is any nonnegative real number. A limit

$$\lim_{\Delta x \to 0} \frac{1}{\Delta x} N_1(x, x + \Delta x, t) = \xi(x,t),$$

$$\lim_{\Delta x \to 0} \frac{1}{\Delta x} N_2(x, x + \Delta x, t) = \eta(x,t).$$

will be called the stochastic densities $\xi(x,t)$ and $\eta(x,t)$ of the number of applications (female and male population) at age x at time t. In this paper, the problem of finding

the joint probability distribution of the values of the number of applications of ages x serviced in the system at time t is solved.

3. Autonomous system with PH-distriburion of service time

From the study of the autonomous queuing system, we turn to the consideration of a system whose structure coincides with the original, and the service time τ of each application is composed of the durations of a random number of phases

$$\tau = \tau_1 + \tau_2 + ... + \tau_\nu,$$

where τ_i – the duration of the i^{th} service phase. The values τ_i are independent exponentially distributed random variables with the parameter μ_1 for applications of the first type and with parameter μ_2 for applications of the second type, where i=1,2... v. Here v is a random variable and v=1,2.... This system is called an autonomous system with phase distribution or PH distribution of service time.

Let us explain the service process using the example of applications of the first type. Service for each new application of the first type begins in the first phase. The application, having completed the service at the i^{th} phase, with probability q_i proceeds to the service at the $i + 1^{th}$ phase, and with the probability 1-q_i completes the full service and leaves the system.

We determine the transition probability of application to the next phase as

$$q_i = S_1\left(\frac{i}{\mu_1}\right) / S_1\left(\frac{i-1}{\mu_1}\right), \tag{1}$$

and from (1) it is easy to show that

$$\lim_{\mu_1 \to \infty} Me^{-\alpha\tau} = \int_0^\infty e^{-\alpha x} dB_1(x).$$

Thus, for $\mu_1 \to \infty$, the duration of the service with the PH-distribution, by virtue of the choice of (1) probability values q_i, converges to the service time determined by the distribution function $B_1(x)$=1-$S_1(x)$ of the original autonomous queuing system.

Assuming that the application of the second type, having completed the service in the i^{th} phase, with probability s_i goes to the service in the $i + 1^{st}$ phase, and with the probability 1- s_i it completes the full service and leaves the system, it is not difficult to carry out similar reasoning.

The method of approximating the application service time by the sum of a random number of independent and equally exponentially distributed random variables and the limit transition with an unlimited increase in the number of phases and a

proportional decrease in the duration of each phase was called in [1] the virtual phase method.

We will assume that each application of the first type served at the i^{th} phase at time t with intensity $b_i(t)=b(i/\mu_1,t)$ generates a new application. We denote $n(i,t)$ – the number of applications of the first type and $l(i,t)$ – the number of applications of the second type, served at the appropriate i^{th} phase at time t . Then a random process

$$\overline{n}(t) = \{n(1,t), n(2,t), \ldots, \ l(1,t), l(2,t), \ldots\}^T$$

is multidimensional the continuous-time Markov chain. Its probability distribution is

$$P(n_1, n_2, ...l_1, l_2, \ldots, t) = P\{n_1(t) = n_1, ..., l_1(t) = l_1, l_2(t) = l_2, \ldots\}.$$

We write the system of Kolmogorov differential equations

$$\begin{aligned}
\frac{\partial}{\partial t}\{P(n_1, n_2, \ldots l_1, l_2, \ldots, t)\} = -P(n_1, n_2, ...l_1, l_2, \ldots, t)\left\{\sum_{i=1}^{\infty} n_i(\mu_1 + b_i(t)) + l_i\mu_2\right\} + \\
+P(n_1 - 1, n_2, ..., l_1, l_2, \ldots, t) r\{(n_1 - 1)b_1(t) + \sum_{i=2}^{\infty} n_i\, b_i(t)\} + \\
+P(n_1, n_2, \ldots, l_1 - 1, l_2, \ldots, t)(1 - r)\sum_{i=2}^{\infty} n_i b_i(t) + \\
+\mu_1 \sum_{i=1}^{\infty}(n_i + 1)\{P(n_1, n_2, \ldots, n_i + 1, n_{i+1}, \ldots, t)(1 - q_i) + \\
+P(n_1, n_2, \ldots, n_i + 1, n_{i+1} - 1, n_{i+2} \ldots, t) q_i\} + \\
+\mu_2 \sum_{i=1}^{\infty}(l_i + 1)\{P(n_1, \ldots, l_i + 1, l_{i+1}, \ldots, t)(1 - s_i) + \\
+P(n_1, \ldots, l_i + 1, l_{i+1} - 1, l_{i+2} \ldots, t) s_i\}.
\end{aligned} \tag{2}$$

We denote a characteristic function of the number of occupied servers at time t in the form

$$H(y,t) = \sum_{n_1, n_2, \ldots, l_1, l_2, \ldots} P(n_1, n_2, \ldots, l_1, l_2, \ldots, t)\, exp\left\{j\sum_{i=1}^{\infty}(u_i n_i(t) + z_i l_i(t))\right\},$$

where $j = \sqrt{-1}$ – the imaginary unit.

We multiply (2) by $exp\left(j\sum_{i=1}^{\infty}(u_i n_i(t) + z_i l_i(t))\right)$, sum and easily obtain the equation for $H(y,t)$. The solution $H(y,t)$ determines the problem of researching the autonomous system with the PH-distribution of the service time.

4. The basic equation for an autonomous non-Markov system

We use the presentation

$$u_i = u(i/\mu_1),\ n_i(t) = n(i/\mu_1, t), z_i = z(i/\mu_2),\ l_i(t) = l(i/\mu_2, t),$$

and the characteristic function $H(y,t)$ is written as

$$H(y,t) = M\left\{\exp\left[j\sum_{i=1}^{\infty}\left(\frac{1}{\mu_1}u\left(\frac{i}{\mu_1}\right)\mu_1 n\left(\frac{i}{\mu_1},t\right)+\frac{1}{\mu_2}z\left(\frac{i}{\mu_2}\right)\mu_2 l\left(\frac{i}{\mu_2},t\right)\right)\right]\right\}.$$

Let an expression $i/\mu_1 \to x$ and $i/\mu_2 \to x$ for $\mu_1 \to \infty$, $\mu_2 \to \infty$, $i \to \infty$, then we assume that the following limits exist

$$\lim_{i/\mu_1 \to x} u(i/\mu_1) = u(x), \quad \lim_{i/\mu_1 \to x} \mu_1 n(i/\mu_1, t) = \zeta(x,t),$$

$$\lim_{i/\mu_2 \to x} z(i/\mu_2) = z(x), \quad \lim_{i/\mu_2 \to x} \mu_2 l(i/\mu_2, t) = \eta(x,t),$$

$$\lim_{\substack{i/\mu_1 \to x \\ i/\mu_2 \to x}} H(u,z,t) = M\left\{\exp\left[j\int_0^{\infty}(u(x)\zeta(x,t)+z(x)\eta(x,t))dx\right]\right\} = F(y,t).$$

The function $F(y,t)$ is called the characteristic functional of the random functions $\xi(x,t)$ and $\eta(x,t)$ of two arguments x and t. The random function $\xi(x,t)$ is called the stochastic density of the number of applications of the first type of age x served in the system at time t, for all $x \geq 0$, and the random function $\eta(x,t)$ is called the stochastic density of the number of applications of the second type of age x served in the system at time t, for all $x \geq 0$.

With this in mind and for $i/\mu_1 \to x$, $i/\mu_2 \to x$, $\mu_1 \to \infty$, $\mu_2 \to \infty$, we easily rewrite equation for $H(y,t)$ in the form

$$\frac{\partial F(y,t)}{\partial t} = j\int_0^{\infty}\frac{\partial F(y,t)}{\partial u(x)}\left\{\left(1-re^{ju(0)}\right)b(x,t)-(1-r)e^{jz(0)}b(x,t)+\right.$$

$$\left.+\left(e^{-ju(x)}-1\right)\frac{S_1'(x)}{S_1(x)}-ju'(x)\right\}+ \qquad (3)$$

$$+j\int_0^{\infty}\frac{\partial F(y,t)}{\partial z(x)}\left\{\left(e^{-jz(x)}-1\right)\frac{S_2'(x)}{S_2(x)}-jz'(x)\right\}dx.$$

The equation (3) is called the basic equation for the considered non-Markov queuing system.

We will solve the equation (3) by the modified method of asymptotic analysis [1], assuming that the random functions $\xi(x,t)$ and $\eta(x,t)$ take sufficiently large values proportional to some infinitely large value of N, that is $N \to \infty$.

5. Solution of the equation by the method of asymptotic analysis

The method of asymptotic analysis is realized by a sequence of asymptotics of increasing order [4] . For the equation (3), we find the asymptotics of the first and second orders of its solution.

5.1. First-order asymptotics. Denote ε=1/N, where ε – a small positive parameter, in the equation (3) we replace

$$u(x) = \varepsilon w_1(x), z(x) = \varepsilon w_2(x), F(y,t) = F_1(w_1, w_2, t, \varepsilon),$$

and we find the solution of the resulting equation in the form of a characteristic functional

$$F_1(w_1, w_2, t) = exp\left\{ j \int_0^\infty (w_1(x)\, g(x,t) + w_2(x)\, m(x,t)) dx \right\},$$

that determines the average values $g(x,t)$ and $m(x,t)$ of the random functions $\xi(x,t)$ and $\eta(x,t)$.

It can be shown that the functions $g(x,t)$ and $m(x,t)$ are determined by a system

$$\frac{\partial g(x,t)}{\partial t} + \frac{\partial g(x,t)}{\partial x} = g(x,t)\frac{S_1'(x)}{S_1(x)}, \tag{4}$$

$$\frac{\partial m(x,t)}{\partial t} + \frac{\partial m(x,t)}{\partial x} = m(x,t)\frac{S_2'(x)}{S_2(x)}. \tag{5}$$

5.2. Second-order asymptotics. We note that the first-order asymptotics determines only the average values of the stochastic densities $\xi(x,t)$ and $\eta(x,t)$. Therefore, naturally, the need arises to find second-order asymptotics, which allows one to obtain more detailed characteristics.

In order to find the second-order asymptotics, in the basic equation (3) we replace

$$F(y,t) = H_2(\mathrm{y},t)\, exp\left\{ jN \int_0^\infty (\mathrm{u}(x)\, g(x,t) + \mathrm{z}(x)\, m(x,t))\, dx \right\}. \tag{6}$$

From a prior designation and (6) follow, that $H_2(y,t)$ is the characteristic functional for the quantity ($\xi(x,t)+\eta(x,t)$) –$N(g(x,t)+m(x,t))$. The mathematical expectation of {($\xi(x,t)+\eta(x,t)$) – $N(g(x,t)+m(x,t)$} equal to zero.

In the obtained equation we denote ε^2=1/N and make replacements

$$u(x) = \varepsilon w_1(x), z(x) = \varepsilon w_2(x), H_2(y,t) = F_2(w_1, w_2, t, \varepsilon),$$

obtain some equality and we find the solution of this equation in the form of a characteristic functional of the Gaussian distribution

$$F_2(w_1,w_2,t)=exp\Bigg\{-\frac{1}{2}\Bigg[\iint_0^\infty w_1(y)\,w_1(z)\,R_{11}(y,z,t)\,dydz+$$

$$+2\iint_0^\infty w_1(y)\,w_2(z)\,R_{12}(y,z,t)\,dydz+\iint_0^\infty w_2(y)\,w_2(z)\,R_{22}(y,z,t)\,dydz\Bigg]\Bigg\},$$

were $R_{11}(y,z,t)$, $R_{12}(y,z,t)$ and $R_{22}(y,z,t)$ are cross-correlation functions of the normalized numbers of applications of ages y and z, serviced in the system at time t, can be found in the form

$$R_{11}(y,z,t)=\sigma_1^2(t)\,\delta(y)\,\delta(z)+r_1(y,t)\,\delta(z)+r_1(z,t)\,\delta(y)+\sigma_1(y,t)\,\sigma_1(z,t)\,\delta(y-z),$$

$$R_{22}(y,z,t)=\sigma_2^2(t)\,\delta(y)\,\delta(z)+r_2(y,t)\,\delta(z)+r_2(z,t)\,\delta(y)+\sigma_2(y,t)\,\sigma_2(z,t)\,\delta(y-z),$$

$$R_{12}(y,z,t)=\sigma_{12}^2(t)\,\delta(y)\,\delta(z)+r_{12}(y,t)\,\delta(z)+r_{12}(z,t)\,\delta(y)+\sigma_{12}(y,t)\,\sigma_{12}(z,t)\,\delta(y-z),$$

where $\delta(.)$ – δ-Dirac function, and the summands of the cross-correlation functions are found from the system of partial differential equations

$$\frac{\partial\sigma_1^2(t)}{\partial t}=2S_1'(0)\,\sigma_1^2(t)+2r\int_0^\infty r_1(x,t)b(x,t)\,dx+rg(0,t),$$

$$\frac{\partial\sigma_2^2(t)}{\partial t}=2S_2'(0)\,\sigma_2^2(t)+2(1-r)\int_0^\infty r_{12}(x,t)b(x,t)\,dx+(1-r)g(0,t),$$

$$\frac{\partial\sigma_1^2(x,t)}{\partial t}+\frac{\partial\sigma_1^2(x,t)}{\partial x}=\left(2\sigma_1^2(x,t)-g(x,t)\right)\frac{S_1'(x)}{S_1(x)},$$

$$\frac{\partial\sigma_2^2(x,t)}{\partial t}+\frac{\partial\sigma_2^2(x,t)}{\partial x}=\left(2\sigma_2^2(x,t)-m(x,t)\right)\frac{S_2'(x)}{S_2(x)},$$

$$\frac{\partial r_1(x,t)}{\partial t}+\frac{\partial r_1(x,t)}{\partial x}=r_1(x,t)\left(S_1'(0)+\frac{S_1'(x)}{S_1(x)}\right)+r\sigma_1^2(x,t)\,b(x,t),$$

$$\frac{\partial r_2(x,t)}{\partial t}+\frac{\partial r_2(x,t)}{\partial x}=r_2(x,t)\left(S_2'(0)+\frac{S_2'(x)}{S_2(x)}\right)+(1-r)\sigma_{12}^2(x,t)\,b(x,t),\quad(7)$$

$$\frac{\partial\sigma_{12}^2(t)}{\partial t}=(S_1'(0)+S_2'(0))\sigma_{12}^2(t)+(1-r)\int_0^\infty r_1(x,t)b(x,t)\,dx+r\int_0^\infty r_{12}(x,t)b(x,t)\,dx,$$

$$\frac{\partial\sigma_{12}^2(x,t)}{\partial t}+\frac{\partial\sigma_{12}^2(x,t)}{\partial x}=\sigma_{12}^2(x,t)\left(\frac{S_1'(x)}{S_1(x)}+\frac{S_2/(x)}{S_2(x)}\right),$$

$$\frac{\partial r_{12}(x,t)}{\partial t}+\frac{\partial r_{12}(x,t)}{\partial x}=r_{12}(x,t)\left(S_2'(0)+\frac{S_1'(x)}{S_1(x)}\right)+(1-r)\sigma_1^2(x,t)\,b(x,t),$$

$$\frac{\partial r_{12}(x,t)}{\partial t}+\frac{\partial r_{12}(x,t)}{\partial x}=r_{12}(x,t)\left(S_1'(0)+\frac{S_2'(x)}{S_2(x)}\right)+r\sigma_{12}^2(x,t)\,b(x,t).$$

The solutions of the systems (4), (5) and (7) completely determine the parameters of the Gaussian distribution, which is satisfied by the numbers of age groups of applications serviced in the system at time t, that is, they completely solve the task of researching an autonomous non-Markov queuing system with two types of applications.

6. Conclusion

The article proposes a stochastic model of demographic growth in the form of an autonomous non-Markov queuing system with an unlimited number of devices and two types of applications. Its research was carried out using the virtual phase method and the modified method of asymptotic analysis. We were found the main probabilistic characteristics of the number of served applications in the system and was proofed that their asymptotic distribution is Gaussian. To apply the obtained results to the study of demographic processes, it is necessary to choose the explicit form of unknown functions (survival function and a fertility function) and set the parameter values.

Created mathematical model has quite wide possibilities for generalization and modification. This mathematical model of human population growth can be applied to predict the demographic situation in any country and in the world as a whole, including for forecasting the population size taking into account the age structure, marital structure and social status, for forecasting migration processes.

REFERENCES

1. Nosova M. G. A mathematical model of population growth as a queuing system, arXiv preprint arXiv:2005.10518, 21 May 2020.
2. Newelli C. Methods and models in demography, Belhaven, 1988.
3. Staroverov O. V. Models of Population Movement. Nauka, Moscow 1979.
4. Nazarov A. A., Terpugov A. F. Queuing theory. Publishing house of NTL, Tomsk, 2004.

UDC: 004.94

Reliability Analysis of Finite-Source Retrial Queueing Systems With Two-Way Communications to the Orbit and Blocking Using Simulation

János Sztrik[1], Ádám Tóth [1], Ákos Pintér[1], Zoltán Bács[1]

[1]University of Debrecen, Debrecen 4032, Hungary

{toth.adam,sztrik.janos}@inf.unideb.hu, bacs.zoltan@econ.unideb.hu,
apinter@science.unideb.hu

Abstract

A two-way communication, retrial queueing system is considered with a single server which from time to time is subject to random breakdowns. The investigated model is a M/M/1//N type of system where the number of sources is finite. After the service unit becomes idle it is able to call in customers residing in the orbit (outgoing call or secondary customers). Distribution of the service time of primary and secondary customers is exponential with rates μ_1 and μ_2, respectively. Every used random variable is assumed to be totally independent of each other in the model. Each time the server becoming in faulty state the operation of the system is blocked resulting that throughout this period customers can not enter the system. The novelty of this analysis is to study the effect of blocking in such system on the main performance measures using different distributions of failure time. Results are illustrated graphically with the help of a simulation program developed by the authors.

Keywords: simulation, blocking, sensitivity analysis, finite-source queueing system, unreliable server, retrial queue

1. Introduction

Because of the increasing number of users and devices mainly due to the rapid development of technology it is not an easy task to cope with the question of designing communication systems or redesigning an existing pattern or scheme. Nowadays, every company possesses some kind of network infrastructure so it is unavoidable that the exchange of information would not take place therefore developing mathematical and simulation models and algorithms play quite an important role to deal with

The research was financed by the Higher Education Institutional Excellence Programme of the Ministry of Human Capacities in Hungary, within the framework of the NKFIH-1150-6/2019 thematic programme of the University of Debrecen.

traffic growth. Applying retrial queues in such scenarios are useful and powerful tools to describe real-life problems emerging from main telecommunication systems like telephone switching systems, call centers, computer networks, and computer systems. Many researchers are dedicated to investigating this topic, some examples are mentioned which study retrial queueing systems with repeated calls like in [1],[2]. The applicability of these models is utilized in many areas of science like improving the efficiency of systems for example in the case of local-area networks with random access protocols and with multiple access protocols [3],[4].

The characteristics of two-way communication have a beneficial effect on most of the systems consequently its popularity is quite well-founded in recent years. This can be explainable by the fact that the operation of certain real-life systems can be matchable with models based on a two-way communication scheme. In terms of call-centers, this is especially appropriate considering that the service unit (or agent) apart from handling incoming calls may carry out other activities including selling, promoting, and advertising products. In this paper whenever the server gets to idle state after some random time it is capable of calling customers residing in the orbit. In such scenes, the utilization of the service unit (or workload of agents) is crucial and extensively examined by many papers like [5],[6].

Scrutinizing the available literature on the internet relatively quite a high number of papers are found where the service facilities are presumed to be available all the time. Reliable operation is quite optimistic and an unrealistic approach because deterioration, power supply failure, or unforeseen circumstances can happen anytime modifying moderately the system characteristics. Regarding wireless communication, several components affect the transmission rate resulting in interruptions that can arise at any time throughout transmitting the packets. It is always a key question of how the property of unreliable operation alters the performance measures and the characteristics of the system. Recently published works about retrial queuing systems with a non-reliable server can be found for example in [7],[8],[9].

The main aim of this work is to explore the mechanism of blocking of the investigated system and to compare various distributions of failure time on main performance measures like the mean waiting time of an arbitrary customer or the total utilization of the server. The present paper is a natural continuation of [10] and we want to compare the achieved results with each other. Our self-developed simulation program is used to obtain every important performance measure using SimPack [11], which contains C/C++ libraries and executable programs for computer simulation. In this class, numerous algorithms can be found in connection with discrete-event, continuous,

and combined (multi-model) simulation. Because of using other distributions apart from exponential and the fact that providing exact formulas is almost impossible we selected stochastic simulation to approximate the desired performance measures and to freely integrate any distribution in our code. The novelty of this paper is to present a sensitivity analysis of failure time on the main measures besides blocking using various distributions. Graphical illustrations are provided depicting an interesting phenomenon of sensitivity problems and comparison with the non-blocking system.

2. Model description and notations

We considered a finite-source queueing system with the help of two-way communication with retrials which contains a non-reliable server. The source contains N customers and each of them produces requests (primary or ingoing customers) with rate λ/N resulting exponentially distributed inter-arrival time with parameter λ/N. Our model does not comprise queues thus in case of an idle server the service of an incoming customer starts immediately. The distribution of the service time of these customers is exponentially distributed with parameter μ_1. After being successfully served the customers return to the source. Alternatively, arriving customers from the orbit or source finding the server in a busy state are forwarded instantly to the orbit. Waiting an exponentially distributed time with parameter γ/N in this virtual waiting room customers launches another attempt to occupy the service unit. From time to time failure of the server may arise according to gamma, hypo-exponentially, hyper-exponentially, Pareto, and lognormal distribution with different parameters but with the same mean value. During this period customers can not enter the system because they are rejected in that instant, this is the so-called blocking. The recovery process begins instantaneously upon the failure of the server, which is also an exponentially distributed random variable with parameter γ_2. If the service unit breaks down during the service of a customer then that customer is transferred to the orbit immediately. Whenever the server becomes idle it may perform an outgoing call (secondary customers) towards the customers located in the orbit after an exponentially distributed random time with rate ν. The service of these customers is executed according to an exponential distribution with a rate of μ_2. Rates λ/N and σ/N are used because in [12],[13] very similar systems are evaluated by an asymptotic method where N tends to infinity and was proved that the number of customers in the system follows a normal distribution. All the random variables in the model creation are assumed to be totally independent of each other.

3. Simulation results

Our self-written simulation program includes a statistic package that was developed by Andrea Francini in 1994 [14]. Basically, this statistical analysis tool

is suitable to make a quantitative estimation of the mean and variance values of the desired variables using the method of batch means. In each batch, there are n observations and the useful run is divided into numerous batches. The batches should be long enough and approximately independent in order that the estimation would work correctly. This method belongs to one of the most popular confidence interval techniques for a steady-state mean of a process. In more detailed information about this method is included in the following works [15],[16]. The simulations are performed with a confidence level of 99.9%. The relative half-width of the confidence interval required to stop the simulation run is 0.00001.

To realize the sensitivity analysis four different distributions of failure time are selected to compare the performance measures with each other. The parameters are chosen in such a way that the mean value and variance would be equal, so we applied a fitting process that is necessary to be done. [17] contains a detailed description of the whole process characterizing every used distribution. We differentiated two main scenarios from each other. In the first one, the squared coefficient of variation is greater than one so I utilized hyper-exponential, gamma, Pareto, and lognormal distributions. Table 2 quantifies all the used input parameters of the various distributions of failure time while Table 1 shows the values of other parameters. Results in connection with the squared coefficient of variation are less than one was also investigated and will be published in the extended version of the paper because it is less interesting.

Table 1. Used numerical values of model parameters

$\mathbf{N}$	$\boldsymbol{\lambda/N}$	$\boldsymbol{\gamma_2}$	$\boldsymbol{\sigma/N}$	$\boldsymbol{\mu}$	$\boldsymbol{\mu_2}$	$\boldsymbol{\nu}$
100	0.01	1	0.01	1	1.2	0.02

The steady-state distributions are represented on Figure 1 when λ/N is $= 0.01$ comparing the effect of all four applied distributions of failure time. It shows the probability that exactly i customers are located in the system. Averagely the same number of customers resides in the system, slight differences can be perceivable especially in the case of Pareto. Taking a closer look at the graphs all the curves correspond to normal distribution despite the characteristics of the various distribution.

In Figure 2 the mean arbitrary response time is demonstrated as the request generation increases. Results clearly illustrate the effect of various distributions which is quite significant even though the first two moments are equal. The highest values are experienced at Pareto distribution while the lowest values at gamma distribution.

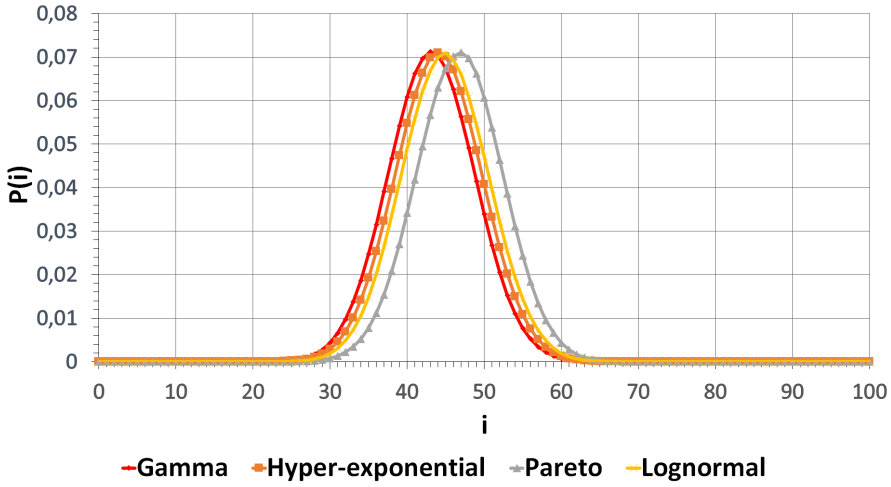


Fig. 1. Comparison of steady-state distributions when $\lambda/N = 0.01$

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			

The maximum property characteristic of a finite-source retrial queueing system arises which under suitable parameter setting occurs in spite of increasing arrival intensity.

Figure 3 shows how the total utilization of the server escalates applying intensifying arrival intensity. Under total utilization, we mean every single service including the service of primary, secondary customers, and the interrupted ones, too. By examining closely the figure the received values are almost identical but the tendency is counteractive as we have seen in Figure 2. As more and more customers enter the system the total utilization of the service unit increases.

Figure 4 emphasizes the effect of blocking on the mean waiting time versus arrival intensity. It is observable that in case of blocking the customers spend less time on average because during server failure the incoming customers go back to the source instead of waiting in the orbit. Besides the higher failure rate, the difference is more significant as well. At Figure 4 the distribution of service time of the incoming

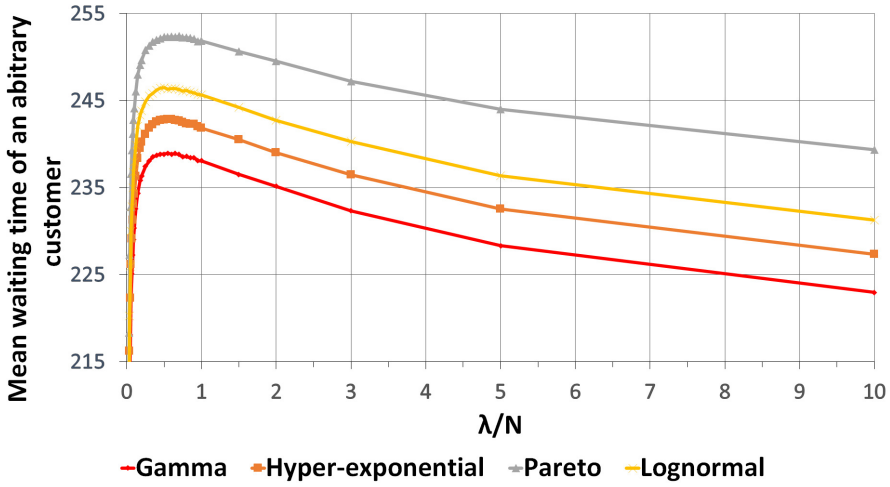


Fig. 2. Mean waiting time vs. arrival intensity

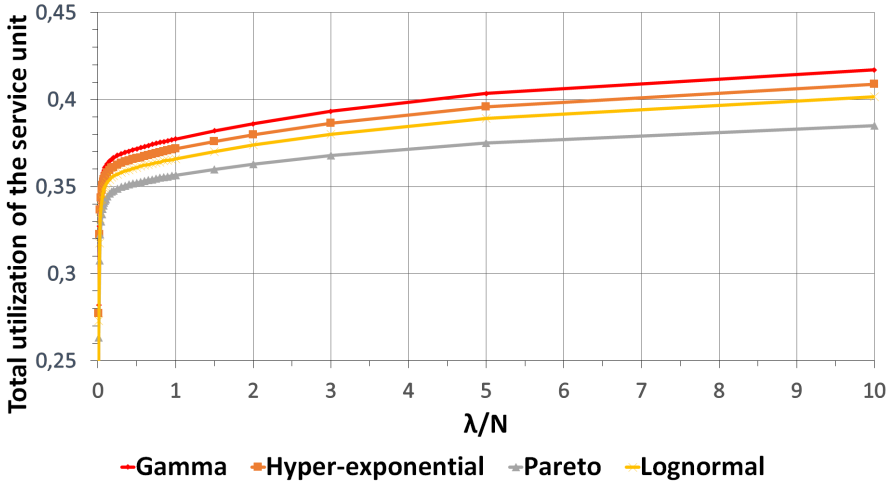


Fig. 3. Total utilization of the server vs. arrival intensity

customer is gamma, but the same tendency can be found in the case of the other distributions, too.

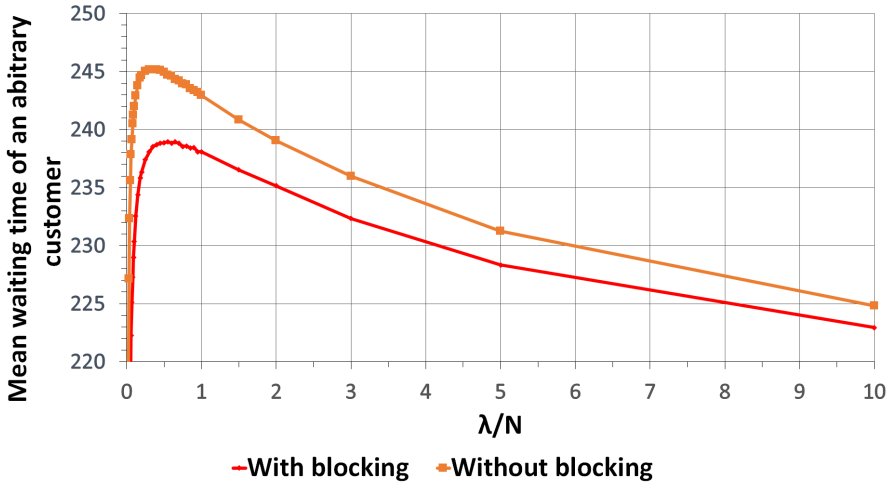


Fig. 4. The effect of blocking on the mean waiting time

4. Conclusion

A finite-source retrial queueing system with the help of two-way communication is introduced with applying blocking and an unreliable server which can make outgoing calls towards the customers of the orbit. The effect of the used distributions and blocking is illustrated by several figures on the mean arbitrary waiting time and the total utilization of the server. With the aid of stochastic simulation, the obtained results clearly revealed that in case the squared coefficient of variation is greater than one the disparity among the values of displayed performance measures is significant having the same mean and variance. In the future we would like to complete this system with other features like experimenting with more distributions, introducing some kind of impatience of the customers, or including more capacity of service.

REFERENCES

1. G. Falin, J. Artalejo, A finite source retrial queue, European Journal of Operational Research 108 (1998) 409–424.
2. D. Fiems, T. Phung-Duc, Light-traffic analysis of random access systems without collisions, Annals of Operations Research (2017) 1–17.
3. J. Artalejo, A. G. Corral, Retrial Queueing Systems: A Computational Approach, Springer, 2008.
4. J. Kim, B. Kim, A survey of retrial queueing systems, Annals of Operations Research 247 (1) (2016) 3–36.

5. 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.
6. 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.
7. N. Gharbi, C. Dutheillet, An algorithmic approach for analysis of finite-source retrial systems with unreliable servers, Computers & Mathematics with Applications 62 (6) (2011) 2535–2546.
8. 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.
9. A. Krishnamoorthy, P. K. Pramod, S. R. Chakravarthy, Queues with interruptions: a survey, TOP 22 (1) (2014) 290–320.
10. J. Sztrik, Á. Tóth, Á. Pintér, Z. Bács, Simulation of finite-source retrial queues with two-way communications to the orbit, in: A. Dudin, A. Nazarov, A. Moiseev (Eds.), Information Technologies and Mathematical Modelling. Queueing Theory and Applications, Springer International Publishing, Cham, 2019, pp. 270–284.
11. P. A. Fishwick, Simpack: Getting started with simulation programming in c and c++, in: In 1992 Winter Simulation Conference, 1992, pp. 154–162.
12. A. Nazarov, J. Sztrik, A. Kvach, A survey of recent results in finite-source retrial queues with collisions, in: Information Technologies and Mathematical Modelling. Queueing Theory and Applications, Springer, 2018, pp. 1–15.
13. A. Nazarov, J. Sztrik, A. Kvach, T. Bérczes, Asymptotic analysis of finite-source M/M/1 retrial queueing system with collisions and server subject to breakdowns and repairs, Annals of Operations Research 277 (2) (2019) 213–229.
14. A. Francini, F. Neri, A comparison of methodologies for the stationary analysis of data gathered in the simulation of telecommunication networks, in: Proceedings of MASCOTS '96 - 4th International Workshop on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, 1996, pp. 116–122.
15. 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.
16. A. M. Law, W. D. Kelton, Simulation Modeling and Analysis, McGraw-Hill Education, 1991.
17. A. Toth, J. Sztrik, A. Kuki, T. Berczes, D. Efrosinin, 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.

UDC: 519.218

Sensitivity Analysis of Characteristics of a k-out-of-n:F System to Shapes of Life and Repair Times Distributions of Its Components

Rykov V.V.[1,2], Ivanova N.M.[1,3], Kozyrev D.V.[1,3]

[1]Department of Applied Probability and Informatics, 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]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, kozyrev-dv@rudn.ru

Abstract

The problem of sensitivity of a redundant system's probability characteristics to shapes of the input distributions is considered. In some previous works, closed-form representations have been found for stationary characteristics of hot redundant systems with exponential lifetime distribution of their elements and general distribution of their repair time. In the current paper we carry out the sensitivity analysis of a k-out-of$-n$:F system with the help of simulation approach. Comparison of analytic and simulation results is presented.

Keywords: k-out-of-n:F system, steady state probabilities, sensitivity analysis, mathematical modeling and simulation, AnyLogic Environment

1. Introduction

Any system that can be considered as a functioning unit has many useful properties, among which stability, reliability, and flexibility can be especially highlighted. These properties are of key importance both from practical and research points of view.

The term "sensitivity analysis" can be understood differently in civil engineering than in basic sciences [1]. In operations research, sensitivity analysis is developed as a method of critical assessment of decisional variables, and is capable to identify

The publication has been prepared with the support of the "RUDN University Program 5-100" (problem setting and simulation model development) and funded by RFBR according to the research projects No. 20-01-00575 (recipient Vladimir Rykov, review and numerical results) and No. 19-29-06043 (recipient Dmitry Kozyrev, formal analysis, validation).

those sensitive variables that influence the final desired result [2]. Another suitable complement to probabilistic reliability analysis is structural sensitivity analysis [3, 4]. In stochastic systems stability often means insensitivity or low sensitivity of their output characteristics to the shapes of some input distributions [5].

The first results of studies on the insensitivity of systems' characteristics to output parameters dealt with systems with Poisson arrivals and fixed mean value of the service time. Thus, in 1957, Sevastyanov in [6] proved the insensitivity of Erlang formulas to the shape of service time distribution for systems with losses. Kovalenko continued the study of the insensitivity of the stationary characteristics of a renewable system. In [7] he found a necessary and sufficient condition for the reliability characteristics of such a system in the case of an exponential distribution of the lifetime and the general distribution of the recovery time. The sufficiency of this condition for the case of general lifetime and repair time distributions has been found by Rykov [8] with the help of multi-dimensional alternating processes theory.

The insensitivity of the characteristics of systems has also been the subject of several recent studies that investigated the problem of the stability of stationary characteristics of systems in the case when one of the input distributions (lifetime or repair time) is exponential [9, 10].

In the field of reliability and stability research, the k-out-of-n:F systems are very popular. These systems consist of n components, which fail, when at least k of them fail [11]. The k-out-of-n systems are widely used in practice, for example, in telecommunications, the oil and gas industry, data transmission, production management [12]. Therefore, it seems very important to investigate various characteristics of such models.

Carrying on the research in the reliability field, the current paper considers a k-out-of-n system using the so called markovization method, which consists in introduction of supplementary variables [13] that allows to describe the system's behavior by a two-dimensional Markov process. This approach allows to find analytical formulas for the system steady state probabilities (s.s.p.) that will be used in this paper.

The main idea of the paper consists in investigation of sensitivity properties of the system's s.s.p. in the case when components' life- and repair times are generally distributed with the help of the simulation method. Simulation results are compared with each other and with analytical results for the special case of exponentially distributed lifetimes of components.

2. Problem Setting and Notation

Consider a hot standby k-out-of-n:F ($k < n$) repairable system. The system fails when at least k of its components fail. In the model the lifetimes of different components and the same component after their repair are supposed to be independent

identically distributed (i.i.d.) random variables (r.v.'s) $A_i, (i = 1, 2, \dots)$ and their common cumulative distribution functions (c.d.f.) are denoted by $A(x) = \mathbf{P}\{A_i \leq x\}$ $(i = 1, 2, \dots)$.

The system is repairable which means that failed components are repaired with the help of a single repair facility, and upon repair completion they are as good as new. When dealing with the repairable model we need to consider some procedures of the system's restoration after failure. In this paper we assume that any component, having failed, is repaired during random time B_i $(i = 1, 2, \dots)$, where the r.v.'s B_i are supposed to be i.i.d. r.v.'s with common c.d.f. $B(x) = \mathbf{P}\{B_i \leq x\}$. But after the system failure the renewal of the whole system begins that requires another random time unit F for the system to be repaired and become as good as new, i.e. return to state 0. Furthermore, we consider a special case when the random time F has the same distribution as the partial repair time. For simplicity it is supposed that all distributions are absolute continuous and their probability density functions (p.d.f.) are denoted by $a(x)$, $b(x)$ correspondingly.

The system state space can be represented as $\mathbb{E} = \{0, 1, 2, ..., k-1, k\}$, where

- 0 means that all n components operate,
- i means that i components out of n $(1 \leq i \leq k-1)$ have failed, one of them is being repaired, and the other $(n-i)$ operate,
- k means that k components have failed, thus, the whole system has failed and is being repaired.

To describe the system's behavior we introduce a random process $J = \{J(t), t \geq 0\}$ on a phase space $\mathbb{E}$:

$$J(t) = j \text{ if at time } t \text{ the system is in state } j \in \mathbb{E}.$$

The current paper deals with the system's s.s.p. $\pi_j = \lim\limits_{t\to\infty} \mathbb{P}\{J(t) = j\}$ and properties of their asymptotic insensitivity to the shapes of life- and repair times of its components.

3. Steady State Probabilities. Analytical Results for $\langle M_{3<6}|GI|1\rangle$ system

We use in the paper a little bit modified Kendall's notation $\langle GI|GI|1\rangle$, where symbol GI in first position means general distributions for independent lifetimes of the components, this symbol in second position means general distributions of independent repair times. Number 1 in the last position means the amount of repair facilities. Symbols "$\langle\ \rangle$" mean that the considered system is a closed one.

In this section analytical results will be represented for the k-out-of-n:F system for the case when its components' lifetimes are exponentially distributed. To calculate the

s.s.p. we use the supplementary variable method [13]. For our case as supplementary variables we use the elapsed repair time of the failed component and the whole system. Thus, we consider a two-dimensional process $Z = \{Z(t), t \geq 0\}$, with $Z(t) = \{J(t), X(t)\}$ where $J(t)$ is the system state at time t, and $X(t)$ represents the elapsed repair time of the failed component or the whole system. Due to the introduction of the supplementary variables the process Z is a Markov one with a state space $\mathbb{E} = \{0, (1,x), (2,x), ..., (k-1,x), (k,x)\}$. Figure 1 shows the transition graph of the considered k-out-of-n:F system.

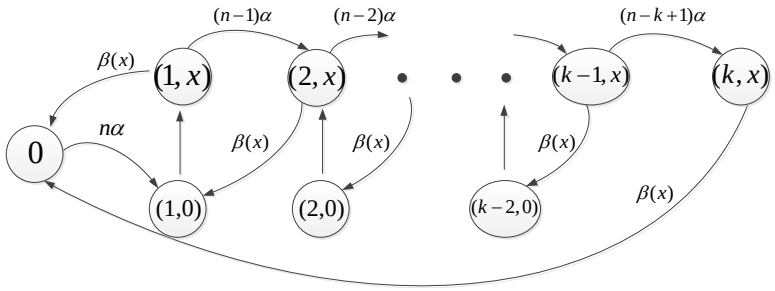


Fig. 1. Transition graph of the k-out-of-n:F system with full repair.

Here and further we will use the following notations:

- α is the failure rate of the system's components;
- $\lambda_i = (n-i)\alpha$, $(i = \overline{0, k-1})$ is the system's partial failure rate when i of n components fail;
- $b = \int\limits_0^\infty (1 - B(x))dx$ is the mean time to repair;
- $\beta(x) = (1 - B(x))^{-1} b(x)$ is the conditional partial and full repair rate, given the elapsed repair time is x;
- $\tilde{b}(s) = \int\limits_0^\infty e^{-sx} b(x)dx$ is the moment generation function (m.g.f.) of the repair time or Laplace transform (LT) of its p.d.f.

The state probabilities of the process are denoted by

$$\begin{aligned} \pi_0(t) &= \mathbb{P}\{N(t) = 0\}, \\ \pi_i(t,x)dx &= \mathbb{P}\{N(t) = i,\ x < X(t) \leq x + dx\} \quad (i = \overline{1,k}) \end{aligned}$$

and the corresponding s.s.p. are

$$\pi_0 = \lim_{t\to\infty} \pi_0(t), \quad \pi_i(x) = \lim_{t\to\infty} \pi_i(t;x) \quad (i = \overline{1,k}).$$

Using the method of comparison of the input and the output flows of failures and repair, we obtain the following system of balance equations for the system's s.s.p.:

$$\begin{aligned}
\lambda_0\pi_0 &= \int_0^\infty \pi_1(x)\beta(x)dx + \int_0^\infty \pi_k(x)\beta(x)dx, \\
\frac{d\pi_1(x)}{dx} &= -(\lambda_1+\beta(x))\pi_1(x), \\
\frac{d\pi_i(x)}{dx} &= -(\lambda_i+\beta(x))\pi_i(x) + \lambda_{i-1}\pi_{i-1}(x), \ \ (i=\overline{2,k-1}), \\
\frac{d\pi_k(x)}{dx} &= -\beta(x)\pi_k(x) + \lambda_{k-1}\pi_{k-1}(x)
\end{aligned} \tag{1}$$

with corresponding boundary conditions

$$\begin{aligned}
\pi_1(0) &= \lambda_0\pi_0 + \int_0^\infty \pi_2(x)\beta(x)dx, \\
\pi_i(0) &= \int_0^\infty \pi_{i+1}(x)\beta(x)dx \ i=\overline{2,k-2}, \\
\pi_{k-1}(0) &= \pi_k(0) = 0.
\end{aligned} \tag{2}$$

Remark 1. Note that the last boundary condition follows from the fact that the process never enters state $k-1$ with the elapsed time x equal to zero since the process enters this state only as a result of failure of another element and the transition from state $(k-2,x)$ with the same elapsed repair time.

Theorem 1. The macrostate s.s.p. of the "3-out-of-6:F" system in terms of LT in case of full repair and same distributions of partial and full repair time have the following form:

$$\begin{aligned}
\pi_0 &= \frac{1+5\tilde{b}(5\alpha)-5\tilde{b}(4\alpha)}{1+6\alpha b+5\tilde{b}(5\alpha)-5\tilde{b}(4\alpha)}, \\
\pi_1 &= \frac{6}{5}\cdot\frac{1-\tilde{b}(5\alpha)}{1+5\tilde{b}(5\alpha)-5\tilde{b}(4\alpha)}\pi_0, \\
\pi_2 &= \frac{3}{2}\cdot\frac{1+4\tilde{b}(5\alpha)-5\tilde{b}(4\alpha)}{1+5\tilde{b}(5\alpha)-5\tilde{b}(4\alpha)}\pi_0, \\
\pi_3 &= \frac{3}{10}\cdot\frac{20\alpha b-16\tilde{b}(5\alpha)+25\tilde{b}(4\alpha)-9}{1+5\tilde{b}(5\alpha)-5\tilde{b}(4\alpha)}\pi_0.
\end{aligned} \tag{3}$$

4. Numerical Results

In this section we use a simulation approach to show the asymptotic insensitivity of the system's s.s.p. to the shapes of its components' life- and repair time distributions for a special case of a "3-out-of-6:F" system. For the simulation of the system we use the multi-method modeling environment AnyLogic and the following notations (all parameters are considered in the same time scale):

- $\mathbf{E}A = a = \alpha^{-1}$ is the mean lifetime of all components,
- $\mathbf{E}B = b$ is the mean time to repair of all components and the whole system,
- $T = 10^6$ is the total simulation time,
- the parameters of all distributions are chosen in such a way that the coefficient of variation (the ratio of the standard deviation $\sigma = \sqrt{\mathbf{D}B}$ to the mean $b = \mathbf{E}B$) $c = \sigma/b$ takes a fixed value, and the mean lifetime a increases,
- $\rho = a/b = \mathbf{E}A/\mathbf{E}B$ is the relative recovery rate of the system's components.

It is shown that as $\rho \to \infty$ the sensitivity of the system s.s.p. to the shapes of its components' life- and repair time distributions becomes negligible. In our experiments the following distributions are used for the repair time: Gamma (Γ), Gnedenko-Weibull (GW), Pareto (P). For the lifetimes we used Exponential and Gamma distributions.

In the first numerical example we compare analytical (using formula (3)) and simulation results for the availability of the system $1 - \pi_3$ with the following distributions: Exponential for lifetime and Γ and GW for the repair time. In this case the mean lifetime $\alpha^{-1} = 1$, the coefficient of variation $c = 0.5$ and the relative recovery rate of the components runs from 0.00001 to 10, thus the mean full and partial repair time $\mathbf{E}B$ lies within the appropriate limits of the definition $\rho = \mathbf{E}A/\mathbf{E}B$.

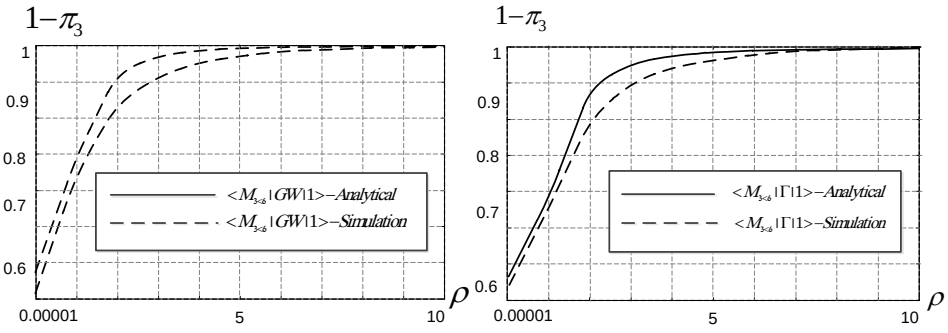


Fig. 2. The comparison of analytical and simulation results for the system availability $1 - \pi_3$, when $B(x) \sim GW$ and $B(x) \sim \Gamma$

As it is can be seen in Fig. 2, the system availability values for these cases are very close to each other. For all values of ρ the difference between analytical and

simulation values does not exceed 1%. Increasing of the relative recovery rate of the system's components provides a rapid increase of probability $1 - \pi_3$ (both analytical and simulation) for all examined distributions. This result shows the veracity of simulation tools and insensitivity of the system to the shape of its components' repair time distributions.

The second experiment presents only simulation results (Fig. 3). The curves represent the system availability $1 - \pi_3$, when repair times for partial and full failures have Γ, GW and P distributions and when lifetime has Γ distribution. In this case, the mean lifetime $\mathbf{E}A = 0.5$, the coefficient of variation $c = 10$ and ρ, as well as $\mathbf{E}B$, have the same values as in the previous example. The availability $1 - \pi_3$ tends rapidly to 1 and shows its asymptotic insensitivity to the shape of life- and repair time distribution of the system's components.

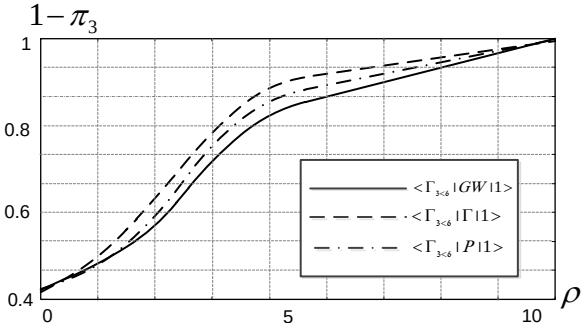


Fig. 3. System availability $1 - \pi_3$ for the case of generally distributed life- and repair times (simulation results).

5. Conclusion

The analytical expressions for the s.s.p. of the "k-out-of-n:F" system with exponential lifetime and general repair time distributions have been found. For the same system the simulation model has been created in AnyLogic environment and applied for calculation of the system's stationary characteristics when both life- and repair times have general distributions. It was shown that as the relative recovery rate of the system's components increases, the system s.s.p. become asymptotically insensitive to the shapes of the life- and repair time distributions of the system's components. Study of more complex systems with dependent failures of their components is the subject of our further research.

REFERENCES

1. Kala, Z.: Sensitivity analysis in probabilistic structural design: A comparison of selected techniques // Sustainability, 12(11), p.19, 2020. DOI:10.3390/su12114788

2. Kala, Z.: Quantile-oriented global sensitivity analysis of design resistance // Journal of Civil Engineering and Management, 25(4), 297-305, 2019. ISSN 1392-3730, E-ISSN 1822-3605. DOI: 10.3846/jcem.2019.9627
3. Kala, Z.: Estimating probability of fatigue failure of steel structures // Acta et Commentationes Universitatis Tartuensis de Mathematica, 23(2), 245-254, 2019. ISSN 1406-2283, E-ISSN 2228-4699. DOI: 10.12697/ACUTM.2019.23.21
4. Kala, Z.: Global sensitivity analysis of reliability of structural bridge system // Engineering Structures, 194, 36-45, 2019. ISSN 1644–9665. DOI: 10.1016/j.engstruct.2019.05.045
5. Rykov, V., Zaripova, E., Ivanova, N., Shorgin, S.: On sensitivity analysis of steady state probabilities of double redundant renewable system with Marshall-Olkin failure model // CCIS, vol 919, 2018, 234-245. DOI: 10.1007/978-3-319-99447-5_20
6. Sevast'yanov, B.A.: An ergodic theorem for markov processes and its application to telephone systems with refusals. Theory Probab. Appl. 2(1), 104-112 (1957)
7. Kovalenko, I.N.: Investigations on Analysis of Complex Systems Reliability, p. 210. Naukova Dumka, Kiev (1976). (in Russian)
8. Rykov, V.: Multidimensional alternative processes reliability models // Communications in Computer and Information Science, vol 356, pp. 147-157. Springer,2013. DOI: 10.1007/978-3-642-359804_17.
9. Efrosinin, D., Rykov, V., Vishnevskiy, V.: Sensitivity of reliability models to the shape of life and repair time distributions. // In: 9th International Conference on Availability, Reliability and Security (ARES 2014), pp. 430-437. IEEE (2014) doi: 10.1109/ARES.2014.65.
10. Rykov, V., Kozyrev, D.: On Sensitivity of Steady-State Probabilities of a Cold Redundant System to the Shapes of Life and Repair Time Distributions of Its Elements, Statistics and Simulation // Springer Proceedings in Mathematics and Statistics vol. 231, Chapter 28, 2018. Springer, Cham. P. 391-402. DOI: 10.1007/978-3-319-76035-3_28.
11. Trivedi, K. S.: Probability and Statistics with Reliability, Queuing and Computer Science Applications. John Wiley & Sons, New York, 2002.
12. Kozyrev, D.V., Phuong, N.D., Houankpo, H.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, 1141 CCIS, pp. 646-656. 2019. DOI: 10.1007/978-3-030-36625-4_52
13. Rykov, V. V., Kozyrev, D. V.: Analysis of renewable reliability systems by Markovization method // Analytical and Computational Methods in Probability Theory (ACMPT 2017), Lecture Notes in Computer Science, Vol. 10684, 2017. Springer, Cham. Pp. 210-220. DOI: 10.1007/978-3-319-71504-9_19.

UDC: 004.75

Timeliness of Redundant Service of a Heterogeneous Request Flow by a Sequence of Nodes of the Info-communication System

V. A. Bogatyrev[1,2,3], A. V. Bogatyrev [1], S. V. Bogatyrev [1]

[1]NEO Saint Petersburg Competence Center, Saint Petersburg, Russia
[2]ITMO University, Saint Petersburg, Russia
[3]Saint Petersburg State University of Aerospace Instrumentation, Saint Petersburg, Russia

vladimir.bogatyrev@gmail.com, anatoly@nspcc.ru, stanislav@nspcc.ru

Abstract

The possibilities are investigated and analytical models of redundant multi-way servicing of a heterogeneous request flow with their replication rate depending on the maximum permissible waiting time for replicas accumulated in the queues of nodes that make up the path for real-time information and communication systems are proposed. Two options are considered for redundant servicing of a heterogeneous flow during the sequential passage of copies of requests through parallel-connected nodes grouped in groups. For the first option, when generating a request, a certain number of copies are created, for each of which a path is predefined as a sequence of nodes of different groups involved in servicing this copy. For the second option, the paths are formed dynamically at each stage, and a copy of the request, executed first at some stage of the sequential passage of groups of redundant nodes, is transferred for redundant service to the next group of nodes. At various stages of service, the redundancy ratio can vary.

Keywords: redundant service; latency; heterogeneous flow; distribution of requests copies; real-time

1. Introduction

Increased requirements for reliability [1,2], continuity, and timeliness of data processing and transmission processes are imposed on real-time systems, including cyber-physical systems [3-4].

In the well-known works [5–6] related to ensuring the reliability of distributed computer systems including cluster ones [7–8], questions of assessing the reliability of real-time cluster systems are not addressed, for which strict requirements are

imposed on service request delays and on the continuity of the computing process, including when the recovery time after failures of redundant resources can exceed the maximum allowable time of interruption of the computing process [9,11].

To reduce the average network transmission delays, transport coding allows [9, 10], in which message fragments are transmitted along different routes, and in case of loss or error of frame transmissions, the entire message can be restored without retransmissions.

Multi-path routing allows for increased network availability and reduced reconfiguration time[11, 12]. With multi-path transmission, the main and several backup routes (paths) are formed, and in case of failure of the nodes making up the path, a switch to a backup, the pre-registered path is performed, which allows to speed up reconfiguration, and in some cases to ensure that the permissible delays in real-time systems are not exceeded. The efficiency of multi-path transmissions is achieved with prioritization of traffic and load balancing, including network reconfiguration after failures [11, 12].

Reliability and timeliness of query execution in a redundant information and communication system (server cluster or switching nodes) can be improved as a result of redundant servicing of copies of requests with the issuance of one completed copy (for example, the first issued in time) [13, 14].

The direction of redundant multipath service with query replication [14-15], researched in this article, is the development of the concepts of multipath routing [11, 12], multicast transmissions [16], broadcast service [17, 18] and dynamic distribution of requests [19].

The effectiveness of redundant maintenance for multi-level cluster systems (sequentially connected groups of redundant nodes, multicluster) is estimated by the probability of the timeliness of multi-stage servicing of at least one copy of the request [13].

For requests serviced in real-time, it is important not only to maintain the operability of the nodes that make up the path, but also the timeliness of the service and requests passing through it. The task is complicated by the heterogeneity of the flow when different requests may have different criticality to service time.

The aim of the work is to investigate the possibility of increasing the functional reliability of a distributed system while increasing the probability of timely execution of time-critical requests of a heterogeneous stream as a result of their replication and redundant servicing by a sequence of nodes included in the path, taking into account the accumulation of latency on all nodes of the path.

2. Options for the formation of the path for phased redundant service of requests in a multi-level cluster

As the object of research consider m levels computer cluster comprising at *i-th* level of n_i parallel-connected servers, each of them can be represented as single-channel queuing systems of the M/M/1 type with infinite queues.

With redundant maintenance at each level, multiple copies of the request are executed.

The following options for organizing redundant services in a multi-level cluster are researched:

Option S_1: k copies of request (replicas) are created in the node of the request source, and for each copy, the service path (route) is specified with the servers that execute the request (copy of the request) at each stage (level) [15]. For requests of different criticality to the total waiting time in the nodes making up the path, their number (paths' redundancy ratio) can be set various.

Option S_2: When a request is generated by a source, it's k_1 copies are created, distributed for nodes in k_1 first-level servers. When one of the copies is serviced at the first level, k_2 copies of requests are created, which are transferred to services in selected k_2 second-level nodes, and so on, until the request is serviced all m levels of the cluster [15]. The redundancy ratio of requests is set depending on their criticality to the total waiting time in the nodes that make up the path and can be different for different levels.

In this paper, we set the task of constructing a new model that allows, for a heterogeneous request flow, of different criticality, to wait delays in the nodes making up the path, to take into account the requirements for not exceeding the maximum allowable accumulated waiting time for sequential redundant request servicing by nodes included in the path with an inhomogeneous flow of requests with allocation of z types of requests by the allowable waiting time in queues t_1, t_2, ..., t_z. The fractions of flows of heterogeneous requests are equal to g_1, g_2, ..., g_z respectively, and their intensities λg_1, λg_2, ..., λg_z, and $\sum_{i=1}^{z} g_i = 1$.

3. Redundant service with preliminary formation of paths

In the case of a heterogeneous request flow for a three-level cluster, the probability of not exceeding the maximum permissible total waiting time t_i of requests of the i-th type in queues of two levels t_0 for one (any) of the k_j assigned paths of the redundant execution of one copy of the request is calculated as:

$$p_{1i} = \sum_{i_1=0}^{N-1} \sum_{i_2=0}^{N-1-i_1} \{[1 - \frac{\Lambda_0}{n_1} v_1 exp(-\frac{i_1 t_i}{N}(v_1^{-1} - \frac{\Lambda_0}{n_1}) - b_{i_1}] \times$$
$$\times [1 - \frac{\Lambda_0}{n_2} v_2 exp(-\frac{i_2 t_i}{N}(v_2^{-1} - \frac{\Lambda_0}{n_2}) - b_{i_2}] \times$$
$$\times [1 - \frac{\Lambda_0}{n_3} v_3 exp(-(t_i - (i_1 + i_2)\frac{t_i}{N})(\frac{1}{v_3} - \frac{\Lambda_0}{n_3}))]\},$$

$$b_{i_1} = \begin{cases} 1 - \frac{\Lambda_0}{n_1} v_1 exp(-\frac{t_i(i_1-1)}{N}(v_1^{-1} - \frac{\Lambda_0}{n_1}), & \text{if } i_1 \geq 1, \\ 0, & \text{if } i_1 = 0. \end{cases}$$

$$b_{i_2} = \begin{cases} 1 - \frac{\Lambda_0}{n_2} v_2 exp(-\frac{t_i(i_2-1)}{N}(v_2^{-1} - \frac{\Lambda_0}{n_2}), & \text{if } i_2 \geq 1, \\ 0, & \text{if } i_2 = 0. \end{cases}$$

$$\Lambda_0 = \sum_{i=1}^{z} k_i g_i \Lambda.$$

The probability of servicing a heterogeneous flow with the requirement to ensure the probability of timely execution of all flows (types of requests), taking into account the accumulation of their waiting time in the queues, can be found as:

$$P_1 = \prod_{i=1}^{z} [1 - (1 - p_{1i})^{k_i}],$$

4. Redundant service of heterogeneous flow with the formation of copies of requests

The probability of a request flow of different criticality for service delays being evaluated will be evaluated. z types of requests according to the criticality of the total waiting time at nodes sequentially receiving a service request will be allocated, while for the i-th type of requests, the maximum allowable accumulated waiting time by all nodes sequentially serving a request is set to t_i.

For redundant service according to option S_2, the formation of a given number of copies of nodes transmitted to the next level is carried out by a copy of the request executed first in time.

As a result, the intensity of requests arriving at the j-th level, taking into account the multiplicity of reservation of k_{ji} requests of the i-th type will be equal to

$$\Lambda_j = \sum_{i=1}^{z} g_i k_{ji} \Lambda.$$

For a three-level cluster, the probability of timely execution of at least one copy of the i-th type request, taking into account the total delay at all levels, is calculated as:

$$P_{2i} = \sum_{i_1=0}^{N-1} \sum_{i_2=0}^{N-1-i_1} \{[1 - \{(\Lambda_1/n_1)v_1 exp(-i_1 \frac{t_i}{N}(v_1^{-1} - \frac{\Lambda_1}{n_1}))\}^{k_{1i}} - b_{i_1}] \times$$
$$\times [1 - \{(\Lambda_2/n_2)v_2 exp(-i_2 \frac{t_i}{N}(v_2^{-1} - \frac{\Lambda_2}{n_2}))\}^{k_{2i}} - b_{i_2}] \times$$
$$\times [1 - [(\Lambda_3/n_3)v_3 exp(-(t_0 - (i_1 + i_2)\frac{t_i}{N})(\frac{1}{v_3} - \frac{\Lambda_3 k_3}{n_3})]^{k_3}]\},$$

$$b_{i_1} = \begin{cases} 1 - \{\frac{\Lambda_1}{n_1} v_1 exp(-\frac{t_i}{N}(i_1 - 1)(v_1^{-1} - \frac{\Lambda_1}{n_1}))\}^{k_{1i}}, & \text{if } i_1 \geq 1, \\ 0, & \text{if } i_1 = 0. \end{cases}$$

$$b_{i_2} = \begin{cases} 1 - \{\frac{\Lambda_2}{n_2} v_2 exp(-\frac{t_i}{N}(i_2 - 1)(v_2^{-1} - \frac{\Lambda_2}{n_2}))\}^{k_{2i}}, & \text{if } i_2 \geq 1, \\ 0, & \text{if } i_2 = 0. \end{cases}$$

For option S_2 of redundant service, the efficiency of servicing an inhomogeneous flow with the requirement to ensure the probability of timely execution of all types of requests taking into account the accumulation of their waiting time in queues is defined as

$$P_2 = \prod_{i=1}^{z} P_{2i}.$$

For option S_2 of the redundant service, the total average residence time of the request of the i-th flow (type) at all m levels of the system calculated as

$$T_i = \sum_{j=1}^{m} \{\int_0^{\infty} [\frac{\Lambda_j v_j}{n_j} e^{(\frac{\Lambda_0}{n} - \frac{1}{v})t}]^{k_{ji}} dt\}.$$

For a heterogeneous flow, the efficiency criterion can be taken as the mathematical expectation of the total delays in the request queues of all z types calculated as:

$$T = \sum_{i=1}^{z} g_i \sum_{j=1}^{m} \{\int_0^{\infty} [\frac{\Lambda_j v_j}{n_j} e^{(\frac{\Lambda_0}{n} - \frac{1}{v})t}]^{k_{ji}} dt\}.$$

The implementation of multi-path redundant service with the formation of copies of requests at each stage requires the interaction of group nodes, which can be quite

simply organized in a server cluster, but the implementation of such interaction in a group of communication nodes is difficult, since it requires interaction through a network, which can significantly slow down the maintenance process.

5. Comparison of reserved service options

Consider a two-level cluster. In the calculations, we assume that, the number of reserved servers at each level is the same and equal to $n = 8$ pcs, the average query execution time by the servers of the first and second level is $v_1 = v_2 = 0.4$ s, and the total allowable wait time is $t_0 = 0.2$ s .

In fig. 1 shows the dependencies of the probabilities of timely service on the intensity of the input request flow Λ. In Fig. 1 a), curve 1 corresponds to non-redundant service $k = 1$, curves 2-3 correspond to service option S_1 with a request redundancy ratio of $k = 2, 3$, and curves 4, 5 correspond to option S_2 with a redundancy ratio of $k = 2, 3$.

In fig. 1 b) Curves 1-3 represent service option S_1, and curves 4-6 option S_2 for request flow intensities corresponding to $\Lambda = 3.3; 3; 3.5$ 1/s.

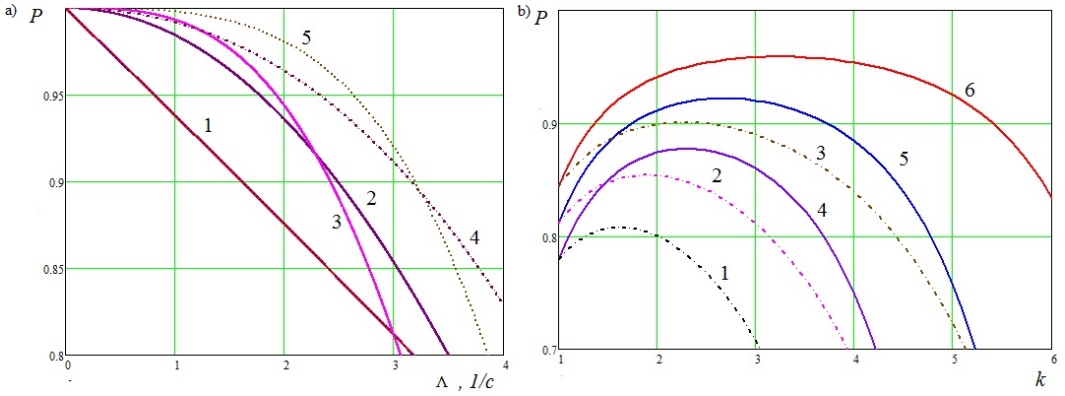


Fig. 1. Dependencies of the probabilities of timely service.

These graphs allow to conclude that there is an optimal multiplicity of redundant service, and the smaller the system load and the permissible total waiting time, the greater the redundancy ratio at which the maximum probability of multistage service timeliness is achieved.

Calculations show that the S_2 redundant service option improves the probability of timely service requests, however, its implementation is more complicated and requires additional research on the organization of interaction between nodes included in the redundant group (cluster).

6. Conclusion

For multilevel info-communication systems involving requests made by a sequence of redundant nodes at each level, an analytical model is proposed and the effectiveness of the options for redundant servicing of a heterogeneous request flow of various criticality to the total waiting delay in the queues is determined.

The proposed model allows to take into account the requirements of not exceeding the maximum allowable accumulated waiting time for sequential redundant request servicing at all levels of the system.

The influence of the redundant service multiplicity on the probability of the timely execution of a heterogeneous request flow taking into account sequential redundant execution at nodes at all levels of the system is analyzed.

The efficiency of redundant service of a heterogeneous request flow with the formation of the number of copies of requests and their distribution in the queue of servers at each system level, taking into account the criticality to the delay of requests of different flows, is shown.

REFERENCES

1. Sorin D. Fault Tolerant Computer Architecture. Morgan&Claypool, 2009. P. 103
2. Koren I. Fault tolerant systems. Morgan Kaufmann publications, visit our San Francisco 2009. P. 378.
3. Zakoldaev D. A., Shukalov A. V., Zharinov I. O., Zharinov O. O. Designing technologies for the interaction of cyber-physical systems in smart factories of the Industry 4.0 // Journal of Physics: Conference Series. 2020. V. 1515, No. 2.
4. Poymanova E. D., Tatarnikova T. M. Models and Methods for Studying Network Traffic. // 2018 Wave Electronics and its Application in Information and Telecommunication Systems (WECONF). 2018. P. 1–5. doi:10.1109 / WECONF.2018.8604470.
5. Seontae Kim, Young-ri Choi. Constraint-aware VM placement in heterogeneous computing clusters // Cluster Computing 23 (SI). March 2020. P. 71–85.
6. Yang C., Liu J., Hsu C. et al. On improvement of cloud virtual machine availability with virtualization fault tolerance mechanism. // The Journal of Supercomputing 69. 2014. P. 1103—1122.
7. Jo C., Cho Y., Egger B. A machine learning approach to live migration modeling. In: Proceedings of the 2017 Symposium on Cloud Computing, vol. 17, pp. 351–364. SoCC (2017).
8. Keller G., Lutfiyya H. Dynamic management of applications with constraints in virtualized data centres. In: Proceedings of IFIP/IEEE International Symposium on Integrated Network Management (IM) (2015).

9. Kabatiansky G., Krouk E., Semenov S. Error Correcting Coding and Security for Data Networks. // Analysis of the Superchannel Concrete ept. Wiley. 2005. P. 288.
10. Krouk E., Semenov S. Application of Coding at the Network Transport Level to Decrease the Message Delay // Proc. of 3rd Intern. Symp. on Communication Systems Networks and Digital Signal Processing. Staffordshire University, UK, 2002. P. 109–112.
11. Prasenjit Chanak, Tuhina Samanta, Indrajit Banerjee. Fault-tolerant multipath routing scheme for energy efficient wireless sensor networks // International Journal of Wireless & Mobile Networks (IJWMN). April 2013. V. 5, No. 2. P. 33–45.
12. Rajeev V., Muthukrishnan C.R. Reliable backup routing in fault tolerant real-time networks. Proceedings. Ninth IEEE International Conference on Networks, ICON 2001.
13. Bogatyrev V. A., Bogatyrev A. V. "Functional reliability of a real-time redundant computational process in cluster architecture systems" // Automatic Control and Computer Sciences. 2015. V. 49, No. 1. P. 46–56.
14. Bogatyrev A. V., Bogatyrev V. A., Bogatyrev S. V. Multipath Redundant Transmission with Packet Segmentation // (2019) 2019 Wave Electronics and its Application in Information and Telecommunication Systems, WECONF 2019, art. no. 8840643. doi: 10.1109/WECONF.2019.8840643.
15. Bogatyrev V. A., Bogatyrev S. V., Bogatyrev A. V. Model and Interaction Efficiency of Computer Nodes Based on Transfer Reservation at Multipath Routing (2019) // 2019 Wave Electronics and its Application in Information and Telecommunication Systems, WECONF 2019, art. no. 8840647 doi: 10.1109/WECONF.2019.8840647.
16. Samuylov A., Moltchanov D., Kovalchukov R., (...), Koucheryavy Y., Samouylov K. Characterizing Resource Allocation Trade-Offs in 5G NR Serving Multicast and Unicast Traffic. // IEEE Transactions on Wireless Communications. 2020. V. 19(5),9003488. P. 3421–3434.
17. Lee M. H., Dudin A. N., Klimenok V. I. The SM/V/N queueing system with broadcasting service // Math. Probl. in Engineer. 2006. V. 2006. Article ID 98171. P. 18.
18. Dudin A. N., Sun B. A multiserver MAP/PH/N system with controlled broadcasting by unreliable servers. // Automatic Control and Computer Sciences. 2009. V. 5. P. 32–44.
19. Bogatyrev V. A. Protocols for dynamic distribution of requests through a bus with variable logic ring for reception authority transfer // Automatic Control and Computer Sciences. 1999. V. 33(1). P. 57–63.

UDC: 123.456

Computational Aspects of Modelling Performance Characteristics for Polling Models with Semi-Markov Switching and Priorities

G.K. Mishkoy[1,2] and L.M. Mitev [1,2]

[1]V. Andrunachievici Institute of Mathematics and Computer Science,
5 Academiei street, Chisinau, Republic of Moldova
[2]Free International University of Moldova,
52 V. Pircalab street, Chisinau, Republic of Moldova

gmiscoi@ulim.md, liliamitev@gmail.com

Abstract

Exhausting polling models with priorities and semi-Markov switching are considered. Some of performance characteristics, such as analog of Pollaczek-Khintchin virtual and steady state transform equations, analog of Kendall functional equation and analog of Gnedenko system's busy period are presented. Elaboration of numerical algorithms and computational aspects of numerical modelling are discussed.

Keywords: Polling model, semi-Markov switching, priority, k-busy period, Kendall functional equation, Pollaczek-Khintchin transform equation, Laplace-Stieltjes transform.

1. Introduction

Polling models were studied by many researchers and till now were obtained many outstanding results (see for example [1]). However, the impetuous development of contemporary practice, particularly network technologies, has posed new challenges and requirements, in particular, to develop new mathematical models, which should be more flexible and more adequate to real processes, compared with known classical ones. Polling models with semi-Markov switching and priorities mostly meet these requirements. Really, these models take into account time losses on various switching on changing priority classes, auxiliary jobs, etc. Consideration of diverse priority lows allows to greatly contributes to the increase in the efficiency of their functioning. More than that, the performance characteristics for such systems despite their complex structure, include classical characteristics as special cases. As

The publication has been prepared with the partially support of State Programs and Grants according to the research project No.20.80009.5007.13.

examples we will present the analog of Pollaczek-Khintchin virtual and steady state transform equations, the analog of Kendall functional equation and the analog of Gnedenko system's busy period.

2. The k-busy period

We consider a queueing system of polling type with semi-Markov delays. Handling mechanism for this system is given by polling table $f : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, r\}$, where the function shows that at the stage $j,\ j = 1, \ldots, n$, user number $k,\ k = 1, \ldots, r,\ r \leq n$ is served (see [1]). The items (messages) of the user k, according to Poisson distribution with parameter λ_k arrive. The service time for the items of class k is a random variable B_k with distribution function $B_k(x) = P\{B_k < x\}$. Duration of the switching from one user to user k is a random variable C_k with distribution function $C_k(x) = P\{C_k < x\}$. Thus C_k can be interpreted as a the loss of time in preparing the service process for user of class k.

The k-busy period is a measure of the time that expires from when a server begins to process, after an empty queue, to when the k-queue becomes empty again for the first time [2].

Denote by $\Pi_k^\delta(x)$ distribution function (d.f.) of the k-busy period, and by $\pi_k^\delta(s)$ it's Laplace-Stieltjes transform.

Theorem 1. Function $\pi_k^\delta(s)$ is determined from equation

$$\pi_k^\delta(s) = c_k(s + \lambda_k - \lambda_k \pi_k(s))\pi_k(s) \tag{1}$$

where

$$\pi_k(s) = \beta_k(s + \lambda_k - \lambda_k \pi_k(s)). \tag{2}$$

Remark 1. If we consider that $C_k = 0$ and $k = 1$, then from formula (1) it follows that $\pi_k^\delta(s) = \pi_k(s)$ and $\pi_1^\delta = \pi_1(s) = \beta_1(s + \lambda - \lambda\pi_1(s))$, respectively.

Remark 2. If $\lambda_k\beta_{k1} < 1,\ \lambda_k c_{k1} < 1$, then first moment of k-busy period is determined from:

$$\pi_{k1}^\delta = \frac{\beta_{k1}}{1 - \lambda_k\beta_{k1}} + \frac{c_{k1}}{1 - \lambda_k c_{k1}},$$

where β_{k1} and c_{k1} are the first moments of $C_k(x)$ and $B_k(x)$.

Thus, expression (1) can be viewed as an analog of Kendall equation obtained for classical $M|G|1$ system.

3. The Pollaczek-Khintchin virtual analog

Let $P_m(x)$ be the probability that at the instant x there are m - messages in the k-queue. Denote

$$P_k(z,x) = \sum_{m=1}^{\infty} P_m(x) z^m, 0 \le z \le 1,$$

the generating function of queue length distribution and

$$p_k(z,s) = \int_0^{\infty} e^{-sx} P_k(z,x) dx,$$

the Laplace transform of function $P_k(z,s)$.

Theorem 2.

$$p_k(z,s) = \frac{1 + \lambda_k \pi_k^{\delta}(z,s)}{s + \lambda_k - \lambda_k z} \tag{3}$$

$$\pi_k^{\delta}(z,s) = \frac{1 - c_k(s + \lambda_k - \lambda_k z)}{s + \lambda_k - \lambda_k z} + \frac{\beta_k(z,s)}{z - \beta_k(s + \lambda_k - \lambda_k z)} \times$$
$$\times [z c_k(s + \lambda_k - \lambda_k z) - \pi_k^{\delta}(s)] \tag{4}$$

$$\beta_k(z,s) = \frac{1 - \beta_k(s + \lambda_k - \lambda_k z)}{s + \lambda_k - \lambda_k z}. \tag{5}$$

Remark 3. In the next section it is shown that from the Theorem 2 it follows Theorem 3, where formula (7) results from (3). Thus, the result from Theorem 2 can be viewed as a virtual analogue of the Pollaczek-Khintchin equation.

4. The Pollaczek-Khintchin steady state analog

Theorem 3. If $\lambda_k \beta_{k1} < 1$, $\lambda_k c_{k1} < 1$, then

$$P_k(z) = \lim_{s \downarrow 0} \; s p_k(z,s),$$

and

$$P_k(z) = \frac{1 + \lambda_k \pi_k^{\delta}(z,0)}{1 + \lambda_k \pi_{k1}^{\delta}}. \tag{6}$$

Function $\pi_k^{\delta}(z,0)$ is determined from (4) for $s = 0$ and π_{k1}^{δ} from Remark 2.

Remark 4. If $C_k = 0$ and $k = 1$, then

$$P_{k1}(z) = P_k(z) = \frac{\beta(\lambda - \lambda z)(z-1)(1 - \lambda \beta_1)}{z - \beta(\lambda - \lambda z)} \tag{7}$$

where $\beta_1(\cdot) = \beta(\cdot)$ and $\beta_{11} = \beta_1$.

Formula (7) is referred to in most text-books on queueing analysis and it is known as the Pollaczek-Khintchin transform equation (Pollaczek (1961); Khintchin (1963)).

5. System's busy period $M_r|G_r|1|\infty$

Denote by $B_i(x)$-d.f. of service of the requests of the i-th priority class, $C_j(x)$-d.f. of switching for service of the requests of the j class, λ_i-parameter of the Poisson flow of priority i, $\Pi(x)$-d.f.of the busy period; $i, j = 1, \ldots, r$; $i \neq j$, $\sigma_k = \lambda_1 + \cdots + \lambda_k$, $\sigma = \sigma_r$, $\beta_i(s) = \int_0^\infty e^{-sx} dB_i(x), c_j(s), \pi(s)$-the Laplace-Stieltjes transform of d.f. $B_i(x)$, $C_j(x)$, $\Pi(x)$.

Theorem 4. *(Priority policy P12: "resume", "repeat again")*

The Laplace-Stieltjes transform $\pi(s) = \pi_r(s)$ of the d.f. of the busy period is determined (at $k = r$) from the system of recurrent functional equations:

$$\sigma_k \pi_k(s) = \sigma_{k-1}\pi_{k-1}(s + \lambda_k) + \sigma_{k-1}\{\pi_{k-1}(s + \lambda_k[1 - \overline{\pi}_k(s)]) -$$

$$-\pi_{k-1}(s + \lambda_k)\}\nu_k(s + \lambda_k[1 - \overline{\pi}_k(s)]) + \lambda_k \pi_{kk}(s) \tag{8}$$

$$\pi_{kk}(s) = \nu_k(s + \lambda_k[1 - \overline{\pi}_k(s)])\overline{\pi}_k(s) \tag{9}$$

$$\overline{\pi}_k(s) = h_k(s + \lambda_k[1 - \overline{\pi}_k(s)]) \tag{10}$$

where

$$\nu_k(s) = c_k(s + \sigma_{k-1}[1 - \pi_{k-1}(s)]) \tag{11}$$

$$h_k(s) = \beta_k(s + \sigma_{k-1}) \left\{ 1 - \frac{\sigma_{k-1}}{s + \sigma_{k-1}} [1 - \beta_k(s + \sigma_{k-1})]\pi_{k-1}(s)\nu_k(s) \right\}^{-1} \tag{12}$$

From Theorem 4, functions $\pi_k(s)$, $\pi_{kk}(s)$, $\overline{\pi}_k(s)$, $\nu_k(s)$ and $h_k(s)$ are the Laplace-Stieltjes transforms of d.f. of the important auxiliary periods Π_k, Π_{kk}, $\overline{\Pi}_k$, N_k and H_k [3]. Thus, $\pi_k(s), \ldots, h_k(s)$ are the distribution (in terms of Laplace-Stieltjes transforms) of busy period for k - priority request and higher, ..., distribution of total time of k - switching, total time of service of k - priority request.

Remark 5. *Gnedenko system's busy period.*

If $C_j = 0$, $j = 1, \ldots, r$, $r > 1$ from relations (8)-(12) follow the result published by Gnedenko et al in monograph "Priority Queueing Systems" (1973)

$$\sigma_k \pi_k(s) = \sigma_{k-1}\pi_{k-1}(s + \lambda_k(1 - \pi_{kk}(s))) + \lambda_k \pi_{kk}(s),$$

$$\pi_{kk}(s) = h_k(s + \lambda_k(1 - \pi_{kk}(s))),$$

$$h_k(s) = \beta_k(s + \sigma_{k-1}) \left\{ 1 - \frac{\sigma_{k-1}}{s + \sigma_{k-1}} [1 - \beta_k(s + \sigma_{k-1})]\pi_{k-1}(s) \right\}^{-1}.$$

6. Computational aspects of elaboration of numerical algorithms

The analytical results formulated above, although they are of interest from fundamental theoretical point of view, are quite complicated for numerical modelling. Indeed, for example, $\pi_k^{\delta}(s)$ given by the Theorem 1 is present in expression (4). The same situation is true for other characteristics of polling model. But for determining this function it is necessary to solve the functional equation (2), which does not have the exact analytical solution, but which effectively can be solved numerically. As an example, we will present a numerical algorithm of successive approximations.

Input: $\{\lambda_k\}_{k=1}^r; \{b_k\}_{k=1}^r; \{c_k\}_{k=1}^r; s; r; \varepsilon > 0$.
Output: $k; \{\pi_k(s)\}_{k=1}^r; \{\pi_k^{\delta}(s)\}_{k=1}^r$.
Descriptions:
1. Laplace-Stieltjes transforms of exponential distribution functions $B_k(x)$ and $C_k(x)$, are determined:
$\beta_k(s) = \frac{b_k}{s+b_k}$; $\overline{c}_k = \frac{c_k}{s+c_k}$.

2. Distribution function for k-busy period is determined, using Theorem 1, for $k = 1, ..., r$.
For $n = 0$, $\pi_k^{(0)}(s) = 0$, $\pi_k{}^{(n)}(s) = \beta_k(s + \lambda_k - \lambda_k \pi_k{}^{(n-1)}(s))$,
$\pi_k^{\delta(n)}(s) = \overline{c}_k(s + \lambda_k - \lambda_k \pi_k{}^{(n)}(s))\pi_k{}^{(n)}(s)$.

Stop condition: $|\pi_k^{(n)}(s) - \pi_k^{(n-1)(s)}| < \varepsilon$.

Remark 6. In the presented algorithm d.f. $B_k(x)$ and $C_k(x)$ are considered as exponential distributed. Analogical algorithms can be elaborated for other distributions of $B_k(x)$ and $C_k(x)$.

The same it can be mentioned about Theorem 4. The functions $\pi_k(s)$, $\nu_k(s)$ and $h_k(s)$ are involved in analytical expressions of most performance characteristics. To numerically model some of the performance characteristics of polling model with priorities it is necessary to solve numerically the system of functional equations (8)-(12). This system can be solved numerically. Some examples of algorithms are presented in [3, 4].

7. Conclusion

As can be seen from the analytical results presented in sections 1-4, the performance characteristics of polling models with priorities and semi-Markov switching are quite complicated from an analytical point of view. Their direct application for modelling purposes is impossible. But the application of numerical methods and

the elaboration of numerical algorithms open remarkable possibilities for modelling performance characteristics for these models.

The presented results and examples, as well as the consequences of these results, which coincide with the known classical results (Kendall functional equation, Pollaczek-Khintchin transform equation, etc.) validates the research methodology and confirms the continuity and concordance of the obtained results with the known classical results.

REFERENCES

1. Vishnevsky V. V, Semenova O. V. Polling Systems: Theory and Applications for Broadband Wireless Networks. London, Academic Publishing, 2012.
2. Mishkoy Gh. K., Bejenari D. D., Mitev L. M., Ticu I. Numerical solutions of Kendall and Pollaczek-Khintchin equations for exhaustive polling systems with semi-Markov delays // Computer Science Journal of Moldova. 2016. N2(71). P. 255–272.
3. Mishkoy Gh. K. Priority systems with orientation. Analytical and numerical results // Analytical and Computational Methods in Probability Theory. Springer. 2017. P. 109–120.
4. Mishkoy Gh. K., Mitev L. M. Modeling of busy period in polling models with semi-Markov switching of states // Proceedings of the International Conference IMCS-55. 2019. P. 215–221.

UDC: 621.375

Scaling error suppression in small signal preamplifiers for vibration monitoring networks

V.P. Morozov[1] and K.A. Alikin[1]

[1]V.A. Trapeznikov Institute of Control Sciences of RAS, 65 Profsoyuznaya street, Moscow, Russia

morbe36@mail.ru, ak-evmt@yandex.ru

Abstract

Energy consumption reduction is of utmost importance for Internet of Things devices. This pertains especially to wireless sensor networks that have to run uninterruptedly for long periods of time, such as jerks and oscillation measurement systems used for earthquake registration or large building vibration monitoring. In these systems electronic data processing equipment is abundant. Lowering of the operating current and voltage in electronic equipment tends as a rule to increase the processing error. It is true particularly of the scaling amplifiers widely used in the course of data processing for preamplification of weak signals from analog sensors. In this paper, we propose an efficient method and corresponding circuit design for reproduction error suppression in scaling amplifier by the additional feedback loop. The implementation of additive negative feedback by error signal supports substantial total scaling error reduction. Experiments and simulation carried out demonstrated manifold error reduction by the proposed structure.

Keywords: internet of things, vibration sensors, energy consumption, scaling amplifier, error suppression, additive negative feedback

1. Introduction

Autonomous wireless sensors serve as the backbone of the Internet of Things (IoT) [1]. These sensors do not have an external power supply and their lifespan is typically limited by the lifespan of their battery nodes [2]. In [3] authors discuss a network of sensors that can extract (harvest) the energy from the surrounding environment (wind, solar, etc) and highlight the importance of energy consumption optimization.

Wireless networks consisting of a great number of autonomous vibration sensors are widely used for various IoT applications (industrial arrangements monitoring,

The reported study was funded by RFBR, project number 19-29-06043.

earthquake registration [4], structural health monitoring for critical infrastructure [2], etc.). While the energy consumption of the micro-controller unit and the wireless communication modules accompanying the sensors may be significantly reduced by employing triggered wake-up and data accumulation, the analog preprocessing equipment (preamplifiers, filters) has to operate uninterruptedly. Therefore, lowering of the energy consumption of the analog electronic parts is at the moment an important problem for vibration monitoring systems designer.

Most kinds of vibration sensors produce relatively weak signals. Due to this the signal amplification by the factor of 50–100 before filtration and conversion into digital form is necessary. Consequently, the power consumption rises significantly, because amplifiers are a major part of the analog preprocessing hardware in the systems discussed.

Advances in semiconductor technology have led to a substantial decrease of micro-circuit amplifiers power consumption. Unfortunately, it attended with a substantial rise of frequency dependent error occurred by the signal amplification [5].

In this paper, we propose a method for lowering processing errors by signals amplification in inverting scaling amplifiers.

2. Structures of scaling amplifiers

Preamplifiers for sensor signals are based as a rule on operational amplifiers (OA) with negative feedback circuit. The commonly used structure of scaling amplifiers (SA) consisting of the operational amplifier OA with negative feedback resistors R_{in}, R_{f} is shown. Such a configuration, shown in fig. 1 has the given amplification factor of $K_{\text{g}} = R_{\text{f}}/R_{\text{in}}$.

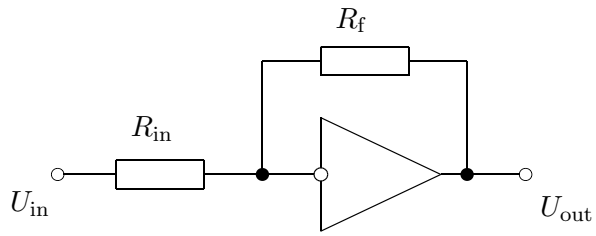


Fig. 1. Circuit of a commonly used scaling amplifier

In spite of rather steady performance, variable error components occur in this structure due to the signal processing, namely:

1) An amplitude error caused by the inaccuracy of the amplification factor and nonlinearity of the transfer function.

2) A frequency dependent error leading to high frequency components weakening in the output signal spectrum.

In some applications a zero shift in the preamplifier output signal has a negative impact on system performance.

To mitigate the first of the aforementioned errors it is enough to hold the ratio $R_{\mathrm{f}}/R_{\mathrm{in}}$ to the highest possible precision. The choice for precise resistive components on the market is sufficient. Additionally, regulated weak DC signal insertion to the SA input eliminates the output voltage shift.

Next we go on considering the frequency dependent error. Let us define the amplification factor in the structure shown in fig. 1 for the case when OA1 has a limited bandwidth in the high frequency region. On the OA in this and similar circuits, requirements of stability by deep degrees of negative feedback are imposed. Consequently, the transfer function for OA in Laplace operator form ought to be of first order: $K_0/(1+sT)$, where K_0 — direct current amplification factor of OA, T — time constant for OA as first order aperiodic section. Therefore, on the base of well-known relations for control systems with negative feedback the real amplification factor of the preamplifier (fig. 1) would be expressible as a transfer ratio for a system with negative feedback coefficient equal $R_{\mathrm{in}}/R_{\mathrm{f}}$:

$$K_r = \frac{K_0/(1+sT)}{1+K_0/(1+sT)(R_{\mathrm{in}}/R_{\mathrm{f}})}, \tag{1}$$

differing from the rated meaning K_{g} by the value:

$$\Delta K = K_{\mathrm{g}} - K_{\mathrm{r}}. \tag{2}$$

After the substitution $K_{\mathrm{g}} = R_{\mathrm{f}}/R_{\mathrm{in}}$ and some simplifications we can get from the equations (1) and (2) the equation for the amplification factors difference:

$$\Delta K = \frac{{K_{\mathrm{g}}}^2(1+sT)}{K_0}. \tag{3}$$

The difference defined above causes the occurrence of the error component in the output signal of the preamplifier. This component grows with the amplified signal frequency as follows:

$$\Delta U_{\mathrm{out}} = U_{\mathrm{in}}\Delta K = U_{\mathrm{in}}\frac{{K_{\mathrm{g}}}^2(1+sT)}{K_0}. \tag{4}$$

The ratio of the retrieved error component to the full output voltage value U_{out} equals

$$\frac{\Delta U_{\mathrm{out}}}{U_{\mathrm{out}}} = \frac{K_{\mathrm{g}}(1+sT)}{K_0}. \tag{5}$$

It is possible that the error voltage ΔU_{out} calculated from (4) is out of the allowable range defined by the system requirements, the type of amplifier chosen and the frequency needed. In the present state of electronic components, this is very possible because a tendency toward lowering amplifiers power consumption [5].

Let us consider as an example a low-power operational amplifier LTC2063 with supply current 2 µA, operating supply voltage 1.7... 5.25 V, gain-bandwidth product $GBP = 20$ kHz, $K_0 = 140$ dB [6].

Before the error calculation in accordance with the formula (4) it is necessary to find the time constant T for the chosen amplifier. For this purpose a well-known relation: $T = K_0/(2\pi GBP)$ may be used, of which after the substitution of the data presented above follows that $T = 79.6$ sec. Further calculation shows that the frequency dependent relative error for scaling unit with $K_{\text{g}} = 100$ based on the chosen OA is equal to 0.05 or 5% by the frequency 10 Hz and further grows with the frequency roughly linearly according to (5). Taking into account that in many cases the frequencies of registered vibration signals are up to 100 Hz, the value of the frequency dependent error calculated above is completely unsuitable.

The method for the extraction of the error component from the total output voltage of an inverting scaling amplifier was described in [7]. According to the method additive resistors equal to R_{in} and R_{f} are connected in series between the input and the output of the SA. The common point of the resistors is connected with the input of the additional non-inverting amplifier, that has an amplification factor of $K_{\text{g}} + 1$ (fig. 2).

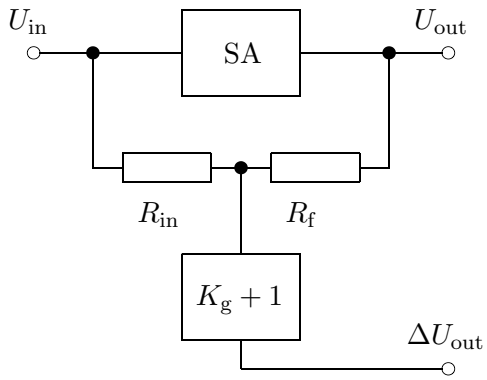


Fig. 2. Circuit for the scaling error extraction

On the output of the structure shown in fig. 2 occurs the error signal of the SA containing the error component ΔU_{out} defined in (4). It is clear that if the extraction

of this component is possible, a design of an additive negative feedback loop for the error suppression will be realizable — see the figure 3a.

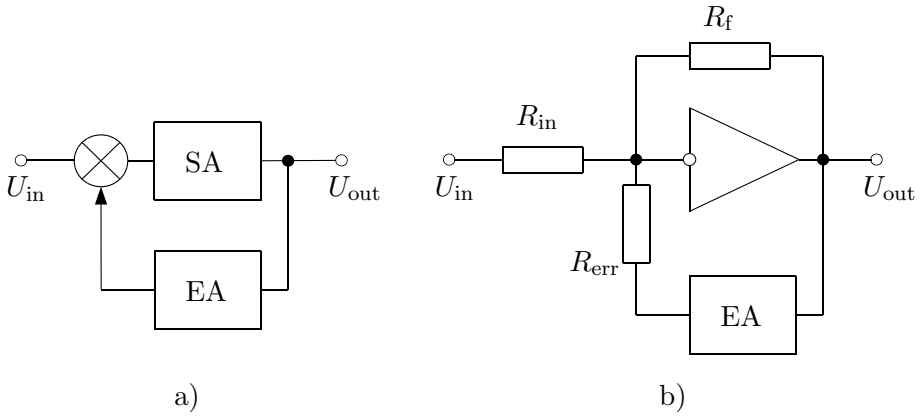


Fig. 3. Scaling amplifier with an error suppression loop

In this structure EA is an amplifier of the error signal in terms as used in control systems theory [8]. Additive resistors chain R_{in}–R_{f} in fig. 2 produces a reference signal for error component extraction. Resistor chain deteriorates this component ΔU_{out} by $K_{\text{g}} + 1$ times additive amplifier regains it. Therefore, the amplification factor of the EA is equal to one and as the only amplifying element in the loop, we can use a SA with an additive input.

The summing input for error signal may be achieved by the connection of additive resistor R_{err} to the summing point of the main OA (fig. 3b).

It is evident that in the structure presented in fig. 3b the amplification factor in the additive negative feedback loop is equal to $K_{\text{err}} = K_{\text{EA}} \cdot R_{\text{f}}/R_{\text{err}}$. As mentioned above the EA is intended only for the extraction of the error signal ΔU_{out} without amplification so that $K_{\text{EA}} = 1$ and the equation $K_{\text{err}} = R_{\text{f}}/R_{\text{err}}$ is valid. Application of the well-known expression for the error signal suppression in systems with a negative feedback to the structure in fig. 3b gives:

$$\frac{\Delta U'_{\text{out}}}{\Delta U_{\text{out}}} = \frac{1}{K_{\text{err}}} = \frac{R_{\text{err}}}{R_{\text{f}}}, \tag{6}$$

where $\Delta U'_{\text{out}}$ is the error component value in the presence of the additive negative feedback.

Consequently, controlling the value of the error signal suppression is possible by the choice of a proper R_{err} value.

3. Simulation

The modeling was carried out using LTspice XVII software to determine the frequency dependent error in the circuit shown in fig. 3b with and without a negative feedback. The manufacturer provided model of the micropower twin operational amplifier AD8502 [9] was used with the other components values as follows $R_{\text{in}} = 100$ kΩ, $R_{\text{f}} = 10$ MΩ, so that the rated amplification factor is $K_{\text{r}} = 100$. The operating supply voltage was set to ±2.75 V and the input signal amplitude to $U_{\text{in}} = 20$ mV.

Scaling error — the relative difference between the given value of the output voltage equal to 2 V and the actual voltage for varying frequencies we can see in fig. 4, curve a.

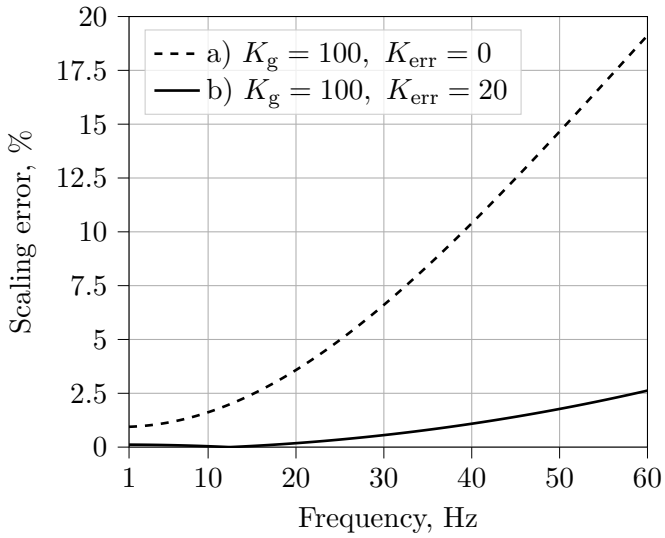


Fig. 4. Scaling error without and with error suppression loop

The similarly defined difference in presence of an additional negative feedback loop ($R_{\text{err}} = 500$ kΩ) is shown in fig. 4, curve b. Hence it follows that a negative feedback by the error voltage with depth mentioned above reduces the scaling error at 50 Hz frequency by the factor of 8.

4. Conclusion

The method of the error suppression presented in this paper allows a substantial reduction of the processing error in inverting scaling amplifiers. It should be mentioned that not only the frequency dependent error is suppressed by this method but also the total difference of the output signal from the rated value.

The main additional element required in order to realize the loop of error suppression described above is another operational amplifier. An important point is that this additional component may be equal in power consumption to the main amplifier. Therefore, an effective application of readily available twin micropower operational amplifiers is possible.

REFERENCES

1. Atzori L., Iera A., Morabito G. The internet of things: A survey //Computer networks. 2010. V. 54(15). P. 2787-2805.
2. Noel A. B. et al. Structural health monitoring using wireless sensor networks: A comprehensive survey //IEEE Communications Surveys & Tutorials. 2017. V. 19(3). P. 1403-1423.
3. Dudin A.N., Kim C.S., Dudin S.A. Optimal control by a node of wireless sensor network with quality of transmission depending on the amount of harvested energy //Distributed computer and communication networks: control, computation, communications (DCCN-2019). 2019. P. 29-36.
4. Kinoshita S. Kyoshin Net (K-Net), Japan //INTERNATIONAL GEOPHYSICS SERIES. 2003. V. 81(B). P. 1049-1056.
5. Kolombet E.A. Microelectronic means of analog signal processing. – Moscow: Radio and Communication. 1991. (in Russian)
6. https://www.analog.com/en/products/ltc2063.html
7. Babayan R.R., Morozov V.P. Bandwidth increase of an electromagnetic sensor signal amplifier. //Sensors & Systems. 2017. No. 3. P. 62-65 (in Russian)
8. Goodwin G. C. et al. Control system design. – Upper Saddle River, NJ: Prentice Hall. 2001.
9. https://www.analog.com/en/products/ad8502.html

UDC: 004.051

Data migration rate of the CRUSH-based distributed object storage with dynamic topology

A.B. Vanin[1,2], V.A. Bogatyrev[3], S.V. Bogatyrev[1]

[1]NEO Saint Petersburg Competence Center, Saint Petersburg, Russia
[2]ITMO University, Saint Petersburg, Russia
[3]Saint Petersburg State University of Aerospace Instrumentation, Saint Petersburg, Russia

alexey@nspcc.ru, vladimir.bogatyrev@gmail.com, stanislav@nspcc.ru

Abstract

Distributed systems are widely used to solve problems that require large computational or storage resources. The scalability of such systems makes them cheaper. However, the overhead for maintaining the system's performance or operation ability may be significant. This paper considers a distributed P2P storage system in uncontrolled dynamic environment. The change of the structure or the topology in such system can lead to data migration that may consequently cause system overload. This paper examines the intensity and the amount of these migrations for different CRUSH-based data placement approaches.

Keywords: Distributed system, storage system, simulation modeling, data migration, CRUSH, DHT, P2P

1. Introduction

The amount of generated and stored data is increasing every year [1]. This allows to develop new big data processing algorithms and train accurate neural networks. These research areas continues to improve because there is a place where all this big data can be stored. Due to technical limitations and economic reasons, storage systems virtualize their resources in hardware level: drives or magnetic tapes combined in clusters - as well as at operation level when data is spread across several storage nodes. These storage nodes may be geographically distributed [2, 3] to increase reliability [4].

Data should be securely stored and accessible to a variety of agents, that interact with it. Such agents are databases, web applications, raw data parsers, neural networks, etc. In order to unify interaction, modern storage systems store data

in the form of the objects. These objects contain data as a payload and provide additional meta-information, some context to it. Amazon S3 object storage interface has become de facto standard for clients working with such data. Other cloud storage providers [5] implementing this interface to be competitive on the market.

This paper considers a distributed object storage system as a set of storage peers. This approach allows to make cheap, horizontally scaling P2P systems [6, 7]. However, this imposes additional costs for maintaining data consistency in the storage. It must remain available, replicated and workload on storage nodes should be uniformly distributed.

The data load is controlled by a distribution algorithm (or distribution function) that selects nodes on which data must be stored. If network topology changes, data should be migrated in order to remain available for clients. DHT-based algorithms are widely used, in P2P network [8]. CEPH[9] object storage uses Controlled, Scalable, Decentralized Placement algorithm of Replicated Data (CRUSH)[10]. This paper evaluates the effectiveness of these approaches in a distributed decentralized object store. Efficiency criteria are the rate of data migration. Migration is a necessary overhead to maintain the consistency of the storage system. A large number of migrations can significantly reduce the efficiency of the entire system by reducing data availability or increasing the probability of storage node overload. The more migrations are in the system, the less likely client meets declared quality of service.

2. Program simulation of data migration

The intensity and amount of migrated data are examined in a simulation model. The model defines a set of nodes that have different attributes, or buckets in CRUSH terminology. Nodes are connected with a logical topology that determines how the system places objects in the storage nodes. Hash-ring[11] topology was used to examine DHT approach. In CRUSH, nodes are connected in a graph, where the vertices correspond to the attributes of the nodes, and the leaves are to the nodes themselves, see figure 1. Implementation of the CRUSH algorithm called network map [12].

Simulation is performing in two stages. The whole procedure is briefly described in algorithm 1. At the first stage, the model generates a stream of objects. Each object proceeds through a placement function that produces the set of storage nodes. These nodes have a finite capacity. Simulation stops at first storage failure: when there is no free space for the object in the storage node.

At the second stage, the topology is expanded with a few nodes, filled gray in figure 1. Each uploaded object from the first stage proceeds through placement function once again. If the locations at the second stage have changed, then the object migrated in the storage system.

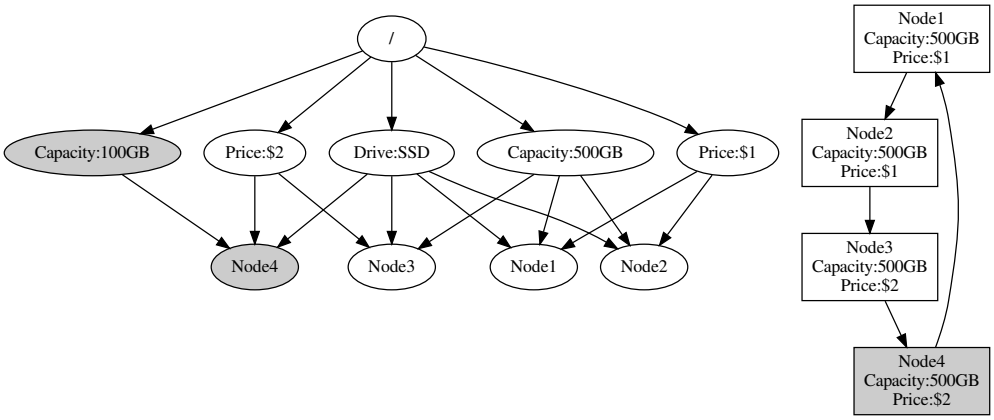


Fig. 1. Topology representation for CRUSH and hash ring

All experimental results presented in this paper provided with confidence intervals based on Student's T distribution with $\alpha = 0.05$.

3. CRUSH-based storage with dynamic topology

3.1. Topology in open storage systems. Paper [10] describes the CRUSH. This algorithm places objects in a distributed object storage system in a controlled and predictable way, achieving a uniform load distribution over nodes. Paper [12] proposes the use of the CRUSH in open object storage systems. Unlike proprietary closed systems, such systems work in untrusted dynamically changing environment without a single point of control over the network. Nodes could be provided by different unrelated authorities, therefore storage system becomes a platform for node owners to provide storage as a service.

Provide efficient data storage in such environment can be really tricky, but CRUSH defines weight coefficients for the storage nodes. In equation 1, CRUSH uses both hash distance from node i to the stored object x and the weight of node w_i itself to calculate a numeric characteristic of node c_i. Nodes with lowest c store the object.

$$c_i(x, r) = f(w_i)hash(x, r, i) \tag{1}$$

Storage system can determine weight function on its own. It could be represented as reputation value based on P2P interaction, node capacity parameter, or even stor-

Algorithm 1 Migration simulation

```
1: procedure MIGRATION(O, T_1, T_2)        ▷ Set of objects and two topologies
2:     Counter ← 0
3:     N ← length(O)
4:     for k ← 1 to N do
5:         P_1 ← Placement(T_1, O_i)        ▷ Set of storage nodes
6:         P_2 ← Placement(T_2, O_i)
7:         if P_1 ≠ P_2 then                ▷ If the placement has changed,
8:             Counter ← Counter + 1        ▷ increase the counter
9:         end if
10:    end for
11:    return Counter ÷ N
12: end procedure
```

age pricing [13]. For the experiments we defined weight function as a multiplication of normalized node capacity c and storage price p, see equation 2.

$$f(w_i) = \bar{c}_i \bar{p}_i \tag{2}$$

With this weight function, we placed objects in the model and expanded the topology. Results presented in figure 2.

While all nodes are the same, 2x topology expansion leads to the 50% object migration. However, node weights influence migration ration: if there are more profitable nodes in the system, objects will try to migrate there. Deviation from the non-weighted expansion can be controlled by normalization function: capacity weight is less significant than price weight in this example.

3.2. Migration ratio at different system load. In the first experiment, the model generated objects until storage failure: if the amount of stored objects is more than a capacity parameter. Overall system load was about 80% at the end of the first simulation stage. Model went from the transient state to the steady state. If we stop this stage at any specific system load limit, migration rate will be the same, see table 1. This allows to exclude system load parameter from the model and make it more robust.

3.3. General and specific storage policies. With the graph topology CRUSH allows to define flexible placement rules for the objects. Since every node has attributes, the data owner can filter some nodes by its attributes. We call such placement rules as *specific policies*. In figure 3 there is a specific policy to store data on nodes with attribute A. There are also *general policies*. They do not define any specific node attributes.

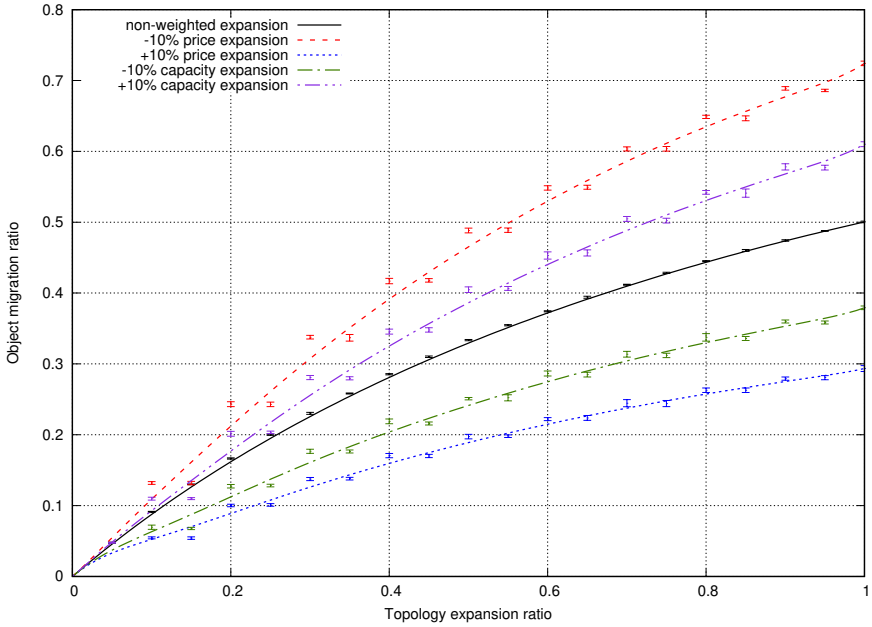


Fig. 2. Object migration ratio by nodes expansion with different weights

System Load	Migration ratio		
	25 % expansion	50 % expansion	100 % expansion
0.1	0.201 ± 0.003	0.331 ± 0.002	0.499 ± 0.004
0.2	0.201 ± 0.002	0.332 ± 0.003	0.500 ± 0.003
0.3	0.201 ± 0.002	0.332 ± 0.001	0.499 ± 0.002
0.4	0.200 ± 0.001	0.334 ± 0.001	0.499 ± 0.002
0.5	0.199 ± 0.001	0.333 ± 0.002	0.499 ± 0.002
0.6	0.200 ± 0.001	0.334 ± 0.002	0.500 ± 0.002
0.7	0.200 ± 0.001	0.333 ± 0.002	0.500 ± 0.001
0.8	0.200 ± 0.001	0.333 ± 0.001	0.500 ± 0.001

Table 1. Migration rate at different system load

Previously we expanded CRUSH topology in a symmetric way: at the 1.0 expansion ratio every node had a pair with the same attributes. But this is not very accurate in the real-world environment. We add new model parameter $\phi = \frac{E}{A}$ where E is a set of attributes in expanding nodes and A is a set of all attributes in the

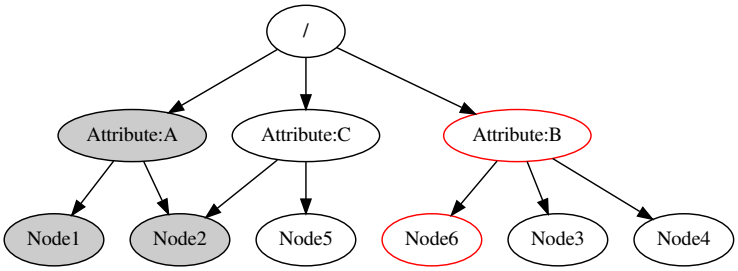


Fig. 3. CRUSH topology with specific policy (gray) and expanded node(red)

graph. In figure 3 there is a one expanded node with one out of three attributes, therefore $\phi = \frac{1}{3}$.

If the storage system store data by general policies, migration ratio does not affected by ϕ, see figure 4. However specific policies may significantly lower migration rate as soon as ϕ tends to zero. With $\phi = 1$ migration ratio will met worst case scenario with up to 50% migration if the topology graph size doubled.

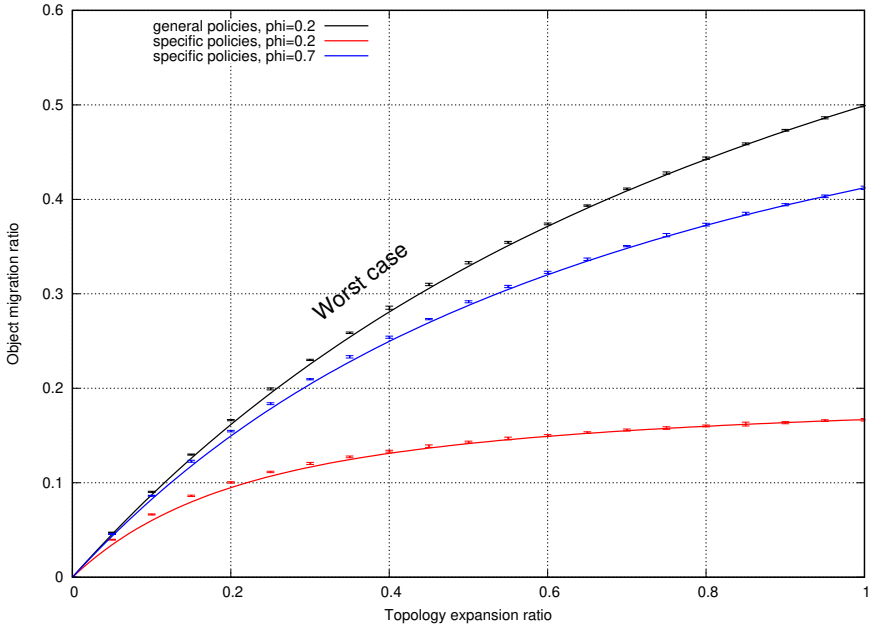


Fig. 4. Object migration ratio by nodes expansion with different ϕ and storage policies

3.4. DHT approach. DHT and CRUSH are based on the hash function that used to decide the location of the object. Despite the fact that CRUSH is designed to

predictably reduce randomness in the placement function, P2P networks still actively use DHT as a storage cache. To place the data, each node calculates a hash distance between hash of the data and the hash of the node (3). Nodes P with minimal hash distance will store the data (4).

$$distance(x, y) = |hash(x) - hash(y)| \tag{3}$$

$$N = \{N_1, N_2, \ldots, N_m\}, \min_{n \in N} distance(n, obj) = P \tag{4}$$

This called consistent hashing. CRUSH also uses consisting hashing after filtering nodes and attributes. Therefore DHT has the migration ratio results as CRUSH with general policies only. It is enough to build efficient distributed caches. They do not need active replication of stored data because cache miss is a regular situation. But flexible work with node attributes is mandatory for distributed object storage.

4. Conclusion

In this work, we built a simulation model for assessing the migration ration in a CRUSH-based distributed object data storage. With this model, the one can evaluate the effects of node weight coefficients to find the optimal weight and normalizing functions for a specific data storage system.

With this model we found out that the migration ratio does not depend on the system load. The more specific policies there are in the system, the less migration there could be. However, it depends on the structure of the network topology change, which in this paper defined by ϕ parameter. In practice the larger the system becomes, the less migration occurs in it. If the system has a small number of nodes, then it is recommended to apply new topology in several steps to avoid large migration waves.

Further research will be aimed at studying the influence of different normalization methods of node parameters for the CRUSH weight function. Performance evaluation will be done using this model.

REFERENCES

1. J. Paulsen, Enormous Growth in Data is Coming — How to Prepare for It, and Prosper From It (accessed June 1, 2020).
   URL `https://blog.seagate.com/business/enormous-growth-in-data-is-coming-how-to-prepare-for-it-and-prosper-from-it`
2. K. Spirovska, D. Didona, W. Zwaenepoel, Optimistic causal consistency for geo-replicated key-value stores, 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS) (2017) 2626–2629.

3. S. A. Noghabi, S. Subramanian, P. Narayanan, S. Narayanan, G. Holla, M. Zadeh, T. Li, I. Gupta, R. H. Campbell, Ambry: Linkedin's scalable geo-distributed object store, in: Proceedings of the 2016 International Conference on Management of Data, 2016, pp. 253–265.
4. V. Bogatyrev, Fault tolerance of clusters configurations with direct connection of storage devices, Automatic Control and Computer Sciences 45 (2011) 330–337. doi:10.3103/S0146411611060046.
5. Yandex Cloud. How to use API (accessed June 8, 2020).
URL https://cloud.yandex.ru/docs/storage/s3/
6. S. Ratnasamy, P. Francis, M. Handley, R. Karp, S. Shenker, A scalable content-addressable network, in: Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications, 2001, pp. 161–172.
7. S. Muralidhar, W. Lloyd, S. Roy, C. Hill, E. Lin, W. Liu, S. Pan, S. Shankar, V. Sivakumar, L. Tang, et al., f4: Facebook's warm blob storage system, in: 11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14), 2014, pp. 383–398.
8. I. Stoica, R. Morris, D. Karger, M. Kaashoek, H. Balakrishnan, Chord: A scalable peer-to-peer lookup service for internet applications, ACM SIGCOMM Computer Communication Review, vol. 31 31 (12 2001). doi:10.1145/964723.383071.
9. S. Weil, S. Brandt, E. Miller, D. Long, C. Maltzahn, Ceph: A scalable, high-performance distributed file system., 2006, pp. 307–320.
10. S. A. Weil, S. A. Brandt, E. L. Miller, C. Maltzahn, Crush: Controlled, scalable, decentralized placement of replicated data, in: SC '06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing, 2006, pp. 31–31.
11. D. Karger, E. Lehman, T. Leighton, R. Panigrahy, M. Levine, D. Lewin, Consistent hashing and random trees: Distributed caching protocols for relieving hot spots on the world wide web, in: Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing, STOC '97, Association for Computing Machinery, New York, NY, USA, 1997, p. 654–663. doi:10.1145/258533.258660.
URL https://doi.org/10.1145/258533.258660
12. A. Bogatyrev, S. Liubich, F. Wahle, S. Bogatyrev, A. Vanin, The model of network map and data placement in the distributed decentralized storage platform, in: Proceedings of the 10th Majorov International Conference on Software Engineering and Computer Systems, CEUR-WS, 2018.
13. Research Plan for Distributed Decentralized Blockchain-based Storage Platform (2018 (accessed November 5, 2019)).
URL https://github.com/nspcc-dev/research-plan/blob/master/research_plan.pdf

UDC: 004.75

Comparison of different methods for smoothing initial 2D data of the DSN-PC system's weather prediction algorithm

Ádám Vas[1] and László Tóth[2]

[1]Faculty of Informatics, University of Debrecen, Kassai str. 26, Debrecen 4028, Hungary

[2]Scitech Műszer Kft, Debrecen, Hungary

vas.adam@inf.unideb.hu, laszlo.toth@scitechmuszer.com

Abstract

Our Distributed Sensor Network for Prediction Calculations (DSN-PC) is a surface-based observational and computational network which is currently capable of calculating the change of an upper-air atmospheric parameter. The number of actually installed stations is limited thence we include data from the NOAA GFS database to create the 2D field of initial values for the prediction calculations. This hybrid application leads to numerical instability because some grid points get values from DSN-PC stations while others from NOAA GFS data. Previously we applied a smoothing algorithm based on moving average calculations. As presented in this paper, we applied two other methods. Each algorithm improved the numerical stability and prediction reliability. The type of the algorithm and the adjacency distance had a significant impact on the goodness of the forecasts. Below we compare these new results with the moving average method and the raw, unsmoothed case.

Keywords: sensor network, distributed computing, weather prediction, data assimilation, data smoothing, DSN-PC

1. Introduction

Since the 1990s, distributed sensor networks have been an actively researched and developed field. There have been a trend of moving from centralized to distributed systems consisting of cheap nodes which together are often capable of more complex tasks compared to the centralized ones. Such distributed systems are displacing the traditional systems at a significant rate [1].

A distributed sensor network is a collection of nodes that are distributed by a logical, spatial or geographical aspect and are connected to each other through

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

wired or wireless networks. These nodes can be equipped with many types of sensors (temperature, wind, air pressure, sound, light, magnetic field, acceleration etc.) and always contain a central unit used for signal processing, communication control and computational tasks. With the emergence and wild availability of high-speed networks, distributed sensor networks are used extensively in aerospace industry, automation, medical imaging, geology, weather prediction etc. The purpose of interconnecting these nodes can be the improvement of the data collection speed and reliability or the coverage of wider areas where central network availability is limited.

Recent technological advances have enabled the development of low-cost, low-power and multi-functional sensor devices. These are autonomous nodes with sensing, processing, and communication capabilities. Looking at a weather sensor network, the processing capabilities usually do not include mathematical computations, only signal processing and communication tasks. However, they can be used for distributed calculations of differential equation systems which generally are executed on central supercomputers of meteorological agencies.

Our goal with the DSN-PC system is to build a large distributed sensor network not only for the measurement of certain atmospheric parameters but also for weather prediction calculations [2]. This way the central supercomputer can be omitted from the whole system. Previously we followed a mixed approach by integrating our own DSN-PC nodes and NOAA GFS data into a hybrid sensor- and computational network [3]. After that we increased the involvement of our own measurements by directly inserting DSN-PC node measurements instead of the simultaneous interpolation [4]. This lead to problems due to significant differences (spikes) between adjacent grid points which caused numerical instability and incorrect results. This is a known issue in meteorology and several data assimilation [5, 6, 7] and data smoothing [8, 9, 10] techniques have been developed to address it. The spline methods have been the most widely used [11, 12, 13, 14, 15, 16] and distributed algorithms have been developed based on them [17]. Their applications in atmospheric and geosciences have shown their viability [18, 19, 20]. Before moving to these advanced methods we tried one simpler approach based on 2-dimensional moving average calculations [4] to see whether smoothing is enough to maintain numerical stability.

In this paper we compare the previous smoothing method with two others: a distributed moving median algorithm [8] and a Savitzky-Golay filter [10] - the latter not yet implemented in a distributed way thence executed in MATLAB. Below the results of the numerical weather prediction calculations are shown.

2. System and model description

2.1. Geographical properties. Currently our system implements a 20×20 size hybrid distributed sensor network which consist of 5 pieces of DSN-PC weather

stations with the rest being simulated as Java threads on a server computer. This network forms a regular grid over Europe using polar stereographic projection [21]. Fig. 1 shows the locations of the grid points. The detailed properties of the grid are:

- lower-left grid point coordinates: 39°N, 2.6°W
- upper-right grid point coordinates: 54.1371°N, 38.6715°E
- grid step at North Pole: 150 km
- central angle of the map: 0°

2.2. Input data sources and their assimilation. The initial values for the forecast calculations are taken from 2 sources: our 5 pieces of existing DSN-PC weather stations and publicly available data from the NOAA GFS-ANL database [22]. The DSN-PC weather stations in their present state can measure temperature, pressure and relative humidity [23]. The currently used forecast variable (500 hPa geopotential height) can be approximated based on these parameters using the hypsometric equation [24, 3]. From the NOAA GFS-ANL database we downloaded and applied the 0.5° resolution dataset which already contains the 500 hPa geopotential height values.

To calculate the initial data of the 20×20 grid, as a first step, natural neighbor interpolation [25, 3] was performed considering only the GFS-ANL grid points.

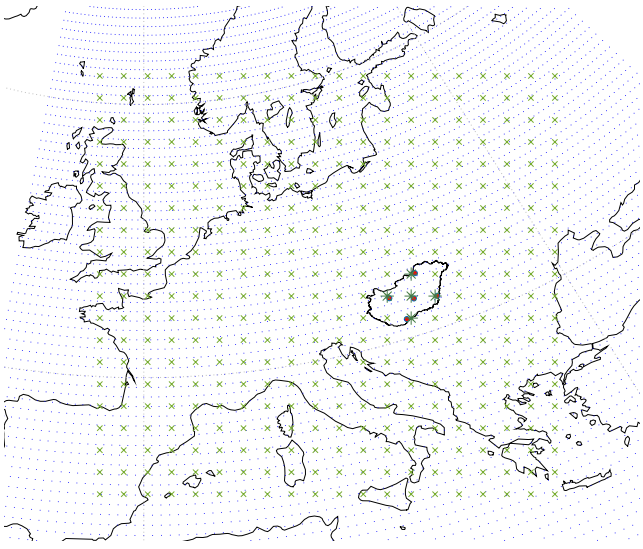


Fig. 1. The regular grid of the 20×20 computational network (x), the locations of the NOAA GFS dataset points (·), the locations of our DSN-PC weather stations (o) and the 5 grid points whose data were replaced with the nearest DSN-PC stations' measurements (*)

Before the interpolation the latitude($\varphi[^\circ]$) and longitude($\lambda[^\circ]$) coordinates of the grid points and the GFS-ANL points were converted to (x,y) coordinates based on polar stereographic map projection [21]:

$$r = \frac{\cos(\varphi)}{1 + \sin(\varphi)} \cdot 2a, \tag{1}$$

where

$$a = \frac{4 \cdot 10^7}{2\pi} \tag{2}$$

is the radius of the Earth (m). Then

$$x = r \cdot \sin(\lambda) \tag{3}$$

$$y = -r \cdot \cos(\lambda) \tag{4}$$

After the first step 5 grid points' initial values were replaced by their nearest DSN-PC stations' measurements. Table 1 shows the locations of the affected grid points and their respective DSN-PC stations.

2.3. The applied numerical weather prediction algorithm. In its present state the DSN-PC runs a relatively simple weather forecast model that is based on the barotropic vorticity equation originally developed by Charney, Fjørtoft, and von Neumann (CFvN) [21]. The original algorithm was refactored to a distributed form so that the nodes of the DSN-PC network are able to solve the equations in a fully distributed way [2]. In this article we cover the period between 21 March 2019 and 27 March 2019. Each day the 00:00 UTC measurements were chosen as initial values for the forecasts. We investigated the goodness of the forecasts by calculating the Mean Absolute Error (MAE):

ID	$\varphi[^\circ\text{N}]$	$\lambda[^\circ\text{E}]$	grid point $\varphi[^\circ\text{N}]$	grid point $\lambda[^\circ\text{E}]$
1	48.17	20.42	48.18	20.14
2	46.92	19.67	47.08	19.55
3	46.65	21.29	46.67	21.15
4	47.31	18.01	47.46	17.91
5	46	18.68	45.98	18.99

Table 1. The locations of our currently operational DSN-PC stations and the grid points whose values were replaced by DSN-PC measurements

$$MAE = \frac{1}{18 \cdot 18} \sum_{i=1}^{18} \sum_{j=1}^{18} |z_{500,i,j} - z'_{500,i,j}|, \tag{5}$$

where z'$_{500,i,j}$ is the predicted and z$_{500,i,j}$ is the measured 500 hPa geopotential height (m) 24 hours later.

2.4. Smoothing the initial data. During our forecast calculations numerical instability can occur. This happens because we use GFS-ANL and DSN-PC data simultaneously and GFS-ANL contains data that are already initialized, smoothed and interpolated onto a grid. However, DSN-PC measurements contain raw data. This hybrid approach leads to large differences between adjacent grid points which our simple CFvN model is not able to handle. Another reason is that it was originally designed for a larger geographical area, larger-scale atmospheric movements and larger distance between grid points. Also, DSN-PC measurements may contain measurement errors and may be affected by local weather phenomena.

Trying to address this problem we previously applied a 2-dimensional moving average algorithm which could smooth those large differences between adjacent points [4]. That approach lead to satisfactory results in terms of numerical stability and the goodness of the predictions. Presented in this paper, we applied two other methods: moving median and Savitzky-Golay filter.

We implemented the moving median algorithm in a distributed form. First, each node queries the initial values from its adjacent nodes over TCP/IP. Then based on the queried values they calculate the median of the whole dataset that also includes their own measurements. The adjacency distance (Np) varies between 1-3 hops. On Fig. 2 these adjacency distances are visualized on a 20×20 grid. On Fig. 3 the flowchart diagram of the moving median algorithm is shown.

The Savitzky-Golay filter [10] is yet to be refactored to a distributed form which will be applicable to DSN-PC nodes. In the meantime we used MATLAB's sgolayfilt function to calculate the filtered matrix from the raw grid data. As the function is not capable of 2-dimensional filtering, we applied two consequent steps by first going row-by-row then column-by-column:

```
matrixOut = sgolayfilt(matrixIn', degree, Np*2+1);
matrixOut = sgolayfilt(matrixOut', degree, Np*2+1);
```

Np=2 and Np=3 cases were applied in the case of the Savitzky-Golay filter. The degree of the fitted polynomial varied between 2 and (2*Np)-1.

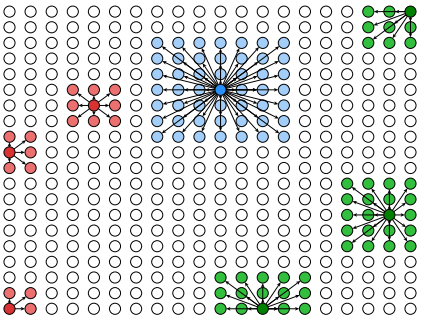


Fig. 2. Examples of moving median smoothing areas for different nodes with Np=1 (red), Np=2 (green) and Np=3 (blue)

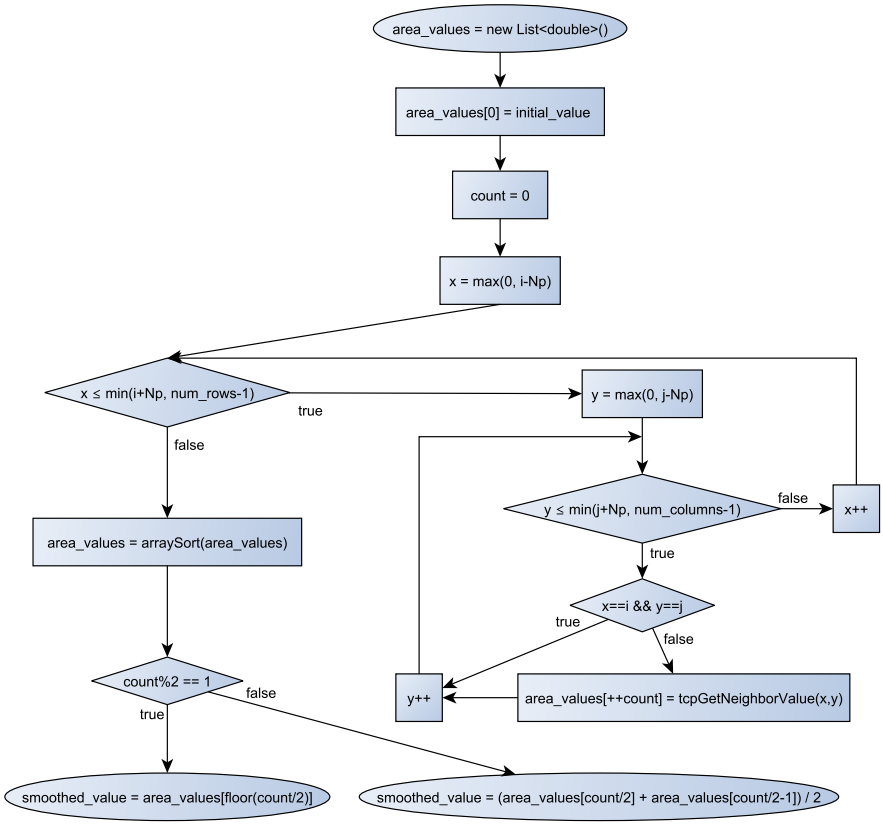


Fig. 3. The moving median algorithm as executed on a node at position (i,j)

3. Results

The MAE values of the forecast calculations are summarized in Table 2 and Table 3. Smoothing seems to be necessary at our current state of development to maintain numerical stability. The adjacency distance highly affects the goodness of the prediction calculations. A minimum of Np=2 seems to be necessary in the case of moving median. The Savitzky-Golay method produced satisfactory results in both cases (Np=2 and Np=3). Increasing the degree of the fitted polynomial doesn't result in significantly better forecasts in the investigated time period.

Comparing the moving average, moving median and Savitzky-Golay methods show the advantage of the Savitzky-Golay filter as it produces the smallest MAE values generally. However, this needs more review in the future by involving much more datasets as the current results containing 7 cases are not definite.

date	no smoothing	Np=1	Np=2	Np=3
21 March	NaN	NaN	25.55	54.63
22 March	NaN	191.65	43.44	65.23
23 March	81.19	60.83	95.01	NaN
24 March	NaN	64.47	111.53	71.29
25 March	89.95	210.41	145.99	58.42
26 March	NaN	194.47	76.89	100.65
27 March	NaN	199.53	78.78	50.23

Table 2. Mean Absolute Error (m) values of the forecast calculations performed by the CFvN algorithm on initial data smoothed using moving median method

date	no smoothing	Np=2		Np=3			
		2nd	3rd	2nd	3rd	4th	5th
21 March	NaN	22.65	48.71	20.36	34.21	48.94	56.13
22 March	NaN	45.66	44.84	46.07	44.75	44.02	44.58
23 March	81.19	64.95	70.72	65.22	61.12	72.77	73.17
24 March	NaN	59.96	31.38	59.24	42.00	37.38	51.07
25 March	89.95	158.96	81.24	158.37	113.86	79.86	98.40
26 March	NaN	38.85	47.39	37.05	42.39	48.53	50.03
27 March	NaN	56.72	44.26	50.39	47.03	35.40	44.18

Table 3. Mean Absolute Error (m) values of the forecast calculations performed by the CFvN algorithm on initial data smoothed using Savitzky-Golay filter with different polynomial degrees values

4. Conclusion

During our efforts for creating a highly autonomous weather prediction network we succeeded in moving one important step forward as now we are able to utilize DSN-PC station measurements as data sources for a weather prediction model by applying fully distributed smoothing algorithms to the nodes. Although the spatial coverage of our current network is not optimal as of now, involving public databases like NOAA GFS-ANL has proven to be a viable solution to address that problem. Since our measurements are not initialized and assimilated together with the GFS-ANL data it was necessary to apply 2-dimensional data smoothing methods to get reasonable results with the CFvN algorithm. Applying three different methods show the advantage of the Savitzky-Golay filter in terms of prediction reliability but its computational needs are the highest. Further investigation and comparison of these methods seems to be necessary to draw a general conclusion. Also, it's worth investigating whether there are some particularities in the certain initial fields that make one smoothing method more useful than the others.

5. Acknowledgements

We wish to thank Ficsor Endre, Perlaki Csaba, Szabó Sándor, Vas Ferenc and the Baptist Church of Kecskemét for providing place for our DSN-PC weather stations and thus supporting our research. We also thank SciTech Műszer Kft. for providing the possibility to use their facilities and for supporting our work financially; and Gábor Nagy for his contribution in designing, manufacturing and testing our weather stations' electronic circuits.

REFERENCES

1. S. S. Iyengar, R. R. Brooks, Distributed Sensor Networks : Sensor Networking and Applications (Volume 2), Chapman and Hall/CRC, 2016.
2. Á. Vas, Á. Fazekas, G. Nagy, L. Tóth, Distributed Sensor Network for meteorological observations and numerical weather Prediction Calculations, Carpathian Journal of Electronic and Computer Engineering 6 (1) (2013) 56–63.
3. Á. Vas, L. Tóth, Investigation of a Hybrid Sensor- and Computational Network for Numerical Weather Prediction Calculations, in: V. M. Vishnevskiy, D. V. Kozyrev (Eds.), Communications in Computer and Information Science, Vol. 1141, Springer International Publishing, Cham, 2019, pp. 510–523.
4. Á. Vas, O. J. Owino, L. Tóth, Improving the simultaneous application of the DSN-PC and noaa GFS datasets, Annales Mathematicae et Informaticae 51 (2020) 77–87.

5. W. Lahoz, B. Khattatov, R. Ménard, Data Assimilation and Information, in: Data Assimilation, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 3–12.
6. P. Lynch, X.-Y. Huang, Initialization, in: Data Assimilation, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 241–260.
7. E. Andersson, J.-N. Thépaut, Assimilation of Operational Data, in: Data Assimilation, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 283–299.
8. B. I. Justusson, Median Filtering: Statistical Properties, in: Two-Dimensional Digital Signal Prcessing II, Springer-Verlag, Berlin/Heidelberg, 1981, pp. 161–196.
9. C. H. Woodford, An algorithm for data smoothing using spline functions, BIT 10 (4) (1970) 501–510.
10. A. Savitzky, M. J. E. Golay, Smoothing and Differentiation of Data by Simplified Least Squares Procedures., Analytical Chemistry 36 (8) (1964) 1627–1639.
11. W. Freeden, M. Schreiner, Special Functions in Mathematical Geosciences: An Attempt at a Categorization, in: Handbook of Geomathematics, Springer Berlin Heidelberg, Berlin, Heidelberg, 2015, pp. 2455–2482.
12. W. Van Assche, W. Freeden, T. Gervens, and M. Schreiner, Constructive Approximation on the Sphere, with Application to Geomathematics, Journal of Approximation Theory 112 (2) (2001) 324–325.
13. W. Freeden, W. Törnig, On spherical spline interpolation and approximation, Mathematical Methods in the Applied Sciences 3 (1) (1981) 551–575.
14. V. Baramidze, M. J. Lai, C. K. Shum, Spherical Splines for Data Interpolation and Fitting, SIAM Journal on Scientific Computing 28 (1) (2006) 241–259.
15. W. Freeden, M. Z. Nashed, M. Schreiner, Spherical Harmonics Interpolatory Sampling, 2018, pp. 267–300.
16. L. L. S., G. Wahba, Spline Models for Observational Data., Mathematics of Computation 57 (195) (1991) 444.
17. Z. Shang, G. Cheng, Computational limits of a distributed algorithm for smoothing spline, Journal of Machine Learning Research 18 (108) (2017) 1–37.
18. M. F. Hutchinson, Interpolation of rainfall data with thin plate smoothing splines. Part I: Two dimensional smoothing of data with short range correlation, Journal of Geographic Information and Decision Analysis 2 (2) (1998) 139–151.
19. M. F. Hutchinson, Interpolation of Rainfall Data with Thin Plate Smoothing Splines -Part II: Analysis of Topographic Dependence, Journal of Geographic Information and Decision Analysis 2 (2) (1998) 152–167.
20. M. Kiani, Template-based smoothing functions for data smoothing in Geodesy, Geodesy and Geodynamics 11 (4) (2020) 300–306.

21. J. G. Charney, R. Fjørtoft, J. V. Neumann, Numerical Integration of the Barotropic Vorticity Equation, Tellus 2 (4) (1950) 237–254.
22. Global Forecast System (GFS) — National Centers for Environmental Information (NCEI) formerly known as National Climatic Data Center (NCDC).
URL `https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-forcast-system-gfs`
23. Á. Vas, G. Nagy, L. Tóth, Networkable Sensor Station for DSN-PC System, Carpathian Journal of Electronic and Computer Engineering 8 (2) (2015) 37–40.
24. J. M. Wallace, P. V. Hobbs, Atmospheric Thermodynamics, in: Atmospheric Science, 2006, pp. 63–111.
25. R. Sibson, A Brief Description of Natural Neighbour Interpolation, in: Interpreting multivariate data, 1981, pp. 21–36.

UDC: 519.872

Group Polling Method for Sensors Detecting in Unsynchronized Structured Wireless Monitoring Networks

I.I.Tsitovich[1]

[1]Institute for Information Transmission Problems (Kharkevich Institute) RAS,
Bolshoy Karetny per. 19, build.1, Moscow, Russia

cito@iitp.ru

Abstract

It is investigated the problem of detecting alarming sensors in large monitoring networks when there are objects where sensors can activate alarming signals simultaneously. We propose a generalization of the method of a sensor signal coding for an alarm signalization when a sensor signal can not be synchronized in time. This method bases on the method of group polling for alarming sensors identification for a synchronized network and has similar characteristics of complexity. Our methods has more complicated structure than previous ones. For a network with very large object we propose to use a sub-network of sensors at the object based on the Wi-Fi HaLow technology. For a network with middle size objects it is proposed the method with synchronized in time sensors at every object.

Keywords: wireless sensor network, sensor for an alarm signalization, group polling, unsynchronized time, heterogeneous sensors

1. Introduction

Studies of wireless sensors monitoring networks (WSNs) are widely conducted in different directions. One of the most interesting is the network organization which provides quick identification of a sensor that has an urgent message for sending to the situation center (SC). The lack of common communication protocols that ensure the interests of all users suggests that networks are too diverse and can be fundamentally different in their characteristics. One such aspect is covered in this paper, where it is suppose that probabilities of possible emergencies are very unlikely.

We examine the interaction of sensors with the SC via the WSN, where three nature sensor stages of activity are possible:

The publication has been prepared with the support of ... according to the research project No.AAAA-A19-119022590088-5.

I. the sensor transmits the information about its current state according to a given schedule sharing information with the SC;

II. the sensor detects an emergency and sends the alarm signal into the communication radio channel which is common to all sensors;

III. the sensor transmits information about the emergency at the request of the SC through the dedicated communication channel.

It is evident that for sensors in the first stage of activity we have a stable information flow such as the timetable of sensors activity is known. Number of sensors in the third stage of activity is small and their information flow is determined by the protocols of an emergency information.

The main interest consists in constructing such sensors and organizing the WSN in such way that it is possible to ensure energy efficiency of sensors, reliability, and timeliness of communications in case of a large number of sensors connected to a common wireless channel when sensors can be in the second stage of activity. The main feature is the following: most of the sensors does not send signal of the second stage for the rest of their lives. Therefore, it is not energy efficient to maintain continuous communication with such sensors. Since the number of active sensors (sending alarm signal) at the same time is small, it is advisable to allocate a common radio channel for them and, therefore, it is the problem to identify such sensors based on their common signal in the channel such as sensor signals are mixed.

In the paper [1], it was proposed the method of group polling for detecting of alarming sensors in the WSN network where properties of this method were investigated under the assumption of independent activity of the alarming sensors. It is supposed that the WSN is very large and contains thousands of sensors but all sensors synchronize in time their alarm signals. The last demand is difficult for its practical realization. But proposed in [1] method ensures the fulfilment of a short time of an alarming sensors detection, i.e. if t is a number of sensors in the WSN then the detection time is $O(\log t)$. In the paper [2] and [3], it is proposed a generalization of this method onto a case of unsynchronized in time alarming signal sending. It is showed that the group polling method for alarming sensors identification is applicable at this case but its computation complexity is such that it is difficult to use the method for online detection or it is very expensive. In [5] it is proposed a more complex profiles of sensors which give us possibility to detect alarming sensors in time similar to a WSN with synchronized in time alarming signal sending.

But it was supposed that the WSN consists with homogeneous sensors and only one of them is active in the case of an emergency at the object. Really we have WSNs where there are many objects with sensors for emergencies detection. If there is an emergency at the object then several sensors at this object begin to send the alarm signal for a real WSN. Such as a number of simultaneously sending sensors is

a critical for the stable detecting active sensors it is necessary to investigate WSNs with heterogeneous sensors. Sense of sensor heterogeneity consists in their dependent activity for the sensors at the same object where it has an emergency. This problem is investigated in [4] where it is supposed that all sensors at the object are active in the case of emergency at this object. The motivation of such approach lies in the fact that the analysis of the data from all sensors makes the possibility of more accurately characterize of the emergency. If it is transferred the data from the sensors that do not detect the emergency then it only leads to additional traffic in the communication channels and can be considered as an insignificant increase in the spending by the emergency detection. This assumption is valid only in the case when number of sensors at the object is small (this restriction was introduced in [4]). Now we suppose more real situation when number of sensors at the object can be large: a few tens or even thousands.

2. Setting of the problem

For the mathematical model description we follow the notations from [4]. The WSN consists of B objects, such that n_j sensors are mounted at the jth object, $j = 1, \ldots, B$. We assume that the number B is relatively large and $n_j \ll B$ for all j. Thus, we have in the WSN $t = \sum_{j=1}^{B} n_j$ sensors and develop the polling strategy aimed at the fastest identification of objects $j_1, \ldots, j_r$, whose sensors are ready for data transmission, and corresponding sensors.

The objects with numbers $j_1, \ldots, j_r$ are named as active objects and, therefore, it is necessary to find the set $S^o = \{j_1, \ldots, j_r\}$.

It was supposed in [4] that the data from all sensors at the object are received in the case of emergency at this object. This assumption is valid only in the case when n_j is small (this restriction was introduced in [4]). Such as we suppose now that n_j can be large we need to find alarming sensors also. Such sensors we name active ones. Their set is denoted by $S^s = \{i_1, \ldots, i_s\}$. Therefore, we have a more complicate problem than in [4].

In this problem we assume that the emergency probability at the jth object p_j is unknown but is relatively small:

$$p_j \leq p$$

where p is the predetermined probability, which is similar to the quantity $p = \frac{s}{t}$ from [5]. Now we suppose that $p = \frac{s}{B}$ and $s \leq 5$.

We assume that z_j sensors are simultaneously activated at the jth object in the presence of the emergency, so that z_j is a discrete random value distributed on the set $\{1, \ldots, n_j\}$. Therefore we have the set

$$S_j = \{i_1^j, \ldots, i_{z_j}^j\}$$

of active sensors at the jth object and the set

$$S^s = \cup_{j \in S^o} S_j$$

of all active sensors in the WSN.

The distribution of z_j depends on the object, the reasons for emergency, the positions of sensors, etc. These parameters are either unknown or difficult to take into account. But we assume that the probability of the activation more than one sensor at one object is relatively high. For example, if it is supposed that any sensor can be active with a conditional probability p_j^a independently of another sensors at the object when the object is active one then it is natural to suppose that

$$p_j^a \gg p. \tag{1}$$

In contrast with [4] when all sensors begin to send their signal simultaneously, now they start to send signals separately.

Let us outline in brief profiles of the sensor alarm signal as a combination of profiles from [4] and [5]. Every sensor has the unique code in [5] as a sensor at the object $\mathbf{o} = (o^1, \ldots, o^{M_o})$, and for the jth object its code is denoted by $\mathbf{o}_j$. Coordinates of the j-th code have the values 1 or 0. For creating the codes we construct the Boolean matrix $\mathbf{A} = (\mathbf{a}_i, i = 1, \ldots, B)$, where a_i^j are independent random numbers 0 or 1 with a proper probability p^0 for 1 in the matrix. The value p^0 will be specified by the formulas (5) and (6) in [5]. Also the sensor at the jth object has the unique code $\mathbf{s} = (s^1, \ldots, s^{M_j})$ where M_j depends on n_j. We propose several ways of such codes constructing in dependence of M_j and types of sensors communication.

The total code of the jth object's sensor is $\mathbf{a} = (o^1, \ldots, o^{M_o}, s^1, \ldots, s^{M_j})$. This code gives us the possibility to determine the number of sensor's objects and its number at this object.

The code $\mathbf{a}_i$ generates the profile for a signal of the ith sensor $a_i(u)$ by the formula (3) in [5]. The length of signal depends on the vector $\mathbf{o}_i$ and is denoted by $L_i\Delta$ where Δ is the length one time sampling interval. Under the conditions of the vector $\mathbf{o}_i$ constructing L_i is a random value with the mean $18 + (3 + 4p^0(1 - p^0))M_o$ (if we base on the recommended values of parameters in [5].

In the common channel signals from active sensors are mixed by the formula

$$f(u) = a_{i_1}(U_{i_1} + u) \vee, \ldots, \vee a_{i_{s^s}}(U_{i_{s^s}} + u),$$

where U_i is the time of the ith sensor activation and $\vee$ is the Boolean sum.

The resulting output continuous signal is dropped onto short time intervals with the length Δ; as a result, the continuous function $f(u)$ drops onto the group of

observations $(f_1, f_2 \dots)$, that can be 0, 1, or *nil* as in [5]), where *nil* means that the output signal can not be interpreted correctly.

We assume that data transmission errors are possible in the channel. This means that the value f_j is known with a certain error. Therefore, results 0 or 1 of f_j can be transformed in accordance with the matrix

$$\mathbf{W} = \begin{pmatrix} 1-\beta_0, & \beta_0 \\ \beta_1, & 1-\beta_1 \end{pmatrix}$$

and we get the vector of observation $\mathbf{g} = (g_1, g_2, \dots)$. If $f_j = nil$ then g_j equals 0 or 1 with probability 0.5 and, therefore, the observations $f_j = nil$ can not help for an alarming sensor detection.

Based on the vector $\mathbf{g}$ we detect active objects as in [5]. For detecting of the active sensors we propose coding methods and outline their in the next section.

Let $\hat{j}_1, \dots, \hat{j}_{\hat{s}^o}$ be detected emergency object numbers, $\hat{s}^o$ be the number of identified emergencies as in [4] and additionally $\hat{i}_1, \dots, \hat{i}_{\hat{s}^s}$ be detected active sensor numbers and $\hat{s}^s$ be the number of detected sensors.

The quality of algorithm is characterized by the probability of correct identification of emergencies. Let $\hat{S}^o = \{\hat{j}_1, \dots, \hat{j}_{\hat{s}^o}\}$ then P_1 is the probability of missing of the emergency object, i.e. the probability of the event $S^o \not\subseteq \hat{S}^o$, and P_2 is the probability of missing of the active sensor, i.e. the probability of the event $S^s \not\subseteq \hat{S}^s$.

3. Coding sensors in structured WSN

It is followed from (1) that the proposed in [5] group polling method is not applicable for coding sensors at the object if the object is relatively large, for example $n_j > 5$. For this cases we propose two different methods of communication sensors with the SC.

If objects of the WSN have little number of sensors then we can use the method from [5] applied for the active objects detection. All sensors on the detected objects are considered as active ones. This method is effective if $\mathbf{E}z_j \approx n_j$.

3.1. Structured WSN with middle size of objects. If $n_j < 100$ we propose binary orthogonal code design with sharing access time for sensors at the jth object. It means the following. All sensors at the object are synchronized in time. This means that a sensor can start to send signal in fixed times and the sequence of possible start times is common for all sensors at the object. In contrast with a synchronized in time WSN ([1], [4]) we have no large number of sensors and it is technically feasible requirement.

All sensors have the same part of the code that corresponds to the object and have different codes of the part that corresponds to the sensor. The length of the last part of the code is $2n_j$. Every sensor has its number at the objects $(0, \dots, n_j - 1)$;

let, for example, it is i. Then $\mathbf{s}_i = (s^1, \ldots, s^{2n_j})$ has two 1 in positions $2i+1$ and $2i+2$ and 0 in different positions.

The transformation of codes $\mathbf{s}_i$ into the sensor signal differs from (3) in [5]. This part of the signal is the following

$$a_i(u) = \begin{cases} 0 & \text{for } 0 \le u < 4i\Delta, \\ 1 & \text{for } 4i\Delta \le u < u_i + 4(i+1)\Delta, \\ 0 & \text{for } 4(i+1)\Delta L \le u < 4n_j\Delta. \end{cases} \tag{2}$$

It is followed from (2) that is case of the unmistakable signal transmission we detect 2 times symbol 1 in positions that corresponds to the ith sensor. Signals of another active sensors do not mix. We detect as active all sensors at the detected objects that the output signal $\mathbf{g}$ has at list one symbol 1 in the positions which correspond to the symbols 1 in the second part of the sensor's signal $a(u)$ in (2). Therefore,

$$\mathbf{P}(S^s \not\subseteq \hat{S}^s | S^o \subseteq \hat{S}^o) \le \beta_1^2 \sum_{j \in S^o} n_j$$

and the conditional mean number of detected sensors has the following inequality

$$\mathbf{E}(\hat{s^s} | S^o \subseteq \hat{S}^o) \le 4\beta_0 \sum_{j \in \hat{S}^o} n_j.$$

Using this formulas we can estimate P_2 if P_1 is known. Therefore the problem of estimating the quality properties of active sensors detection may be reduced to the problem of active objects detection that can be solved as in [5] if its method is applied to the active objects detection.

B	n_j	L	δ
1000	50	3546	5.6
5000	50	4387	4.6
5000	100	4387	9.2
10000	100	1260	8.5
25000	100	1395	7.7
100000	100	1620	6.6
100000	150	1620	9.9

Table 1. Increasing the alarm signal length of B and n_j when $\beta_0 = \beta_1 = 0.01$, $s_0^o = 2$

The proposed method, in addition to synchronizing sensors, increases the length of the alarm signal. The Table 1 shows the results of the the alarm signal length

calculation by numbers of objects B and sensors at the object n_j when $s_0^o = 2$. Here δ is the additional length of alarm signal generated by $\mathbf{s}$ as a percentage of L. It is followed from this data that for $n_j \leq 100$ or $n_j \leq 150$ for very large WSN the additional signal length and the time of active sensors identification grow less then 10%.

3.2. Structured WSN with large objects. One of possible way of the problem solving consists in using the Wi-Fi HaLow technology ([6]), based on the IEEE 802.11ah sensorndard. This way gives a possibility to organize a communication between sensors and the SC when an object has hundreds and even some thousands of sensors. The communication is set up via a Wi-Fi access point (AP). An AP is used for sending the alarm signal of any sensor at the object. If a sensor detects a danger then it sends to the AP a signal that it is ready to send an information to the SC. The AP send its profile signal in the common wireless radio channel. If the AP is detected by the SC then it sends an information from the active senors at the object to the SC.

For energy loss minimizing a sensor set-up a link with the AP at the moment of a danger's detection only. It is followed from (1) that a number of such sensors may be relatively large. In [7] it is studied the link set-up process in Wi-Fi HaLow networks, which consists of two main handshakes: authentication and association. Both handshakes are performed using Wi-Fi random channel access, the performance of which significantly degrades in case of a high number of contending sensors.

Such situation is typical for Wi-Fi HaLow networks because this technology has been designed as a version of Wi-Fi for the Internet of Things scenarios, so the Wi-Fi HaLow has two possible solutions to limit the contention for channel access, namely, Centralized Authentication Control and Distributed Authentication Control. The properties and comparisons of these methods are given, for example, in [7].

4. Conclusion

The main result consists in that for unsynchronized in time structured WSN can be constructed a group polling with $O(\log t)$ time for alarming sensors detection. A time of sensors detection may be similar to one for a WSN with independent activity of sensors as in [5].

It is proposed the method with partially synchronized sensors when objects in the WSN have less then 100 sensors and its effectiveness is investigated.

For a WSN with large objects it is proposed the method of sensors detection which is based on the Wi-Fi HaLow technology.

The computational complexity is growing no so much as in the case of [3].

The decoding procedure for unsynchronized in time structured WSN can be similar to one for a WSN with independent activity of sensors as in [5].

For a WSN with heterogeneous objects and sensors the emergency probability p_j can vary in a wide range. For this case preferably use codes with the signal profile length in dependence of p_j and this problem will be investigated later.

REFERENCES

1. Malikova E. E., Tsitovich I. I. Group Polling Upon the Independent Activity of Sensors in the Monitoring Networks // Journal of Communications Technology and Electronics. 2011. Vol.56. P. 1556–1563.
2. Tsitovich I. I., Shtokhov A. N. Method of Group Polling Upon the Independent Activity of Sensors in Nonsynchronized Monitoring Networks // Information Processes. 2016. Vol. 16. P. 237-245.
3. Shtokhov A., Tsitovich I., Poryazov S. On the Method of Group Polling upon the Independent Activity of Sensors in Unsynchronized Wireless Monitoring Networks // Communications in Computer and Information Science. 2016. Vol. 678. P. 266–278.
4. Malikova E.E., Tsitovich I.I. Analysis of the Efficiency of Group Polling Upon the Dependent Activity of Sensors in the Monitoring Networks. // Journal of Communications Technology and Electronics.2011. Vol. 56. P. 1552–1555.
5. Tsitovich I. Group Polling Method Upon the Independent Activity of Sensors in Unsynchronized Wireless Monitoring Networks // Distributed Computer and Communication Networks. 22nd International Conference, DCCN 2019, Moscow, Russia, September 23–27, 2019, Revised Selected Papers /Editors: Vishnevskiy, Vladimir M., Samouylov, Konsensorntin E., Kozyrev, Dmitry V. (Eds.). Communications in Computer and Information Science. 2019. Vol. 1141. P. 436–448.
6. Khorov E., Lyakhov A., Krotov A., Guschin A. A survey on IEEE 802.11 ah: An enabling networking technology for smart cities // Comput. Commun. 2015. Vol. 58. P. 53–69.
7. Bankov D., Khorov E., Lyakhov A. et al. What is the Fastest Way to Connect sensortions to a Wi-Fi HaLow Network? // Sensors. 2018. Vol. 18, no. 9. P. 1–23. URL: http://www.mdpi.com/1424-8220/18/9/2744.

UDC: 123.456

Regenerative estimation of $M/G/2$-type system with simultaneous service and speed scaling

R. S. Nekrasova[1,2]

[1]Institite of Applied Mathematical Research KarRC RAS, st. Pyshkinskaya 11, Petrozavodsk, Russia

[2]Petrozavodsk State University, Lenina 33, Petrozavodsk, Russia

ruslana.nekrasova@mail.ru

Abstract

We deal with a simultaneous service system under speed scaling policy. Arriving customer occupies a random number of servers, speed regimes are switched at arrival/departure instants according to the corresponding transition probability matrices. The paper concentrates on the particular case of 2 server system with general distribution of service times. As stability results were obtained for only for Markovian case, we rely on monotonicity properties to establish steady-state regime and apply regenerative confidence estimation.

Keywords: regenerative estimation, non work-conserving model, simultaneous service, speed scaling

1. Introduction

Simultaneous service models have a wide sphere of modern applications like high performance clusters, distributed computing or parallel computing systems, etc. The model under consideration admits idle servers, while the queue is not empty. Thus, service discipline is non work-conserving. That makes the analysis much more complicated (see, at instance [1], [2]). Stability criterion of a model with simultaneous service policy (under exponential assumptions) has been obtained in [3], authors based on a matrix-analytic approach. In more general cases we have to rely on assumptions or simulation.

Regenerative method is a powerful instrument in stochastic modelling and performance analysis. In recent work [4] regenerative confidence estimation was applied for mean queue size in $M/M/2$-type model with simultaneous service and speed scaling policy. Another application of regeneration theory for analysis of a simultaneous service model was presented in [5], where hypoexponential distribution of service time was considered.

In this paper we construct 2-server model with a general distribution of service time and speed scaling policy: both servers simultaneously switches speed regimes in arrival/departure instants. Under stability condition for dominating single server queueing model, we apply regenerative confidence estimation for original system with simultaneous service and illustrate confidence intervals for mean number of customers in the system, considering particular cases of Pareto and Weibull distributions of service time.

2. Description of a model

We construct 2-server queueing model with infinite buffer. Arrivals join the system at instants $\{t_n;\, n \geq 1\}$ according to Poisson input with a rate λ. The n-th customer is characterized by a pair of random parameters: $(S_n,\, C_n)$ which define the amount of work and a number of required servers, respectively.

Note, that current customer tries to capture $C_n \in \{1,\, 2\}$ servers simultaneously. If the resources are not available at instant t_n, the customer joins a queue (organised according to FIFO discipline) until the service is possible. Then after completion of the required work amount S_n, all C_n servers are seized and released simultaneously. The sequences $\{S_n;\, n \geq 1\}$ is independent and identically distributed (iid) and independent of an iid sequence $\{C_n;\, n \geq 1\}$. Denote the corresponding generic elements of such sequences by S and C, respectively. We consider *general* distribution of S and discrete distribution of C such, that $\mathsf{P}(C = k) := p_k$, $k = 1,\, 2$, where p_k are given probabilities.

Other significant feature of the system under consideration is a *speed scaling* technique: the servers can process L distinct speeds with "rates" $\mu_1 < \cdots < \mu_L$, and speed switching simultaneously affects to both servers. Namely, if the servers work at rate μ_i (in the i-th mode) a work amount S is completed in S/μ_i time interval.

We consider a particular case of $L = 2$ modes, and assume, that speeds is altered only at arrival/departure epochs according to (corresponding) Markov chains with transition matrices $||a_{ij}||$ and $||d_{ij}||$, $i,\, j = 1,\, 2$. Namely, speed μ_i may be switched to speed μ_j at arrival or departure instant with probability a_{ij} or d_{ij}, respectively. We assume $a_{12} = a$, $a_{22} = 1$, $d_{11} = 1$, $d_{21} = d$, where a, d are given probabilities. Thus, the system tries to keep the second (the faster) mode at arrival instants and the first mode at departure instants.

3. Regenerative approach

Regenerative method is a strong instrument in stochastic analysis. In this section we describe the main features of regeneration approach. First, define by $\nu_n \in \{0,\, 1,\, 2\}$ and $Q_n \in \{0,\, 1,\, \dots\}$ the number of customers on service and the queue size just before time instant t_n, $n \geq 1$, respectively. Note, that ν_n is an actual number of

customers on servers, which may not coincide with the number of busy servers. In particular, if at instant t_n^- both servers are occupied by the customer, that captured 2 servers, $\nu_n = 1$. The process, associated with total number of customers, is defined by $X_n = \{\nu_n + Q_n;\, n \geq 1\}$. We additionally assume, that system has zero initial state ($X_0 = 0$) and construct the following sequence:

$$\beta_k = \min_n\{n > \beta_{k-1} : X_n = 0\}, \quad k \geq 1,\, \beta_0 = 0. \tag{1}$$

Namely, $\{\beta_k, k \geq 1\}$ indicates actual numbers of customers, that arrive into totally empty system: $\nu_{\beta_k} = Q_{\beta_k} \equiv 0$. The system at instants t_{β_k} starts over in stochastic sense or *regenerates*. Random segments $\{X_n;\, \beta_{k-1} \leq n < \beta_k\}, \quad k \geq 1$ of a process X are iid, and $\{X_n;, n \geq 0\}$ is called a *regenerative* process.

Define by

$$\alpha_k = \beta_k - \beta_{k-1},\, k \geq 1 \tag{2}$$

the sequence of iid regeneration cycles length (with a generic length α).

Next define iid accumulated numbers of customers over each regeneration cycle by

$$Y_k = \sum_{i=\beta_{(k-1)}}^{(\beta_k)-1} X_i,\, k \geq 1.$$

Under the condition

$$\mathsf{E}\alpha < \infty, \tag{3}$$

the process X is called *positive recurrent* [6, 7], and the following limit exists:

$$r_n := \frac{\sum_{k=1}^n Y_k}{\sum_{k=1}^n \alpha_k} \to \frac{\mathsf{E}Y}{\mathsf{E}\alpha} = r, \qquad n \to \infty, \tag{4}$$

where Y is a generic element of a sequence $\{Y_k,\, k \geq 1\}$.

Note, that r_n coincides with an average number of customers in a system within interval $[0,\, t_{\beta_n})$:

$$r_n = \frac{1}{\beta_n} \sum_{i=1}^{\beta_n} X_i. \tag{5}$$

Actually, the result (4) means, that with a growth of cycle number, time average value of regeneration process converges to the ratio of mean cumulative value over cycle to mean cycle length. Namely, in case of positive recurrence, the behavior of regenerative process could be described by its cycle characteristics.

By other significant result from regeneration theory [8], we obtain, that if α is aperiodic, the following weak convergence holds

$$X_i \Rightarrow X^{(S)}, \qquad i \to \infty, \tag{6}$$

where $X^{(S)}$ is a *stationary analogue* of regenerative process X. From (6), we easily obtain

$$\lim_{n\to\infty} r_n = \mathsf{E}X^{(S)}. \tag{7}$$

By Proposition 4.1 from [9] the estimator r_n satisfies the following Central Limit Theorem

$$\sqrt{n}\big(r_n - r\big) \Rightarrow \mathbf{N}(0,\, \sigma^2), \qquad n \to \infty, \tag{8}$$

where

$$\sigma^2 = \frac{\mathsf{E}[Y - r\alpha]^2}{\big(\mathsf{E}\alpha\big)^2}.$$

Hence, if limits (4) and (6) exist, then weak convergence (8) holds and implies the following $100(1-\gamma)\%$ confident interval:

$$\mathsf{E}X^{(S)} \in \Big[r_n - \Delta_n,\, r_n + \Delta_n\Big], \qquad \Delta_n = \frac{z_\gamma \overline{\sigma}_n}{\sqrt{n}}, \tag{9}$$

where γ is a given reliability and

$$\overline{\sigma}_n^2 = \frac{n^2}{n-1}\frac{\sum_{i=1}^n \big(Y_i - r_n\alpha_i\big)^2}{\big(\sum_{i=1}^n \alpha_i\big)^2}.$$

(The value z_γ defines $(1-\gamma/2)$-quantile of the standard normal law and also could be obtained via Laplace function Φ as $z_\gamma = \Phi^{-1}(1-\gamma)/2$.)

A point estimator r_n coincides with mean average, which could be obtained without regenerative approach. Our goal is to apply regenerative method (RM) to build more informative interval estimator (9) for mean number of customers in considered system with speed scaling policy and simultaneous service.

First, we construct an auxiliary infinite buffer queueing system $M/G/1$ as follows: an input stream is Poisson with the same rate λ as in original system with speed scaling, and service times are iid and stochastically equivalent to S/μ_1 (time, demanded to serve a work amount S on the "slowest" regime in original system). By monotonicity properties [10] such a new system dominates the original system in a sense of load. Thus, its stability implies the stability of considered system with speed scaling.

Define a random generic inter-arrival time by τ and assume, that the following conditions hold

$$\rho := \lambda \mathsf{E}S/\mu_1 \quad < \quad 1, \tag{10}$$

$$\mathsf{P}\Big(\tau > S/\mu_1\Big) \quad > \quad 0. \tag{11}$$

Note, that $\mathsf{E}\tau = 1/\lambda$ and ρ defines a load coefficient in dominating queuing system. Remind considered zero initial state in original system: $X_0 = 0$. Under condition $\rho < 1$ the dominating system is stable. Hence, we obtain the stability of original model, which also yields, that the process $\{X_n\}$ is stochastically bounded. Then basing on results from regeneration theory and under (11) we derive the positive recurrence: $\mathsf{E}\alpha < \infty$, where α is a generic regenerative cycle length, defined in (2) (see [6] for details).

The condition (11) also implies, that with a positive probability there exist a regeneration cycle based the only arrival ($\mathsf{P}(\alpha = 1) > 0$), which means that α is aperiodic. Thus, (10)–(11) allow to apply RM for confident estimation of mean number of customers.

Note, that $\mathsf{E}X^{(S)}$ is bounded above by the mean number of customers in dominating system $M/G/1$ (denoted by $\mathsf{E}L$). The value $\mathsf{E}L$ is obtained by Pollaczek–Khinchine formula [11] as

$$\mathsf{E}L = \rho + \frac{\rho^2 + \lambda^2 \mathsf{D}S/\mu_1^2}{2(1-\rho)}. \tag{12}$$

4. Simulation

In this section we present simulation results for $M/G/2$-type system with simultaneous service and speed scaling for a few distributions of S under conditions (10)–(11). Note, that as input is Poisson, $\mathsf{P}(\tau > S/\mu_1) > 0$ automatically holds.

4.1. Pareto service. Consider Pareto distribution of S with a scale parameter fixed and equal to 1 and shape parameter denoted by $\mathcal{K}$. Thus,

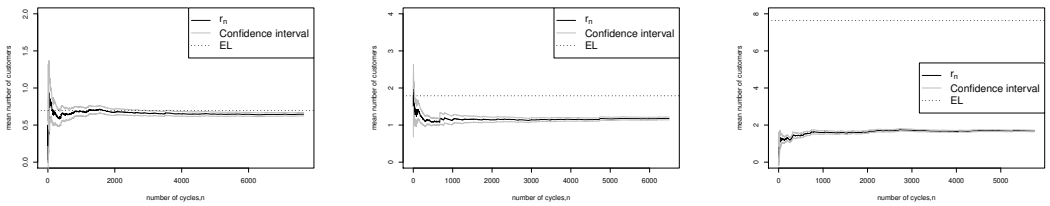
$$\mathsf{P}(S \le x) = 1 - x^{-\mathcal{K}}, \qquad x \ge 1, \tag{13}$$

and corresponding characteristics are defined for $\mathcal{K} > 2$ by

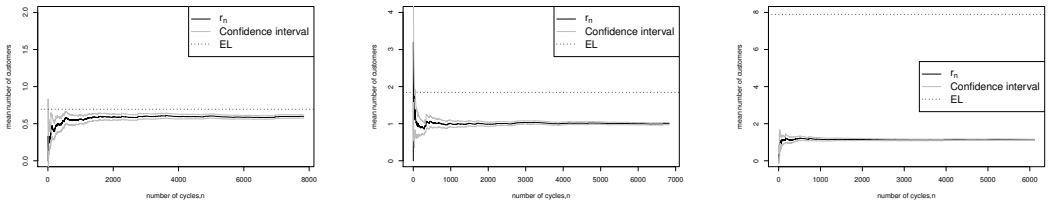
$$\mathsf{E}S = \frac{\mathcal{K}}{\mathcal{K}-1}, \qquad \mathsf{D}S = \frac{\mathcal{K}}{(\mathcal{K}-1)^2(\mathcal{K}-2)}. \tag{14}$$

Note, that in this case, dominating queuing system $M/G/1$ has Pareto distribution of service time with a scale parameter $1/\mu_1$ and shape parameter $\mathcal{K}$, while load coefficient is defined by

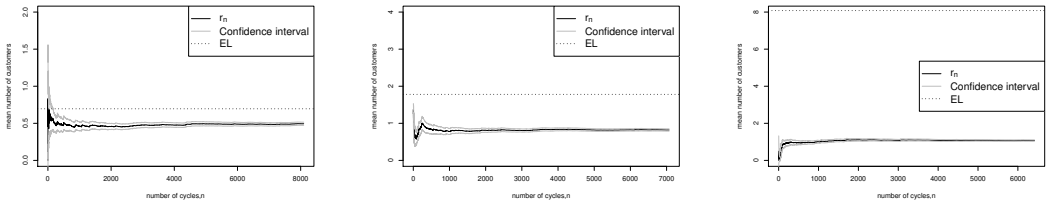
$$\rho = \frac{\lambda \mathcal{K}}{\mu_1(\mathcal{K}-1)}. \tag{15}$$



$\mu_1 = 0.9,\ \rho = 0.45$ $\mu_1 = 0.9,\ \rho = 0.70$ $\mu_1 = 0.9,\ \rho = 0.92$



$\mu_1 = 0.7,\ \rho = 0.45$ $\mu_1 = 0.7,\ \rho = 0.70$ $\mu_1 = 0.7,\ \rho = 0.92$



$\mu_1 = 0.5,\ \rho = 0.45$ $\mu_1 = 0.5,\ \rho = 0.70$ $\mu_1 = 0.5,\ \rho = 0.92$

Fig. 1. Confidence estimation of mean number of customers in $M/Pareto/2$-type speed scaling system, $\mathcal{K} = 3$.

The results for confidence estimation of $\mathsf{E}X^{(S)}$ in $M/Pareto/2$-type system with speed scaling and simultaneous service are illustrated on figure 1. Note, that presented experiments were done for fixed probabilities $a = 0.4$, $d = 0.6$, $p_1 = 0.5$, other results had shown that values of a, d, p_1 do not seriously affect to the accuracy of obtained intervals.

The first row on figure 1 corresponds to the system with "light" speed scaling: $\mu_1 = 0.9$, $\mu_2 = 1.1$. We based on 10 000 arrivals and obtained $n = 7672$ regenerations for the "light" load $\rho = 0.45$. In this case the accuracy of regenerative estimation $\Delta_n = 0.02185$, while the accuracy of estimation, based on monotonicity properties $\mathsf{E}L - r_n = 0.09176$. With a growth of load (the greater λ) the difference between dominating queueing system and original speed scaling system increases. Thus, for

the case $\rho = 0.95$ we obtained $\mathsf{E}L - r_n = 6.39651$, while regenerative estimation is much more accurate: $\Delta_n = 0.06655$.

The second and the third rows on figure 1 correspond to "medium" ($\mu_1 = 0.7$, $\mu_2 = 1.3$) and "large" ($\mu_1 = 0.5$, $\mu_2 = 1.5$) speed scaling, respectively. Smaller values of μ_1 force the difference in load between dominating and original systems, and estimation r_n by $\mathsf{E}L$ became less accurate. This fact is easily explained by statement (12). The increasing of such a difference is more notable in light load case (the first column on figure 1). Meanwhile, varying of parameters μ_1, μ_2, ρ does not strongly affect to the accuracy of interval estimator, obtained by RM. In particular, for $\mu_1 = 0.5$, $\mu_2 = 1.5$, $\rho = 0.92$, we obtained $\Delta_n = 0.03734$, $n = 6303$. Such a result illustrates the advantages of regeneration estimation (at least in comparison with bounds build basing on monotonicity properties).

4.2. Weibull service. For Weibull distribution of S we define a scale parameter fixed and equal to 1 and shape parameter denoted by w. Thus,

$$\mathsf{P}(S \le x) = 1 - e^{-x^w}, \qquad x \ge 0. \tag{16}$$

and corresponding parameters are defined via Gamma-function as

$$\mathsf{E}S = \Gamma\Big(1 + \frac{1}{w}\Big), \qquad \mathsf{D}S = \Gamma\Big(1 + \frac{2}{w}\Big) - (\mathsf{E}S)^2. \tag{17}$$

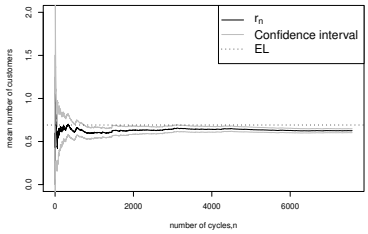
The dominating queuing system $M/G/1$ has also Weibull distribution of service time with shape parameter w, but scale parameter is equal to $1/\mu_1$ and. Its load coefficient is defined by

$$\rho = \frac{\lambda}{\mu_1}\Gamma\Big(1 + \frac{1}{w}\Big). \tag{18}$$

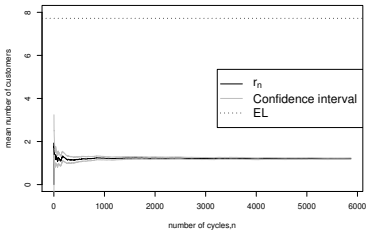
Numerical results for regenerative estimation of mean number of customers in $M/Weibull/2$-type model under consideration (for fixed values $w = 2$, $a = 0.4$, $d = 0.6$, $p_1 = 0.5$) are presented on figure 2. The first and the second rows correspond to light ($\mu_1 = 0.9$, $\mu_2 = 1.1$) and strong ($\mu_1 = 0.5$, $\mu_2 = 1.5$) speed scaling, respectively. Similar to the case of Pareto service, we obtained, that regenerative confidence estimation provides more accurate results (in comparison with theoretical bound $\mathsf{E}L$), specially with a growth of load coefficient.

5. Conclusion

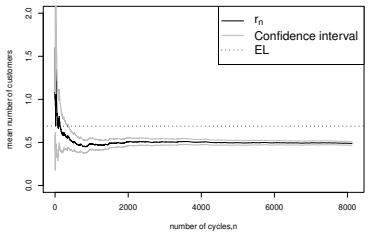
In this paper we considered a system with simultaneous service under speed scaling policy. Such a model could be useful in simulation of modern parallel computing systems. As stability results were obtained only in Markovian case, we relied on stability condition for the dominating queuing system. Basing on regenerative



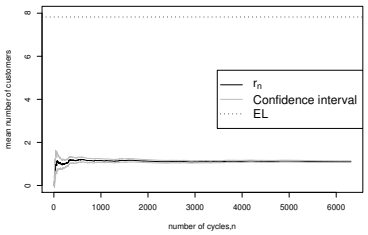
$\mu_1 = 0.9,\ \rho = 0.45$



$\mu_1 = 0.9,\ \rho = 0.45$



$\mu_1 = 0.5,\ \rho = 0.92$



$\mu_1 = 0.5,\ \rho = 0.92$

Fig. 2. Confidence estimation of mean number of customers in $M/Weibull/2$-type speed scaling system, $w = 2$.

structure of a system under consideration, we applied confidence estimation for mean number of customers in the $M/G/2$-type model in stable regime. Numerical results, obtained for different configurations, had shown the advantages of applied method in comparison with interval estimation, based on monotonicity properties.

6. Acknowledgement

The study was carried out under state order to the Karelian Research Centre of the Russian Academy of Sciences (Institute of Applied Mathematical Research KRC RAS).

The research is partly supported by Russian Foundation for Basic Research, projects 18-07-00147, 18-07-00156, 19-07-00303, 19-57-45022.

REFERENCES

1. Brill, P., Green, L. Queues in which customers receive simultaneous service from a random number of servers: A system point approach // Management Science. 1984. V. 30(1). P. 51—68.

2. Fletcher, G. Y., Perros, H., Stewart, W. A queueing system where customers require a random number of servers simultaneously // European Journal of Operational Research. 1986. V. 23. P. 331-–342.
3. Rumyantsev A., Morozov E. Stability criterion of a multiserver model with simultaneous service //Ann Oper Res.Jun. 2015. P. 1–11.
4. Nekrasova R., Rumyantsev A. Regenerative estimation of a simultaneous service multiserver system with speed scaling // Proceedings of the 26th FRUCT Conference. 2020. P. 346–351.
5. Afanaseva, L., Bashtova, E., Grishunina, S. Stability Analysis of a Multi-server Model with Simultaneous Service and a Regenerative Input Flow // Methodol Comput Appl Probab. 2019.
6. Morozov E. Weak regeneration in modeling of queueing processes // Queueing Systems. 2004. V. 46. P. 295–315.
7. Sigman K., Wolff R. W. A review of regenerative processes // SIAM Review. 1993. V. 35. P. 269–288.
8. Crane M. A., Lemoine A. J. An Introduction to the Regenerative Method for Simulation Analysis. Berlin, Springer-Verlag. 1977.
9. Asmussen S., Glynn, P. W. Stochastic Simulation: Algorithms and Analysis. Springer-Verlag New York. 2007.
10. Morozov E., Rumyantsev A., Peshkova I. Monotonicity and stochastic bounds for simultaneous service multiserver systems // 2016 8th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). 2016. P. 294–297.
11. Haigh J. Probability Models. Springer. 2002.

УДК:004.032.26:550.34.064

Нейросеть в составе станции сейсмического мониторинга

В.М.Воробьев[1] and Р.А.Дягилев[2]

[1]Научно-производственное объединение"Информационные и сетевые технологии",Москва
[2]Фиц Единая геофизическая служба РАН,
[1]svetazarobn@bk.ru
[2]dra@gsras.ru

Аннотация

В работе представлено исследование возможности обработки сигнала с выхода сейсмоприемника станции с помощью нейронной сети для оптимизации мониторинга сейсмических событий техногенного характера.Сверточная нейросеть обучалась на выборке записей сейсмичеких сигналов техногенного характера.Показана возможность нейросети различать сейсмические сигналы от взрывов в двух каменных карьерах с вероятностью 96 процентов.

Ключевые слова: *распределенные системы,нейронные сети,сейсмический мониторинг,сейсмограммы,глубокое обучение,взрывы в каменных карьерах*

1. Введение

Сейсмический мониторинг обеспечивает наблюдение за сейсмической активностью Земли, имеющей как природное, так и техногенное происхождение. В стране на базе ФИЦ ЕГС РАН [1]функционируют 12 региональных филиалов, расположенных во всех сейсмоопасных регионах РФ и обеспечивающих круглосуточную работу более чем 330 современных цифровых сейсмических станций.

Мониторинг сейсмичности многих горнодобывающих регионов Российской Федерации показал наличие мощной техногенной составляющей в её проявлениях. Специальные детальные наблюдения в районе шахт, рудников и карьеров позволили выявить ряд характерных особенностей в проявлении техногенной сейсмичности. В районах крупных горнодобывающих предприятий зафиксированы два типа техногенных сейсмических активизаций: 1) в виде распределенной по площади и по глубине сейсмичности, не имеющей прямой связи с горными выработками; 2) в виде более слабой сейсмичности, приуроченной непосредственно к горным выработкам. В[1]разработана технология локального мониторинга природно-техногенной сейсмичности, сопровождающей разработку различных

месторождений полезных ископаемых. Технология предусматривает установку локальной сети наземных и (или) подземных сейсмических станций в пределах исследуемого объекта, обеспечивающей регистрацию в контролируемом массиве широкого спектра сейсмических событий малых магнитуд (–2<ML<4):взрывы, подвижки, обрушения и т.п. Результатами наблюдений являются карты эпицентров сейсмических очагов, графики динамики развития сейсмичности в пределах объекта, параметры сейсмического режима и прогнозные зоны повышенного сейсмического риска и др. Использование современного регистрирующего оборудования, высокоскоростных средств связи и эффективных методов обработки данных позволяет осуществлять мониторинг и передачу его результатов Заказчику в режиме, близком к реальному времени. Например, технология позволяет по непрерывным записям сейсмостанций, расположенных в окрестностях ГЭС, решать важные задачи диагностики объекта: осуществлять дистанционный вибрационный контроль состояния работающих гидроагрегатов, выявлять режимы повышенных вибраций, выполнять мониторинг технического состояния плотин по изменениям собственных частот сооружений, получать дополнительную объективную информацию, необходимую для расследования причин возникновения нештатных ситуаций на гидроэлектростанциях. Эффективность работы такой распределенной системы напрямую зависит от плотности размещения сейсмостанций. Требуемая плотность тем больше, чем меньшее значение магнитуды, связанной с той или иной нештатной ситуацией на объекте наблюдения. Наряду с увеличением плотности станций для решения задач оперативной обработки информации,что сопряжено с большими расходами на создание новых сейсмостанций, предлагается использовать дополнительную обработку сигналов на станциях в оперативном режиме на основе применения алгоритмов нейронных сетей. Для апробирования такого подхода была создана и обучена нейросеть глубокого обучения на выборке сейсмограмм от взрывов в двух каменных карьерах.Показана возможность эффективного обучения нейронной сети на сравнительно небольшой выборке(около сотни оригинальных сейсмограмм)с генерацией искусственных данных. Кроме того,для различения техногенного источника сейсмичности от других сейсмограмма преобразовывалась в двумерное изображения. Различение изображений сейсмограмм реализовывалось с помощью алгоритмов компьютерного зрения.Показана возможность получения большей полезной информации по данным одной сейсмостанции.Это может в дальнейшем уменьшить требования к плотности наблюдательной сети при той же эффективности мониторинга.

2. Постановка задачи

Целью дополнительной обработки данных на станции является опознание каких-то признаков нештатной ситуации (в случае с гидростанциями),вызвавшей

возникновения сейсмического сигнала. Особенности такого сигнала могли бы свидетельствовать о том, какая именно ситуация на каком-то из объектов. В работе предлагается использовать возможности в преобразовании сигнала с выхода сейсмоприемника с помощью нейронных сетей глубокого обучения. Особенностью этой задачи в нашем случае является то, что необходимо провести обучение нейросети на небольшом объеме данных. Как известно для хорошо работающей нейросети ее обучают на большой выборке с десятками тысяч и более данных.Перед авторами встал вопрос о расширении данных добавлением новых, полученных из оригинальных данных некоторым преобразованием(так называемая аугментация на языке нейронных сетей глубокого обучения). Такой подход применялся в работе[2],в которой генерировались искусственные звуковые сигналы для обучения нейросети.В [3], исходя из того, что оригинальные сигналы представляют собой сумму сигнала от взрыва и шума, искусственные данные получали как сумму из оригинальных значений сигналов ,уменьшенных на величину “a”, и случайно сгенерированного числа с средним нулевым и квадратичным отклонением "a". Для определения особенностей сигналов от взрывов на сейсмической станции можно рассчитать спектральный состав сейсмического волнового пакета от взрыва на основе натурной записи скорости сейсмосмещений в породном массиве при массовом взрыве групп скважинных зарядов. Анализ полученных таким образом спектров показывает существенный разброс в них от взрывов в одном и том же карьере.

3. Данные для обучения нейросети

Данные для обучения нейросети предоставлены по результатам регистрации сейсмических волн от взрывов сейсмостанции Романово. При проведении взрывных работ в каменных карьерах возникает сейсмическая волна. Сигнал от взрыва приходит на станцию сильно зашумленный.Были предоставлены цифровые трехкомпонентные(x,y,z) записи сигналов с выхода сейсмоприемника за несколько лет.Всего в базе 93 записей, которые отмечены как сигналы от взрывов Заготовкинского и 89 записей Утесовского карьеров. Заказчик предоставил время взрывов, рассчитанное по его методике обработки записей сигналов.Зная время в очаге и скорость распространения P и S волн можно ориентировочно определить окно данных, связанное с “полезным” сигналом.Всего было отобрано по 50 окон для train,20 для валидации и 15 для тестирования. Из формата .W данные преобразованы в формат .txt.Для обучения использовались записи вертикальной компоненты z.

4. Подход к решению

Для классификации принадлежности сигнал сигналу от взрыва необходимо каким-то образом преобразовать входные сигналы. С этой целью используют получение спектра с помощью преобразования Фурье[4]. Работы по анализу спектров сигналов от взрывов в карьерах [5]показывают большой разброс для взрывов в одном карьере. Существует другой подход к преобразованию сейсмического сигнала с помощью вейвлет-преобразований (или в русско-язычной терминалогии – всплеск-преобразований). Базисы всплесков имеют ряд преимуществ по сравнению с другими базисами, используемыми в качестве аппарата приближения функций. Они обладают так называемой время-частотной локализацией, т. е. быстро убывают на бесконечности как сами базисные функции, так и их преобразования Фурье. Благодаря этому свойству при разложении по базису сигналов, частотные характеристики которых меняются по времени или по пространству (таковыми являются, в частности, речевые или музыкальные сигналы, сейсмические сигналы, а также изображения), много коэффициентов разложения при ненужных на данном пространственном или временном участке гармониках оказываются малыми и могут быть отброшены, что обеспечивает тем самым сжатие информации. Допустимость такого отбрасывания объясняется другим важным свойством: всплеск-разложения являются безусловно сходящимися рядами. Кроме того, существуют эффективные алгоритмы, позволяющие быстро вычислять коэффициенты всплеск-разложений. При применении всплеск преобразований входного сейсмического получается двумерное изображение, что позволяет использовать потенциал сверточных нейронных сетей. Идея использования моделей машинного обучения в нашем случае заключается в поиске соответствующего представления входных данных, которое сделают данные более пригодными для решения такой задачи, как классификация. Известно[6], что глубокие нейронные сети превращают исходные данные в результат, выполняя длинную последовательность простых преобразований (слоев), и обучаются этим преобразованиям на входных данных. В данном случае, обучение предполагалось проводить на записях сейсмических сигналов от взрывов в карьерах. Модели глубокого обучения могут обучаться на дополнительных данных без полного перезапуска, что делает их пригодными для непрерывного и продолжительного обучения — очень важное свойство для промышленных моделей. Кроме того, обучаемые модели глубокого обучения можно перенацеливать и, соответственно, использовать многократно: например, модель, обученную классификации изображений, можно включить в конвейер обработки видео. Это позволяет использовать предыдущие наработки для создания все более сложных и мощных моделей. Это также дает возможность применять глубокое обучение к очень маленьким объемам данных.Поставленная задача связана с процедурой бинарной

классификации. Исходя из исследований параметров сигналов от взрывов[5] была выдвинута гипотеза о том, что для различения сигналов от взрывов необходимо получить оценку изменения спектрального состава сигнала во времени. С этой целью применен аппарат wavelet преобразований. Рассмотрены пять типов вейвлетов библиотеки pywavelet [6]: scg.set$_d efault_w avelet('cmor1 - 1.5')$

scg.set$_d efault_w avelet('cgau5')$

scg.set$_d efault_w avelet('cgau1')$

scg.set$_d efault_w avelet('shan0.5 - 2')$

scg.set$_d efault_w avelet('mexh')$

После проведения визуального просмотра преобразований в качестве основного выбран default$_w avelet('cmor1 - 1.5')$

Для расширения dataset генерировались дополнительные окна событий, добавлением к реальным данным окон случайных данных с нулевым средним с отклонением k *std , где k коэффициент ,std –значение амплитуды сигнала взрыва в данный момент времени. В работе использовался коэффициент к=0.25. В целях еще большего эффекта от расширения данных производилось подача .png файлов в директориях train, test, val в ImageDataGenerator, который настраивает генераторы Python для автоматического преобразования файлов с изображениями в пакеты готовых тензоров на вход модели нейронной сети Модель представляет собой глубокую сверточную сеть, которая принимает результат wavelet преобразования сейсмограмм z компоненты от взрывов двух карьеров в качестве входных данных и прогнозирует их метку либо как сигнал от взрыва на одном карьере, либо как событие от взрыва на другом карьере.

Параметры сети оптимизированы для минимизации расхождений между прогнозируемыми метками и истинными метками в обучающем наборе. Сеть принимает на входе тензоры с формой (высота изображения, ширина изображения, каналы) (не включая измерение, определяющее пакеты). В данном случае сеть настраивалась на обработку входов с размерами (64, 64, 3).Размер задавался в экземпляре класса ImageDataGenerator. Значения, помещаемые в тензоры, масштабировались и приведены к меньшим величинам в диапазоне [0, 1]; Сначала была сконструирована модель, обладающая эффектом переобучения. Для чего добавлялись слои. Задавалось большое количество параметров в слоях. Модель тестировалась на большом количестве эпох. Тестировались разные архитектуры: добавлялись и удалялись слои; разные гиперпараметры (число нейронов на слой ,шаг обучения оптимизатора, параметр Dropout (от 0.2 до 0.6), количество фильтров, разные оптимизаторы: RMSProp, Adam,SGD и их параметры(lr,например), чтобы найти оптимальные настройки; дополнительно выполнялся цикл конструирования признаков : удалялись синтетические данные (т.е . S –случайный процесс со средним нулевым) и обучение проводилось с изменением параметров

экземпляра класса ImageData Generator для автоматического преобразования файлов с изображениями в пакеты готовых тензоров.

5. Заключение

Дополнительная обработка сигналов с сейсмоприемника одной станции распределенной системы с использованием сверточной нейронной сети позволила с вероятностью 96 процентов отличить сейсмические сигналы от взрывов в двух близлежащих друг от друга карьерах. Преобразование входного сигнала в двумерное изображение и расширение выборки обучения за счет генерирования дополнительных данных позволяет использовать последние достижения в распознавании образов сетями глубокого обучения. При этом сама обученная сеть не требует больших вычислительных мощностей и может в дальнейшем использоваться в обработке сейсмического сигнала на сейсмостанции. Это может помочь в увеличении эффективности мониторинга без увеличения плотности станций распределенной системы наблюдения за техногенной обстановкой.

6. ЛИТЕРАТУРА

1.Единая геофизическая служба РАН 2017 [http://www.gsras.ru]

J. 2.Salamon, J. P. Bello, Deep convolutional neural networks and data augmentation for environmental sound classification. https://arxiv.org/abs/1608.04363 (2016).

3.Thibaut Perol, Michaël Gharbiand Marine Denolle Convolutional neural network for earthquake detection and location https://advances.sciencemag.org/content/4/2/e1700578

4.Методы и средства изучения быстропротекающих процессов В. Шкуратник, А. Вознесенский, И.

5.О влиянии массового взрыва в карьере строительного камня на формирование спектра сейсмических волн В. Н. Опарин и др.

https://pywavelets.readthedocs.io/en/latest/

UDC: 004.7

A methodology for adapting wireless channel resources to the load by switching between medium access protocols

M.A. Rudenkova[1], H. Khayou [2], L.I. Abrosimov[3]

[1,2,3]National Research University "Moscow Power Engineering Institute",Krasnokazarmennaya 14, Moscow, Russia

RudenkovaMA@mpei.ru, hussein.khayou@gmail.com, AbrosimovLI@mpei.ru

Abstract

The enterprise WLANs networks use a wireless channel to transmit data from network applications. The most popular enterprise applications are multimedia applications, such as Voice over IP, videoconferencing, telepresence and remote desktop. The multimedia applications create challenges and demands on wireless channel resource management. It is very difficult to manage contention of network application flows for wireless channel resources in order to improve the end-user experience.

The purpose of this paper is to provide a methodology for obtaining a specific metric that can describe end-user perception of the network performance. This metric allows us to compare the performance of a wireless channel with various media access protocols under certain conditions. Switching between different media access protocols can improve wireless performance, resulting in a better end-user experience.

Keywords: PCF, DCF, quality, methodology for compare shared media access protocols, wireless networks, WLAN

1. Introduction

Wireless local area network (WLAN) is the most popular technology for providing access to Internet services to various user devices in the Enterprise. However, users may experience some inadequacy while connecting their network applications to the WLAN especially while using multimedia applications. This problem is due to the shared nature of the wireless media, which leads to the performance decrease under heavy load. The load depends on the number of users and the intensity of network applications traffic. The statistical parameters of the packet rate passing through the wireless channel does not show how much the user wants to send, so it is

difficult to predict the guaranteed service rate of network application traffic. When the load increases, the time needed to transmit one network packet from the network application is changed. This time is one of the criteria for QoS estimation [1].

There are many publication on estimating the delay, throughput or performance of the IEEE 802.11 wireless channel for a set of configuration parameters and WLAN characteristics [2, 3, 4, 5, 6]. The authors of [7] [8] examine the performance and delay of the wireless channel for a set of configuration parameters and different WLAN characteristics, but it is very important to develop a methodology and a metric for comparing the performance for different sets of WLAN characteristics and wireless channel configuration. Therefore, we provide a specific metric, which is the guaranteed rate of network application traffic in a wireless channel with a specific set of configuration parameters and WLAN characteristics. This metric makes it possible to compare the wireless channel under different conditions using an analytical model, statistics or simulation results. In this paper the results are obtained using time-driven modeling with network simulator ns-3. Also, the methodology to compare different shared media access protocols under different conditions is provided.

2. Characteristics of Wireless Local Area Network

We study WLAN in infrastructure mode for Enterprise networks which contains an access point AP , wireless stations STA ($k = \overline{1, K}$) and an internal corporate LAN as Switch and corporate LAN resources as Server (Fig. 1). Each STA has its own set of network applications and application traffic intensity. AP has a specific wireless channel with shared media access control protocol CF ($CF = [DCF, PCF]$).

The WLAN under study has the following characteristics. The network applications utilize the network by sending packets with an average length l [bit/packet]. The intensity of network application traffic determines the time between network packets θ_k which is assumed to have an exponential distribution. Total network application traffic intensity Λ in the wireless channel is:

$$\Lambda = \sum_{k=1}^{K+1} \lambda_k \tag{1}$$

The main feature of processing network packets in a wireless channel is the service intensity u_k. Network application has a guaranteed time for network packet delivery T_g and the traffic which has a waiting time higher than T_g will be dropped. The intensity of this traffic is $\Delta\lambda_k$ and thus we have the following equation:

$$\lambda_k = u_k + \Delta\lambda_k \tag{2}$$

The total network application traffic intensity Λ can be separated into two parts: the flow of network packets which is serviced with an intensity u and the flow of network

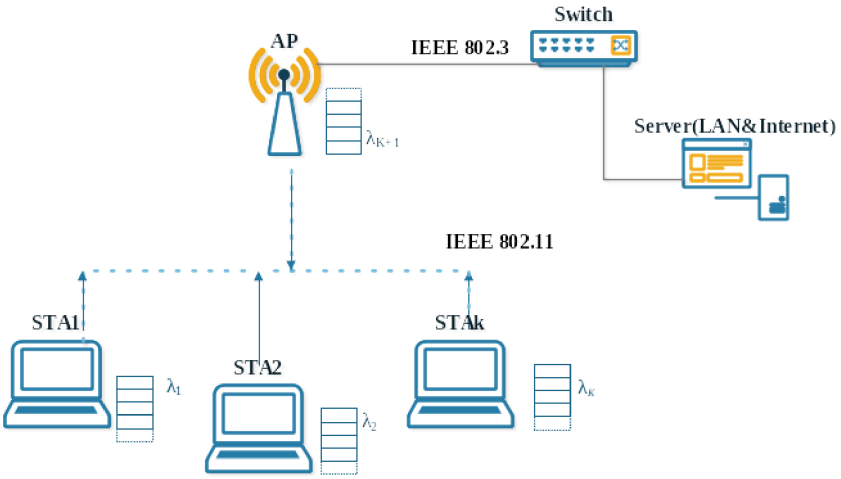


Fig. 1. The wireless local area network

packets which is dropped with an intensity $\Delta\Lambda$:

$$\Lambda = u + \Delta\Lambda \tag{3}$$

The intensity of dropped network packets depends $\Delta\Lambda$ on the denial of service probabilities p_d and the total network application traffic intensity Λ:

$$\Delta\Lambda = p_d \cdot \Lambda \tag{4}$$

3. Performance Parameters of Wireless Local Area Network

The wireless channel is used to service network packets. The main resource in the wireless channel is throughput C [bit/s]. However, a more important parameter is the throughput in network packets C_p[packet/s] with average length l_p [packet/s] which can be sent through the wireless channel. The time to transfer one network packet is:

$$\tau^1 = \frac{l_p}{C_p} \tag{5}$$

Let's assume that μ^1 is the intensity of network packet service in a wireless channel without a specific shared medium access control protocol. We have:

$$\tau^1 = \frac{1}{\mu^1} \tag{6}$$

If the wireless channel uses a specific shared medium access control protocol CF, let the service time be T^{CF} , which is increased by the time t^{CF} because of the introduced overhead for the specific functions of CF. Then

$$T^{CF} = \tau^1 + t^{CF} \tag{7}$$

The intensity of network packet service in a wireless channel with a specific CF is:

$$\mu^{CF} = \frac{1}{T^{CF}} \tag{8}$$

To estimate the functioning capacity of the wireless channel we introduce a coefficient α which depends on the total intensity of network application traffic and the service intensity:

$$\alpha = \frac{\Lambda}{\mu^{CF}} \tag{9}$$

The load in the wireless channel ρ depends on intensity of serviced network packets u in the wireless channel using a specific CF, therefore:

$$\rho = \frac{u}{\mu^{CF}} = u \cdot T^{CF} \tag{10}$$

4. Shared medium access protocol for wireless channel

The specific CF introduces overheads to transmit each network packet. In this paper, we developed a program for Network Simulator ns-3 [10] to obtain the service time value for a specific scenario such as set of WLAN characteristic (table 1). The results of the experiment for $\overline{T^{DCF}}(\Lambda)$, $\overline{T^{PCF}}(\Lambda)$, $\sigma^2_{DCF}(\Lambda)$, $\sigma^2_{PCF}(\Lambda)$ will be presented during the conference presentation.

5. Model formulation

We introduce a metric for wireless channel quality, in which we estimate the guaranteed rate U of network applications traffic, which is serviced by the wireless channel. For this purpose, we developed an analytical model for the wireless channel using a modification of $M/G/1/s$ to obtain the guaranteed time of network packet transmission in WLAN for a set of characteristic in section 2.

First, let's assume $M/M/1/s$ model as the basic model of wireless channel where s is the queue size. We assume the number of places in the queue as: $s = 1+(k+1)\cdot 2$, because we have $k+1$ packets of k STA ($k = \overline{1, K}$) and AP which compete for the wireless channel, and $k+1$ packets are waiting in STA and AP buffers. Also, there

is the case when one packet can get access without competing and there is one place for it.

The system has i states ($i = \overline{0, s+1}$):

when $i = 0$ - the wireless channel is empty, there are 0 packets waiting the service;

when $i = 1$ - the wireless channel services 1 packet, there are 0 packets waiting the service;

when $i = 2$ - the wireless channel services 1 packet, there is 1 packet waiting the service;

when $i = x$ - the wireless channel services 1 packet, there are $x-1$ packets waiting the service;

when $i = s+1$ - the wireless channel services 1 packet, there are s packets waiting the service.

The probabilities $P_i(\ i = \overline{0, S+1})$ of these states are:

$$\begin{aligned} P_0 &= \frac{1}{1+\alpha+\alpha\cdot\sum_{k=1}^{K+1}\alpha^k} \\ P_1 &= \frac{\alpha}{1+\alpha+\alpha\cdot\sum_{k=1}^{K+1}\alpha^k} \\ &\dots \\ P_{s+1} &= \frac{\alpha\cdot\alpha^s}{1+\alpha+\alpha\cdot\sum_{k=1}^{K+1}\alpha^k} \end{aligned} \tag{11}$$

To switch to M/G/1/s model let's apply the Khintchine-Pollaczek rule. Let the multiplier coefficient be γ:

$$\gamma = \frac{\left(1 + \frac{\sigma^2_{CF}}{\left(\overline{T^{CF}}\right)^2}\right)}{2} \tag{12}$$

where $\overline{T_{CF}}$ is the average service time for the specific CF. σ^2_{CF} is the variance of service time for the specific CF

When the new probabilities of system states for M/G/1/s model of wireless channel are:

$$\begin{aligned} \widehat{P_{i\ni 1I}} &= \frac{\alpha^i\cdot\gamma^{\frac{i-1}{2}}}{\sum_{i\ni 1I}\alpha^i\cdot\gamma^{\frac{i-1}{2}}+\sum_{i\ni 2I}\alpha^i\cdot\gamma^{\frac{i}{2}}} \\ \widehat{P_{i\ni 2I}} &= \frac{\alpha^i\cdot\gamma^{\frac{i}{2}}}{\sum_{i\ni 1I}\alpha^i\cdot\gamma^{\frac{i-1}{2}}+\sum_{i\ni 2I}\alpha^i\cdot\gamma^{\frac{i}{2}}} \end{aligned} \tag{13}$$

where $I = 1I + 2I$ is the set of states, 1I – is the subset of odd states, 2I – is the subset of even states.

The average number $\overline{n}$ of packets in the system is:

$$\overline{n} = \sum_{i=1}^{s+1} i\cdot\widehat{P}_i \tag{14}$$

The guaranteed time of network packet transmission in the wireless channel with a specific *CF* is determined as:

$$T_g^{CF} = \frac{\overline{n}}{u} \tag{15}$$

The guaranteed rate U^{CF} of network applications traffic passing through the wireless channel with a specific *CF*:

$$U^{CF} = \frac{1}{T_g^{CF}} \tag{16}$$

or

$$U^{CF} = \Lambda \cdot \left(1 - \frac{\alpha^i \cdot \gamma^{\frac{i-1}{2}}}{\sum_{i \ni 1I} \alpha^i \cdot \gamma^{\frac{i-1}{2}} + \sum_{i \ni 2I} \alpha^i \cdot \gamma^{\frac{i}{2}}}\right) + \Lambda \left(1 - \frac{\alpha^i \cdot \gamma^{\frac{i}{2}}}{\sum_{i \ni 1I} \alpha^i \cdot \gamma^{\frac{i-1}{2}} + \sum_{i \ni 2I} \alpha^i \cdot \gamma^{\frac{i}{2}}}\right) \tag{17}$$

6. Methodology of adaptation wireless channel resources by switching shared medium access protocol

This methodology allows us to discover the range of WLAN characteristics for a better performance of the wireless channel with a specific CF. The numerical results are obtained using NS-3 which is a very popular tool and is used in a lot of research work [11, 12]. The methodology includes the following steps:

Step 1. Determine the configuration parameters of the wireless channel and CF and determine the range of WLAN characteristics set.

Step 2. Develop the ns-3 program which describes the topology of the WLAN of interest, the set of characteristics and configuration parameters.

Step 3. Obtain $u^{CF}(\Lambda)$, $\overline{T_{CF}}(\Lambda)$, $\sigma^2_{CF}(\Lambda)$ using ns-3 modeling and plot the numerical results.

Step 4. Obtain μ^{CF} (8), $\alpha^{CF}(\Lambda)$(9) and $\gamma^{CF}(\Lambda)$(12) using the numerical results from step 3.

Step 5. Obtain $U^{CF}(\Lambda)$ using the numerical results in step 4. and plot $U^{CF}(\Lambda)$.

To demonstrate this methodology, let's obtain $U^{DCF}(\Lambda)$ and $U^{PCF}(\Lambda)$ for WLAN with IEEE 802.11g wireless channel using the steps described earlier:

Step 1. The configuration parameters of the wireless channel. The parameters of CF and WLAN characteristics are specified for the Orthogonal frequency division multiplexing (ERP-OFDM) PHY and C = 54 Mbps (see table 1)[13]

Step 2. The ns-3 program is developed using the parameters from table 1 and the topology in fig.1.

σ_c	9 us	CW_{max}	1023	CW_{min}	15
t_{SIFS}	16 us	t_{DIFS}	34 us	l_{ack}	14 B
C	54 Mbit/s	Λ	250 - 12000 packet/s	K	4
$overlinel$	1500 B	$l_{CF-POLL}$	20 B		

Table 1. The configuration parameters and WLAN characteristics

Step 3. The fig. 2 shows dependence of service intensity u and total intensity of network application traffic Λ for CF = DCF and CF = PCF.

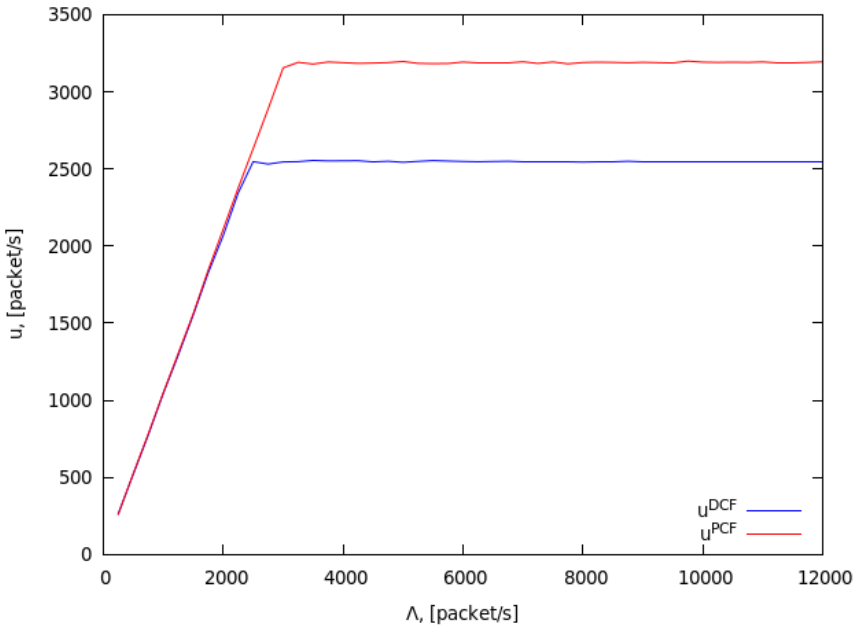


Fig. 2. $u^{DCF}(\Lambda), u^{PCF}(\Lambda)$

Step 4. The values for $\mu^{CF}(\Lambda)$, $\alpha^{CF}(\Lambda)$,$\gamma^{CF}(\Lambda)$ are obtained using parameters from table 1 and results of experiments from section 4.

Step 5. The fig. 3 shows $U^{DCF}(\Lambda)$ and $U^{PCF}(\Lambda)$.

We present here the numerical results of comparison between U^{DCF} of DCF and U^{PCF} of PCF. We show the plot which estimates the range of Λ for $K = 4$. Then, we obtain range of Λ which estimates the working zone for the protocols DCF and PCF. Using this result it is possible to switch from DCF to PCF or vice versa if Λ is changed, and thus it is possible to provide higher intensity of network application traffic in WLAN.

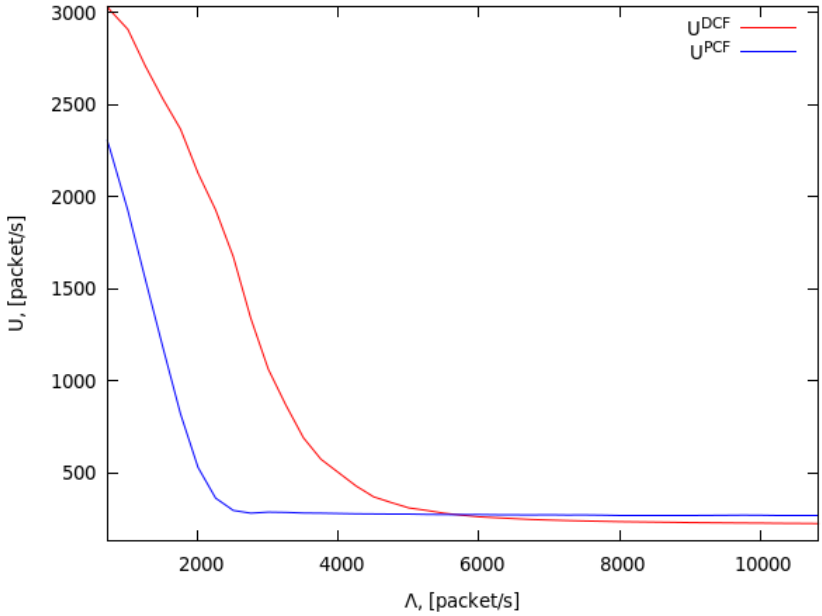


Fig. 3. $U^{DCF}(\Lambda)$, $U^{PCF}(\Lambda)$

The fig. 3 shows that the DCF provides higher guaranteed rate of network application traffic when $\Lambda = [250, 6000]$ and the PCF provides higher guaranteed rate of network application traffic when $\Lambda = [6000, 12000]$ and PCF provides better performance over 19% than DCF.

7. Conclusion

The main contributions in this research as follows. The wireless channel service time for DCF and PCF shared media access protocols under certain conditions (wireless channel configuration parameters and WLAN feature set) is computed using NS-3. We presented a new metric and a methodology for comparing wireless channel performance under certain conditions with different medium access control protocols. We gave an example of applying the methodology for WLAN with a wide range of network application traffic and K=4 wireless stations with IEEE 802.11g wireless channel. The examples show how to determine the working zone of the DCF and PCF protocols. Using this methodology it is possible to adapt the wireless channel resources to the increasing load and estimate the possible guaranteed rate for network applications in the Enterprise WLAN.

REFERENCES

1. Tim Szigeti, Robert Barton, Christina Hattingh, Kenneth Briley, Jr.. End-to-End QoS Network Design. Second Edition: Cisco Press, 2014.
2. G. Bianchi, Performance analysis of the IEEE 802.11 distributed coordination function, IEEE J. Sel. Areas in Commun. 18 (March 2000), pp. 535–547.
3. E. Ziouva and T. Antonakopoulos, "CSMA/CA performance under hightraffic conditions: throughput and delay analysis", Computer Communications, vol. 25, 2/15/2002, pp. 313-321.
4. X. Yang, "Performance analysis of priority schemes for IEEE 802.11 andIEEE 802.11e wireless LANs," IEEE Transactions on Wireless Communications, vol. 4, 2005, pp. 1506-1515.
5. D. Malone, K. Duffy, D. Leith. Modeling the 802.11 distributed coordination function in nonsaturated heterogeneous conditions IEEE/ACM Trans. Netw., 15 (1) (2007), pp. 159-172
6. F. Daneshgaran, M. Laddomada, F. Mesiti, and M. Mondin, "Unsaturated Throughput Analysis of IEEE 802.11 in Presence of Non Ideal Transmission Channel and Capture Effects", IEEE Transactions on Wireless Communications, vol. 7, 2008, pp. 1276-1286.
7. Sarah Shaaban et al., Performance Evaluation of the IEEE 802.11 Wireless LAN Standards - World Congress on Engineering 2008 Vol. I.
8. Ali, Qutaiba. (2009). Performance Evaluation of WLAN Internet Sharing Using DCF & PCF Modes. International Arab Journal of e-Technology.
9. I. Tinnirello, G. Bianchi, and Y. Xiao, "Refinements on IEEE 802.11 Distributed Coordination Function Modeling Approaches," IEEE Transactions on Vehicular Technology, vol. 59, 2010, pp. 1055-1067.
10. Nsnam. n.d., ns-3 — a discrete-event network simulator for internet systems. Available from: `https://www.nsnam.org/`.
11. Yin Y., Gao Y., Hei X. (2019) Performance Evaluation of a Unified IEEE 802.11 DCF Model in NS-3. In: Song H., Jiang D. (eds) Simulation Tools and Techniques. SIMUtools 2019. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 295. Springer, Cham, pp 395-406
12. Patricia Deutsch, Leonid Veyster and Bow-Nan Cheng LL SimpleWireless: A Controlled MAC/PHY Wireless Model to Enable Network Protocol Research
13. IEEE Standard for Information technology – Telecommunications and information exchange between systems Local and metropolitan area networks–Specific requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," in IEEE Std 802.11-2016 (Revision of IEEE Std 802.11-2012) , vol., no., 14 Dec. 2016.

UDC: 519.688

Distribution of Computing Load by using a P2P Network

Mamonov Anton Alekseevich[1], Varlamov Ruslan Aleksandrovich[1], Salpagarov Soltan Ismailovich[1]

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

anton.mamonov.golohvastogo@mail.ru, salpagarov-si@rudn.ru, ravarlamov@mail.ru

Abstract

There is a matter of performance problems in modern calculation tasks, one that cannot be solved only by increasing the quantity and quality computers. In this work we careful scrutiny and develop methods of distributed computing.

By applying knowledge from different areas of science, such as queueing theory and theory of computation, we create our own distributed computing system. The main idea behind our approach is equality hierarchy and decentralization. Usage of peer-to-peer architecture requires a thoughtful design but offers several advantages, which are discussed in this work.

In order to compose our system, we set a mathematical model of the target situation, draw attention to its weaknesses, and develop software principles to resolve them. After solving a number of design issues, we create simulation models to analyze the effectiveness of the future system.

After analyzing the models, we select next steps to improve and finalize our system.

Keywords: p2p, distributed computing, queueing theory, theory of computation, program design

1. Introduction

The past decade has shown a major change in the nature of computing. The only thing that can outpace growth of hardware performance, is increasing requirements of software. With each new generation of computing or networking devices, a dozen more high-performance technologies appear. In other words, there is ongoing race between providing and consuming in IT-sphere [1].

In the result of this prolonging situation, we have countless computing devices of various performance, major part of which have long ceased to meet the requirements.

Providers constantly upgrading and increasing quantity of their products. As such, there is a large pool of only part-time used computing devices, not suitable for modern high-performance tasks. Such tendencies was noticed and studied for a time now [2]. That grants a perfect case for using distributed computing, for example [3].

Despite this, many devices are not connected to any of them and a significant amount of computing power is wasted. The amount of these computing power can reach the performance of several supercomputers. One of the main reasons for not using these capacities is the complexity of configuring distributed systems. The ratio between costs and benefits of deploying distributed computing systems is too high for common usage. In our opinion, the reason for this is the common usage of the centralized architecture which is not the optimal method for distributing calculations. A possible analog for this is the peer-to-peer architecture, already used to some extent for distributed computing, for example [4]. As such, we chose to develop new, non-centralized method for distribution.

2. High-performance computing overview

The problem of high-performance computing is not a new one. It has a long history of searching for solutions. And finds those in various scientific areas. Before starting to develop our own solution, we reviewed the most significant trends in this direction.

Recently it became popular to use graphics processing units for General-purpose computing since GPU has good computing power, which is usually idle [5]. But this is not suitable for all tasks, the system architecture becomes more complicated and not as Energy efficient as CPU and FPGAs.

You can also use special high-level, numerically oriented programming languages such as Matlab, Scilab etc. A lot of scientific and developing efforts are focused on their constant upgrading, which helps computer algebra systems stay up-to-date [6]. Although this approach can save a lot of time, it is not usable for any task, and despite the fact that tasks will be solved faster, this time gain is not satisfactory for some of them.

This article specifically delves into distributed computing technologies. Work on distributed computing started with a very applied purpose - for military needs, and namely, the automation of secret communication processes and intelligence processing information has been conducted intensively in the United States since the 1960s. This technology is currently used in many volunteer and commercial projects [7].

For example, BOINC - The Berkeley Open Infrastructure for Network Computing, is a an open-source middleware system for volunteer and grid computing [8]. Currently, in mid-2020, it has an average daily performance of about thirty thousand petaflops, from eight hundred thousand of computers. But to use even a fraction of this,

newcomers will need to learn how to build up their own project, and then draw attention to it. On the other end, while participants, donating they performance to the system, are not burdened with such difficulties, but also has no benefits from it.

In case you want to set up more equal system, you will need either learn how to modify one of existing solutions or, more likely, build it up yourself. Which, even with usage of specific libraries and utilities, is no easy task. JPPF [9] – an open source Grid Computing platform written in Java, written with sole propose of easing said process, still requires well honed programmer's skills.

Although all those solutions are valid in their own areas, there is a lack of user-friendly technologies, ones that would not require additional studies or financial investment. So, we present PDCS - Peer-to-peer distributed computing system, a network which use knowledge from teletraffic engineering and computational complexity analysis to manage beneficial calculation distribution for all participants.

3. Concept

3.1. Mathematical basis. In order to design an efficient distributed computing system, we raise the issue of load balancing. To do this, we're using models from the queueing theory [10]. It is a science field that study queueing systems, and a computing system can be easily interpreted as such.

Consider the work of the office with v employees, each of whom uses a personal computer for periodic calculations. In the terminology of the queuing theory, each employee is a stream of applications and personal computers are the service devices. With the flow rate of applications λ, the processing speed μ, and with the maximum number of applications in the queue of the device r, it is possible to compose an Erlang model and calculate the time characteristics [11].

However, for the standard situation, when each employee uses only his personal computer, not one common system, but a set of v separate systems is formed. In a situation like this, when λ is small and μ is large, there is an imbalance between the individual devices - while half of them are free, the other half are in the queue and vice versa.

But if you use middleware for communication between devices, it is possible to combine the threads into a single thread with flow rate λ and the total queue capacity $R = r * v$.

Even with the additional time spent on transferring information between devices, the benefit of using a distributed system is obvious. The shorter total time spent in the queue reflects a more efficient use of the available computing resource.

To implement such a model in practice, it is necessary to organize a distributed computing system. To do this, you need to write software that will read user requests, generate applications and distribute them over the network. The main principles

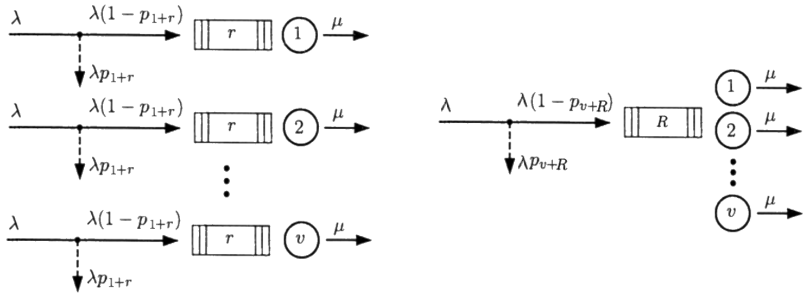


Fig. 1. Model on divided(left) and distributed(right) system

that should be met by the software are the generalization of calculations, the equality between network nodes and the transparency of working with the system.

3.2. Generalization. Generalization of computation stands for the reduction of requests received by the system to a set of generalized operations.

In the terminology of Post, which was set in one of the first articles concerning the theory of algorithms [12], the concept of a general problem, which is consisting of a class of specific problems, is introduced. Thus, giving an algorithm for solving one general problem, we obtain a solution for a whole set of specific problems. This is fairly obvious for simple cases such as arithmetic operations, but the boundaries of generalization are easily extended if recursion is allowed.

So, for example, calculating the factorial is essentially a special case of multiplying a natural number by the factorial of the previous one. The calculation of the number of combinations is several divisions of certain factorials. This means that the calculation of one or another probability can be reduced to a set of arithmetic basic operations. By giving the system the ability to compute primitives nested into each other, all that remains is to provide a generalization process — the generation of these primitives from the user's request.

3.3. Equalization. Equality between network nodes stands for the implementation of Peer-to-peer architecture during system creation.

Peer-to-peer technology which has been developing over the past decade, continues to find new uses. Nowadays many initially centralized platforms and distributed computing tools are gradually integrating the ability to implement partial or full peer-to-peer. This is due to the fact that peer-to-peer networks have many advantages over classic client-server counterparts. Among them, scalability, high fault tolerance, simple moderation, etc.

In the situation under consideration, the most valuable will be the fact that the organization of a network using a peer-to-peer architecture will allow us to avoid the

difficulties of installing and separately maintaining a server. This should provide a way to popularize peer-to-peer computing, which, in turn, will provide peer-to-peer systems with the necessary capacities.

3.4. Transparency. Transparency of the system means that its internal workflow is invisible to users.

The main source of resources for any peer-to-peer system is users. A large number of nodes means more memory, higher speeds for file sharing and more computing power for distributed computing. In order to provide our system with users, we make system's transparency one of the main development goals.

It's not enough to provide significant time gains in time-consuming calculations. The system also should not force users to learn local syntax or make complicated queries. Any advantage in the speed of actual calculations might be lost if a comparable amount of time is required to organize them. In such a situation, the audience of the system will significantly narrow, which will reduce the total number of nodes and, as a result, the output of the system.

In order to provide a user-friendly interface and low entry threshold, it was decided to change the focus of development from creating a domain-specific language in favor of using the well-known general-purpose language of mathematical formulas. There is no need to have a special education in order to know that "=" means "equal to", that the dependent parameters are listed in brackets after the name of the formula. Just having an education is enough.

4. The Algorithmic analysis

During the construction of the system, there was a question about the efficient allocation of resources, but for this it was first necessary to decide how the complexity of a given task would be assessed. Fortunately we are not the first to encounter such a problem and there is a very large range of works devoted to the algorithmic analysis.

Our system should be able to calculate the complexity of the task, regardless of the class of algorithms used in it. In mathematics, asymptotic notation is used to analyze the complexity of algorithms. This assessment is called the complexity of the algorithms, and by using it, our system determines which algorithm to use.

The most straightforward and at the same time effective way of estimating the complexity of the algorithm is the operational analysis. It splits the algorithm into a set of basic operations. For each basic operation, the execution time is calculated experimentally, taking into account the type of input data. Then overall time complexity of the algorithm is calculated as sum of its parts.

Since the system can be used in various fields of scientific research, calculations can use a huge range of elementary operations. Therefore, at the beginning of its

work, the computer will calculate the approximate time for all these operations. Also, the computer that sends a request for solving the problem will count the number of simple operations involved in it. Based on these data, the task will be most effectively divided using the distribution mechanism.

5. The Mechanism of Distribution

The overall efficiency of any system depends to a large extent on the distribution of the workload. Especially so in purposefully distributed systems, and ones there request-report process on one node unavoidably engages others nodes as well. Distributed computing is a vivid example of said systems.

As was mentioned earlier, PDCS is working with natural mathematical distribution. For practical example, let expression F be the formula for the probability that the number of successes for a set of independent experiments N equals i. With each independent success or failure respectively equal to p and q, probability of at least i_0 successes will be (1).

$$Sum(N) = \sum_{i=i_0}^{N} p^i q^{N-i} C_N^i \tag{1}$$

Such request can be entered by user in string form and will look like :

input : sum(i = i0, N)(p ∧ i * q ∧ (N-i) * comb(N, i)

Which would be fractured by parser and then would be used for forming secondary network request, one for each member of sum.

solve : (i = i0)(p ∧ i * q ∧ (N-i) * comb(N, i)

solve : (i = i0+1)(p ∧ i * q ∧ (N-i) * comb(N, i) ...

As it is common for calculation of combinations, the bigger difference between N and i values is, the bigger result will be, which system will determine by evaluating the complexity of each request. More complex request will be send on more powerful nodes, and vice versa.

When all secondary requests are fulfilled and sent back on origin node, it compose them in final sum and notify user with numeric output.

6. Practical results

Before developing the first prototype of the system, it was necessary to check superiority of our distribution method over the others, as well as measure its approximate effectiveness: the time saved by the system, stability of the system etc. For this, it was decided to develop a system models using a suitable simulation program. The first model should have been based on separate computers that solve tasks on their own, while second one is the representation of our approach.

AnyLogic was a good choice for our purposes. It is a multimethod simulation modeling tool. It supports agent-based, discrete event, and system dynamics simulation methodologies. It is suitable for our task, as it has a fairly extensive functionality that allows you to create dynamic models with different amounts of input and output data. At the same time, it allows you to track each task individually and has built-in statistics tools. In addition, the program itself is also written in Java, just like our system, which saved our time.

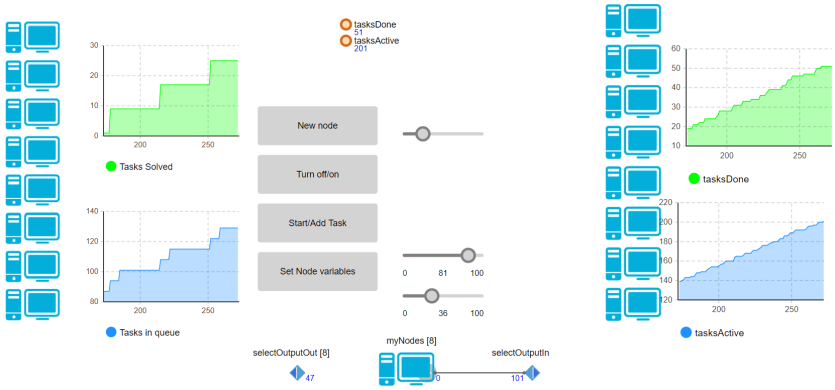


Fig. 2. Model of separately working computers on left; P2P model on right.

The P2P model showed overall high level of reliability and validity while the comparison of models was used to measure the potential time gains.

7. Conclusion

In this article we have studied a problem with high-performance computing, and various ways of it's solution. Concluding that most solutions are not user-friendly or easily accessible, we have developed our own paradigm of computing system, and created a mathematical model for it. With usage of different scientific fields, we have improved our design and created simulating models to analyze resulting system

In further research we aim to use obtained models to create a software template of PDCS - peer-to-peer distributed computer system, and conduct a comparative analysis between our system and others solutions for high-performance computing problem.

8. Acknowledgment

The publication has been prepared with the support of the “RUDN University Program 5-100”.

REFERENCES

1. A. Trattner, L. Hvam, C. Forza, Z. N. L. Herbert-Hansen, Product complexity and operational performance: A systematic literature review, CIRP Journal of Manufacturing Science and Technology, Volume 25, (2019) 69–83 doi: 10.1016/j.cirpj.2019.02.001
2. D. Post, The Future of Computing Performance, Computing in Science & Engineering, vol. 13, no. 4, pp. 4-5, (2011) doi: 10.1109/MCSE.2011.69
3. G. Liu, Z. Xiao, G. Tan, K. Li, A. T. Chronopoulos, Game Theory-Based Optimization of Distributed Idle Computing Resources in Cloud Environments, Theoretical Computer Science 806, (2020) 468–488 doi: 10.1016/j.tcs.2019.08.019
4. K. M. Khachumov, S. I. Salpagarov, A. A. Mamonov, V. A Varlamov, Combinatorial problem solving method by allocating resources, Selected Papers of the IX Conference "Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems", Moscow, Russia, 19-Apr-2019 (2019) http://ceur-ws.org/Vol-2407/paper-07-165.pdf
5. S. Mittal, J. S. Vetter, A Survey of Methods for Analyzing and Improving GPU Energy Efficiency. ACM Comput. Surv. 47, 2, Article 19 (2015), 23 doi: 10.1145/2636342
6. Kulyabov, D.S., Korolkova, A.V., Sevastyanov, L.A. New Features in the Second Version of the Cadabra Computer Algebra System. Program Comput Soft 45, 58–64 (2019) doi: 10.1134/S0361768819020063
7. E. Ivashko, I. Chernov, N. Nikitina, A Survey of Desktop Grid Scheduling, IEEE Transactions on Parallel and Distributed Systems (2018) 2882–2895. doi: 10.1109/TPDS.2018.2850004
8. Anderson, D.P., BOINC: A Platform for Volunteer Computing, J Grid Computing 18, (2020) 99—122 doi: 10.1007/s10723-019-09497-9
9. Pengembangan Prototipe Sistem Network Rendering Alternatif Berbasis JPPF, (2018) doi: 10.22146/jnteti.v7i1.395
10. Naumov, V.A.; Gaidamaka, Y.V.; Samouylov, K.E. Computing the Stationary Distribution of Queueing Systems with Random Resource Requirements via Fast Fourier Transform. Mathematics (2020), 8, 772. doi: 10.3390/math8050772
11. Basharin G.P. Lectures on the mathematical theory of teletraffic
12. E. Post, Finite Combinatory Processes-Formulation 1, J. Symbolic Logic vol. 1 no. 3 Sep. (1936) 103–105. doi: 10.2307/2269031

UDC: 004.7

Identification of devices in a mesh networks based on Digital Object Architecture

D.D. Sazonov[1] and R.V. Kiricheck[1]

[1]The Bonch-Bruevich Saint-Petersburg State University of Telecommunications, Saint-Petersburg, Russian Federation

dim-saz@yandex.ru

Abstract

The analysis of the possibility of building a system for identifying Internet of Things devices based on the Digital Object Architecture in mesh networks is given. The basic principles of the LPWAN networks and mesh networks are described. A brief overview of Digital Object Architecture technology is given. Schemes for integration of the Digital Object Architecture platform into mesh network are proposed. Various configuration options for this system in order to increase productivity are considered. The model of the Handle resolution system for identifiers of digital objects in mesh networks as a queuing system is proposed. An analysis of the results is given.

Keywords: Internet of Things, Mesh networks, LPWAN, LoRaWAN, Digital Object Architecture, Handle System, Identification

1. Introduction

The rapid growth of the Internet of Things (IoT) technology has led to the proportional growth of the market for various applications that use this concept. The most popular areas are the following [1]:

- augmented reality applications;
- smart home applications;
- smart cities application.

For IoT applications, power consumption, low latency, edge device density, and communication security are important parameters.

One of the answers to these requirements for new generation networks was the creation of the concept of energy-efficient long-range networks (LPWAN). LPWAN is

The publication has been prepared with the support of the grant from the President of the Russian Federation for state support of leading scientific schools of the Russian Federation according to the research project NSh-2604.2020.9.

a specification that allows Internet of Things devices to communicate and transmit data over significant distances (kilometers) with low power consumption [2].

One of the implementation of the LPWAN concept is LoRaWAN technology. LoRaWAN based on LoRa modulation technology (Long Range) at the physical level [3, 4]. LoRaWAN protocol provides a significant coverage area compared to analogs.

LoRaWAN solutions based on star or star-of-stars topology, where edge devices interact through multiple gateways with the base network server [4-6]. The obvious disadvantage of this topology is the instability of the network to possible failures of the central node.

One of the possible solution to this problem is the switching from the star topology to mesh topology for LoRa-based applications.

There are many IoT solutions based on mesh topology [6]. A key feature of the mesh topology compared with the star topology or star-of-stars topology of LoRaWAN is the resilience of the network to the failure of individual components [5, 6].

Switching to mesh topology for LoRa devices gives us the possibility to achieve the fault tolerance of mesh topology and low power consumption and long distance transmitting of LoRaWAN.

The idea of organization low power long range mesh networks was considered in [5]. This article presented a model showing the possibility of organizing LoRaWAN devices in a mesh topology. In our work, we will focus on the problem of identifying edge devices in such hybrid network.

2. Digital Object Architecture Based Identification

The International Telecommunication Union (ITU) has the list of recommendations for identification systems for Internet of Things devices in narrow-band wireless communication networks [7]. Digital Object Architecture (DOA) fully meets this requirements.

Digital Object Architecture was created by Corporation for National Research Initiatives - CNRI two decades ago [8].

DOA has proven itself worldwide as a system for identifying academic, professional and government information: it is the well-known DOI identifiers of articles and book publications [9].

The main structural elements of DOA are a digital object, digital object registry and repository and resolution system [8, 9].

Handle system as a implementation of the DOA concept [9, 10] provides programming interface for clients (rest api) for managing records of digital objects and obtaining information (resolution) as well as a set of tools and libraries for writing

custom applications. It is possible to organize various schemes of client interaction with the Handle system to organize various identification scenarios for mesh networks.

Consider the following architecture for identification system in a mesh network based on Handle System. The administrator (device owner) sends information about the IoT devices to the Handle System, providing the basic description and various metadata that will be required in the future to complete the identification process. In addition, it is possible to implement a role-based access model for devices, which can then be used at the stages of network configuration to split it into subnets. Next, the Handle System sends unique identifiers of the registered devices to the owner. Administrator flashes the identifiers to the device memory. When the devices starts to configure in a mesh network, identifiers are embedded in a message body from the edge devices and transmitted over the network. Through the gateway, such messages with identifiers are sent to the cloud for further processing and identification.

In the figure 1, the basic version of the identification system for the mesh network is shown, according to what we described earlier.

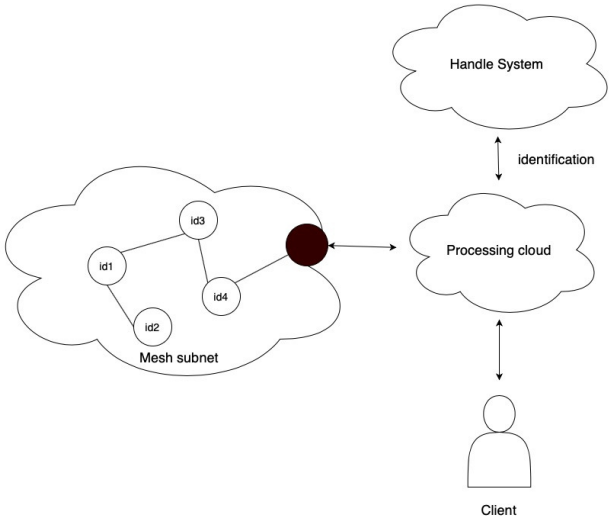


Fig. 1. Basic mesh network identification process

Mesh subnet consisting of several edge devices and a gateway device (black). The authorization subsystem in the cloud processes the message flow, sends information to the Handle subsystem (or its local cache if it is already filled with previously processed identifiers) to obtain identification information. To speed up flow processing, authorization subsystem can be configured to periodically scan the message stream.

After checking the authorization information, the data from edge device is sent to further processing. If the request is not authorized, messages from the device are rejected. The authorization subsystem can send a notification request to the network coordinator to block an unauthorized device. The subsystem also can send notification to the network administrator about the presence of an unauthorized device. Figure 2 shows a diagram of the interaction with filtering an unauthorized device.

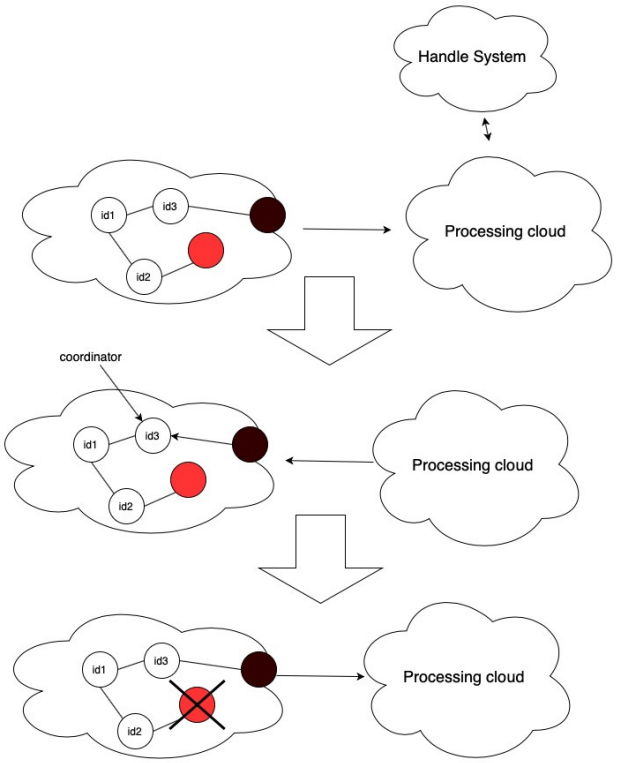


Fig. 2. Unauthorized device filtering

Next, we consider a simulation model of a simplified identification system for a mesh network, based on the DOA architecture and those concepts for constructing such systems that were described earlier.

3. Identification system simulation model for mesh network

The queuing model is based on the model of the DOA system considered earlier in [11, 12]. We introduce a couple of assumptions into our model to be able to simulate it and further analyze it.

It is possible to represent a mesh network edge device in our model as a single entity with a buffer and a delay block. The total latency of the transmission from the edge device to the processing server is defined as the product of the delay in one device by the number of hops.

Within our model, we will consider the Handle System as a separate server consisting of a buffer and a delay block.

Figure 3 shows a system model built in the AnyLogic package.

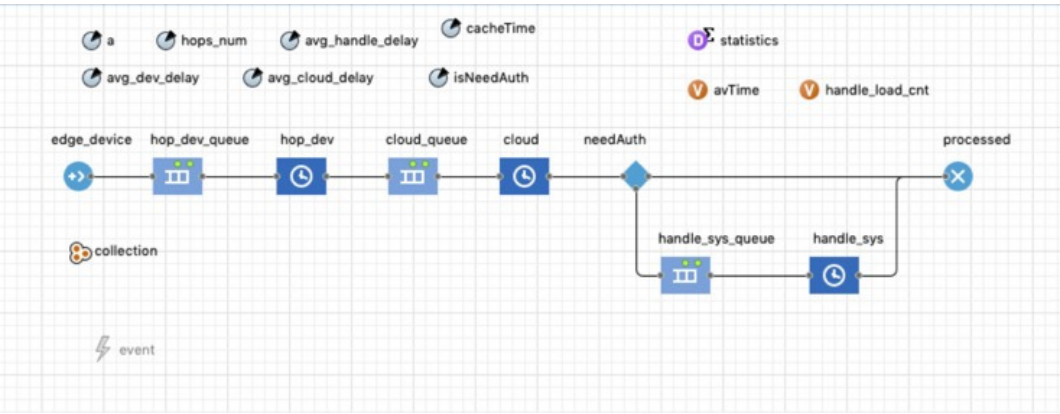


Fig. 3. Model

The model consists of the following components:

- edgeDevice - single edge device of the mesh network, transmitting data through intermediate devices to the gateway and further to the cloud;
- hopDevQueue and hopDev - model of intermediate edge devices (also include gateway). The number of devices in the chain is determined by the hopsNum parameter. In our model, it was 3 intermediate devices. HopDelay block in our model has the following parameters: capacity = hopsNum (each edge device in the mesh network usually can process one request at a time); processing time = hopsNum * avgDevDelay (where avgDevDelay is the average processing time of the request on the edge device, the value of 100 ms was selected in our model);
- cloudQueue + cloud - model of remote server that process all incoming request from mesh network. The average processing time of one request in our model

is 200 ms. Server in our model can handle 15 requests in parallel (we assume that the service processes requests in 15 threads);
- handleSysQueue + handleSys - model of the Handle System remote service that processes resolution requests from the cloud. The average processing time (with the network delay) is 400 ms;
- needAuth - a switch block that control data flow that needed to resolve in Handle System.

The time between incoming requests from edge device is distributed exponentially with load parameter a. In our model, it varies discretely from 1 to 200 with step 0.5.

In this model we analyze several parameters: average processing time of the request from the edge device, the load of the Handle System and the total load of the identification system.

We considered several configurations of the identification system:

- every incoming request from the edge device will send to the Handle System;
- the cloud server periodically switch to identification mode and start to send all incoming requests to the Handle System. Scan period is 10 minutes;
- the cloud server switch to caching mode by adding all responses from Handle System to local cache which is then used to identify subsequent request. The cache lifetime in out model is 10 minutes.

The following figures 4 and 5 show the results of the analysis and experiments with our model which was configured in different modes that was described previously. Figures characterize our model system according to the parameters of the average request processing time, the number of requests to the Handle System to the total number of requests (percentage). The simulation time was 100 minutes.

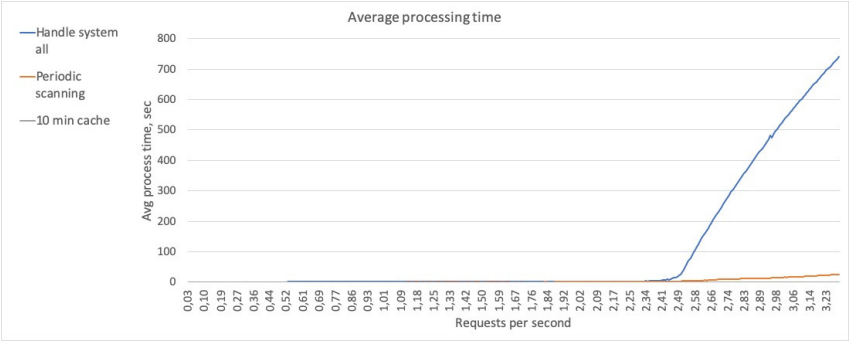


Fig. 4. Processing average time

From figure 4 it can be seen that when our model was process every incoming request with identification request to the Handle System, even at low load level (3

requests per second), the average processing time for requests is around 700 seconds. When model switched to the periodic scan mode with a period of 10 minutes, it showed performance improve (about 25 sec). When our model worked with the cache with 10 minutes lifetime, it showed the best results (0.5 sec).

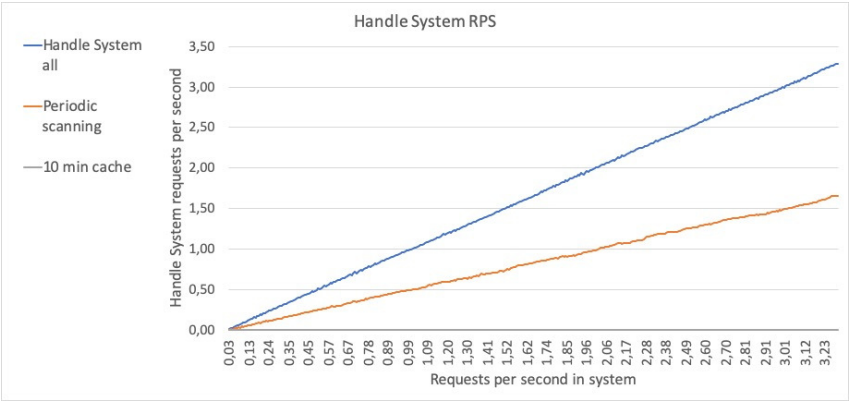


Fig. 5. Handle System loading

The figure 5 shows the dependence of the rps load (requests per second) received by the Handle System from the total number of rps that incomes to our system. It can be seen that when working with the cache, the load level on the Handle System stay constant (4 devices, cache 10 minutes). In other configurations, the dependence is linear.

4. Conclusion

In this paper, the possibility of implementing a DOA-based identification system in a hybrid network of LoRa devices combined in mesh topology was considered. An attempt to move from the classic star topology, which is typical for LoRaWAN to mesh, is caused by the desire to solve the problem with a single point of failure, typical for the star topology, and move to a distributed mesh network topology, which has a number of advantages listed in this work.

To build the identification system, the DOA platform was chosen, since it is quite promising, there are already solutions for identification based on DOA in other areas.

Possible schemes of the identification system in the described mesh network are considered. To analyze the system performance, a queuing model was built in the AnyLogic package.

Even an analysis of the simple model shows that the implementation of an identification system based on the Handle System for mesh networks is possible.

Handle System is a client-server type system, and interaction and integration of third-party systems with it can occur via http requests (rest api).

Analysis of the described model shows that the operating mode of the identification system using the query cache shows the best performance. This operating mode of the system is possible, since device identifiers usually do not change throughout the life cycle of a system.

As part of the development of the described concept in future work, it is planned to build a test mesh network system with DOA-based identification and analyze the configuration of the system.

REFERENCES

1. Kucheryavy, A.: Internet of things. Elektrosvyaz (1), 21–24 (2013)
2. Kumaritova, D., Kirichek, R.: Review and comparative analysis of LPWAN network technologies. Information technologies and telecommunications vol. 4, 33–48 (2016)
3. Khutsoane, O., Isong, B., Abu-Mahfouz, A.: IoT devices applications based on lora/lorawan. Electronics Society, IECON 2017-43rd Annual Conference of the IEEE, 6107-–6112 (2017)
4. History of LoRa technology, `https://nekta.tech/en/technology/`. Last accessed 9 Sep 2020
5. Pham, V., Gallyamov, D., Vorozheykina, O., Kitichek, R.: Model of energy-efficient long range mesh network. Elektrosvyaz (5), 33–41 (2020)
6. Wireless Mesh Network, `https://www.eot.dk/Files/Images/Elektronikmesse/2017/Konferencerne/Pre-fra-SPEAKERE/Wireless-Mesh-Network-a-well-proven-alternative-to-LPWAN.pdf`. Last accessed 9 Sep 2020
7. Recommendation X.660 (07/11) X.660 : Information technology - Procedures for the operation of object identifier registration authorities: General procedures and top arcs of the international object identifier tree, `https://www.itu.int/rec/T-REC-X.660-201107-I/en`. Last accessed 9 Sep 2020
8. Kahn, R.E.: A framework for distributed digital object services. International Journal on Digital Libraries vol. 6, 115–123 (2006)
9. Borodin, A., Kirichek, R., Sazonov, D., Rozhkov, M., Kolesnikov, A., Birman, A., Rogdev, A.: Identification of devices and systems of narrowband wireless networks of the Internet of things. Part I. Elektrosvyaz (5), 24–33 (2020)
10. The Handle System, `https://www.dona.net/handle-system`. Last accessed 9 Sep 2020

11. Kirichek, R., Sazonov, D.: Digital Object Architecture as an Approach to Identifying Internet of Things Devices. Distributed Computer and Communication Networks, 597–611 (2019)
12. Al-Bahri, M., Kirichek, R., Sazonov, D.: Modeling of the Internet of things device identification system based on the architecture of digital objects. In: Proceedings of educational institutions of communication (1), 42–47 (2019)

UDC: 519.718.2

Evaluation of Network Reliability and Element Importance Metrics Using the R Software Package

Aleksandr Moshnikov[1]

[1]ITMO University, 49 Kronverksky Pr., St. Petersburg, Russia

moshnikov.alex@gmail.com

Abstract

The article considers an approach to assessing the importance metrics and reliability of networks. The Monte Carlo method is used to estimate Birnbaum metrics and failure probability with determination of the confidence interval. To conduct a computational experiment, the R software package is used. A description is given of the representation of the control system reliability model in the iGraph package, which provides visualization of the results. The model of a three-level network structure is considered as an example.

Keywords: Monte Carlo method, importance metrics, reliability estimation, network connectivity, R language.

1. Introduction

Currently, there is a rapid development of information technologies and their implementation in various areas of human activity [1]. Control transmission networks have become an integral part of people's lives, without which information exchange is practically unthinkable. In such a situation, the analysis of the technical characteristics of existing data transmission networks and the design of new networks, taking into account the given characteristics, remains one of the urgent tasks in the field of information technology.

In addition to such technical characteristics of computer networks as: performance, latency, security, scalability, extremely important characteristics are complex reliability indicators: availability factor, average unavailability time per year [2]. The reliability of the network also indirectly depends on the safety of the operation of control systems for any objects in which the untimely response (due to failures and failures in the data transmission network) of the control system to any critical changes in the control object can lead to serious consequences. In this situation, the

analysis of reliability indicators of distributed control systems is a particularly relevant problem. Reliability is defined as the probability of a system or a sub-component functioning correctly under certain conditions over a specified interval of time [3].

Issues of reliability of systems with a network structure are still relevant [4, 5].

For instance, the reliability of network nodes, termed as the terminal reliability, is the probability that a set of operational edges provides communication paths between every pair of nodes. Another closely related concept with reliability is availability, which can be defined as the probability that a component will be available when demanded [3].

Importance measures (IMs) are used to evaluate the effect of component reliability on system reliability. IMs are useful tools in reliability engineering [14], risk analysis [15], and system reliability optimization. These measures can help reliability engineers to find a better solution rapidly because they can identify the weakest links of the system, which are the premise and foundation of system design, maintenance, and resource configurations. During the system design period, component importance can help designers determine cost effective design ideas with relatively high system reliability and low cost rapidly.

Note that the significance of a system element according to Birnbaum I_{BIM} reflects the degree of influence of changes in the element's readiness coefficient on changes in the system's readiness coefficient. The significance of a system element according to Barlow-Proshan I_{BP} reflects the probability that a system failure that occurred at a certain point was caused by this element. The significance of a Vesely-Fassel system element reflects the probability that this element is one of the failed elements, provided that the system failed. Vesely-Fassel significance I_{VF} characterizes those elements that are most often involved in system failures. The cost of increasing the risk for a system element I_{RAW} reflects the importance of maintaining the current level of reliability of this element. The cost of reducing the risk for a system element I_{RRW} reflects the degree to which the system's availability coefficient increases if this element is replaced with a flawless element. The critical significance of a system element I_C reflects the probability that this element is critical for the system at a given time. The potential for improvement of a system element I_{IT} reflects the gain in system reliability if this element is replaced with a completely reliable one [16].

2. Reliability and importance assessment

An unbiased estimate is used as a reliability estimate using Monte-Carlo simulation:

$$P(t) = \frac{m}{n} \tag{1}$$

where m is number of success simulation, in which the system did not fail (in the case of networks, connectivity is not broken) for the specified time t, and n total number of simulation cycles. It is natural to assume that different elements affect the system's behavior in terms of reliability in different ways. The ability of the researcher to quantify the nature of the elements influence on the behavior of the system is of particular importance in the analysis of systems. This makes it possible to identify system weaknesses, select optimal redundancy, and make a rational impact on the reliability of the system as a whole.

The importance of the element e_i in the system is defined as a private derivative of the availability factor (the probability of) the system availability (the probability of) the element for which an analysis of its significance:

$$I_{BIM}(i,p) = \frac{\partial h(p)}{\partial p_i} \tag{2}$$

This characteristic is called Birnbaum significance (BIM-significance) [6]. The significance is estimated by the number of times the system availability coefficient increases when the element availability coefficient increases. BIM-significance does not depend on the readiness coefficient of the element p, but depends only on p_j for all $i = j$ in satisfies the inequalities $0 \leq I_{BIM}(i) \leq 1$. For the BIM-significance indicator, you can get an expression in the following form:

Birnbaum importance considers the relationships between the system performance when component i is perfect, the system performance when component i fails, and the current system performance.

$$I_{BIM}(i,p) = h(1_i,p) - h(0_i,p) \tag{3}$$

where $h(i,p)$ is reliability function, $h(1_i,p)$ for absolutely reliable component, $h(0_i,p)$ for absolutely unreliable component;

The performance measure of the importance the Birnbaum in this form greatly simplifies the calculations. To calculate it, it is enough to calculate the value of the readiness coefficient once under the assumption that the i-th element is absolutely reliable, and once that it is absolutely unreliable. Based on the Birnbaum significance index, the concept of element contribution to system reliability is formulated. Many researches have been devoted to computational aspects of significance estimation, including [7, 8].

3. Modeling using R

R is a programming language for statistical data processing and working with graphics, as well as a free open-source computing environment for the GNU project.

The R language contains tools that allow you to create multiple parallel threads of calculations (due to simultaneous loading of several processor cores) and reduce the time spent on modeling several times. To assess the accuracy of the results obtained, the bootstrap method is proposed. The essence of the method in this case is that on the basis of one available sample (obtained using the graph traversal algorithm), a series of pseudo-samples of the same size is formed, consisting of random combinations of the original set of elements. In this case, the "random selection with return" algorithm is used, i.e. the extracted element is returned to the original set and has a chance to be selected again. For each random sample to estimate the probability of failure (or probability of failure) and thus formed the sample probabilities of system failure (or probability of failure-free operation), which further evaluated the necessary statistical data (standard deviation or confidence limits). To calculate the number of iterations and estimate the confidence interval, a standard approach is used in accordance with [3]. Various techniques can also be used to improve accuracy, with the most widespread sampling by significance [10].

To search the graph for paths between certain vertices, use the width traversal algorithm (an implementation of this algorithm in the iGraph library is used). To generate random numbers with an exponential distribution law, the basic functions of the R language are used [9].

Reliability modeling includes N iterations. At each iteration, a random operating time before failure is generated for each system element (vertex) (this time is generated based on the specified failure rate of the system element). After that, the elements are sorted in ascending order of uptime to failure, and the element (vertex) with the lowest time is selected. This vertex is removed from the graph and the existence of paths between certain vertices of the graph is checked (between which vertices the presence of a path should be checked is listed in the description of system failure criteria). If all necessary vertexes are found, the current iteration continues and the next element in increasing time to failure is selected and the corresponding vertex is removed from the graph. Next, it checks again whether there are paths between certain vertexes. If no paths are found between the specified vertexes, the system is considered to have failed. The failure time T_i is fixed and a new iteration begins.

4. Numerical example - SCADA system

Automated process control system is a group of technical and software solutions designed to automate technological processes in industrial enterprises. As a rule, automatic process control systems are understood as a complete solution that ensures the automated execution of the basic operations of the technological process of production. Components of automatic process control systems can be separate automatic control systems and automated devices connected in a single complex.

Such as Supervisory control and data acquisition systems (SCADA), distributed control systems (DCS), emergency protection systems.

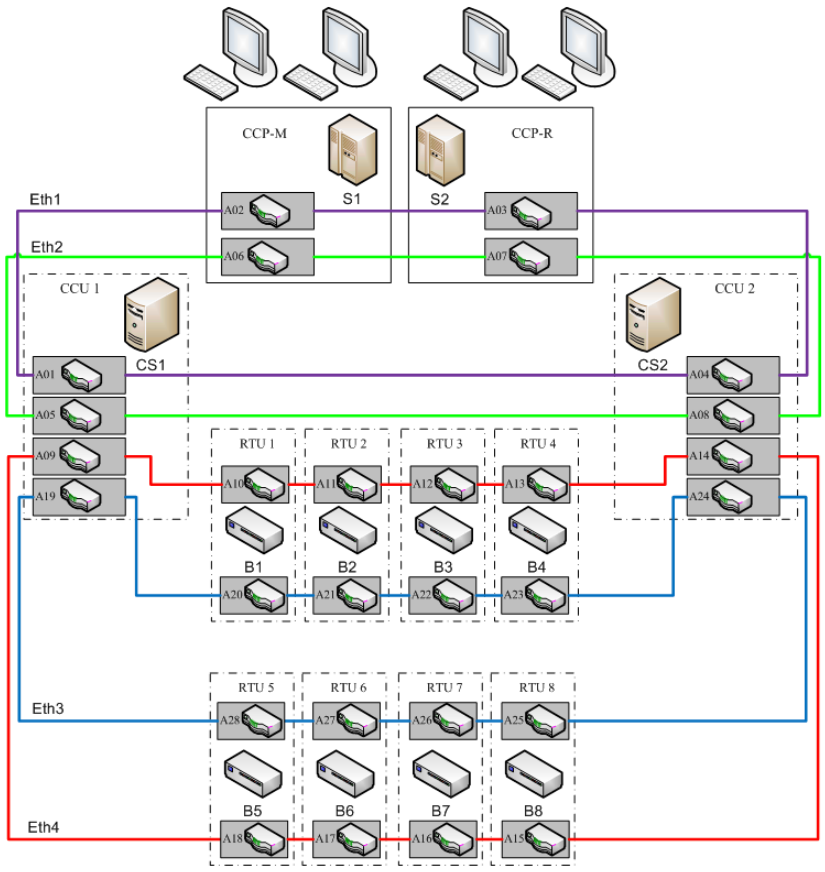


Fig. 1. Three-level SCADA, with duplicated data transfer rings

SCADA is a complex of equipment, distributed across three levels of the hierarchy, depending on the functional purpose: upper level: process operator panels; mid-level: server racks, central computing server, lower level: remote control terminal [12]. The architecture of the process control system takes into account the requirements for the implementation of the principle of a single failure and has structural redundancy [2], the structure of the process control system is shown in Fig. 3.

The system consists of the following units:

1. The Hardware of the main computing resources (S1, S2);

2. Control servers (CS1, CS2) is designed for collecting, processing and storing information about the operation of system equipment, as well as information interaction.

3. Remote terminal unit (B1-B8) is designed for control field equipment.

4. Four independent Ethernet line;

The representation in SCADA of the system in the form of a graph and its representation in the package R is shown in Fig. 4. The standard reliability data (Failure rate: Central control panel main, redundant - $50 \cdot 10^{-6}h^{-1}$, Central control unit - $30 \cdot 10^{-6}h^{-1}$, Remote control terminal - $20 \cdot 10^{-6}h^{-1}$, Commutator - $10 \cdot 10^{-6}h^{-1}$) are considered as initial data [12]. It is assumed that issues related to the process of ensuring computational reliability are provided by the necessary capacities [13].

A system failure is considered to be the loss of communication between the fictitious vertex F1 (SCADA system operator) and the field equipment control subsystems B1-B8.

According to the results of the Monte Carlo simulation, it can be argued that the probability of the SCADA functioning in 5000 hours will be no less than 0.992 with a confidence probability of 0.90. To improve accuracy, methods of reducing the variance of a sample estimate, for example, the Cross-Entropy Monte-Carlo method [11], can be used. Increasing the reliability of the elements with the biggest significance will allow achieving the required failure probability. As a result of the BIM assessment, CPUs of CS unit make the greatest contribution to system reliability. If further reliability improvements are needed, these elements should be considered. Possible ways to improve reliability can be considered: the introduction of continuous monitoring, the choice of more reliable components, using the of redundancy by reserving.

5. Conclusion

The reliability models of three-level networks based on the model systems with independent elements are also considered, a method for assessing reliability is proposed. The Monte Carlo method is used to estimate failure probability and BIM metric with determination of the confidence interval. To conduct a computational experiment, the R software package is used. A description is given of the representation of the control system reliability model in the iGraph package, which provides visualization of the results. R language was primarily created and is continuing to evolve as a statistical data processing tool. The value of the BIM metric is determined for further system improvement.

REFERENCES

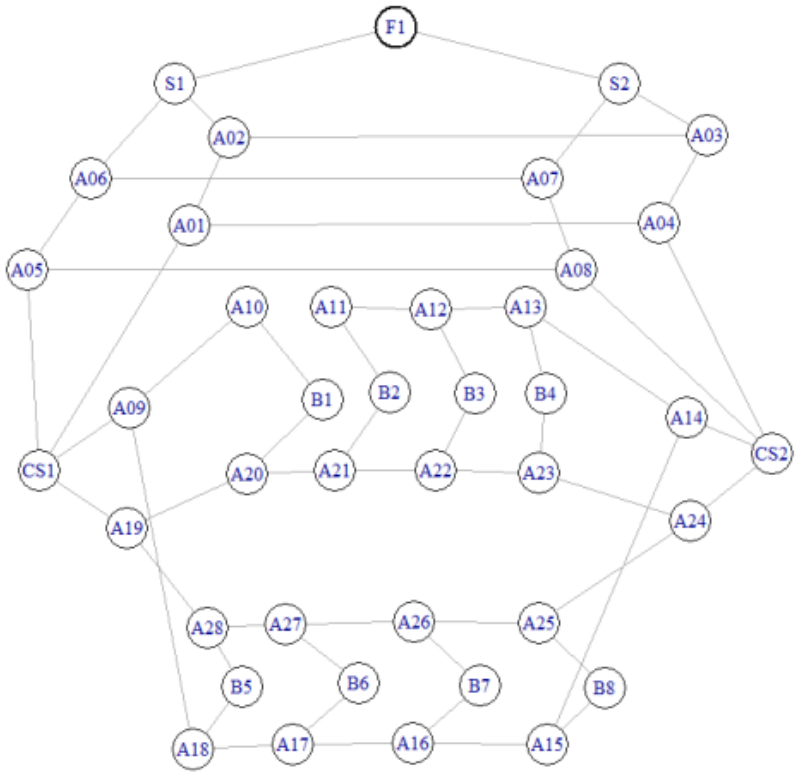


Fig. 2. Graph describing the topology of the process control system (representation in the R language)

1. Krieger U.R., Markovich N. Modeling and Reliability Analysis of a Redundant Transport System in a Markovian Environment. Distributed Computer and Communication Networks. DCCN 2019. Lecture Notes in Computer Science, vol 11965. Springer, Cham
2. Bogatyrev, A.V., Bogatyrev, V.A., Bogatyrev, S.V. Multipath Redundant Transmission with Packet Segmentation (2019) 2019 Wave Electronics and its Application in Information and Telecommunication Systems, WECONF 2019, art. no. 8840643. doi: 10.1109/WECONF.2019.8840643
3. Gertsbakh B., Shpungin Y. Models of Network Reliability: Analysis, Combinatorics, and Monte Carlo// Boca Raton, FL, USA: CRC, 2009
4. Andronov A., Jackiva I., Santalova D. Estimation of the Parameters of Continuous-Time Finite Markov Chain. Distributed Computer and Commu-

nication Networks. DCCN 2019. Lecture Notes in Computer Science, vol 11965. Springer, Cham
5. Nikiforov I. Detection and Detectability of Changes in a Multi-parameter Exponential Distribution. Distributed Computer and Communication Networks. DCCN 2019. Lecture Notes in Computer Science, vol 11965. Springer, Cham
6. Birnbaum Z. W. "On the importance of different components in a multicomponent system," in Multivariate Analysis II. New York, NY, USA: Academic, 1969, pp. 581–592.
7. Y. Du, S. Si, T. Jin. Reliability Importance Measures for Network Based on Failure Counting Process. IEEE Transactions on reliability, vol. 65, no. 1, pp. 267-279
8. Kamalja K. K., Amrutkar K. P. Reliability and Reliability Importance of Weighted-r-Within-Consecutive-k-out-of-n. IEEE Transactions on reliability, vol. 67, no. 3, pp. 951-969
9. Crawley MJ. The R Book. 2nd ed. Wiley Publishing; 2012.
10. Blanchet J., Rudoy D. Rare event simulation and counting problems, in Rare Event Simulation Using Monte Carlo Methods. 1st ed. New York, NY, USA: Wiley, 2009
11. Vaisman R., Kroese D.P., Gertsbakh I.B. Improved Sampling Plans for Combinatorial Invariants of Coherent Systems. IEEE Transactions on reliability, vol. 65, no. 1, pp. 410-424
12. Moshnikov A.S., Bogatyrev V.A. Risk Reduction Optimization of Process Systems under Cost Constraint Applying Instrumented Safety Measures // Computers - 2020, Vol. 9, No. 2, pp. 50
13. Bogatyrev V.A., Bogatyrev S.V., Golubev I.Yu. Optimization and the process of task distribution between computer system clusters. Automatic Control and Computer Sciences 2012, 46(3), pp. 103-111
14. Compare M, Bellora M, Zio E. Aggregation of importance measures for decision making in reliability engineering. Proceedings of the Institution of Mechanical Engineers 2017, Part O: Journal of Risk and Reliability, 231(3): 242–254
15. Fang C, Marle F, Xie M. Applying importance measures to risk analysis in engineering project using a risk network model. IEEE Systems Journal 2017, 11(3): 1548–1556
16. Kuo W, Zhu X (2012). Importance Measures in Reliability, Risk and Optimization: Principles and Applications. Chichester: John Wiley and Sons

UDC: 123.456

SARSA based method for WSN transmission power management

A.Alexandrov [1] and V. Monov [1]

[1]Institute of Information and Communication Technologies - Bulgarian Academy of Sciences, Akad. G.Bonchev 1113, Sofia, Bulgaria

akalexandrov@iit.bas.bg, vmonov@iit.bas.bg

Abstract

The scope of this research is to propose an adaptive machine learning approach which can help the WSN's nodes to manage their transmission power and to improve the internode wireless communications. The optimized transmission power has benefits in terms of WSN energy consumption and RF interlink interference. The paper proposes an adaptive method of a wireless sensor node based on Multi-Layer Perceptron (MLP) network representation and machine learning. The presented in the paper approach, uses the SARSA (State-Action-Reward-State-Action) algorithm which is a form of reinforcement machine learning. The aim of the new method is to improve the sensor nodes Transmission Power Management (TPM) process. This inspires many practical solutions that maximize resource utilization and prolong the shelf life of the battery-powered wireless sensor networks.

Keywords: ANN, MLP, neural networks, wireless sensor network, transmission power control, SARSA, energy efficiency, quality of service.

1. Introduction

Inter node communications are usually the most energy consuming event in Wireless Sensor Networks (WSNs). One way to significantly reduce energy consumption is by applying an adaptive transmission power management techniques. This approach dynamically adjusts the transmission power in which depends on factors as wireless link Quality of Service (QoS) and the wireless node Received Signal Strength (RSS) value. As is illustrated on the Fig.1 the reliable connection between sensor nodes depends on the distance between nodes, received signal strength and the level of the existing RF noise. In the real environment, the deviation between the needed transmission power for reliable communication between WSN nodes at one and the same distance can reach dramatically high values because of the mentioned above factors.[1]

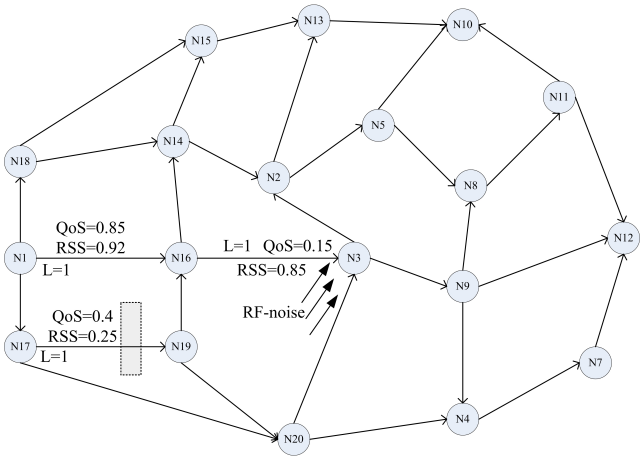


Fig. 1. WSN with RF barrier between Note17 and Node19 and RF noise source near Node3

The task of the WSN power management (WSN-PM) stays more complex when the propagation of the RF transmission signal is influenced by the factors which are time changeable as cyclic sources of RF noise, interference RF sources with variable sizes and etc. Therefore, one of the possible ways to solve the complexity problem is to use an adaptive method of a wireless sensor node based on a self-learning Artificial Neural Network (ANN). The machine learning is a set of algorithms and statistical models that software application use to perform a specific task without using explicit instructions, relying on patterns and inference instead. The method takes place when the problem is too complicated to be solved in real time, or in case that is not impossible the problem to be solved in a classical way. One of the machine learning methods is the Reinforcement Learning (RL) [2]. The RL method uses an agent executor - environment approach and is based on the concept of reward. Reinforcement Learning involves two main entities: an agent and the environment (Figure 2). The agent plays as a learner and decision-maker at the same time, while the environment is unpredictable and unknown which influences the agent's performance.

Where:

S_k - represents the status of the environment;

a_k – actions – decisions of the Agent. It is noted by default that the agent can choose among a predefined list of possible actions.

r_k - feedback called reward which evaluates the effect of the actions a_k

In our case the target of the objective function for the reinforcement learning approach is to maximize the cumulative reward r_k as follows:

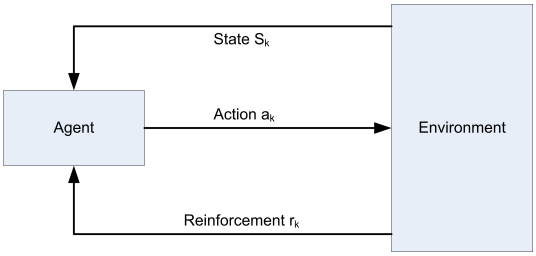


Fig. 2. Reinforcement Learning model - agent environment interaction

$$MAX \sum_{k=0}^{k=\infty} \lambda^k \mathbb{E}[r_k(a_k, s_k)] \tag{1}$$

Where λ^k - is the probability distribution of the k reward

A possible implementation of the reinforcement learning methods is the Temporal Difference (TD) Learning approach. The TD approach refers to a class of the model-free reinforcement methods which learns by the state of the current estimate of the value function. The possible options of the TD methods are the Q-Learning, SARSA, Rescorla-Wagner, PVLV and etc. In the current research, we are focused mainly on the SARSA method and the algorithm as relatively the most adaptive and flexible for the needs of WSN power management. SARSA is part of the group of Temporal Difference (TD) algorithms used in Reinforcement Learning and it was proposed in 1994 by Rummery & Niranjan [3]

2. Related work

To overcome the disadvantages of the proactive and reactive techniques, machine learning represents an attractive solution [4] to reach a defined goal by learning the dynamics of the WSNs, predicting and adapting the transmission power values in different conditions. The objective is making WSNs autonomous without the intervention of developers and users to set the transmission power.

To the best of our knowledge, only a few contributions have applied machine learning in TPC, mainly fuzzy logic and Reinforcement Learning (SARSA, Q-Learning and etc.)[5].

Q-Learning in WSNs has been used as WSN management approach in the literature but mainly for path selection in routing protocols and sleeping techniques, maintaining constant learning factors [6]. The static values would either bring the system slowly to convergence or make the system too reactive if the learning factor is constantly low or high respectively.

SARSA [7,8]as a reinforcement learning method is similar to Q-Learning. The main difference between SARSA and Q-Learning is that SARSA uses an on-policy algorithm which means that SARSA calculates the Q-value based on the action executed by the current policy and is contrary to the off-policy used in Q-Learning.

3. Proposed model and SARSA based method

In the current development, we represent the WSN as a set of Multi-Layer Perceptron's (MLPs). Every wireless node can be represented as a perceptron consisting of four components, i.e. inputs, weights, activation function, and output.

The proposed MLP model of the wireless sensor node is shown in Fig. 3.

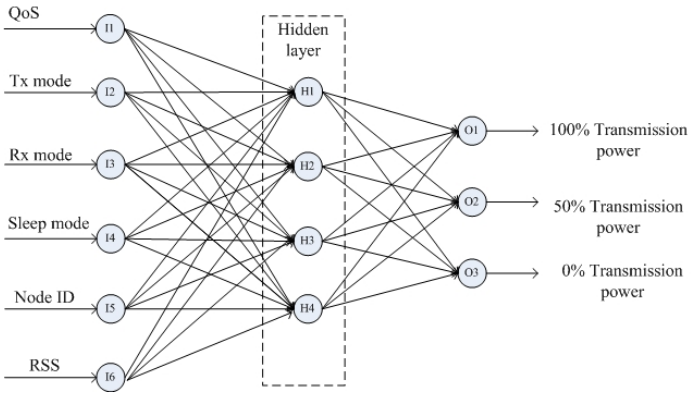


Fig. 3. MLP power management model with a single hidden layer of WSN node

The considered mechanism takes as input the mode as well as the energy consumption in a specific interval of time. The captured weights are processed in the hidden layer. The three possible outputs are 100% Power transmission, 50% Power transmission and No transmission (0% Power transmission).

According to the MLP diagram above the inputs of the MLP model of a sample WSN node are as follows:

- QoS – Quality of Service. This parameter is calculated by the wireless sensor node and depends on the bandwidth, packet delay, and packet loss real-time measurement;
- Tx mode – the mode when the wireless sensor device transmits RF packets. In the current research is considered only the Tx mode parameter transmission power which can vary from 0 to 100% of the existing RF device transmission capacity;
- R_x mode – mode of the receiving data packets from the sensor node;
- Sleep mode – mode when the sensor node doesn't transmit or receive any data;
- Node ID – the unique ID assigned of the node during the WSN forming and configuration.

- RSS – Received Signal Strength. Also referred to as RSSI (Received Signal Strength Indicator) is the parameter calculated on the basis of the RF power presented in the received radio signal. RSS is measured in dB and typically vary between 0dBm (excellent) and -110 dBm (very poor).

The output of the proposed MLP model is related to the level of power transmission as the main energy utilization parameter and the key factor for transmission power management.

In the active mode (Tx or Rx mode), the node fall in the following active patterns: 100% transmit power, 50% transmit power and 0% transmit power.

The proposed method, considered in this paper consists of three main phases – data collection, learning, and results in phases.

Data collection – in this phase, every wireless sensor node keeps track of the RF packets received from neighbors, the packets forwarded to neighbors and the average energy consumption in a specific period of time. This phase may take time to get a few optimal values of the system usage over a specific period of time.

The following parameters are collected and calculated during the initial data collection phase:

- the average value of the RSS signal for the last 10 received RF packets;
- the average value of the QoS parameter calculated for the last 10 received RF packets;
- Tx mode is set to 100% RF transmission power;
- Rx mode is set to state 0 (wait);
- Sleep mode is set to state 0 (wait);
- NodeID is assigned and fixed during the WSN configuration;

Learning phase – in this phase, the modified MLP is trained to identify different communication sources which are located in its environment.

In this stage, the sensor node learns the parameters of packets which the neighbor wireless node receives and sending over a specific time and the related energy consumption. The learning process uses SARSA algorithm as a form of implementation of Reinforcement Learning.

Based on the SARSA based Reinforcement Learning function definition described in details in [7], we have:

$$Q_k = r_k(s_k, a_k) + \gamma \, maxQ(S_k, a_k, w_k) \quad Q_k \in (0, 0.5, 1) \tag{2}$$

Where:

r_k – reinforcement value

s_k - state of the environment in stage k

a_k – action state

γ - discount factor $0 \leq \gamma \leq 1$

w_k – weights in stage k

$maxQ(s_k, a_k, w)$ - is the function value state/action

A diagram which describes the practical implementation of the proposed method is shown on Fig.4.

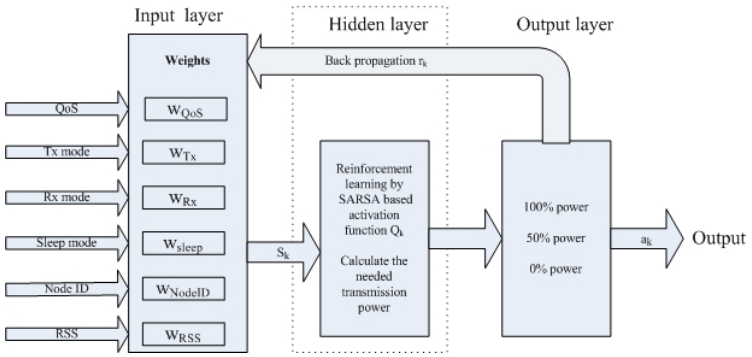


Fig. 4. The working mechanism of the proposed method

In the proposed method the learning process starts with a preliminary fixed combination of weights (w_k), s_k=0 and a_k=0, forming the initial function value state/action $Q(s_k, a_k, w)$. During the iterations stages of the learning process, Q-function generates a corrective signal and send it to the input of the system.

The reinforcement value r_k provides for every new state S, a signal that can be reward or punishment. In our case, r_k takes values -1, 0 or 1.

The discount factor γ role is to mark the importance of future reward. In our case,γ takes only two values: 0 or 1.

The captured weights (w_{QoS}, w_{Tx}, w_{Rx}, w_{sleep}, w_{NodeID} and w_{RSS}) are multiplied with the input values and processed in the hidden layer for an activation function generation.

As part of the reinforcement learning process the SARSA algorithm follows the main actions shown on Fig.2:

- the environment sends his state to the agent;
- the agent takes action in response;
- the environment sends a pair of next state and rewards back to the agent;
- the agent updates its knowledge with the feedback from the environment to evaluate its last action;
- the cycle continues until the environment sends a termination signal.

The updated equation, related to the modified SARSA algorithm is:

$$QS_t, a_t \leftarrow QS_t, a_t + \alpha[r_{t+1} + \gamma QS_t, a_{t+1} - QS_t, a_t] \tag{3}$$

Where:
Q – action value which refers to state S and action a in moment t and t+1;
S_t – State of the environment in moment t;
a_t – action of the agent in moment t;
r_{t+1} - reinforcement value in moment t+1;
α – learning rate level $\alpha \in [0, 1]$;
γ – discount factor $\gamma \in [0, 1]$;

Results – In this phase, the accuracy of the learning process is reviewed. The sys-tem is capable to calculate the needed transmission energy amount depend on the current topology of the neighbor nodes and their current status.

Based on the described above method were prepared two simulations to compare SARSA and Q-Learning algorithms.

The executed simulation experiments based on NS2 network simulator shows that the learning phase has a typical length between 1 and 10 hours for 1000 sensor node based WSN and depends on the needed accuracy of the wireless system.

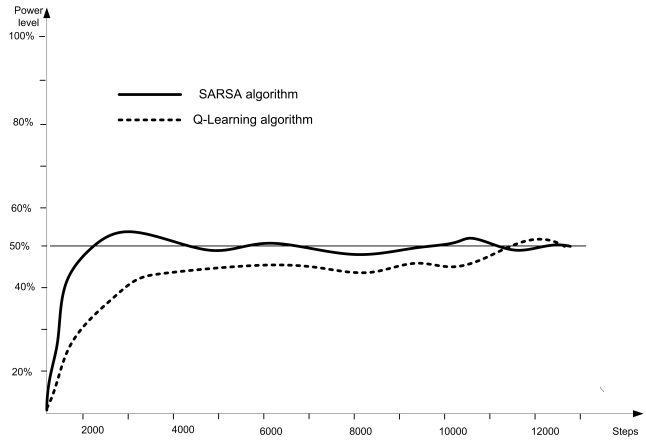


Fig. 5. The performance of SARSA compared to the Q-Learning algorithm

The diagram from Fig. 5 shows that the proposed modification of SARSA algorithm is sensitively faster and reach the level of the predefined power level of the simulated sensor node. As is shown from the diagram after approximately 6000 itera-tions the SARSA algorithm achieved the predefined level of 50 percent power level com-pared to the 11000 iterations of the Q-Learning algorithm.

4. Conclusions

We have proposed a new model and adaptive method for wireless sensor node power management based on the SARSA algorithm which uses an AI Reinforcement Learning approach. The proposed sensor node model is developed on the basis of Multi-Layer Perceptron (MLP) network representation and machine learning.

The simulation results show that the implemented in a wireless sensor node SARSA algorithm has sensitively better performance compared to the related Q-Learning algorithm.

At the same time, the software implementation of the SARSA algorithm is more compact and uses less sensor node microcontroller resources compared to the Q-Learning algorithm.

This inspires many practical solutions that can maximize resource utilization and prolong the shelf life of the battery-powered wireless sensor networks.

REFERENCES

1. Gummeson, J.; Ganesan, D.; Corner, M.D.; Shenoy, P. An adaptive link layer for hetero-geneous multi-radio mobile sensor networks. IEEE J. Sel. Areas Commun. 2010, 28, 1094–1104.Balmelli
2. Wiering, M.; van Otterlo, M. (Eds.) Reinforcement Learning. In Adaptation, Learning, and Optimization; Springer: Berlin/Heidelberg, Germany, 2012; Volume 12.
3. Rummery, Gavin and Niranjan, Mahesan (1994), On-line Q-learning using Connectionist systems, technical report no.166, University of Cambridge, Engineering Department.
4. Torrey, L.; Shavlik, J. Transfer learning. In Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques; IGI Global: Hershey, PA, USA, 2009; pp. 242-264.
5. Sung, Y.; Ahn, E.; Cho, K. Q-learning Reward Propagation Method for Reducing the Transmission Power of Sensor Nodes in Wireless Sensor Networks. Wirel. Pers. Commun. 2013, 73, 257–273.
6. A. Forster, "Machine learning techniques applied to wireless ad-hoc networks: Guide and survey," in 3rd International Conference on Intelligent Sensors, Sensor Networks and In-formation. IEEE, 2007, pp. 365–370
7. L. P. Kaelbling, M. L. Littman, and A. P. Moore, "Reinforcement learning: A survey," Journal of Artificial Intelligence Research, vol. 4, pp. 237–285, 1996
8. 11. Van Seijen, H., Van Hasselt, H., Whiteson, S. and Wiering, M. (2009) A Theoretical and Empirical Analysis of Expected Sarsa. 2009 IEEE Symposium on Adaptive Dynamic Pro-gramming and Reinforcement Learning, Nashville, 30 March-2 April 2009, 177-184. http://dx.doi.org/10.1109/ADPRL.2009.4927542

UDC: 519.872

Asymptotic analysis of $M^{[n]}/M/1$ RQ-system with feedback and batch Poisson arrival

A.A. Nazarov[1], S.V. Rozhkova[1,2], E.Yu. Titarenko[1,2]

[1]Tomsk State University, 36, Lenin Avenue, Tomsk, Russia
[2]Tomsk Polytechnic University, 30, Lenin Avenue, Tomsk, Russia

nazarov.tsu@gmail.com, rozhkova@tpu.ru, teu@tpu.ru

Abstract

The paper studies the $M^{[n]}/M/1$ RQ-system with batch Poisson arrival. Customers for system come in groups. Every moment in time no more than one customer is served, others go into orbit. Having been served, the customer leaves the system or goes to re-service or into orbit. An asymptotic analysis method is used to find the stationary distribution of the number of customers in the orbit. A long delay between customers from the orbit is proposed as an asymptotic condition.

Keywords: queuing system, RQ system, batch arrival, feedback, asymptotic analysis

1. Introduction

Quite often in practice, there are queuing systems in which a customer that has already received service requires re-service, depending on the quality of service received, external factors, etc. Similar situations occur in multi-agent systems (MAS), where a customer having been received satisfactory service requires reservice from the same agent. The functioning of such systems is accurately described by queuing systems with feedback. The application of queuing theory to optimize the performance of multi-agent systems is described in detail in [1, 2, 3, 4]. Retrial queuing systems are considered in many works [4, 6]. In this paper, we study a single-channel RQ-system with exponential service, a batch Poisson arrival, with instant and delayed feedbacks.

2. The model description and the problem statement

We consider the $M^{[n]}/M/1$ queuing system with repeated calls and batch Poisson arrival process with parameter λ and given probabilities q_ν of occurrence of customers

in the group ($\nu > 0, q_0 = 0, \sum_{\nu=1}^{\infty} q_\nu = 1$). An incoming customer, which sees a server idle, occupies it, other customers from the group go to the source of the repeated calls (into orbit). Also, if the server is busy, arriving customers go into orbit. Service time is exponentially distributed with parameter μ. Having been served, the customer leaves the system with probability r_0, or goes to re-service with probability r_1, or into orbit with probability r_2, so $r_0 + r_1 + r_2 = 1$. In the orbit each customer independently of others waits for the time exponentially distributed with parameter σ. Then the customer occupies the device if it is idle or remains in the orbit.

We define the Markov process $\{i(t), n(t)\}$ of changing the states of the RQ-system, where $i(t)$ is the number of customers in the orbit at time t, $i(t) = 0, 1, 2, ...$, the process $n(t)$ determines the state of the server at time t, and takes one of two values: $n(t) = 0$, if the server is idle, $n(t) = 1$, if the server is busy.

Our aim is to find the stationary probability distribution of the number of customers in the orbit taking into account the state of the server $P_n(i) = P\{n(t) = n; i(t) = i\}, n = 0, 1, i = \overline{0, \infty}$.

3. Kolmogorov equations

To obtain the probability distribution $P_n(i)$ for the number of customers in the orbit, we derive a system of Kolmogorov equations

$$-(\lambda + i\sigma)P_0(i) + \mu r_0 P_1(i) + \mu r_2 P_1(i-1) = 0; \tag{1}$$

$$(i+1)\sigma P_0(i+1) - (\mu r_0 + \mu r_2 + \lambda)P_1(i) + \sum_{\nu=1}^{i+1} \lambda q_\nu P_0(i-\nu+1) + \sum_{\nu=1}^{i} \lambda q_\nu P_1(i-\nu) = 0.$$

We consider the partial characteristic functions of the number of customers in the orbit

$$H_n(u) = \sum_{i=0}^{\infty} e^{jui} P_n(i)$$

and the number of customers in the group $h(u) = \sum_{\nu=1}^{\infty} e^{ju\nu} q_\nu$, where $j = \sqrt{-1}$. We take into account $\frac{\partial H_n(u)}{\partial u} = \sum_{i=0}^{\infty} ij e^{jui} P_n(i)$,

$$\sum_{i=0}^{\infty} \sum_{\nu=1}^{i} q_\nu e^{jui} P_1(i-\nu) = \sum_{\nu=1}^{\infty} q_\nu e^{ju\nu} \sum_{i=0}^{\infty} e^{jui} P_1(i) = h(u) H_1(u),$$

$$\sum_{i=0}^{\infty} \sum_{\nu=1}^{i+1} q_\nu e^{jui} P_0(i-\nu+1) = e^{-ju} \sum_{\nu=1}^{\infty} q_\nu e^{ju\nu} \sum_{i=0}^{\infty} e^{jui} P_0(i) = e^{-ju} h(u) H_0(u),$$

and rewrite system (1) as

$$\sigma j\frac{\partial H_0(u)}{\partial u} - \lambda H_0(u) + \left(\mu r_0 + \mu r_2 e^{ju}\right) H_1(u) = 0; \tag{2}$$

$$-\sigma j e^{-ju}\frac{\partial H_0(u)}{\partial u} + \lambda e^{-ju} h(u) H_0(u) + \left(\lambda h(u) - \mu r_0 - \mu r_2 - \lambda\right) H_1(u) = 0.$$

The characteristic function $H(u)$ of the number of customers in the orbit for system (2) is expressed in terms of the partial characteristic functions $H_n(u)$ by the following $H(u) = H_0(u) + H_1(u)$.

Applying the inverse Fourier transform to the characteristic function, we can write the probability distribution as $P(i) = \frac{1}{2\pi}\int\limits_{-\pi}^{\pi} e^{-jui} H(u)du$, $i = \overline{0,\infty}$ and mathematical expectation as $E\{i(t)\} = -jH'(0)$. However, it is hardly possible to obtain an analytical expression for these integrals, therefore, it is reasonable to use numerical integration methods. Numerical calculations are computationally expensive, so we consider the asymptotic analysis method which allows us to receive an analytical approximation for the distribution $P(i)$. Using numerical experiments, we analyze the accuracy of the method.

4. Asymptotics of the first order

We solve the equations for the characteristic functions (2) under the asymptotic condition of a growing average waiting time in the orbit, i.e. $\sigma \to 0$. We formulate the result in the following theorem.

Theorem 1. Let $i(t)$ be the number of customers in the orbit in the $\mathrm{M}^{[n]}/\mathrm{M}/1$ RQ-system with a batch Poisson arrival and feedback. Then there is the following equality for a sequence of characteristic functions

$$\lim_{\sigma\to 0} E\left\{e^{jwi(t)\sigma}\right\} = e^{jw\kappa_1}$$

where $\kappa_1 = \lambda\frac{\mu\bar{\nu}(r_0+r_2)}{\mu r_0 - \lambda\bar{\nu}} - \lambda, \bar{\nu} = \sum\limits_{\nu=1}^{\infty}\nu q_\nu$.

Proof. In the system of equations (2) we use the substitutions $\sigma = \varepsilon$, $u = \varepsilon w$, $H_n(u) = F_n(w,\varepsilon)$. Since $\frac{\partial H_0(u)}{\partial u} = \frac{1}{\varepsilon}\frac{\partial F_0(w,\varepsilon)}{\partial w}$, the system (2) can be written as

$$j\frac{\partial F_0(w,\varepsilon)}{\partial w} - \lambda F_0(w,\varepsilon) + \left(\mu r_0 + \mu r_2 e^{jw\varepsilon}\right) F_1(w,\varepsilon) = 0; \tag{3}$$

$$-je^{-jw\varepsilon}\frac{\partial F_0(w,\varepsilon)}{\partial w} + \lambda e^{-jw\varepsilon} h(w,\varepsilon) F_0(w,\varepsilon) + \left(\lambda h(w,\varepsilon) - \mu r_0 - \mu r_2 - \lambda\right) F_1(w,\varepsilon) = 0.$$

Let $\varepsilon \to 0$, $F_n(w) = \lim\limits_{\varepsilon\to 0} F_n(w,\varepsilon)$. Since $\lim\limits_{\varepsilon\to 0} h(w,\varepsilon) = 1$, the system (2) is transformed into an equation

$$j\frac{\partial F_0(w)}{\partial w} - \lambda F_0(w) + (\mu r_0 + \mu r_2)\, F_1(w) = 0.$$

We find a solution to the equation of the form

$$F_n(w) = R_n e^{jw\kappa_1}, \tag{4}$$

then

$$-(\kappa_1 + \lambda)R_0 + (\mu r_0 + \mu r_2)R_1 = 0. \tag{5}$$

Then, summing the equations of system (3), we obtain

$$j\left(1 - e^{-jw\varepsilon}\right)\frac{\partial F_0(w,\varepsilon)}{\partial w} - \lambda\left(1 - e^{-jw\varepsilon}h(w,\varepsilon)\right)F_0(w,\varepsilon) +$$
$$+ \left(\mu r_2(e^{jw\varepsilon} - 1) + \lambda(h(w,\varepsilon) - 1)\right)F_1(w,\varepsilon) = 0,$$

divide it by ε

$$j\frac{1 - e^{-jw\varepsilon}}{\varepsilon}\cdot\frac{\partial F_0(w,\varepsilon)}{\partial w} - \lambda\frac{1 - e^{-jw\varepsilon}h(w,\varepsilon)}{\varepsilon}\cdot F_0(w,\varepsilon) +$$
$$+ \left(\mu r_2\cdot\frac{e^{jw\varepsilon} - 1}{\varepsilon} + \lambda\cdot\frac{h(w,\varepsilon) - 1}{\varepsilon}\right)F_1(w,\varepsilon) = 0$$

and assume $\varepsilon \to 0$

$$j\frac{\partial F_0(w)}{\partial w} + \lambda\,(\bar{\nu} - 1)\cdot F_0(w) + (\mu r_2 + \lambda\bar{\nu})\, F_1(w) = 0,$$

where $\bar{\nu} = \sum\limits_{\nu=1}^{\infty}\nu q_\nu$. We carry out the substitution (4) and obtain the equation

$$-(\kappa_1 + \lambda - \lambda\bar{\nu})R_0 + (\mu r_2 + \lambda\bar{\nu})\, R_1 = 0 \tag{6}$$

Solving the system of equations (5), (6) with the additional condition $R_0 + R_1 = 1$, we obtain

$$R_0 = \frac{\mu(r_0 + r_2)}{\kappa_1 + \lambda + \mu(r_0 + r_2)},\ R_1 = \frac{\kappa_1 + \lambda}{\kappa_1 + \lambda + \mu(r_0 + r_2)} \tag{7}$$
$$\kappa_1 = \lambda\frac{\mu\bar{\nu}(r_0 + r_2)}{\mu r_0 - \lambda\bar{\nu}} - \lambda.$$

Thus, the asymptotic approximation of the characteristic function is $F(w) = e^{jw\kappa_1}$. ■

Asymptotics of the first order determines the average value of the number of customers in the orbit. For a more detailed study of the process $i(t)$, we should consider the second-order asymptotics.

5. Asymptotics of the second order

The main result of the analysis of the second-order asymptotics is presented in the following theorem.

Theorem 2. Let $i(t)$ be the number of customers in the orbit in the $M^{[n]}/M/1$ RQ-system with a batch Poisson arrival and feedback. Then there is an equality as follows:

$$\lim_{\sigma\to 0} E\left\{\exp\left(jw\sqrt{\sigma}\left(i(t)-\frac{\kappa_1}{\sigma}\right)\right)\right\} = \exp\left(\frac{(jw)^2}{2}\kappa_2\right), \tag{8}$$

where $\kappa_2 = \frac{\lambda\mu r_0}{2(\mu r_0-\lambda\bar{\nu})^2}\left(2\lambda\bar{\nu}^2+\mu\nu_2(r_0+r_2)+\mu\bar{\nu}(r_2-r_1)\right)$, $\nu_2 = \sum\limits_{\nu=1}^{\infty}\nu^2 q_\nu$.

Proof. In the system of equations (2), we use the substitutions $H_n(u) = H_n^{(2)}(u)\cdot e^{ju\kappa_1/\sigma}$. Here $H_n^{(2)}(u)$ is the characteristic function of the centered random variable $i(t)-\kappa_1/\sigma$. The system of equations for $H_n^{(2)}(u)$ is

$$\sigma j\frac{\partial H_0^{(2)}(u)}{\partial u} - (\kappa_1+\lambda)H_0^{(2)}(u) + \left(\mu r_0+\mu r_2 e^{ju}\right)H_1^{(2)}(u) = 0;$$

$$-\sigma j e^{-ju}\frac{\partial H_0^{(2)}(u)}{\partial u} + (\kappa_1+\lambda h(u))\,e^{-ju}H_0^{(2)}(u) + (\lambda h(u)-\mu r_0-\mu r_2-\lambda)\,H_1^{(2)}(u) = 0.$$

Let $\sigma=\varepsilon^2$ and use the substitutions $u=\varepsilon w$, $H_n^{(2)}(u) = F_n^{(2)}(w,\varepsilon)$, then we obtain the system

$$\varepsilon j\frac{\partial F_0^{(2)}(w,\varepsilon)}{\partial w} - (\kappa_1+\lambda)F_0^{(2)}(w,\varepsilon) + \left(\mu r_0+\mu r_2 e^{jw\varepsilon}\right)F_1^{(2)}(w,\varepsilon) = 0; \tag{9}$$

$$-\varepsilon j e^{-jw\varepsilon}\frac{\partial F_0^{(2)}(w,\varepsilon)}{\partial w} + (\kappa_1+\lambda h(w,\varepsilon))\,e^{-jw\varepsilon}F_0^{(2)}(w,\varepsilon) + (\lambda h(w,\varepsilon)-\mu r_0-\mu r_2-\lambda)\,F_1^{(2)}(w,\varepsilon) = 0.$$

The solution for the functions $F_n^{(2)}(w,\varepsilon)$ has the following form

$$F_n^{(2)}(w,\varepsilon) = \Phi(w)\cdot(R_n + j\varepsilon w f_n) + O(\varepsilon^2) \tag{10}$$

We substitute (10) into (9), use the approximation for $e^{\pm jw\varepsilon}$ and $h(\varepsilon w) = 1 + j\varepsilon w\bar{\nu} + O(\varepsilon^2)$, take into account (5), (6), and convert the system of equations (9) into an equation

$$\frac{\partial(w)}{\partial w}\frac{1}{w\Phi(w)} = \frac{1}{R_0}\left((\kappa_1+\lambda)f_0 - (\mu r_0+\mu r_2)\,f_1 - \mu r_2 R_1\right). \tag{11}$$

Let us denote

$$\kappa_2 = -\frac{1}{R_0}\left((\kappa_1+\lambda)f_0 - (\mu r_0 + \mu r_2)\, f_1 - \mu r_2 R_1\right) \tag{12}$$

and equation (11) has the form

$$\frac{1}{\Phi(w)}\frac{\partial(w)}{\partial w} = -w\kappa_2, \tag{13}$$

therefore, the function $\Phi(w)$ can be represented in the form $\Phi(w) = exp\left\{-\frac{1}{2}w^2\kappa_2\right\}$ that correlates with (8). To find unknown functions f_0, f_1 and an explicit form for κ_2, we rewrite (12) as

$$-(\kappa_1+\lambda)f_0 + \mu\,(r_0+r_2)\, f_1 = \kappa_2 R_0 - \mu r_2 R_1.$$

Functions f_0, f_1 can be written as the sum of a general solution of the homogeneous equation and two particular solutions:

$$f_n = C\cdot R_n + g_n + \kappa_2\varphi_n,\ n = 0, 1. \tag{14}$$

Here $C\cdot R_n$ are the general solution of the homogeneous equation due to (5), while g_n is the solution of the equation

$$-(\kappa_1+\lambda)g_0 + \mu\,(r_0+r_2)\, g_1 = -\mu r_2 R_1, \tag{15}$$

and φ_n satisfies the equation

$$-(\kappa_1+\lambda)\kappa_2\varphi_0 + \mu\,(r_0+r_2)\,\kappa_2\varphi_1 = \kappa_2 R_0. \tag{16}$$

Differentiating (5) with respect to κ and comparing with (16), we note that $\varphi_0 = \frac{\partial R_0}{\partial\kappa}$, $\varphi_1 = \frac{\partial R_1}{\partial\kappa}$, $\varphi_0 + \varphi_1 = 0$. Then, taking into account (7), we obtain $\varphi_0 = -\frac{\mu(r_0+r_2)}{(\kappa_1+\lambda+\mu(r_0+r_2))^2}$, $\varphi_1 = -\varphi_0$. Similarly, we assume $g_0 + g_1 = 0$ and receive from equation (15) $g_0 = \frac{\mu r_2 R_1}{\kappa_1+\lambda+\mu(r_0+r_2)}$, $g_1 = -g_0$.

In order to find the explicit form κ, we sum the equations of the system (9)

$$-\varepsilon j\left(e^{-jw\varepsilon}-1\right)\frac{\partial F_0^{(2)}(w,\varepsilon)}{\partial w} + \left(\kappa_1(e^{-jw\varepsilon}-1) + \lambda(h(w,\varepsilon)e^{-jw\varepsilon}-1)\right)F_0^{(2)}(w,\varepsilon) +$$

$$+\left(\mu r_2(e^{jw\varepsilon}-1) + \lambda(h(w,\varepsilon)-1)\right)F_1^{(2)}(w,\varepsilon) = 0.$$

We write the solution for $F_n^{(2)}(w,\varepsilon)$ as (10), $e^{\pm jw\varepsilon} = 1 \pm jw\varepsilon + (jw\varepsilon)^2/2 + O(\varepsilon^3)$, $h(\varepsilon w) = 1 + j\varepsilon w\bar{\nu} + \frac{(j\varepsilon w)^2\nu_2}{2} + O(\varepsilon^3)$, where $\nu_2 = \sum\limits_{\nu=1}^{\infty} \nu^2 q_\nu$. After some transformation and using expressions (6), (13), we assume $\varepsilon \to 0$ and obtain

$$R_0\kappa_2 = \frac{1}{2}(\kappa_1 + \lambda + \lambda\nu_2 - 2\lambda\bar{\nu})R_0 + \frac{1}{2}\left(\mu r_2 + \lambda\nu_2\right)R_1 -$$
$$-(\kappa_1 + \lambda - \lambda\bar{\nu})f_0 + (\mu r_2 + \lambda\bar{\nu})\, f_1 \qquad (17)$$

Substituting (14) into (17) and taking into account (6), (15), (16), expressions for $R_0,\ R_1,\ \varphi_1,\ g_1$, we finally obtain $\kappa_2 = \frac{\lambda\mu r_0}{2(\mu r_0 - \lambda\bar{\nu})^2}\left(2\lambda\bar{\nu}^2 + \mu\nu_2(r_0 + r_2) + \mu\bar{\nu}(r_2 - r_1)\right)$. ■

Theorem 2 shows that the asymptotic probability distribution of the number of customers in the orbit in the $M^{[n]}/M/1$ RQ-system with a batch Poisson arrival and feedback is Gaussian with the parameters κ_1/σ and κ_2/σ, which allows us to make the following approximation for the distribution $P(i)$ as

$$P_{apr}(i) = \frac{G(i+0.5) - G(i-0.5)}{1 - G(-0.5)},$$

where $G(x)$ is the normal distribution function with parameters κ_1/σ and κ_2/σ.

6. Numerical results

We consider the system with parameters $\lambda = 1, \mu = 7, r_0 = 0.5, r_1 = 0.3, r_2 = 0.2, q_1 = 0.5, q_2 = 0.3, q_3 = 0.1, q_4 = 0.1, q_5 = 0, q_6 = 0, ...$ Table 1 shows the results of calculating the mathematical expectation of the number of customers in the orbit for various values of σ. For each value of σ, the exact value of $E\{i(t)\}$ obtained with characteristic function, the asymptotic value of the mathematical expectation κ_1/σ, and the relative error $\delta = \left|E\left\{i(t)\right\} - \frac{\kappa}{\sigma}\right| / E\left\{i(t)\right\}$ are shown.

σ	1	0.5	0.1	0.01	0.005
$E\{i(t)\}$	5.439	9.627	43.133	420.074	838.897
κ_1/σ	4.188	8.376	41.882	418.824	837.647
δ	0.2	0.1	0.03	0.003	0.001
Δ	0.1	0.05	0.03	0.009	0.006

Table 1. Accuracy of the approximation

To determine the accuracy of the approximation $P_{apr}(i)$, we use the Kolmogorov distance $\Delta = \max\limits_{0 \le k \le \infty} \left| \sum\limits_{i=0}^{k} \left(P_{apr}(i) - P(i)\right) \right|$ that defines the difference between the

asymptotic probability distribution $P_{apr}(i)$ and the probability distribution $P(i)$ obtained by the matrix method. Table 1 shows the Kolmogorov distance for a given set of parameters and different values of the parameter σ.

Thus, the asymptotic method can be used to find probability distribution and the average number of customers in orbit with a long waiting time in orbit, i.e. when $\sigma < 0.01$.

7. Conclusion

In this paper, we have studied the queuing system $M^{[n]}/M/1$ with feedback and batch Poisson arrival, we have applied the method of asymptotic analysis under the condition of growing average waiting time in the orbit. To determine the range of the obtained approximation with regard to the parameters of the system, the calculations were carried out. The obtained results show the convergence of asymptotic results to prelimit ones which obtained using the matrix method.

REFERENCES

1. Horling B., Lesser V. Using Queueing Theory to Predict Organizational Metrics // AAMAS'06, Hakodate, Japan. 2006. P. 1098–1100.
2. Gnanasambandam N., Lee S. C., Gautam N., et al. Reliable MAS Performance Prediction Using Queueing Models // IEEE First Symposium on Multi-Agent Security and Survivalibility. 2004. P. 55–64.
3. Gnanasmbandam N., Lee S., Kumara S. R. T. An autonomous performance controlframework for distributed multi-agent systems: A queueing theory based approach // AAMAS'05.July 25–29, 2005. Utrecht, Netherlands. P. 1313–1314.
4. Lee M. H., Birukou A., Dudin A. N., Klimenok V. I., Kostyukova O., Choe C-H. Queueing model of a single-level single-mediator with cooperation of the agents // Agent and Multi-agent Systems: Technology and Applications (Nguyen N. T., Ed). Berlin; Heidelberg: Springer. 2007. P. 447–455.
5. Artalejo J. R. Accessible Bibliography on Retrial Queues // Progress in 2000–2009 Mathematical and Computer Modeling. 2010. V. 51. P. 1071–1081.
6. Dudin A. N., Klimenok V. I., Vishnevsky V. M. The Theory of Queuing Systems with Correlated Flows. Springer International Publishing, 2020.

UDC: 004.942

Synchronisation of ISS-OFDM signals

S.V. Dorokhin[1]

[1]MIPT, 9 Institutskiy per., Dolgoprudny, Russian Federation

dorohin.sv@phystech.edu

Abstract

Interleaved Spread Spectrum OFDM (ISS-OFDM) is a new spread-spectrum modulation method which is simular to conventional OFDM. Despite properties of the signal itself are quite well studied, the problem of synchronisation remains unsolved. A novel sampling clock offset tracking algorithm is proposed and analysed in this article together with classical correlation-based acquisition algorithm. Simulation of ISS-OFDM communication system shows that it can operate at negative signal-to-noise ratios with acceptable bit error rate.

Keywords: ISS-OFDM, spread spectrum, synchronisation, SCO tracking, OFDM

1. Introduction

Interleaved Spread Spectrum OFDM (ISS-OFDM) is an OFDM-like spread spectrum method which was introduced by Pingzhou Tu et al. [1] in 2006. Pingzhou Tu et al. [2] also managed to demostrate that ISS-OFDM modulation can be used as PARP (peak-to-average-ratio) reduction technique. The same research group [3] suggested parallel FFT demodulation method, which reduced demodulation time. The subband-like spectral structure of ISS-OFDM was exploited in works featuring adaptive subband filtering. Pingzhou Tu et al. emphasised that it is possible to avoid information loss even if some subbands are not present in the spectrum [3]. The above-mentioned researchers suggested [4] flexible adaptive filtering scheme which can be used by several radiodevices to share the same frequency band. This direction was further explored by Qin Danyang et al. [5], who studied ISS-OFDM assuming fading channel model.

So far ISS-OFDM was studied primary as a technique for cognitive radio. However, ISS-OFDM has an interesting feature: increasing the number of subcarriers four times allows to decrease the minimal acceptable SNR (signal-to-noise ratio) by 3 dB. With sufficient number of subcarriers it is possible to operate at extremely low SNR and the main limiting factor is computational complexity. This property makes it

possible to use ISS-OFDM in low SNR applications, such as communication with low probability of detection in military scenario. As far as the author is concerned, the problem of ISS-OFDM synchronisation was not studied, probably due to low popularity of the method. The main purpose of this article is to cover that gap.

This article includes the following contributions:

- A novel fine time tracking algorithm is proposed.
- Classical pilot-based SCO tracker is adapted to ISS-OFDM
- A novel ISS-OFDM SCO tracker is proposed
- The entire system is simulated

The article is structured as following: in section 2 a brief overview of ISS-OFDM properties is given, section 3 contains a small review of analogs, followed by the detailed description of the proposed tracker (section 4), results of simulations (section 5) and a brief conclusion.

2. ISS-OFDM modulation

In this section the basic properties of ISS-OFDM will be discussed. For detailed explanation one may refer to [3], only theory essential for further understanding is presented in this section. In baseband each ISS-OFDM symbol carrying N_c constellation points $a_i, i = \overline{0, N_c - 1}$ consists of N_c^2 samples $y_m, m = \overline{0, N_c^2 - 1}$:

$$y_m = y_i(n) = a_i e^{2\pi j \frac{in}{N_c}}, i = m \mod N_c, n = \lfloor \frac{m}{N_c} \rfloor \tag{1}$$

where j is a complex unity. Discrete Fourier Transform of one ISS-OFDM symbol is

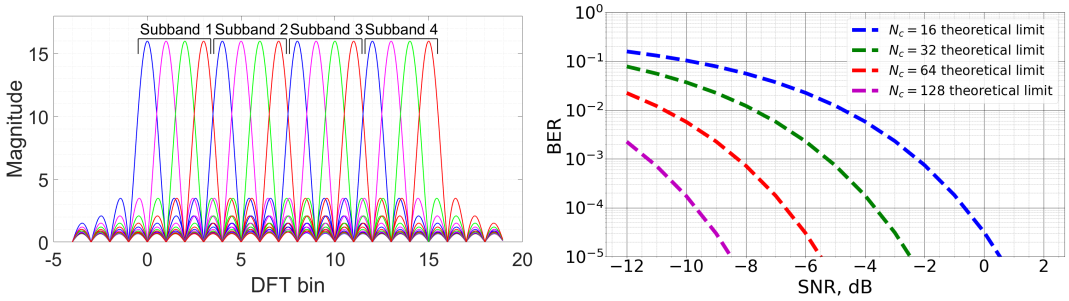
$$Y_k = N a_i e^{-2\pi j \frac{(nN_c+i)i}{N_c^2}}, i = k \mod N_c, n = \lfloor \frac{m}{N_c} \rfloor \tag{2}$$

It is important that the same constellation point a_i is spread on N_c subcarriers with period of N_c. At the same time, N_c subsequent subcarriers form one subband with all a_i present in it, as it is shown in Fig. 1a (subcarriers with the same colour carry the same constellation point). As long as fading is not taken into account, ISS-OFDM BER-SNR dependancy for Gaussian channel is given by

$$BER(SNR) = Q(\sqrt{N_c SNR}), \quad Q(x) = \frac{1}{2\pi} \int_x^\infty e^{-\frac{u^2}{2}} du \tag{3}$$

Doubling N_c inreases the number of ISS-OFDM subcarriers by four and results in 3 dB BER performance improvement, as it is shown in Fig. 1b.

It is important to understand how sampling clock offset (SCO) impacts demodulated constellations. Let us suppose that original signal (1) was received with SCO of ξ.



(a) Structure of spectrum

(b) Bit error rate performance

Fig. 1. ISS-OFDM properties

Constellation points after demodulation:

$$\widetilde{a_k} = \sum_{n=0}^{N_c-1} y_m e^{-2\pi j \frac{in}{N_c}(1+\xi)} = a_i e^{\pi j(i-k(1+\xi))} \frac{sin(\pi j(i-k(1+\xi)))}{sin(\frac{\pi j(i-k(1+\xi))}{N_c})}, i = \overline{0, N_c - 1} \quad (4)$$

If SCO is present, the constellations points are rotated and slightly attenuated. Moreover, constellation points start to influence the demodulation of each other due to inter-carrier interference (ICI). However, when SCO is sufficiently small, ICI can be neglected.

3. Previous work

Since the spectrum of ISS-OFDM resembles one of conventional OFDM, it seems quite rational to adapt OFDM-synchronisation algorithms to ISS-OFDM. The existing algorithms were selected to be tested on ISS-OFDM according to the following criterion:

- Popularity and simplicity. Since it is the first study on ISS-OFDM synchronisation, only classical OFDM synchronisation algorithms are considered.
- Robustness. ISS-OFDM offers acceptable performance at negative SNRs (refer to Fig. 1b), so algorithms must comply with this feature.
- Scalabilty in the number of subcarriers. The main method to improve ISS-OFDM robustness is to increase the number of subcarriers. That is why this parameter must be explicit in the algorithm's structure.

Generally, there are three approaches to initial synchronisation in OFDM systems. The first one is to use cyclic prefix and correlation alongside with estimator (commonly Maximum Likelyhood [6]). Antoher approach is to perform acquisition based on training symbols [7] or 2D-pilot map [8], without utilising guard interval. The third

approach, which Chinese standart DTMB-A relies on, implies using PN-sequences as guard interval [9].

Both PN-sequence and pilot-aided approaches appear to be difficult to scale in the number of subcarriers and to keep robust at the same time. Additional study has to be carried out in order to determine optimal parameters (length of PN-sequence, number of scattered pilots etc.) for low-SNR applications. On the contrary, correlation-based approach is easy to scale as the length of guard interval depends on the number of subcarriers. Averaging over consequent symbols can improve robustness.

For continuous transmission it is essential to track sampling clock offset (SCO). In OFDM systems there is a classical approach based on pilot correlation between subsequent symbols [10]. It was also shown [11] that SCO and carrier frequency offset (CFO) are coupled in OFDM systems and should be estimated jointly. Since it is the first study on ISS-OFDM synchronisation, CFO is left out of scope. In this paper a novel robust SCO tracking scheme is proposed and compared with this classical approach. The new tracker exploits ISS-OFDM signal structure in modulation domain, yielding promising results.

4. Proposed method

ISS-OFDM symbols carrying N_c constellation points consists of N_c^2 points and a guard interval (typically $\frac{1}{8}$ of bare symbol length). For the following analysis it is assumed that correlation-based initital sychronisation algorithm provides the estimation of the next symbol start with an error of no more than $\frac{N_c}{2}$.

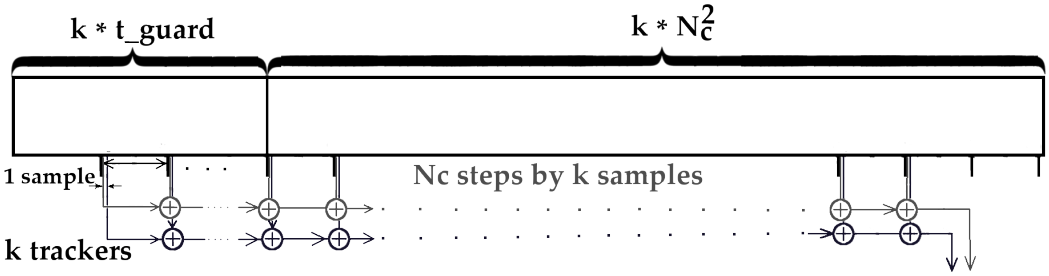


Fig. 2. The proposed clock frequency offset tracker

If the signal is upsampled with the factor of k, the total length of a symbol is $\frac{9}{8}N_c^2k$. From (1) it can be seen that $y_0(n) = a_0\ \forall n = \overline{0, N_c - 1}$, so every N_c sample of non-upsampled signal is the same. The new time tracking algorithm uses this periodicity. The simplified time tracking algorithm consists of the following steps (refer to Fig. 2): 1. jump to the estimated middle of guard interval; 2. Step $kN_c/2$

samples back; 3. Sum up N_c subsequent samples with a distance of kN_c between them; 4. Step k samples forward; 5. Sum up again, until N_c steps are performed

Let us explain the idea behind this tracker, assuming interpolation by DFT-zero padding for simplicity of analysis. The normalisation factor is set to 1 as it does not affect the results. The signal, interpolated by zero-padding with interpolation factor k (see (2)):

$$y_p = \sum_{l=0}^{N_c^2-1}\sum_{m=0}^{N_c^2-1} s_m e^{-2\pi j\frac{ml}{N_c^2}} e^{2\pi j\frac{pl}{kN_c^2}} = \sum_{m=0}^{N_c^2-1} s_m \sum_{l=0}^{N_c^2-1} e^{-2\pi j\frac{l(km-p)}{kN_c^2}} =$$
$$= \sum_{m=0}^{N_c^2-1} s_m e^{-\pi jl(m-\frac{p}{k})(1-\frac{1}{N_c})} \frac{sin(\pi k(m-\frac{p}{k}))}{sin(\frac{\pi k(m-\frac{p}{k})}{N_c^2})} = \sum_{m=0}^{N_c^2-1} s_m e^{-\pi jl(m-\frac{p}{k})(1-\frac{1}{N_c})}\delta_{m,\frac{p}{k}}, \quad (5)$$

where $\delta_{i,j}$ is the Kronecker delta. Tracker $l, l = \overline{0, k-1}$ has starting offset of l samples and provides maximum location estimation m_l from following N_c variables:

$$S_i^l = \sum_{\substack{p=l \\ \text{step } k}}^{kN_c-1} y_p = \sum_{\substack{p=l \\ \text{step } k}}^{kN_c^2-1}\sum_{m=0}^{N^2-1} s_m e^{-\pi jl(m-\frac{p}{k})(1-\frac{1}{N})} \frac{sin(\pi k(m-\frac{p}{k}))}{sin(\frac{\pi k(m-\frac{p}{k})}{N^2})} \quad (6)$$

It can be seen from this formula that for a tracker which has a shift of t from the closet a_0 sample the abscissa of maximum is not an integer, as it is shifted by $\frac{l}{k}$. Interpolation is used to determine a precise location of maximum for every tracker. Taking into account upsampling and start offset for every tracker yields the following formula for time offset m_{offs} measured in samples:

$$m_{offs} = \frac{\sum_{l=0}^{k-1}(m_l - \frac{N_c}{2})k + l}{k} \quad (7)$$

It is important to emphasise that the proposed algorithm gives time offset estimation by modulo N_c, and therefore the error of initial estimation must not exceed $\frac{N_c}{2}$.

If time synchronisation is performed correctly, one can track SCO in modulation domain based on formula (4). The phase difference of pilots in two subsequent symbols is proportional to SCO:

$$\xi = \frac{\Delta\phi}{\pi N_c}, \quad (8)$$

where $\Delta\phi$ is the difference between pilots' phases in subsequent symbols. The estimation can be averaged over many symbols, thus improving robustness.

5. Results

Aforementioned synchronisation algorithms were simulated in Matlab first separately and then as a united system. Gaussian channel model was chosen as the simpliest one to start with. In following simulations only QPSK modulation was used, as the impact of different modulation schemes on synchronisation performance is out of scope of this study. Random data bytes were generated and then scrambled with a PN-sequence of length $2^{14} - 1$. Guard interval of $\frac{1}{8}$ symbol length was chosen. The general structure of baseband model is presented in Fig. 3.

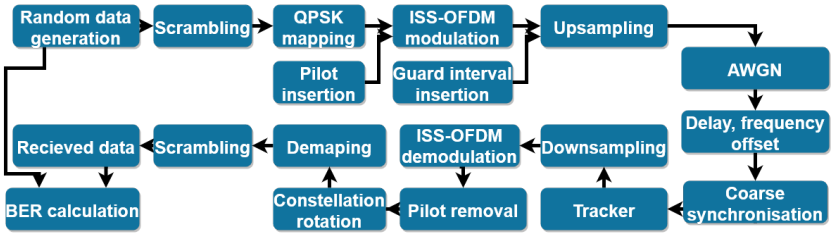
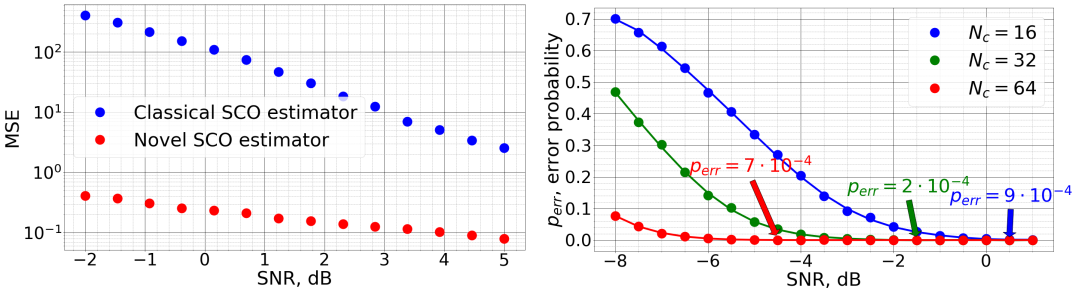


Fig. 3. Simulated baseband communication system

Since it is the first study on ISS-OFDM SCO tracking, the CFO has not been taken into account. One of ISS-OFDM subcarriers was reserved for a constant pilot tone to compensate constellation rotation caused by remained frequency offset (4). SCO was simulated via resampling. Every plot data point was averaged over 4980 symbols, with random time delay being introduced to every group of 60 symbols. Initial synchronisation and tracking algorithm were applied to every group and resulting BER was measured for further averaging.

SCO and time tracking algorithms described in section 4 were simulated separately as well. Time tracking algorithm included averaging over 15 consequent symbols, inital synchronisation algorithm was averaged over 3 symbols. For time tracking algorithm the probability that an error of at least one sample is present was calculated. SCO tracker was compared with the one proposed in [10]. Both SCO trackers were averaged over 80 symbols. The results are presented in Fig. 4.

The optimal parameters of averaging were used in a simulation to evaluate the performance of the entire system. The results of final simulation together with theoretical curves are presented in Fig. 5. Simulation data lay closer to theoretical boundary as SNR increases. It is explained by high error probabilty of synchronisation algorithms at low SNRs, as the averaging parameter was optimsed for SNRs corresponding to acceptable BER only. It can be seen that the proposed SCO tracker performs better than the classical one. It means that time and frequency



(a) SCO algorithms comparison (b) Proposed time tracker error probability

Fig. 4. Synchronisation algorithms simulation

diversity present in ISS-OFDM system has a potential for robust synchronisation algorithms.

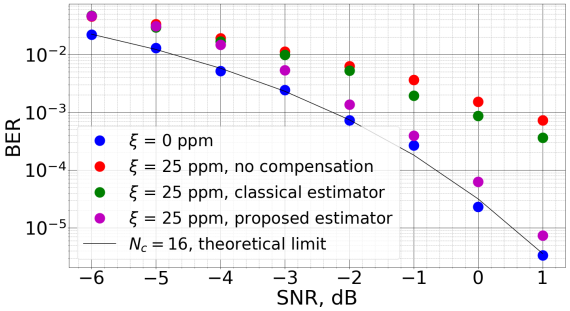


Fig. 5. Performance of the modeled communication system

6. Conclusion

In this article synchronisation algorithms for ISS-OFDM signals were discussed for the first time. A novel robust sampling clock offset tracking scheme was proposed and ISS-OFDM communication system model was simulated. In particular, it was shown that with 16^2 subcarriers presented synchronisations algorithms can operate at SNR as low as 1 dB, with the system keeping the bit error rate at a value of 10^{-5}. The proposed algorithms scale in the number of subcarriers and the scaled system can perform even at lower SNR values. The major drawback of the study is that it does not consider CFO. CFO and SCO should be estimated jointly and the coupling between them should be studied. Moreover, only AWGN model was used without any consideration of fading. Therefore, additional research is needed to determine whether the proposed alogrithms are suitable for fading channels.

REFERENCES

1. P. Tu, X. Huang, E. Dutkiewicz, A novel approach of spreading spectrum in ofdm systems, in: 2006 International Symposium on Communications and Information Technologies, 2006, pp. 487–491.
2. Pingzhou Tu, Xiaojing Huang, E. Dutkiewicz, Peak-to-average power ratio performance of interleaved spread spectrum ofdm signals, in: 2007 International Symposium on Communications and Information Technologies, 2007, pp. 82–86.
3. Pingzhou Tu, Xiaojing Huang, E. Dutkiewicz, Diversity performance of interleaved spread spectrum ofdm signals over frequency selective multipath fading channels, in: 2007 International Symposium on Communications and Information Technologies, 2007, pp. 184–189.
4. P. Tu, X. Huang, E. Dutkiewicz, Subband adaptive filtering for efficient spectrum utilization in cognitive radios, in: 2008 3rd International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CrownCom 2008), 2008, pp. 1–4.
5. Qin Danyang, Ma Lin, Wang Erfu, Ma Hongbin, Ding Qun, An interference suppression mechanism for wsn, in: PROCEEDINGS OF 2013 International Conference on Sensor Network Security Technology and Privacy Communication System, 2013, pp. 28–33.
6. J. van de Beek, M. Sandell, M. Isaksson, P. Ola Borjesson, Low-complex frame synchronization in ofdm systems, in: Proceedings of ICUPC '95 - 4th IEEE International Conference on Universal Personal Communications, 1995, pp. 982–986.
7. T. M. Schmidl, D. C. Cox, Robust frequency and timing synchronization for ofdm, IEEE Transactions on Communications 45 (12) (1997) 1613–1621.
8. M. J. Fernandez-Getino Garcia, J. M. Paez-Borrallo, S. Zazo, Dft-based channel estimation in 2d-pilot-symbol-aided ofdm wireless systems, in: IEEE VTS 53rd Vehicular Technology Conference, Spring 2001. Proceedings (Cat. No.01CH37202), Vol. 2, 2001, pp. 810–814 vol.2.
9. J. Wu, Y. Chen, X. Zeng, H. Min, Robust timing and frequency synchronization scheme for dtmb system, IEEE Transactions on Consumer Electronics 53 (4) (2007) 1348–1352.
10. M. Speth, S. Fechtel, G. Fock, H. Meyr, Optimum receiver design for ofdm-based broadband transmission .ii. a case study, IEEE Transactions on Communications 49 (4) (2001) 571–578.
11. Y. Jung, J. Kim, Y. You, Complexity efficient least squares estimation of frequency offsets for dvb-c2 ofdm systems, IEEE Access 6 (2018) 35165–35170.

UDC: 621.396

On effectiveness of message retransmission in wireless sensor networks

I.M. Nikolsky[1] and K.K. Furmanov[2]

[1]Lomonosov Moscow State University, 119991, Leninskie Gory, Moscow, Russia
[2]Higher School of Economics, 109028, Pokrovsky blvd. 11, Moscow, Russia

oliv_mail@mail.ru, kfurmanov@hse.ru

Abstract

Unrealibility of radio communication is an important issue for wireless sensor networks. Message retransmission is a common remedy for data loss in such networks. Present paper is devoted to research of number of retransmissions necessary for succesful message deliver. We introduce a simple method for determinig this number based on statistical analysis of data loss.Our method was implemented in the form of protocol for data gathering from working node to the main node. The protocol was tested on a pair of nodes built on Arduino board and NRF24 transciever. Our experiments show that proposed method reduces data loss significantly.

Keywords: wireless sensor network, Arduino, NRF24, retransmission

1. Introduction

At present, automation of management processes penetrates many areas of human activity. Automated control relies on the periodic collection of a large amount of data on the controlled process using a set of sensors. For example pipeline monitoring system periodically measures pressure, flow rate, temperature and many other parameters of flowing liquid [1].

Monitoring systems are often based on wireless sensor networks (WSN) [2]. WSN is a set of interconnected devices (*motes*) each of which consists of a microcontroller, a transceiver, a sensor, and a battery. These devices measure one or more physical values (flow rate, density, lighting level etc.) and transmit data over a radio channel.

One of the main problems arising when implementing WSN is the unreliability of the radio channel. Data collected by sensors is partially lost due to signal collisions from different network nodes, reflection of radiowaves from obstacles, etc. Information loss may be reduced using message retransmission. Determining number of duplicates N_d is quite an important and nontrivial problem: choosing small N_d may be not

enough to transmit at least one copy of particular data sample, whereas picking too big N_d leads to energy waste. Various methods have been proposed to optimize the number of retransmissions. For example, in [3] there was introduced a method based on fuzzy logic, and authors of [4] developed a distributed algorithm based on dynamic programming.

In present work a new method for determination of retransmission number is introduced. It is based on a statistic analysis of loss level in radiochannel. The main advantages of this method are absence of timeouts (the duration of which is difficult to determine reasonably) and simplicity of realisation . The method was tested on a pair of motes developed by one of the authors . These motes are based on Arduino Nano microcontroller boards and use NRF24L01 tranceivers for data transmission. It was shown with a serie of experiments that proposed method reduces message loss during data transmission from work node (*sensor*) to accumulator node (*sink*).

The work is organized as follows. In section 2 statistical formulas are given. Protocol for communication between sink and sensor node is presented in section 3. Test motes are described in section 4. Results of experiments are given in section 5. We give a conclusion in section 6.

2. Determining number of message duplicates using a stream of test messages

Consider a pair of nodes one of each measures some physical quantity using a sensor and transmits data to another node. We call the first node simply a sensor and the latter node will be further refered as sink. We assume that this nodes are connected with a radiochannel which is lossy (that is some of messages may not reach the destination). Because of link imperfection sensor node will duplicate messages. The main purpose of presented work is reasonable determination of number of duplicates N_d. It seems obvious that N_d should depend on level of message loss in channel.

Channel properties estimation may be performed on startup of nodes by transmitting a serie of test messages (a *test stream*) from sensor node to sink node. Number of test messages should be chosen so that to ensure that the proportion of successful transmissions is close to the probability of success. More formally we choose the number of transmitted messages so that the difference between the sample proportion $\hat{p}$ and the population proportion p exceeds certain value b with probability not greater than α:

$$P(|\hat{p} - p| > b) \leq \alpha \tag{1}$$

Assume that the successful transmissions occur as a series of Bernoulli trials, i.e. successes occur independently of each other with constant probability. In this case

the number of messages that guarantees the required accuracy when estimating the population proportion is

$$n_{test} = \left\lceil \frac{z^2_{1-\alpha/2}}{4b^2} \right\rceil \tag{2}$$

as shown in [5]. Here, $z^2_{1-\alpha/2}$ is the quantile of order $1-\alpha/2$ of a standard normal distribution. Equation (1) allows to calculate the required length of the test stream n_{test} for given values of the error bound b and the admissible probability of error α. The value of n_{test} may be unknown to sink, it just sends the number of received messages n_{rcv} to the sensor, and the sensor calculates the sample proportion of successful transmissions $\hat{p} = n_{rcv}/n_{test}$.

The probability of receiving at least one of N_d transmitted duplicates is

$$\pi = 1 - (1-p)^{N_d} \tag{3}$$

If the probability of successful transmission is known, inverting (2) we obtain the value of N_d that guarantees the succesful transmission of at least one duplicate for each message with probability π from the following formula:

$$N_d = \lceil log_{1-p}(1-\pi) \rceil \tag{4}$$

In our case the probability p is unknown, so the required number of duplicates is estimated by substituting p with its sample counterpart $\hat{p}$.

3. Protocol for data transmission

Our scheme of interaction between sensor node and sink node will consist of two phases: transmission of test messages (intended for loss level estimation) and sensor data transmission itself. First phase will be referred as *test stream phase* and the second as *data strea phase*.

We assume that the sink node will be an initiator of interaction. It will send a START message to a sensor node. This message may be lost because of channel imperfection. We solve this problem prescribing the sink node to send the START messages cyclically until a test message received.

After the end of test stream sink should pass number of received messages n_{rcv} to the sensor node, enabling the sensor node to compute number of replicas according to (3) . But sink node should be notified of the end of test stream. Sensor node uses a service message FIN to mark the end of test stream, which can also be lost. Similarly to the issue with START message we propose that after sending all test messages sensor will send FIN messages periodically. Sink counts received test messages until FIN receipt. After that sink starts cylcic dispatch of STAT messages. Each STAT

message includes n_{rcv} (number of test messages received by sink). On STAT message receipt sensor node finishes dispatching FIN messages, computes N_d according to (2). After that it starts transmitting DATA messages with samples from sensors. Each sample is replicated N_d times. Sink finishes dispatching of STAT messages on receipt of any message with data of monitoring. Formal descriptions of sink and sensor node algorithmes are given below.

Algorithm of sink node

1) Sink starts a loop of sending START messages, signalling sensor to start a test stream. This loop ends on receipt of a test message.
2) After receiving a test message sink starts counting received test messages.
3) On receipt of an end marker (FIN message) sink starts a loop of sending STAT messages, notifying a sensor node of number of received test messages. This loop ends on receipt of first data message
4) Sink receives data messages from sensor node

Algorithm of sensor node

1) Sensor waits for START message
2) Sensor starts a test stream. Number of messages in this stream is determined according (1).
3) After the end of test stream sensor cyclically sends end-of-stream marker(FIN message). It goes on before receipt of STAT message
4) Using information from STAT message, sensor computes necessary number of message replicas N_d according to (3).
5) Sensor switches to regime of periodical dispatch of information about monitored quantity. Quantity samples are piggybacked into DATA messages. Each of such message is replicated N_d times .

4. Test motes

The proposed protocol was tested on a pair of motes developed by one of the authors. We intentionally used low-cost easy-to-use electronic components so that people with little knowledge in electronics could benefit from our expirience. So Arduino Nano 3.0 microcontroller board was chosen as a core of our motes and NRF24L01 radio module was picked to connect them over the wireless channel.

Arduino Nano 3.0 is one of the cheapest and compact models in the Arduino family. It is built on an Atmel ATmega328 microcontroller with a frequency of 16 MHz. The flash memory capacity is 32Kb, which is quite enough for our purposes. We also note the simplicity and convenience of programming Arduino, the presence of a large number of libraries for utilizing various sensors. This allows the experimenter

to focus on building a WSN for his task, without delving into the intricacies of microcontroller programming.

Arduino family boards are often used in WSN development. The guidelines for constructing Arduino-based motes can be found in [6]. This book (and many other works) suggest using Xbee modules as tranceivers. These modules implement the ZigBee protocol popular in the field of the Internet of Things.

Unfortunately, Xbee modules are quit costy. We use less expensive NRF24L01 radios which is an order of magnitude cheaper than Xbee in some electronic stores. Besides moderate price NRF24L01 transceiver has some other advantages, such as high throughput and availability of high-level libraries for this module (for example, the RadioHead used in this project), which allow avoiding of complex tuning of registers. The module operates at unlicensed 2.4 GHz band.

Broad literature is devoted to analysis of the effectiveness of usage NRF24L01 modules in sensor networks. Authors of [7] state that NRF24L01 module is less energy efficient than Xbee. In another study (see [8]), it was shown that in a star-topology WSN, the radio transmission protocol of the NRF24L01 module is superior to the protocols ESP-Now (a special implementation of Wi-Fi) and Bluetooth Low Energy (BLE) in terms of capacity and current consumption. Despite criticism of the level of energy efficiency of NRF24L01 by some researchers, this transceiver is used in some real-world WSNs (for example, see [9]).

As stated above our testbde consists of two nodes – a work node(sensor) and sink. The work node is equipped with a DHT11 temperature probe. This node is intended to periodically transmit temperature measurements to the sink which is connected to PC. Because of unstability of radio connection some messages are lost, so we have an opportunity to test the above message duplication scheme. The results of our experiments are presented in the following section.

5. Experiment results

In order to test presented protocol we performed a serie of experiments using test motes described above. In all experiments we set desired probability of successful transmission π equal to 0,9. The following values were used to calculate the length of the test stream: $b = 0,06$, $\alpha = 0,05$. According to the equation (1) the corresponding number of test messages is 277. However we have decided to use 300 messages to improve the reliability of our results.

In first experiment we placed sensor and sink nodes at a distance of 10 cm from each other. On the test stream phase of the protocol 122 test messages out of 300 were delivered to the sink. Thus sample proportion of successes was $\hat{p} = 0,41$. According to (3) number of duplicates N_d was set equal to 5. Sensor node duplicated each message with temperature data N_d times.

The results of the experiment were as follows. For each message at least one duplicate was delivered to sink node. For most messages two duplicates were delivered. For only 46 messages first duplicate was delivered (making all other duplicates redundant)

In second experiment we intentionally worsened quality of communication. Motes were located in different rooms (so that there was a concrete wall between them). Moreover a sensor node was placed in a wooden cupboard.

On a test phase only 62 test messages of 300 were delivered to the sink node (i.e. sample proportion of successfull delivers was $\hat{p} = 0,21$). According to (3) sensor node set N_d equal to 10.

After that 100 messages with temperature data was sent from sensor node to the sink. Each message was duplicated N_d times. Three messages (out of 100) were lost (no duplicates received by sink). Average number of recieved duplicates for a message was 2,79. For only 16 messages first duplicate was delivered

To provide a visual representation of our results we display number of duplicates for first 20 messages in second experiment on Fig.1. On Fig.2 sequential numbers of first delivered duplicates for the same 20 messages are shown.

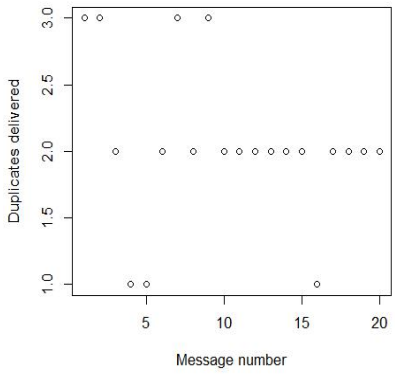


Fig. 1. Number of duplicates fo first 20 messages

6. Conclusion

Wireless sensor network is a tool useful both for academia research and industrial routine tasks. Unfortunately lossy radiochannel may cause instability in WSN operation. Data samples generated by sensors may be lost in transit requiring message retransmission. It is important to rationally define necessary number of

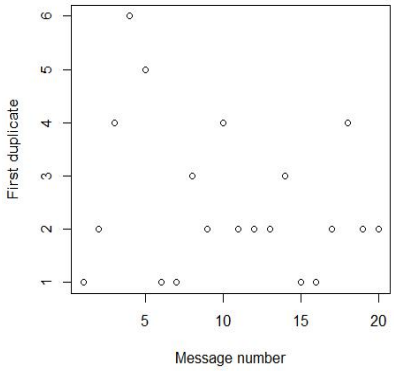


Fig. 2. Numbers of first delivered duplicates for first 20 messages

retransmissions. In present work we suggest to determine this number using statistical estimation of proportion of succesful transmissions. In our method we don't use any timeouts which is a positive point because usually it's difficult to reasonably estimate duration of timeout. Another advantage of our method is simplicity which allows to implement it with little amount of code thus optimizing consumption of precious microcontroller memory.

Our method was implemented as a part of protocol for gathering data from sensor node to sink node. It was tested on real hardware - two test motes consisting of Arduino and NRF24 transciever. Numerous experiments on our testbed showed effectivenes of proposed idea.

REFERENCES

1. Adegboye M.A., Fung W.K., Karnik A. Recent Advances in Pipeline Monitoring and Oil Leakage Detection Technologies: Principles and Approaches. Sensors (Basel). 2019;19(11):2548. Published 2019 Jun 4.
2. W.Dargie, C.Poellabauer Fundamentals of Wireless Sensor Networks: Theory and Practice Wireless Communications and Mobile Computing John Wiley & Sons, 2010, 336 p.
3. I. Umoren , D. Asuquo, O. Gilean, M. Esang Performability of Retransmission of Loss Packets in Wireless Sensor Networks // Computer and Information Science; Vol. 12, No. 2; 2019 pp71-86
4. Bi, R., Li, Y., Tan, G., & Sun, L. (2016). Optimizing Retransmission Threshold in Wireless Sensor Networks. Sensors (Basel, Switzerland), 16.

5. J.L. Devore, K. N. Berk Modern Mathematical Statistics with Applications Thomson Brooks/Cole, 2007, 838 p.
6. R. Faludi Building Wireless Sensor Networks with ZigBee, XBee, Arduino, and Processing. O'Reilly Media, 2010, 322 p.
7. H. Saha, S. Mandal, S. Mitra, S. Banerjee, U. Saha,"Comparative Performance Analysis between nRF24L01+ and XBEE ZB Module Based Wireless Ad-hoc Networks", International Journal of Computer Network and Information Security(IJCNIS), Vol.9, No.7, pp.36-44,
8. N. Sizen, "A Comparative study of Wireless Star Networks Implemented with Current Wireless Protocols" (2019).Masters Theses. 920.
9. A. Rahman, N. Athilah, A. Jambek Wireless sensor node design // 3rd International Conference on Electronic Design (ICED), August 11-12, 2016, Phuket, Thailand 332-336.

UDC: 123.456

Approach to indoor distance measurement in wireless sensor networks by means of Ultra-Wide-Band chaotic radio pulses

Efremova E.V.[1] and Kuzmin L.V.[1]

[1]Kotelnikov Institute of Radioengineering and Electronics (IRE) of Russian Academy of Sciences , Mokhovaya 11-7, Moscow, 125009, Russia

Abstract

A method for wireless distance measurement using ultra-wideband chaotic radio pulses based on statistical analysis is proposed. The approach belongs to methods for determining the distance by the power of the received signal. The method is based on determining the amplitude of the envelope of Ultra-Wide-Band chaotic radio pulses by comparing it with a certain reference value and counting the fraction of pulses whose envelope amplitude exceeds this reference value. Experimental results are discussed. The relative accuracy of measuring the distance of 15% has been experimentally confirmed.

Keywords: Distance measurement, ultrawideband chaotic radio pulses, wireless communication channel

1. Introduction

Indoor localization in the absence of global positioning services is an actual area especially in the era of Internet of Things (IoT), machine-to-machine interaction (M2M), robotics, etc.

To date, a number of indoor positioning systems is proposed and realized. They use different wireless technologies, such as WiFi, Bluetooth, BLE, ZigBee, UWB, acoustics and so on. The main approaches to distance measurement in wireless systems are based on estimation of signal strength (RSSI), time of flight (TOA, TDOA, RTOF, etc.), or phase [1], [2].

Depending on the technology, measurement method and post-processing tools, the distance accuracy is 15 cm to several meters. The best accuracy is achieved in systems based on ultra-wideband (UWB) ultra-short pulses. These systems use time of flight estimation, to measure distance. However, such systems have the most complex hardware.

UWB chaotic radio pulses are one of the UWB signal types [3], [4]. They are practically immune to multipath fading in wireless channels [5]. As is shown in the experiments [6], UWB signals provide smooth dependence of signal power on the distance according to power law $1/d^n$. As for narrow band signals (e.g., used in ZigBee), it is not so [7].

In the narrow-band systems, multipath propagation leads to high variations of the measured signal power in the receiver, which results in large errors of distance estimation. Chaotic (noise-like) radio pulses have an Ultra-Wide power spectrum, which practically coincides with the power spectrum of the continuous (non-modulated) chaotic signal and does not depend on the radio pulse length. The ultra-wide power spectrum along with the noise-like nature of the signal gives a delta-like autocorrelation function, that provides the signal's robustness to multipath fading.

This multipath immunity gives a reason for distance measurement by means of pulse power estimation in the receiver based on the power attenuation law.

From a practical point of view, another reason for the use of chaotic radio pulses for distance measurement is that they are an optional solution in communication systems of IEEE 802.15.4 standard and one of the main solutions of IEEE 802.15.6 standard.

In the existing equipment [8], [9], chaotic radio pulses are used to transmit information, so it is practically interesting to create a method for distance measurement, which would be a part of the process of wireless data exchange between the communication devices.

Here, a possibility of distance measurement with UWB chaotic radio pulses, based on the measurement of relative power of the signal at the receiver input using a statistical analysis of its characteristics is investigated experimentally.

2. Method

The idea of the method is to form and to emit a train of chaotic radio pulses by a transmitter, to detect them by a receiver, to estimate the pulse power and to calculate the number of the pulses whose power exceeds a preassigned level. Based on this information and on the channel power attenuation law, the distance between the transmitter and receiver is evaluated.

Let $P_d \sim 1/d^n$ be the power attenuation rule at distance d between transmitter and receiver, where n is the attenuation rate in a real wireless channel. This dependence follows from the well-known Friis transmission equation, which defines for the receiver point the ratio of the emission and reception power for the given signal frequency, antenna gains and the distance between the transmitter and receiver. Moreover, the same dependence applies to the wireless channel models recommended

by the IEEE profile committees. Among other things, these models differ from each other in the attenuation rate n. Then P_d can be calculated as

$$P_d = P_0 + 10n \lg (d/d_0) \tag{1}$$

where P_0 is power attenuation at distance d_0 [10].

In the experiment, three values can be obtained: P_d, P_0, and d_0. P_d, and P_0 can be measured with a log-detector [9], [11] that forms the output signal with amplitude A_d proportional to the input signal power P_d. d_0 is determined simultaneously with P_0.

Power pathloss $P_{add} = 10n \lg (d/d_0)$ is defined as difference $10n \lg (d/d_0) = P_d – P_0$ by means of comparing amplitude A_d of the pulses with power P_d at the input of log-detector with the amplitude A_0 of the signal with power P_0 at distance d_0 from the emitter. Using the input signal power-to-output voltage dependence [11] we can obtain:

$$P_d – P_0 = (A_d – A_0)/h \tag{2}$$

where h is the slope of the detector characteristics.

Distance d is calculated as

$$d = d_0 10^{(P_d – P_0)/(10n)} = d_0 10^{(A_d – A_0)/(10nh)} \tag{3}$$

In the free space $n = 2$, in a real wireless channel with line of sight $n < 2$, in a no-line-of-sight channel $n > 3$ [12].

Amplitude A_d is measured by means of comparing it with some variable threshold level A_T. In more details, for a fixed threshold level A_T the number of exceeding pulses is counted; and then threshold A_T is varied, in order to find the level at which the detector stops “feeling” the pulses.

Due to chaotic nature of the signal, the power of UWB chaotic pulses varies from pulse to pulse. Distributions of the pulse envelope amplitudes in the detector measured in the experiment at different distances between emitter and receiver is shown in Fig. 1.

Note that the width of the pulse amplitude distribution for different distances d remains approximately the same. Since the distance is calculated from the difference of amplitudes $(A_d – A_0)$, this allows us to fix the error level for a reference distance d_0 and distance d at the reception point. By means of estimating threshold level A_T for these distances at a fixed error level it is possible to calculate distance d.

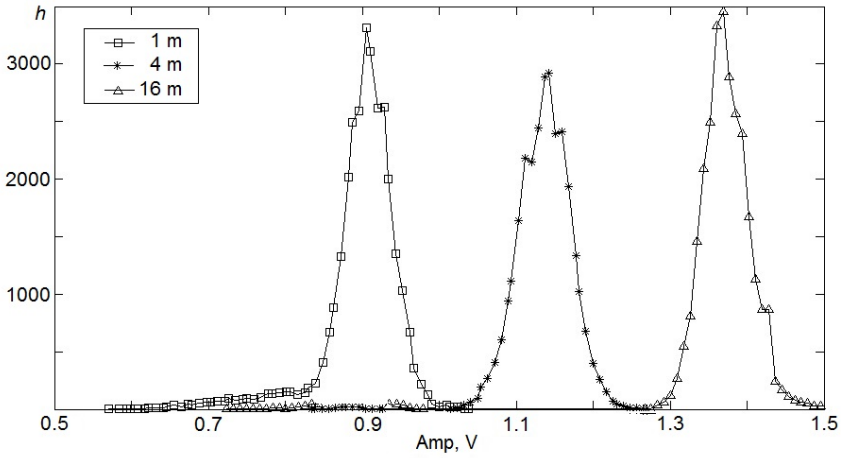


Fig. 1. Distribution of instantaneous amplitudes Amp of the envelope of chaotic radio pulses in the receiver at various distances d between the emitter and the receiver. Squares – 1 m, stars – 4 m, triangles – 16 m.

3. Experiment

To confirm the idea, an experiment was carried out. Experimental setup consists of two UWB chaotic transceivers [9] that play the role of the source and the receiver of chaotic radio pulses. Frequency range of the chaotic signal emitted by the source is $\Delta F = 3...5$ GHz. The devices were located in direct line of sight. The receiver position was fixed, the transmitter was moved by 0.25–0.5 m steps along a straight line. The scheme of the experiment is shown in Fig. 2, Fig. 3.

Transmitter Tx forms chaotic radio pulse packets supplemented with service information (number of pulses, check sum) necessary to control the received packet integrity.

In receiver Rx the signal is amplified, is passed through log-detector [8], [9], [11] and is compared with threshold. If envelope amplitude is higher than the threshold, then the pulse is considered received; otherwise, not.

The number of pulses that meet the condition is counted. The process is equivalent to the scheme of packet reception during data exchange between transceivers [8], [9].

For a given distance d between signal emitter and receiver the fraction of packets $P_B(d)$, received with an error PER (packet error ratio) is calculated. For every d, the threshold level A_d is varied so as to obtain a preassigned level $P_B(d)$ of packet error. The result of the experiment is the dependence of threshold voltage A_d on distance d at a given level of $P_B(d)$.

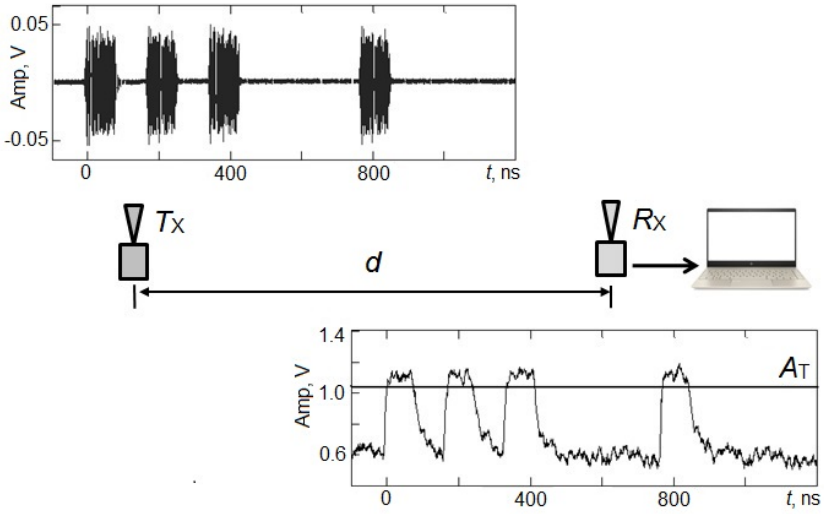


Fig. 2. Scheme of the experiment.

Fig. 3. UWB transceiver.

In the experiment, the threshold value T_d can be set at one of 256 positions (provided by DAC) corresponding to varying A_d in the range [0, 3.3] V. Thus, the threshold value is set with precision $k = 3.3/255 \approxeq 12.9$ mV. Since the detector slope is $h = -22$ mV/dB [11], one threshold step corresponds to a change of the input power by $S = k/h \approxeq -0.59$ dB.

As follows from expression (3), the threshold to distance conversion is given by

$$d = d_0 10^{(T_d - T_0)k/(10nh)} \tag{4}$$

where T_0, T_d are threshold values at the reference point (on the distance d_0) and on the distance d (at given error level), respectively, k is increment of the amplitude (Volts) corresponding to one threshold step ($A_T = T_d k$).

Taking into account theoretical dependence (4), where $d_0 = 1$ m, after conversion to logarithms, we have:

$$(T_d – T_0)S = 10n \lg d \tag{5}$$

With the measured values $10 \lg d$ put along X-axis and the values $(T_d – T_0)S$ along Y-axis, where $S = –0.59$ dB, the dependence $10 \lg d$ on $(T_d – T_0)S$ must be linear with the slope n.

The experiments were set on two locations: in the conference hall (A) and in an office (B). In Fig. 4 and Fig. 5 the results for the conference hall are presented.

The dependence of threshold value T_d on distance d was measured. The threshold value was taken at $PER = 95\%$.

Five measurements were made in each experimental point.

The first measurement was used to determine attenuation rate n. The rate n was determined using mean square method as the slope of the straight line approximating the experimental dependence on $(10 \lg d, (T_d – T_0)S)$ plane (Fig. 4). In the figure, the dependence is well approximated by a straight line with $n = 1.76$. This value was used to evaluate the distances between the emitter and receiver.

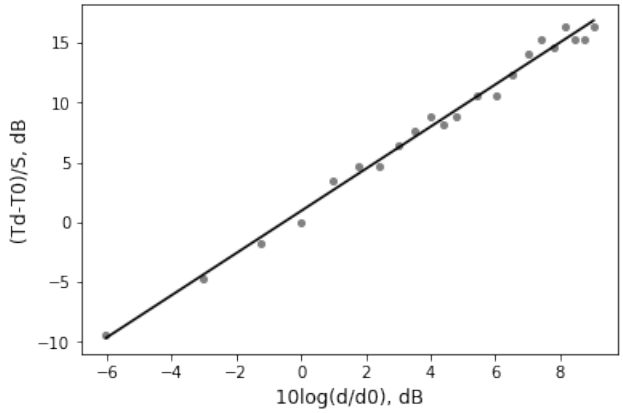


Fig. 4. Threshold difference $(T_d - T_0)S$ as a function of $10 \lg (d/d_0)$ in experiment.

In Fig. 5 the measurement results as functions of the actual distance are depicted, including the measured distance estimate, absolute and relative errors. For each real

distance, the mean value and standard deviation of the measured parameter (over 5 measurements) are shown.

The average relative error for conference hall A at distances up to 8 m was 15%. For room B the results are similar, the average relative error at distances up to 5 m was 13%.

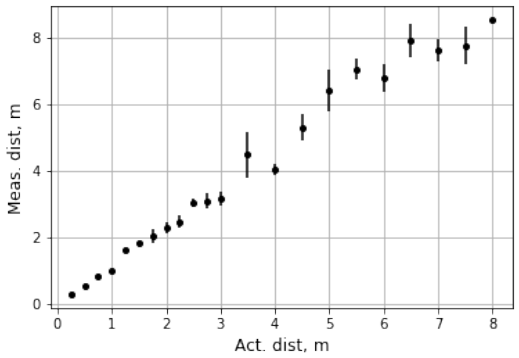
4. Conclusions

A method of distance measurement in wireless channel based on statistical analysis is proposed and experimentally tested. The method is using UWB chaotic radio pulses. Its results are comparable with modern positioning technologies, whereas its technical implementation is much simpler and does not require additional capabilities of the existing equipment above those related with the reception and processing of UWB signal during wireless information transmission.

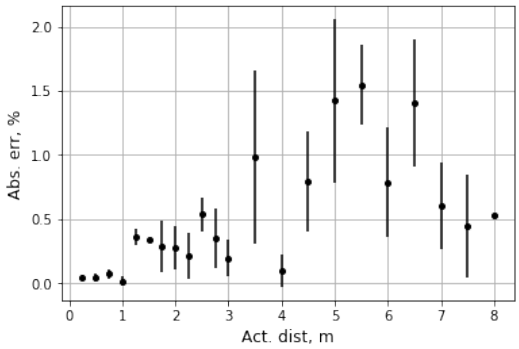
REFERENCES

1. F. Zafari, A. Gkelias, K. K. Leung. A Survey of Indoor Localization Systems and Technologies. // IEEE Communications Surveys & Tutorials. V. 21, No. 3. P. 2568–2599.
2. A. Alarifi, A. Al-Salman, M. Alsaleh, A. Alnafessah, S. Al-Hadhrami, M. A. Al-Ammar, H. S. Al-Khalifa. UltraWideband Indoor Positioning Technologies: Analysis and Recent Advances // Sensors. 2016. V. 16. P. 707–742.
3. Dmitriev A. S. , Efremova E. V., Panas A. I., Maksimov N. A. Generation of Chaos, Tekhnosfera, Moscow, 2012.
4. Dmitriev A. S., Efremova E. V. Radio-frequency illumination sources based on ultrawideband microgenerators of chaotic oscillations. // Tech. Phys. Lett. 2017. V. 43. P. 42–45.
5. Yu. V. Andreev, A. S. Dmitriev, V. A. Lazarev. // Proceedings of the 5th All-Russia Armand's Readings, Murom, 2015, p. 211.
6. Yu. V. Gulyaev, A. S. Dmitriev, V. A. Lazarev, T. I. Mokhseni, M. G. Popov. Interaction and navigation of robots based on ultrawideband direct chaotic communication // J. Commun. Technol. Electron. 2016. V. 61. P. 894–900.
7. A. V. Ponikar, O. V. Evseev, V. E. Antsiperov, and G. K. Mansurov, // Proceedings of the 4th All-Russia Conference on Radiolocation and Radio Communication (IRE RAN, Moscow, 2010), p. 914.
8. Dmitriev A. S., Efremova E. V., Kletsov A. V., Kuz'min L. V., Laktyushkin A. M., Yurkin V. Yu. Wireless ultrawideband communications and sensor networks. // J. Commun. Technol. Electron. 2008. V. 53. P. 1206–1216.

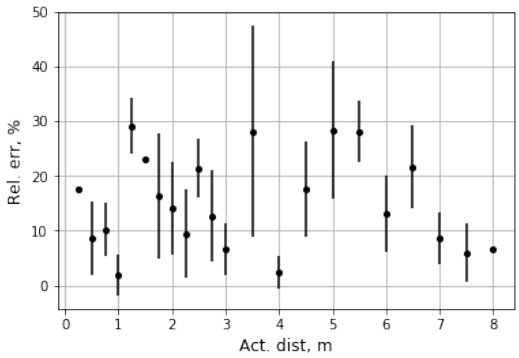
9. Dmitriev A. S., Gerasimov M. Yu., Itskov V. V., Lazarev V. A., Popov M. G., and Ryzhov A. I., Active wireless ultrawideband networks based on chaotic radio pulses. // J. Commun. Technol. Electron. 2017. V. 62. P. 380–388.
10. Sklar B. Digital Communications: Fundamentals and Applications. Prentice Hall, 2017.
11. Analog Device Data Sheet for AD8317 1 MHz-10GHz 50 dB High Precision Logarithmic Detector. http://www.analog.com
12. Molisch A. F. Ultra-Wide-Band Propagation Channels. // Proc. IEEE. 2009. V. 97. P. 353–371.



(a)



(b)



(c)

Fig. 5. Measurement results as functions of the actual distance: (a) measured distance estimate; (b) absolute measurement error; (c) relative error.

UDC: 519.872

Stationary Characteristics of the two-node Tandem Queueing System with Poisson Arrivals and General Renovation

L.A. Meykhanadzhyan[1], I.S. Zaryadov[2], T.A. Milovanova[2]

[1]Financial University under the Government of the Russian Federation 49 Leningradsky Pr, Moscow, 125993, Russian Federation

[2]Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation

lamejkhanadzhyan@fa.ru, zaryadov-is@rudn.ru, milovanova-ta@rudn.ru

Abstract

We consider the two-node tandem queueing system with finite capacity queues in both nodes and Poisson input flows. There is one server in each node and the service times are assumed to be i.i.d. random variables, having Erlang distributions with different parameters. General renovation mechanism is assumed to be implemented in each node. It implies that the queue is controlled upon customers' departure instants. Upon quitting the 1^{st} node a customer pushes out i customers from its queue with the given probability distribution $\{q_i^{(1)}, 0 \le i \le N_1 - 1\}$, with N_1 being the 1^{st} node capacity. Pushed-out customers leave the system and do not have any further effect on it. Upon quitting the 2^{nd} node a customer pushes out customers from its queue with another given probability distribution $\{q_i^{(2)}, 0 \le i \le N_2 - 1\}$, where N_2 is the 2^{nd} node capacity. The analytic method, based on well-known matrix analytic technique, is being briefly discussed, which allows one to compute the main stationary performance characteristics of the model including loss probabilities.

Keywords: tandem queue, renovation, active queue management, loss probability

1. Introduction

In this short note one dwells on the stationary analysis of the tandem $M/E_{n_1}/1/(N_1-1) \to \cdot/E_{n_2}/1/(N_2-1)$ queueing system with the general renovation mechanism implemented in each node. The detailed description of the system and overview of the analysis will be given in the next sections and here one briefly outline the motivation for this research.

The reported study was funded by RFBR, project number 19-07-00739.

Apparently queues with renovation were firstly mentioned in [1] and recently generalized and thoroughly studied in [2, 3]. Roughly speaking renovation implies that each customer, having received service, may remove some additional work from the system (may renovate it). It is quite natural to think of the renovation as an AQM. Active queue mechanisms are mostly often encountered in the communication network context (congestion avoidance) but are not only restricted to it. Most of the AQMs control and manage the queue upon arrivals of packets (customers, jobs etc.). One of the best known examples is the Random Early Detection (RED) scheme. On the contrary, in the renovation scheme the control instants are put off until service completions. Thus it is not evident whether renovation can be indeed called an AQM i.e. whether it can maintain the same performance as well-known AQMs. Some recent numerical results (based on analytic, not simulation solutions of mathematical models) show (see [4, 5, 6, 7]) that at least for single-server systems with deterministic service times it is possible to replace RED with a general renovation scheme and obtain at least the same performance level in the overloaded conditions. This new (numerical) finding motivates further practical and theoretical interest in queues with general renovation. One of the (theoretical) directions which remains explored yet is the analysis of stationary characteristics of queues in tandem with renovation in each node. In the next section consideration is given to one of the instants of the problem and the solution framework, which is based on the well-known matrix-analytic technique, is discussed in short.

2. The model description

Consider the $M/E_{n_1}/1/(N_1-1)$ system i.e. the system with the queue of finite capacity N_1-1 and Poisson arrival rate, say λ_1. There is one server and the service times are i.i.d. having Erlang distribution with n_1 phases and service rate at each phase equal to μ_1. The queue service discipline is FIFO. When an arriving customer sees that the queue is full, it is lost. The renovation mechanism, is implemented in the system. Define N_1 numbers, say $q_i^{(1)} \geq 0$, $0 \leq i \leq N_1-1$, satisfying $\sum_{i=0}^{N_1-1} q_i^{(1)} = 1$. If upon service completion there are i, $1 \leq i \leq N_1-1$, customers waiting in the queue, then the served customer leaves the system and

- with probability $q_0^{(1)} + Q_i^{(1)}$ nothing else happens, where $Q_i^{(1)} = q_i^{(1)} + q_{i+1}^{(1)} + \cdots + q_{N_1-1}^{(1)}$;
- with probability $q_j^{(1)}$, $0 < j < i$, exactly j customers are pushed out from the queue and those customers are chosen successively starting from the head of the queue.

Thus after the renovation (if it happened) the system never becomes empty. Customers, which have received service (but not those which were pushed out) enter

the $M/E_{n_2}/1/(N_2-1)$ system, which also has an independent Poisson flow of customers with rate λ_2. The renovation mechanism is implemented in the second system as well, but the renovation probabilities are different and are further denoted by $\{q_i^{(2)}, 0 \le i \le N_2-1\}$.

3. The joint stationary distribution

Let $(x_i(t), y_i(t))$ denote the total number of customers and the service phase in the i^{th} system. Due to the fact that the underlying processes are exponential, $\xi(t) = (x_1(t), y_1(t), x_2(t), y_2(t))$ is a homogeneous continuous-time Markov chain with the state space

$$\mathcal{X} = \{(0,0)\} \cup \{(i,n,j,m), 1 \le i \le N_1, 1 \le n \le n_1, 1 \le j \le N_2, 1 \le m \le n_2, i+j \ge 1\}.$$

Due to the finiteness of the state space, the stationary regime always exists. Introduce the joint stationary distribution:

$$p_{0,0} = \lim_{t\to\infty} \mathbf{P}\{x_1(t) = 0, x_2(t) = 0\},$$

$$p_{i,n,j,m} = \lim_{t\to\infty} \mathbf{P}\{x_1(t) = i, y_1(t) = n, x_2(t) = j, y_2(t) = m\}.$$

If we denote the infinitesimal generator of $\xi(t)$ by $\mathbb{Q}$, then the stationary distribution is found by solving the system of linear algebraic equations

$$\vec{p}\,\mathbb{Q} = \vec{0}, \quad \vec{p}\,\vec{1} = 1, \tag{1}$$

where $\vec{p}$ denotes the vector of stationary probabilities ordered in the manner specified below:

$$\begin{aligned}
\vec{p}_0 &= (p_{0,0}, p_{0,1,1}, \ldots, p_{0,1,n_2}, p_{0,2,1}, \ldots, p_{0,2,n_2}, \ldots, p_{0,N_2,1}, \ldots, p_{0,N_2,n_2}), \\
\vec{p}_{i,n} &= (p_{i,n,0}, p_{i,n,1,1}, \ldots, p_{i,n,1,n_2}, p_{i,n,2,1}, \ldots, p_{i,n,2,n_2}, \ldots, p_{i,n,N_2,1}, \ldots, p_{i,n,N_2,n_2}), \\
\vec{p}_i &= (\vec{p}_{i,1}, \ldots, \vec{p}_{i,n_1}), \ 1 \le i \le N_1, \\
\vec{p} &= (\vec{p}_0, \vec{p}_1, \ldots, \vec{p}_{N_1}).
\end{aligned}$$

Due to the presence of the renovation mechanism in the both nodes, $\mathbb{Q}$ does not have any special structure, which would allow one to obtain $\vec{p}$ in any other way, except for solving the system of balance equations (1). Analysis of $\mathbb{Q}$ shows that in general the system does not admit decomposition and thus the joint stationary distribution of both nodes is not the product of stationary distribution of the nodes working in isolation. In fact, $\mathbb{Q}$ is of $G/M/1$-type. From the structure of the vectors $\vec{p}_0$ and $\vec{p}_{i,n}$ it can be seen that (in terms of celebrated QBD processes) the number

of customers and the service phase in the first system is the level process, and the number of customers and the service phase in the second system is the phase process. Thus for the solution of (1) one can use any of the huge variety of methods developed so far for $G/M/1$-type infinitesimal generators (see, for example, [8, 9, 10]).

4. Stationary loss probability

Since in general the main motivation behind the analytical study of queues with renovation is the comparison of the renovation mechanism with known active queue mechanisms like RED and its ramifications, one is interested in specific performance characteristics: moments of the queue size, stationary loss probabilities, moments of the waiting/sojourn times, moments of consecutive losses (as introduced in [11]). Once the joint stationary distribution is found as explained in the previous section, moments of the queues' sizes are computed according to the definition.

Computations of the stationary loss probabilities are much more involved. This is due to the fact that once the tagged customers arrived at the system (either in the first node or the second), its loss probability depends on the future arrivals and thus one has to count all possible transitions during the tagged customer sojourn time. Due to the fact that the underlying distributions are assumed to be exponential, these computations can be performed in a recursive manner. For example, the probability that the arriving to the first system customer will enter the queue of the second system is equal to

$$\sum_{n=0}^{N_1} \vec{p}_0 \mathbb{P}_n(0,1,n_1) + \sum_{m=1}^{N_1-1} \sum_{j=1}^{n_1} \sum_{n=0}^{N_1-1} \vec{p}_{mj} \mathbb{P}_n(0,m,j),$$

where $\mathbb{P}_n(0,m,j)$ are certain matrices (of size $1+n_2N_2$), which record the possible transitions in the whole system during the tagged customer sojourn time and which are computed recursively. Such straightforward computations come at price: in order to compute the loss probabilities one has to perform a huge number (at worst of order $n_1^2 N_1^4$) of matrix inversions and for large queue capacities this becomes prohibitive. Yet the stationary analysis of the system under general service time distribution (not of the phase-type) seem to be prohibitive as well.

REFERENCES

1. Kreinin A. Y. Queueing systems with renovation // Journal of Applied Mathematics and Stochastic Analysis. 1997. V. 10. No. 4. P. 431–441.

2. Zaryadov I. S., Pechinkin A. V. Stationary Time Characteristics of the $GI/M/n$ System with Some Variants of the Generalized Renovation Discipline // Automation and Remote Control. 2009. No. 12. P. 2085–2097. doi:10.1134/S0005117909120157
3. Zaryadov I. S. The $GI/M/n/\infty$ queuing system with generalized renovation // Automation and Remote Control. 2010. V. 71. No. 4. P. 663–671. doi:10.1134/S0005117910040077
4. Chydzinski A., Chrost L. Analysis of AQM queues with queue size based packet dropping // International Journal of Applied Mathematics and Computer Science. 2011. V. 21. No. 3. P. 567–577. doi:10.2478/v10006-011-0045-7
5. Chydzinski A., Mrozowski P. Queues with Dropping Functions and General Arrival Processes // PLoS ONE. 2016. V. 11. No. 3. e0150702. doi:10.1371/journal.pone.0150702
6. Konovalov M., Razumchik R. Comparison of two active queue management schemes through the M/D/1/N queue // Informatika i ee Primeneniya (Informatics and Applications). 2018. V. 12. No. 4. P. 9–15. doi:10.14357/19922264180402
7. Zaryadov I., Bogdanova E., Milovanova T., Matushenko S., Pyatkina D. Stationary Characteristics of the GI/M/1 Queue with General Renovation and Feedback // 10th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). 2018. P. 1–6. doi:10.1109/icumt.2018.8631244
8. Bocharov P. P., D'Apice C., Pechinkin A. V., Salerno S. Queueing Theory. Utrecht, Boston: VSP, 2004. 446 p.
9. Dudin A.N., Klimenok V.I., Vishnevsky V.M. The theory of queuing systems with correlated flows. Heidelberg, Germany: Springer, 2019. 447 p.
10. Pechinkin, A. V., Razumchik R. V. Discrete Time Queuing Systems. Moscow: Fizmatlit. 432 p. ISBN 978-5-9221-1791-3 (in Russian)
11. Bonald T., May M., Bolot J. C. Analytic evaluation of red performance // Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. 2000. V. 3. P. 1415–1424.

UDC: 001.57+519.876.5+001.891.57

The multi-model approach to the study of complex systems using the example of the RED active queue management algorithm

Anna V. Korolkova[1,2], Dmitry S. Kulyabov[1,4], Michal Hnatič[2,3,4]

[1]Peoples' Friendship University of Russia (RUDN University), Moscow, Russia
[2]SAS, Institute of Experimental Physics, Košice, Slovakia
[3]Pavol Jozef Šafárik University in Košice, Košice, Slovakia
[4]Joint Institute for Nuclear Research, Dubna, Russia

korolkova-av@rudn.ru, kulyabov-ds@rudn.ru, hnatic@saske.sk

Abstract

Different kinds of models are used to study various natural and technical phenomena. Usually, the researcher is limited to using a certain kind of model approach, not using others (or even not realizing the existence of other model approaches). The authors believe that a complete study of a certain phenomenon should cover several model approaches. The paper describes several model approaches that we used in the study of the random early detection active queue management algorithm. Both the model approaches themselves and their implementation and the results obtained are described.

Keywords: active queue management, mathematical modeling, simulation, surrogate modeling, stochastic systems

1. Introduction

Scientific research is easy to start but difficult to complete. Our study of the Random Early Detection (RED) algorithm stood out from the study of approaches and mechanisms of traffic control in data transmission networks. But the further we went, the less satisfied we were with the results. The originally constructed mathematical model seemed to us somewhat artificial and non-extensible. To build a more natural mathematical model from first principles, we have developed a method for stochastization of one-step processes. To verify the mathematical model, we have built physical and simulation models. To conduct optimization studies, we began to build a surrogate model for the RED algorithm. In the end, we came to an

The publication has been prepared with the support of the. RUDN University Program 5-100" and funded by Russian Foundation for Basic Research (RFBR) according to the research project No 19-01-00645.

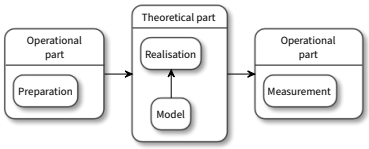


Fig. 1. Generic structure of the model approach

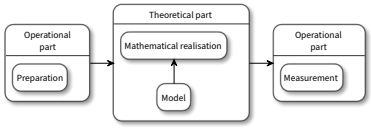


Fig. 2. Generic structure of the mathematical model

understanding that all of our models form some kind of emergent structure, with the help of which we can investigate various phenomena. In particular, stochastic and statistical systems. In this paper, we try to present our understanding of the multi-model approach to modeling.

2. Model approaches

Modeling as a discipline encompasses different types of model approaches. From our point of view, these approaches can be schematically described in a unified manner (see Fig. 1). In this case, the research structure consists of operational and theoretical parts. The operational parts are represented by the system preparation and measurement procedures. It is also common to describe the operational part as input and output data. The theoretical part consists of two layers: a model layer and an implementation layer. The implementation layer describes the specific structure of the evolution of the system. Depending on the type of implementation, different types of models can be obtained: a mathematical model (implementation — mathematical expressions), a simulation model (implementation — an algorithm), a physical model (implementation — an analog system), a surrogate model (implementation — approximation of behavior). Each type of model has its area of applicability, its advantages and disadvantages. The use of the entire range of models allows the most in-depth and comprehensive study of the modeled system.

3. RED active queue management algorithm

Random Early Detection (RED) is at the heart of several mechanisms to prevent and control congestion in router queues. Its main purpose is to smooth out temporary bursts of traffic and prevent prolonged network congestion by notifying traffic sources about the need to reduce the intensity of information transmission. The operation of a module implementing an algorithm of the RED type can be schematically represented as follows. When a packet of transmitted data enters the system, it enters the reset module. The decision to remove the package is made based on the value of the $p(\hat{q})$ function received from the control unit. The function $p(\hat{q})$ depends on the exponentially weighted moving average of the queue length $\hat{q}$, also calculated

by the supervisor based on the current value of the queue length q. The classic RED algorithm is discussed in detail in [1].

The main effort in the design of new algorithms like RED is directed at various modifications of the type of the drop function. Since the complete simulated system consists of interoperable TCP and RED algorithms, it is necessary to simulate the evolution of the TCP source as well. Since the original model was based on the TCP Reno protocol, we simulated this particular protocol.

The general congestion control algorithm is of the AIMD type (Additive Increase, Multiplicative Decrease) — an additive increase in the window size and its multiplicative decrease.

4. Mathematical model

The most rigorous research is usually based on a mathematical model (see Fig. 2). In this case, the model layer is realized through mathematical expressions describing the evolution of the system. There are several approaches to modeling RED-type algorithms. The most famous approach is modeling using the automatic control theory approach [2–4]. To us, this approach seems somewhat artificial and inconsistent. We prefer to do our modeling from first principles. We have developed a method of stochastization of one-step processes, which allows us to obtain models from first principles. Moreover, the resulting model models are immanently stochastic [5–7]. Our model of interaction between the TCP source and the RED algorithm is based on these methods and is mathematically represented in the form of stochastic differential equations with Wiener and Poisson processes [8–10].

5. Physical model

The resulting mathematical model should be compared with experimental data and verified. Unfortunately, we do not have the resources to take data from a working network or build a full-scale test bench on real network equipment. Therefore, we tried to create a virtual experimental installation based on the virtual machines [11]. Virtual machines run images of real routers operating systems. This is what allows us to call this model *physical*. To create the stand, the software package GNS3 (Graphical Network Simulator) [12] was chosen. This allows you to simulate a virtual network of routers and virtual machines. Works on almost all platforms. It is a graphical interface for different virtual machines. To emulate Cisco devices, the Dynamips emulator is used. Alternatively, emulators such as VirtualBox and Qemu can be used. The latter is especially useful when used with a KVM system that allows for a hardware processor implementation. GNS3 coordinates the operation of various virtual machines, and also provides the researcher with a convenient interface for creating and configuring the required stand configuration. Also, the

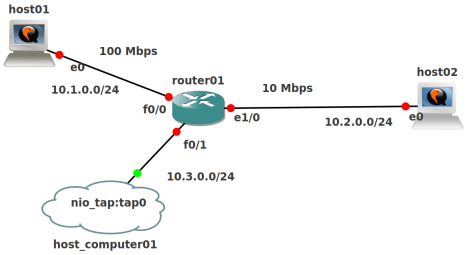


Fig. 3. Virtual stand for studying the functioning of the RED algorithm. `host01` is the packet source; `host02` is the recipient.

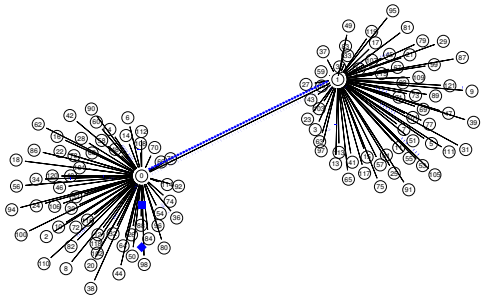


Fig. 4. Visualization of the simulation. Packets drop is shown

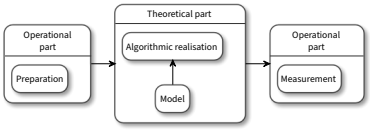


Fig. 5. Generic structure of the simulation model

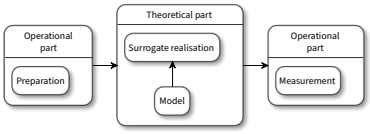


Fig. 6. Generic structure of the surrogate model

developed topology can be linked to an external network to manage and control data packets. The stand consists of a Cisco router, a traffic generator, and a receiver. D-ITG (Distributed Internet Traffic Generator) is used as a traffic generator (see Fig. 3). D-ITG allows us to obtain estimates of the main indicators of the quality of service (average packet transmission delay, delay variation (jitter), packet loss rate, performance) with a high degree of confidence.

6. Simulation model

With the development of computer technology, it became possible to specify a model implementation, not in the form of a mathematical description, but in the form of some algorithm (Fig. 5). This type of model is called simulation models, and the approach itself is called simulation. The simulation model plays a dual role. A simulation model, debugged and tested on experimental data and a physical model, can itself serve the purposes of verifying the mathematical model. On the other hand, the simulation model makes it possible to study the behavior of the modeled system more effectively than the mathematical model for different variants of the input data.

6.1. Simulation model on NS-2. The `ns2`package is a network protocol simulation tool. During its existence, the functionality has been repeatedly verified by data from field experiments. Therefore, this package itself has become a reference modeling tool. This is exactly the case when a simulation model is a replacement

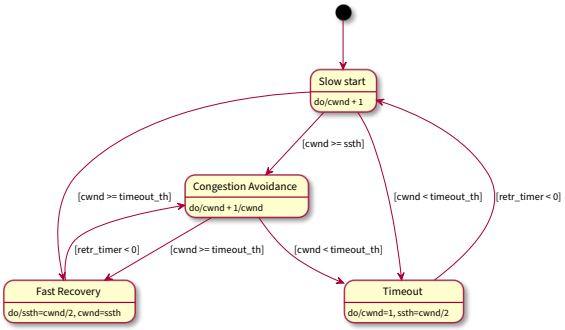


Fig. 7. TCP state diagram

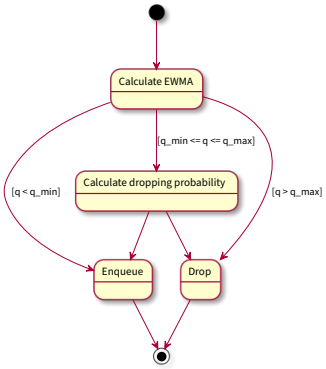


Fig. 8. RED state diagram

for a physical model and a natural experiment. The program for ns2 is written in the TCL language.The simulation results can be represented using visualization tool `nam` (see Fig. 4). The simulator is built on an event-driven architecture. That is, it implements a discrete approach to modeling. On the one hand, this is a plus, since it directly implements the TCP and RED specification (see section 3). On the other hand, the amount of resulting data sharply increases, which makes it difficult to carry out any lengthy simulation experiment. In our works, this software is used precisely as a means of verifying the results obtained [13, 14].

6.2. Hybrid model for RED algorithm. To study the RED algorithm, we developed a prototype of a simulation model. We wanted to avoid the resource intensiveness of discrete modeling approaches. However, it was necessary to take into account the discrete specifics of TCP and RED (see section 3). Therefore, we have chosen a hybrid (continuous–discrete) approach. The model was implemented in the hybrid modeling language Modelica [15, 16]. Since we are building a hybrid continuous–discrete model, then to describe each phase of TCP functioning, we will turn to a model with continuous time. The transition between phases will be described by discrete states. The resulting diagrams are directly converted into a Modelica program [17–19].

7. Surrogate model

Most scientific and technical problems require experiments and simulations to obtain results, to determine the limitations imposed on the result. However, for many real-world problems, simulation alone can take minutes, hours, days. As a result, routine tasks such as decision optimization, decision space exploration, sensitivity analysis, and what-if analysis become impossible as they require thousands or millions

of modeling evaluations. One way to simplify research is to build surrogate models (approximation models, response surface models, metamodels, black box models) (see Fig. 6) that mimic the behavior of the original model so closely as much as possible, while computationally cheap [20]. Surrogate models are built using a data-driven approach. The exact inner workings of the simulation code are not supposed to be known (or even understood), only the input—output (preparation—measurement) behavior is important. The model is built based on modeling the response to a limited number (sometimes quite large) of selected data points *. The scientific challenge for surrogate modeling is to create a surrogate that is as accurate as possible using as few modeling estimates as possible. For some problems, the nature of the true function is a priori unknown, so it is unclear which surrogate model will be the most accurate. Moreover, it is not clear how to obtain the most reliable estimates of the accuracy of a given surrogate. In this case, the model layer (Fig. 6) is replaced by the researcher's guess. In our case, the surrogate model is based on a clearly formulated mathematical model, which allows us to obtain clear, substantiated results of surrogate modeling. At the moment, we are developing a methodology for constructing surrogate models for both algorithms of the RED type proper and arbitrary stochastic one-step processes.

8. Conclusion

The authors tried to outline the concept of a multi-model approach to the study of physical and technical systems using the example of the interacting TCP protocol and the RED-type of active queue management algorithm. This research is in line with the research of stochastic models in science and technology. The multi-model approach makes it possible to increase the efficiency of the study of the phenomenon, to consider it from different angles, and to create effective software systems.

References

1. S. Floyd, V. Jacobson, Random Early Detection Gateways for Congestion Avoidance, IEEE/ACM Transactions on Networking 1 (1993) 397–413. doi:10.1109/90.251892.

2. V. Misra, W.-B. Gong, D. Towsley, Stochastic Differential Equation Modeling and Analysis of TCP-Windowsize Behavior, Proceedings of PERFORMANCE 99 (1999).

*Note that this type of model is known to many researchers. When only one modeling variable is involved, the process of building the surrogate model is called curve fitting

3. V. Misra, W.-B. Gong, D. Towsley, Fluid-Based Analysis of a Network of AQM Routers Supporting TCP Flows with an Application to RED, ACM SIGCOMM Computer Communication Review 30 (2000) 151–160. doi:10.1145/347057.347421.

4. C. V. V. Hollot, V. Misra, D. Towsley, A Control Theoretic Analysis of RED, in: Proceedings IEEE INFOCOM 2001. Conference on Computer Communications. Twentieth Annual Joint Conference of the IEEE Computer and Communications Society (Cat. No.01CH37213), volume 3, IEEE, 2001, pp. 1510–1519. doi:10.1109/INFCOM.2001.916647.

5. M. N. Gevorkyan, A. V. Demidova, T. R. Velieva, A. V. Korol'kova, D. S. Kulyabov, L. A. Sevast'yanov, Implementing a Method for Stochastization of One-Step Processes in a Computer Algebra System, Programming and Computer Software 44 (2018) 86–93. doi:10.1134/S0361768818020044. arXiv:1805.03190.

6. M. Hnatič, E. G. Eferina, A. V. Korolkova, D. S. Kulyabov, L. A. Sevastyanov, Operator Approach to the Master Equation for the One-Step Process, EPJ Web of Conferences 108 (2016) 02027. doi:10.1051/epjconf/201610802027. arXiv:1603.02205.

7. A. V. Korolkova, E. G. Eferina, E. B. Laneev, I. A. Gudkova, L. A. Sevastianov, D. S. Kulyabov, Stochastization Of One-Step Processes In The Occupations Number Representation, Proceedings 30th European Conference on Modelling and Simulation (2016) 698–704. doi:10.7148/2016-0698.

8. A. V. Korolkova, D. S. Kulyabov, T. R. Velieva, I. S. Zaryadov, Essay on the study of the self-oscillating regime in the control system, in: 33 European Conference on Modelling and Simulation, ECMS 2019, volume 33 of *Communications of the ECMS*, European Council for Modelling and Simulation, Caserta, 2019, pp. 473–480. doi:10.7148/2019-0473.

9. T. R. Velieva, D. S. Kulyabov, A. V. Korolkova, I. S. Zaryadov, The approach to investigation of the the regions of self-oscillations, Journal of Physics: Conference Series 937 (2017) 012057.1–8. doi:10.1088/1742-6596/937/1/012057.

10. T. R. Velieva, A. V. Korolkova, D. S. Kulyabov, B. A. Dos Santos, Model Queue Management on Routers, Bulletin of Peoples' Friendship University of Russia. Series "Mathematics. Information Sciences. Physics" 2 (2014) 81–92.

11. T. R. Velieva, A. V. Korolkova, D. S. Kulyabov, Designing Installations for Verification of the Model of Active Queue Management Discipline RED in the

GNS3, in: 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), IEEE Computer Society, 2015, pp. 570–577. doi:10.1109/ICUMT.2014.7002164. arXiv:1504.02324.

12. C. Welsh, GNS3 network simulation guide, PACKT Publisher, 2013.

13. T. R. Velieva, A. V. Korolkova, D. S. Kulyabov, S. A. Abramov, Parametric study of the control system in the TCP network, in: 10th International Congress on Ultra Modern Telecommunications and Control Systems, Moscow, 2019, pp. 334–339. doi:10.1109/ICUMT.2018.8631267.

14. T. R. Velieva, A. V. Korolkova, A. V. Demidova, D. S. Kulyabov, Software Package Development for the Active Traffic Management Module Self-oscillation Regime Investigation, in: Proceedings of the DepCoS-RELCOMEX, 2018, volume 761 of *Advances in Intelligent Systems and Computing*, Springer International Publishing, Cham, 2019, pp. 515–525. doi:10.1007/978-3-319-91446-6_48.

15. P. Fritzson, Principles of Object-Oriented Modeling and Simulation with Modelica 2.1, Wiley-IEEE Press, 2003.

16. P. Fritzson, Introduction to Modeling and Simulation of Technical and Physical Systems with Modelica, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2011. doi:10.1002/9781118094259.

17. T. R. Velieva, E. G. Eferina, A. V. Korolkova, D. S. Kulyabov, L. A. Sevastianov, Modelica-based TCP simulation, Journal of Physics: Conference Series 788 (2017) 012036.1–7. doi:10.1088/1742-6596/788/1/012036.

18. A.-M. Y. Apreutesey, A. V. Korolkova, D. S. Kulyabov, Modeling RED algorithm modifications in the OpenModelica, in: Proceedings of the Selected Papers of the 9th International Conference "Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems" (ITTMM-2019), volume 2407 of *CEUR Workshop Proceedings*, Moscow, 2019, pp. 5–14.

19. A. V. Korolkova, T. R. Velieva, P. A. Abaev, L. A. Sevastianov, D. S. Kulyabov, Hybrid Simulation Of Active Traffic Management, Proceedings 30th European Conference on Modelling and Simulation (2016) 685–691. doi:10.7148/2016-0685.

20. Y. Jin, Surrogate-assisted evolutionary computation: Recent advances and future challenges, Swarm and Evolutionary Computation 1 (2011) 61–70. doi:10.1016/j.swevo.2011.05.001.

UDC: 519.218

Simulation a modified Erlang system with priority customers

S.S. Rogozin

Institute of Applied Mathematical Research Karelian Research Centre RAS,
Petrozavodsk, Russia

ppexa@mail.ru

Abstract

We consider a modified Erlang loss system with two-class priority customers. Class-1 customers are lost if meet all servers busy but may terminate class-2 customers, while class-2 customers may stay in infinite capacity queue. We show the stability condition of this system and conduct discrete event simulation to verify this condition. We conduct simulation for both exponential and Pareto service time distributions. Results include sample mean of queue sizes in cases, when the stability condition is satisfied or not, and completely confirm this condition.

1. Introduction

In recent years, Internet traffic has been explosively increased because of the increased use of tablet and laptop computers, smart-phones, etc. This is the cause of a spectrum shortage problem of wireless networks. A promising technology for solving this problem is the cognitive wireless network [2-5,7]. In wireless networks, secondary users (un-licensed users) have to use the bandwidths in such a way that does not interfere primary users (licensed users). In particular, secondary users can use the bandwidths only if primary users are not present. Motivated by this, we analyse a multiserver queueing system with infinite buffer for secondary users, while primary users have absolute priority over secondary users and are lost if all channels are already occupied by other primary users. From the view point of primary users, the system behaves as an Erlang loss system, while from that of secondary users, the system is an infinite buffer model where secondary users are served when some servers are not occupied by primary users. We assume that the service times of primary and secondary customers have two distinct arbitrary distributions. In this work we present the stability condition for this system using the regenerative analysis approach [1,6] and show some numerical examples verifying the stability condition.

The research is supported by Russian Foundation for Basic Research, project No. 18-07-00156.

2. Description of the model

We consider a modified Erlang system with two classes of customers following Poisson inputs, general class-dependent service times and c identical servers. The service discipline is priority: class-1 (preemptive priority) customers are lost if meet all servers busy, while class-2 (non-priority) customers stay in the system regardless of the state of the system, and in particular may stay in infinite capacity queue. Class-2 customers waiting in the queue follow FCFS (first-come-first-served) service discipline.

We denote by λ_i the input rate of class-i customers, and introduce service rates

$$\mu_1 = \frac{1}{\mathsf{E}S^{(1)}}, \quad \mu_2 = \frac{1}{\mathsf{E}S^{(2)}}, \tag{1}$$

where $S^{(i)}$ is the (generic) service time of class-i customers. We also will use notation $\{t_n\}$ for arrival instants of the superposed (Poisson) input with rate $\lambda = \lambda_1 + \lambda_2$, and $\tau_n = t_{n+1} - t_n$ for the i.i.d. (exponential) interarrival times, with τ being generic time. Class-1 customers have preemptive priority over class-2 customers. In particular, class-1 customer occupies server which is busy by a class-2 customer provided there are no idle servers upon his arrival.

The regeneration instants of continuous-time processes $Q(t)$ (queue size) and $W(t)$ (workload) are defined as follows:

$$T_{n+1} = \min\{t_k > T_n \,:\, Q(t_k) = 0\,\}, \quad n \geq 0,\, T_0 := 0, \tag{2}$$

with generic regeneration period length T. The positive recurrence means that

$$T_1 < \infty \;\text{ w.p.1 and }\; \mathsf{E}T < \infty. \tag{3}$$

Denote $\rho_i = \lambda_i/\mu_i$, the traffic intensity of class-i customers. For the system described above one can prove the following statement.

Theorem 1. The positive recurrence (3) holds if and only if the following condition holds:

$$\rho_2 + \sum_{i=1}^{c} i\mathsf{P}_i < c, \tag{4}$$

It is worth mentioning that the stationary probabilities P_i can be found by the Erlang formula (see [1]) because, to analyse class-1 customers only, we can treat the model as a loss $M/G/c/0$ system.

As a by-product of our analysis, we obtain well-known stability condition of a buffered multiclass system (with no losses):

$$\sum_{i=1}^{N} \rho_i < c, \tag{5}$$

where N is the number of customer classes.

3. Simulation

To check our theoretical results, we conduct discrete event simulation to illustrate the behaviour of the sample mean queue size when stability condition (4) is violated. We denote $Q_i(t)$ the number of class-i customers in the system and note that:

$$\mathsf{E}Q_1 = \sum_{i=1}^{c} i\mathsf{P}_i, \tag{6}$$

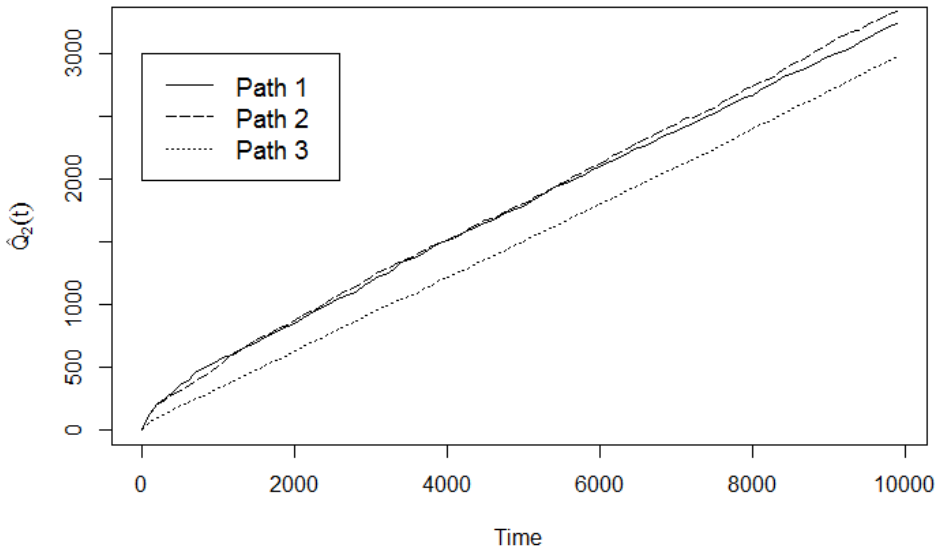
where P_i can be find by the Erlang formula for a fixed ρ_1. We will use this formula below to find $\mathsf{E}Q_1$ for fixed ρ_1. Then the stability condition (4) can be represented as: $\rho_2 + \mathsf{E}Q_1 < c$. We assume that $c = 10$, $\mu_1 = \mu_2 = 10$ and we consider three different values of ρ_1:

$$\rho_1 \approx 1.00\,(\mathsf{E}Q_1 = 1),\ \rho_1 \approx 5.10\,(\mathsf{E}Q_1 = 5)\,\text{and}\,\rho_1 \approx 16.52\,(\mathsf{E}Q_1 = 9). \tag{7}$$

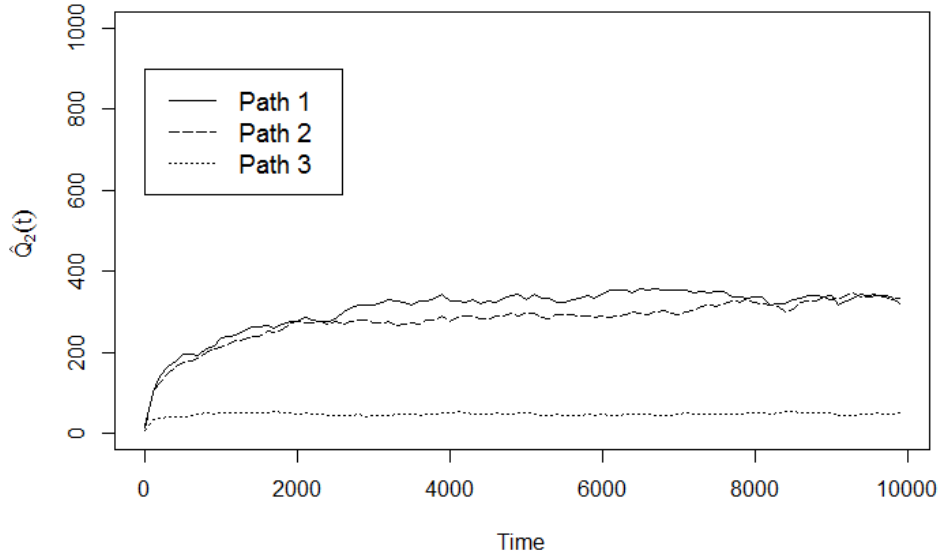
Then for each fixed ρ_1 we can satisfy or violate the stability condition by choosing different values of ρ_2.

First, we construct some paths based on 300 sample-path of $Q_2(t)$. We denote these sample mean paths by $\hat{Q}_2(t)$. Fig.1 presents the estimation results for exponential service times and Fig.2 demonstrates the results for the Pareto service times. It is easy to see that if (4) does not hold, $\hat{Q}_2(t)$ increases linearly to infinity reflecting strong instability. When condition (4) is satisfied, all paths become stationary. Furthermore for the Pareto service times $\hat{Q}_2(t)$ increases faster than for the exponential service times for the same cases. We also note that the last case, where it is low servers occupancy by class-2 customers, is more stable than two others. In addition on Fig.3 we demonstrate more intensity case where $\rho_1 \approx 107.82$ ($\mathsf{E}Q_1 = 9.9$) and $\rho_2 = 0.09$. However it is also stable as the previous cases.

Then we construct the dependence of the stationary mean queue size ($\hat{Q}_2$ and $\hat{Q}_1$) on class-2 customers traffic intensity (ρ_2) with $\rho_1 \approx 5.10$ (see Fig.4). The duration of simulation at each point is about 10000 arrivals. The queue size of class-1 customers does not change because class-2 customers do not affect class-1 customers. When ρ_2 goes to 5, the stationary mean $\hat{Q}_2$ increases to infinity since the stability condition is violated.

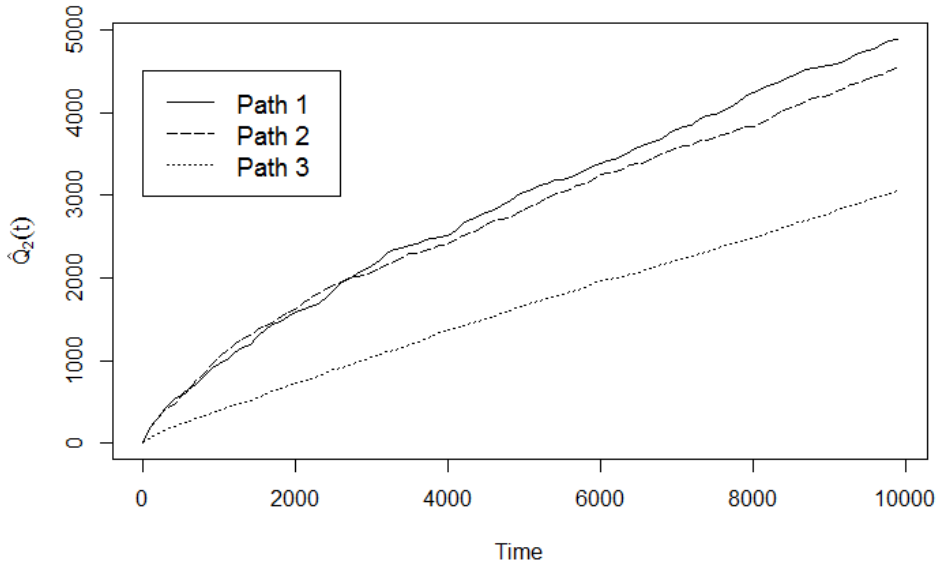


a) Stability condition (4) is $10.03 < 10$ (violated). Path 1: $\rho_2 = 9.03$, $\rho_1 \approx 1.00$; path 2: $\rho_2 = 5.03$, $\rho_1 \approx 5.10$; path 3: $\rho_2 = 1.03$, $\rho_1 \approx 16.52$.

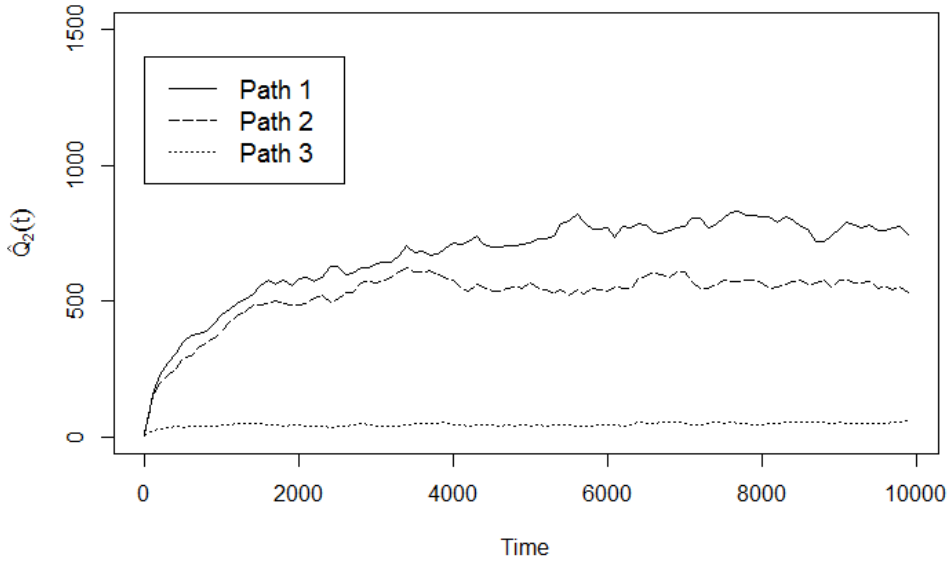


b) Stability condition (4) is $9.97 < 10$ (satisfied). Path 1: $\rho_2 = 8.97$, $\rho_1 \approx 1.00$; path 2: $\rho_2 = 4.97$, $\rho_1 \approx 5.10$; path 3: $\rho_2 = 0.97$, $\rho_1 \approx 16.52$.

Fig. 1. Sample mean of queue size of class-2 customers for exponential service times.



a) Stability condition (4) is $10.03 < 10$ (violated). Path 1: $\rho_2 = 9.03$, $\rho_1 \approx 1.00$; path 2: $\rho_2 = 5.03$, $\rho_1 \approx 5.10$; path 3: $\rho_2 = 1.03$, $\rho_1 \approx 16.52$.



b) Stability condition (4) is $9.90 < 10$ (satisfied). Path 1: $\rho_2 = 8.90$, $\rho_1 \approx 1.00$; path 2: $\rho_2 = 4.90$, $\rho_1 \approx 5.10$; path 3: $\rho_2 = 0.90$, $\rho_1 \approx 16.52$.

Fig. 2. Sample mean of queue size of class-2 customers for Pareto service times.

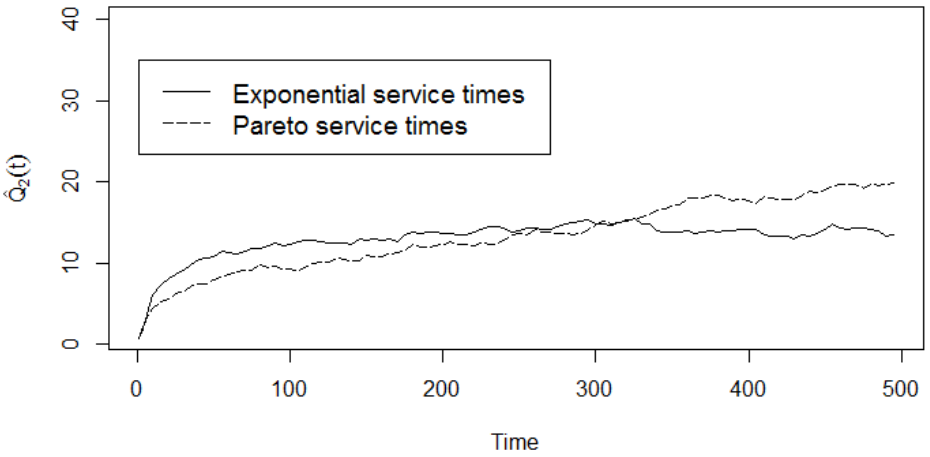


Fig. 3. Sample mean queue size of class-2 customers for exponential and Pareto service times. Stability condition (4) is $9.99 < 10$ (satisfied); $\rho_2 = 0.09$, $\rho_1 \approx 107.82$ ($\mathsf{E}Q_1 = 9.9$).

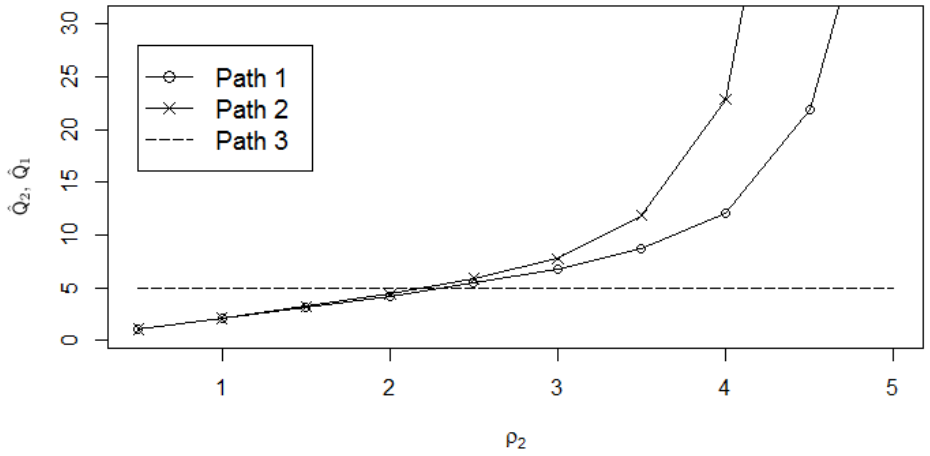


Fig. 4. Dependence of the stationary mean queue size ($\hat{Q}_2$ and $\hat{Q}_1$) on class-2 customers traffic intensity (ρ_2) as $\rho_1 \approx 5.10$. Path 1: $\hat{Q}_2$ for exponential service times; path 2: $\hat{Q}_2$ for Pareto service times; path 3: $\hat{Q}_1$ in both cases.

4. Conclusion

We considered the modified Erlang loss system with two-class priority customers and conducted discrete event simulation of this system. Results include sample mean of queue sizes in cases, when the stability condition is satisfied or not. In conclusion,

we note that all results completely confirm the presented stability condition for considered cases.

5. Acknowledgement

The author thanks his adviser Prof Evsey Morozov for help and useful comments.

REFERENCES

1. Asmussen, S.: Applied probability and Queues. 2nd edn. Springer, Springer-Verlag New York (2003)
2. Akutsu, K. and Phung-Duc, T.: Analysis of retrial queues for cognitive wireless networks with sensing time of secondary users, Lecture Notes in Computer Science, LNCS 11688, pp. 77-91 (2019).
3. Akyildiz, I. F.; Lee, W. Y.; Vuran, M. C.; Mohanty; S. A survey on spectrum management in cognitive radio networks. *IEEE Commun. Mag.* 2008, 46(4), 40-48.
4. Letaief, K. B.; Zhang, W. Cooperative communications for cognitive radio networks. *Proc. IEEE.* 2009, 97(5), 878-893.
5. Ostovar, A.; Keshavarz, H.; Quan, Z. Cognitive radio networks for green wireless communications: an overview. *Telecommun. Syst.* 2020, 1-10.
6. Smith, W.L.: Regenerative stochastic processes. Proceedings of the Royal Society A(232), 6–31 (1955).
7. Wang, B.; Liu, K. R. Advances in cognitive radio networks: A survey. *IEEE J. Sel. Top. Signal Process.* 2010, 5(1), 5-23.

УДК: 004.021:519.2:519.6

Возможности гибридного моделирования систем с управлением на языках Modelica и Julia

А. М. Ю. Апреутесей[1], А. В. Королькова[1], Д. С. Кулябов[1,2]

[1]Российский университет дружбы народов,
ул. Миклухо-Маклая 6, Москва, Российская Федерация, 117198
[2]Объединённый институт ядерных исследований,
ул. Жолио-Кюри 6, Дубна, Московская область, Российская Федерация, 141980
1032193049@pfur.ru, korolkova-av@rudn.ru, kulyabov-ds@rudn.ru

Аннотация

В качестве исследуемой системы выступает система, состоящая из входящего потока, обрабатываемого согласно протоколу Transmission Control Protocol (TCP), а также маршрутизатора, обрабатывающего трафик по алгоритму типа Random Early Detection (RED). Особенностью задачи является то, что при описании её в парадигме непрерывного моделирования возникают сложности в численной реализации. В качестве выхода из этой ситуации предлагается использовать гибридный подход к моделированию, что влечёт за собой проблему выбора языка реализации для численного расчёта. Кроме того, численная реализация усложняется наличием запаздывающего аргумента в математическом описании модели. В работе исследуются возможности языков программирования Modelica и Julia для реализации непрерывно-дискретной парадигмы при моделировании гибридных систем, содержащих как непрерывные, так и дискретные аспекты поведения.

Ключевые слова: активное управление трафиком, имитационное моделирование, Modelica, Julia, Random Early Detection

1. Введение

Среди возможных методов исследования сложных систем можно выделить построение дискретно–событийной модели, построение непрерывной модели, а также гибридное моделирование [1]. В подобных гибридных системах сочетается работа непрерывных и дискретных элементов, например, системы с дискретным устройством управления и объектом управления с непрерывным характером функционирования [2]. В качестве исследуемой системы выступает модель взаимодействия процесса передачи данных по протоколу Transmission Control Protocol

Работа выполнена при финансовой поддержке поддержке Программы РУДН «5-100» и при финансовой поддержке РФФИ в рамках научного проекта № 19-01-00645.

(TCP) и процесса регулирования состояния потока при возникновении перегрузок, в качестве которого рассматривается алгоритм Random Early Detection (RED). При моделировании TCP-подобного трафика можно воспользоваться жидкостным (непрерывным) подходом, однако особенно важно учитывать дискретные переходы между TCP состояниями и функцию сброса пакетов в алгоритмах типа RED, гибридный подход отразит эти важные особенности моделируемой системы. Делается вывод о применимости обоих языков для описания сложных гибридных систем с управлением.

2. Алгоритм активного управления очередью Random Early Detection

Авторами неоднократно описывались проблематика исследования, особенности изучаемого явления, построение математической модели [3, 4].

Алгоритм активного управления очередью с алгоритмом управления типа RED используется для контроля и предотвращения перегрузок в очередях маршрутизаторов [5]. Алгоритмы управления состоянием трафика могут быть представлены как модули управления в сетевом оборудовании. Преимуществом данного алгоритма является его эффективность и относительно простая реализация на сетевом оборудовании.

Математическая модель взаимодействия входящего TCP-потока и маршрутизатора, обрабатывающего трафик по алгоритму управления типа RED, представляет собой автономную систему трёх дифференциальных уравнений [6–9].

3. Моделирование на языке Modelica

Язык Modelica разработан некоммерческой организацией Modelica, которая также разрабатывает на его основе свободно распространяемую библиотеку. Этот язык, позиционируемый как объектно-ориентированный язык физического моделирования, применяется для решения широкого круга задач [10,11]. Modelica хорошо применима для компонентно-ориентированного моделирования сложных систем, состоящих из различных физических компонентов, также имеющих компоненты управления и элементы, ориентированные на отдельные процессы. Продемонстрируем применение данного языка к гибридному моделированию алгоритмов сетей связи [2].

Основу языка составляют имеющие возможность наследоваться классы, которые содержат в себе все элементы наследуемого класса. В Modelica классы содержат методы и поля, которые могут иметь такие типы изменчивости как константа, параметр и переменная. Помимо методов и полей в классе содержатся функции и связывающие переменные уравнения, которые задаются в разделе

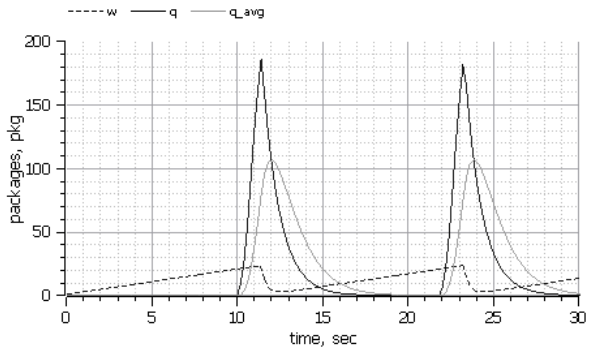


Рис. 1. Поведение параметров $w(t)$, $q(t)$, $\hat{q}(t)$ по результатам моделирования на языке Modelica

`equation`. Одним из обязательных требований программы на Modelica является совпадающее число переменных и уравнений.

Алгоритм контроля перегрузки в очередях маршрутизатора RED на языке Modelica реализован в виде класса `Red` [12].

На языке Modelica оператор `der` задает производную по времени. Запаздывание в Modelica реализуется крайне просто с помощью оператора `delay`, который дает возможность работать с запаздыванием как непрерывных, так и дискретных элементов системы.

В результате моделирования системы была получена динамика изменения $w(t)$, $q(t)$, $\hat{q}(t)$, представленная на рис. 1. График демонстрирует, что при некоторых значениях параметров в системе возникает устойчивый автоколебательный режим функционирования.

4. Моделирование на языке Julia

Язык Julia — это язык высокого уровня, предназначенный для научных и инженерных расчётов [2, 13–15].

Опишем реализацию алгоритма активного управления очередью с алгоритмом управления типа RED на языке Julia.

В данной реализации использовалась библиотека DifferentialEquations [16], предназначенная для эффективного решения дифференциальных уравнений различных видов, таких как обыкновенные дифференциальные уравнения, стохастические обыкновенные дифференциальные уравнения, дифференциально-алгебраические и гибридные уравнения, а также дифференциальные уравнения с запаздыванием.

Для установки пакета используем следующую команду в Julia REPL:

```
using Pkg
Pkg.add("DifferentialEquations")
```

Подключим пакет, используя команду:

```
using DifferentialEquations
```

Зададим вектор начальных параметров системы `p = (T, N, C, wq, q_min, q_max, R, p_max, w_max)`. Переменная `pr`, являющаяся функцией вероятности сброса пакетов, выступает как глобальная переменная.

Так как в исходной системе дифференциальных уравнений присутствуют запаздывающие аргументы, определим функцию истории `h(p, t)`, которая зависит от вектора параметров `p` и времени `t`.

Для предложенной нами задачи динамическая функция `Red`, описывающая поведение дифференциальных уравнений и задающая ограничения для параметров w, q, $\hat{q}(t)$, в DifferentialEquations будет иметь следующий вид:

```
function Red(du, u, h, p, t)
w, q, q_avg = u
hist1 = h(p, t - T)[1]
du[1] = 1.0 / T - (w * hist1 * pr / (2.0 * T))
du[2] = qAdd(q,w,T,C,N,R)
du[3] = -wq * C * q_avg + wq * C * q
end
```

Одним из мощных инструментов пакета DifferentialEquations являются обратные вызовы (callbacks), для работы с которыми определяются две функции. Функция условия (condition function) необходима для проверки того, произошло ли некоторое событие. Воздействующая функция (affect function) будет выполняться, если событие произошло.

Дискретная функция сброса пакетов реализуется как контроллер, который обновляется каждые 0.01 сек до достижения времени моделирования `tf`. Массив `tstops` определяет интервалы выборки проверки условия, функция `condition_control_loop` проверяет является ли `t` одним из экземпляров выборки:

```
tf = 30.0
tstops = collect(0:0.01:tf)
function condition_control_loop(u,t,integrator)
    (t in tstops)
end
```

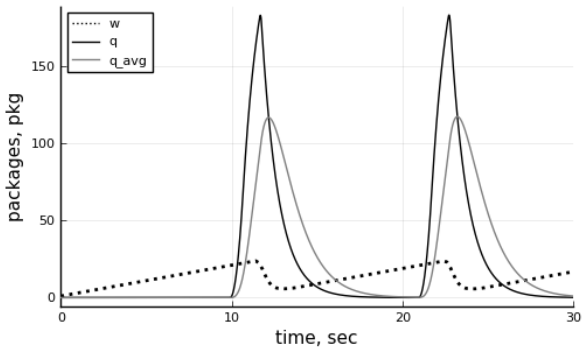


Рис. 2. Поведение параметров $w(t)$, $q(t)$, $\hat{q}(t)$ по результатам моделирования на языке Julia

Далее определим функцию воздействия, которая и является контроллером. Функция `control_loop!` на каждом шаге вычисляет новое значение вероятностной функции сброса пакетов p в зависимости от текущих значений параметров w, q, $\hat{q}(t)$.

Зададим обратный вызов дискретного типа:

```
cb = DiscreteCallback(condition_control_loop, control_loop!)
```

Наконец, определим вектор начального состояния системы, время моделирования и вызовем решатель пакета DDEProblem, в аргументы которого передается функция `Red`, вектор начальных состояний системы, время моделирования и параметры задержки переменных:

```
u0 = [1.0, 0.0, 0.0]
tspan = (0.0, tf)
prob = DDEProblem(Red, u0, h, tspan, p, constant_lags=lags)
alg = MethodOfSteps(Tsit5())
sol = solve(prob, alg, callback = cb, tstops=tstops)
```

В результате моделирования получим график изменения размера окна TCP Reno, отражающий динамику управления перегрузкой TCP, а также график среднего размера очереди, отражающий динамику очереди в маршрутизаторе (или шлюзе) с модулем управления очередью по алгоритму RED (рис. 2).

5. Результаты

В результате исследования получены два программных комплекса, созданных на языках программирования Modelica и Julia. Оба этих комплекса реализуют

одну и ту же математическую модель сети передачи данных с модулем активного управления трафиком, работающим по алгоритму RED.

Численный эксперимент, проведённый в рамках обоих программных комплексов, даёт сопоставимые результаты (см. рис. 1 и рис. 2).

6. Обсуждение

Оба языка программирования, и Modelica, и Julia являются предметно-ориентированными языками. Впрочем, Julia при этом рассматривается как общий язык научных расчётов, а Modelica как специализированный язык моделирования динамических систем, непрерывных и гибридных.

Математическая модель алгоритма RED сформулирована в рамках гибридного подхода. Поэтому в рамках языка Modelica её реализация вышла достаточно простой. На этом языке очень естественно, с помощью высокоуровневых средств производится запись обыкновенных дифференциальных уравнений с запаздывающим аргументом. Кроме того, использование дискретных элементов реализовано весьма наглядно.

Язык Julia направлен на решение более широкого круга задач. Поэтому он не обладает таким количеством синтаксического сахара, как более специализированная Modelica. В частности, реализация гибридной парадигмы в Julia требует более высокой квалификации программиста, чем при работе с языком Modelica.

Для специализированного применения Julia требует большего уровня знаний, нежели специализированные языки, такие, как Modelica. Впрочем, Julia является также и метаязыком, и может служить основой для конструирования других языков. Например, расширение Modia [17,18] было сделано для миграции программ с Modelica на Julia (и, возможно, и в обратном направлении).

7. Заключение

Авторами было продемонстрировано применение непрерывно-дискретного подхода к моделированию нелинейных систем с управлением. В качестве моделируемой системы выступала система, состоящая из входящего потока, обрабатываемого согласно протоколу TCP, а также маршрутизатора, обрабатывающего трафик по алгоритму типа RED.

В результате сравнения программных реализаций на языках программирования Modelica и Julia продемонстрирована простота моделирования гибридных систем в Modelica, где непрерывные элементы системы, реализованные с помощью системы дифференциальных уравнений, хорошо взаимодействует с дискретной функцией сброса пакетов. Ограничения для некоторых параметров системы задается с помощью оператора `when`. Реализация запаздывания также возможна для элементов системы любого характера функционирования.

Julia также дает возможность моделировать системы согласно гибридной парадигме. Пакет DifferentialEquations позволяет решать системы дифференциальных уравнений различных видов, в том числе и дифференциальные уравнения с запаздыванием. Непрерывная вероятностная функция успешно реализована с использованием опции обратных вызовов. Запаздывающий аргумент непрерывной функции был реализован с помощью функции истории `h(p, t)`.

Таким образом, авторами были изучены возможности языков программирования Modelica и Julia при моделировании гибридных систем, содержащих как непрерывные, так и дискретные аспекты поведения. Проведено численное моделирование процесса передачи данных по протоколу TCP и процесса регулирования алгоритмом RED состояния потока при возникновении перегрузок, получены графики, демонстрирующие изменения основных параметров системы.

Литература

1. Korolkova A. V., Velieva T. R., Abaev P. A., Sevastianov L. A., Kulyabov D. S. Hybrid Simulation Of Active Traffic Management // Proceedings 30th European Conference on Modelling and Simulation. — 2016. — 6. — P. 685–691.
2. Färnqvist D., Strandemar K., Johansson K. H., Hespanha J. P. Hybrid Modeling of Communication Networks Using Modelica // The 2nd International Modelica Conference. — 2002. — P. 209–213.
3. Korolkova A. V., Kulyabov D. S., Velieva T. R., Zaryadov I. S. Essay on the study of the self-oscillating regime in the control system // 33 European Conference on Modelling and Simulation, ECMS 2019 / Ed. by M. Iacono, F. Palmieri, M. Gribaudo, M. Ficco. — Vol. 33 of Communications of the ECMS. — Caserta : European Council for Modelling and Simulation, 2019. — 6. — P. 473–480.
4. Kulyabov D. S., Korolkova A. V., Velieva T. R., Eferina E. G., Sevastianov L. A. The Methodology of Studying of Active Traffic Management Module Self-oscillation Regime // DepCoS-RELCOMEX 2017. Advances in Intelligent Systems and Computing / Ed. by W. Zamojski, J. Mazurkiewicz, J. Sugier, T. Walkowiak, J. Kacprzyk. — Cham : Springer International Publishing, 2018. — Vol. 582 of Advances in Intelligent Systems and Computing. — P. 215–224.
5. Floyd S., Jacobson V. Random Early Detection Gateways for Congestion Avoidance // IEEE/ACM Transactions on Networking. — 1993. — Vol. 1, no. 4. — P. 397–413.
6. Misra V., Gong W.-B., Towsley D. Fluid-Based Analysis of a Network of AQM Routers Supporting TCP Flows with an Application to RED // ACM SIGCOMM Computer Communication Review. — 2000. — 10. — Vol. 30, no. 4. — P. 151–160.

7. Misra V., Gong W.-B., Towsley D. Stochastic Differential Equation Modeling and Analysis of TCP-Windowsize Behavior // Proceedings of PERFORMANCE. — 1999. — Vol. 99.
8. Korolkova A. V., Kulyabov D. S., Sevastianov L. A. Combinatorial and Operator Approaches to RED Modeling // Mathematical Modelling and Geometry. — 2015. — Vol. 3, no. 3. — P. 1–18.
9. Hnatič M., Eferina E. G., Korolkova A. V., Kulyabov D. S., Sevastyanov L. A. Operator Approach to the Master Equation for the One-Step Process // EPJ Web of Conferences. — 2016. — Vol. 108. — P. 02027. — arXiv : 1603.02205.
10. Fritzson P. Principles of Object-Oriented Modeling and Simulation with Modelica 2.1. — Wiley-IEEE Press, 2003. — 939 p.
11. Fritzson P. Introduction to Modeling and Simulation of Technical and Physical Systems with Modelica. — Hoboken, NJ, USA : John Wiley & Sons, Inc., 2011.
12. Apreutesey A.-M. Y., Korolkova A. V., Kulyabov D. S. Modeling RED algorithm modifications in the OpenModelica // Proceedings of the Selected Papers of the 9th International Conference "Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems" (ITTMM-2019), Moscow, Russia, April 15-19, 2019 / Ed. by D. S. Kulyabov, K. E. Samouylov, L. A. Sevastianov. — Vol. 2407 of CEUR Workshop Proceedings. — Moscow, 2019. — 4. — P. 5–14.
13. Bezanson J., Karpinski S., Shah V. B., Edelman A. Julia: A Fast Dynamic Language for Technical Computing. — 2012. — P. 1–27. — arXiv : 1209.5145.
14. Bezanson J., Edelman A., Karpinski S., Shah V. B. Julia: A fresh approach to numerical computing // SIAM Review. — 2017. — 1. — Vol. 59, no. 1. — P. 65–98. — arXiv : 1411.1607.
15. Joshi A., Lakhanpal R. Learning Julia. — Packt Publishing, 2017. — 316 p.
16. Rackauckas C., Nie Q. DifferentialEquations.jl – A Performant and Feature-Rich Ecosystem for Solving Differential Equations in Julia // Journal of Open Research Software. — 2017. — Vol. 5, no. May.
17. Elmqvist H., Henningsson T., Otter M. Systems Modeling and Programming in a Unified Environment Based on Julia // Leveraging Applications of Formal Methods, Verification and Validation: Discussion, Dissemination, Applications. ISoLA 2016 / Ed. by T. Margaria, B. Steffen. — Cham : Springer, 2016. — Vol. 9953 of Lecture Notes in Computer Science. — P. 198–217.
18. Otter M., Elmqvist H. Transformation of Differential Algebraic Array Equations to Index One Form // Proceedings of the 12th International Modelica Conference, Prague, Czech Republic, May 15-17, 2017. — Vol. 132. — 2017. — jul. — P. 565–579.

UDC: 123.456

Transient analysis of an M/M/1/N queue with balking, catastrophes, server failures and repairs

M.I.G. Suranga Sampath[1]

[1]Library, University of Kelaniya, Kelaniya-11600, Sri Lanka

Abstract

An M/M/1/N queue with balking, catastrophes, server failures and repairs is considered. The arrivals follow a Poisson distribution and the servers serve according to an exponential distribution. On arrival, a customer either decides to join the queue or may balk the queue based on number of customers in the system. The explicit expressions for the time-dependent system size probabilities are obtained in terms of the modified Bessel function of first kind.

Keywords: M/M/1/N queue, balking, catastrophes, server failure, server repairs.

1. Introduction

Queueing systems with balking can be applied in real life problems, such as computer and communication systems, production line systems, hospital emergency rooms with critical patients and etc. First research which is related to queueing systems subject to balking was conducted by Haight [2]. Abou El-Ata and Hariri [1] investigated multiple servers queueing system with balking and reneging subject to N-policy. Transient analysis of an $M/M/c$ queuing system with balking and retention of reneging customers was investigated by Kumar and Sharma [7]. The property of removing all the customers or some of them in the queueing system is considered as flushing the customers in the queue. Krishna Kumar and Arivudainambi [4] analyzed the transient solution for $M/M/1$ queue with catastrophes. An $M/M/R/N$ queueing system with balking, reneging and server break-downs was analyzed by Wang and Chang [10]. An $M/M/1$ which has N servers with server breakdowns and repairs was analyzed by Neuts and Lucantoni [8]. A single server queueing system with balking, catastrophes, server failures and repairs was analyzed by Tarabia [9] by extending the model of Krishna Kumar and Pavai Madheswari [5] with balking feature. And again, Kalidass et al. [3] derived the explicit expressions for an $M/M/1/N$ queue with catastrophes and a repairable server extending the earlier results which were obtained by Krishna Kumar et al. [6].

The absence of research trying to obtain transient solution for the $M/M/1/N$ queue with balking, catastrophes, server failures and repairs was noted. Thus, the most prominent task of this research was to develop the transient solutions of an $M/M/1/N$ queue with balking, catastrophes, server failures and repairs. This research can be considered as an extension of earlier research, especially, the research done by Krishna Kumar and Pavai Madheswari [5], Tarabia [9] and Kalidass et al. [3]. Sections of this paper is organized as follows; the model is given in section 2. The explicit expressions for the time-dependent system size probabilities are derived in the section 3. Section 4 contains the conclusion of this research.

2. Model Description

An $M/M/1/N$ queueing system with balking, catastrophes, server failure and repairs is considered. Arrivals are allowed to join the system according to a Poisson distribution with rate λ and service takes place according to an exponential distribution with rate μ. If a customer notes that the total number of jobs in the queue is less than or equal to $N-1$, then he is able to join the queueing system. On arrival a customer either decides to join the queue with probability one if the number of customers in the system is less than a threshold value $k(\leq N-1)$. If there are k customers or more ahead of him, then he joins the queue with probability β and may balk with probability $1-\beta$. The capacity of the system is finite. It is assumed that inter-arrival times and service times are mutually independent and the service discipline is First-In, First-Out (FIFO). When the system is idle or busy, catastrophes occur at the service station according to Poisson process of rate γ. Whenever a catastrophe occurs at the busy server, all customers in the system are destroyed immediately and the server gets inactivated. The repair times of failed server are i.i.d, according to an exponential distribution with parameter η. After repair, the server becomes ready to serve new customers. Let $Q(t)$ be the probability that the server is under repair at the instant t with $Q(0)=0$.

Let $\{X(t), t \geq 0\}$ denotes the total number of customers in the system at time t. Let $P_{i,n}(t) = P(X(t)=n|X(0)=i), i,n = 0,1,2,....N$

Then, the set of Chapman-Kolmogorov forward differential difference equations governing the process are given by

$$
\begin{aligned}
Q^{'}(t) &= -\eta Q(t) + \gamma[1-Q(t)], && (1)\\
P^{'}_{i,0}(t) &= \mu P_{i,1}(t) - (\lambda+\gamma)P_{i,0}(t) + \eta Q(t), && (2)\\
P^{'}_{i,n}(t) &= \lambda P_{i,n-1}(t) - (\lambda+\mu+\gamma)P_{i,n}(t) + \mu P_{i,n+1}(t), 1 \leq n \leq k-1, && (3)\\
P^{'}_{i,k}(t) &= \lambda P_{i,k-1}(t) - (\lambda\beta+\mu+\gamma)P_{i,k}(t) + \mu P_{i,k+1}(t), n = k, && (4)\\
P^{'}_{i,n}(t) &= \lambda\beta P_{i,n-1}(t) - (\lambda\beta+\mu+\gamma)P_{i,n}(t) + \mu P_{i,n+1}(t), k < n < N, && (5)\\
P^{'}_{i,N}(t) &= \lambda\beta P_{i,N-1}(t) - (\mu+\gamma)P_{i,N}(t), n = N && (6)
\end{aligned}
$$

with the initial condition $P_{i,m}(0) = P_{i,m}$.

3. Transient Probabilities

3.1. Evaluation of $P_{i,k+n}(t)$. Define the probability generating function $P(z,t)$ for the transient state probabilities $P_{i,n}(t)$ $(|z| \leq 1)$ as follows

$$P(z,t) = Q(t) + r_{i,k}(t) + \sum_{n=1}^{N-k} P_{i,k+n}(t) z^n \tag{7}$$

with $r_{i,k}(t) = \sum_{n=0}^{k} P_{i,n}(t)$, $P(z,0) = \sum_{m=0}^{N} P_{i,m} z^{\tau(m)}$ and $\tau(m) = (m-k)\left(1 - \sum_{n=0}^{k} \delta_{m,n}\right)$.

We will have the following equation after adding the system of equations (1)-(4),

$$Q^{'}(t) + r^{'}_{i,k}(t) = \gamma[1 - Q(t)] - \gamma r_{i,k}(t) - \lambda\beta P_{i,k}(t) + \mu P_{i,k+1}(t). \tag{8}$$

Multiplying the Equations (5) and (6) by appropriate powers of z and summing over the respective ranges of $n, (n > k)$, we can obtain

$$\begin{aligned} \sum_{n=1}^{N-k} P^{'}_{i,k+n}(t) z^n &= \left[\lambda\beta z - \lambda\beta - \mu - \gamma + \frac{\mu}{z}\right] \sum_{n=1}^{N-k} P_{i,k+n}(t) z^n \\ &\quad + \lambda\beta P_{i,k}(t) z - \lambda\beta z P_{i,N}(t) z^{N-k} + \lambda\beta P_{i,N}(t) z^{N-k} - \mu P_{i,k+1}(t) \end{aligned} \tag{9}$$

Adding the two equations (8) and (9) and using the definition of $P(z,t)$, we have

$$\begin{aligned} &\frac{\partial P(z,t)}{\partial t} + \left[\lambda\beta(1-z) + \mu(1-z^{-1}) + \gamma\right] P(z,t) - \gamma \\ &= \left[\lambda\beta(1-z) + \mu(1-z^{-1})\right] [Q(t) + r_{i,k}(t)] + \lambda\beta(1-z) P_{i,N}(t) z^{N-k} + \lambda\beta(z-1) P_{i,k}(t) \end{aligned} \tag{10}$$

with the initial condition $P(z,0) = \sum_{m=0}^{N} P_{i,m} z^{\tau(m)}$ and $Q(0) = 0$.

After some algebra, we have

$$\begin{aligned} P(z,t) &= \int_0^t e^{\left\{-\left[\lambda\beta(1-z) + \mu(1-z^{-1}) + \gamma\right](t-u)\right\}} \left\{ \left[\lambda\beta(1-z) + \mu(1-z^{-1})\right] [Q(u) + r_{i,k}(u)] \right. \\ &\quad \left. + \lambda\beta(1-z) P_{i,N}(u) z^{N-k} + \lambda\beta(z-1) P_{i,k}(u) + \gamma \right\} du \\ &\quad + P(z,0) e^{-\left[\lambda\beta(1-z) + \mu(1-z^{-1}) + \gamma\right] t}. \end{aligned} \tag{11}$$

Let $a = \lambda\beta + \mu + \gamma$, $b = \lambda\beta$ and $c = \mu$, then $\lambda\beta(1-z) + \mu(1-z^{-1}) + \gamma = a - bz - \frac{c}{z}$. Using the Bessel function properties and on account of $\alpha = 2\sqrt{bc}$ and $\nu = \sqrt{\frac{b}{c}}$, we obtain $\exp\left\{\left(bz + \frac{c}{z}\right) t\right\} = \sum_{n=-\infty}^{\infty} (\nu z)^n I_n(\alpha t)$.

Comparing the coefficients of $z^n, n = 1,2,3,....,N$ on the right hand side and left hand side of the Equation (11), we will have

$$\begin{aligned} P_{i,k+n}(t) &= \int_0^t e^{-a(t-u)} \left\{ \left[-b\nu^{n-1} I_{n-1}(\alpha(t-u)) - c\nu^{n+1} I_{n+1}(\alpha(t-u)) \right.\right. \\ &\quad \left.\left. + (a-\gamma)\nu^n I_n(\alpha(t-u)) \right] [Q(u) + r_{i,k}(u)] + \gamma\nu^n I_n(\alpha(t-u)) \right\} du \\ &\quad + \lambda\beta \int_0^t e^{-a(t-u)} \left[\nu^{k+n-N} I_{k+n-N}(\alpha(t-u) \right. \\ &\quad \left. - \nu^{k+n-N-1} I_{k+n-N-1}(\alpha(t-u)) \right] P_{i,N}(u) du + \sum_{m=o}^{N} P_{i,m} \nu^{n-\tau(m)} e^{-at} I_{n-\tau(m)}(\alpha t) \\ &\quad + \lambda\beta \int_0^t e^{-a(t-u)} \left[\nu^{n-1} I_{n-1}(\alpha(t-u)) - \nu^n I_n(\alpha(t-u)) \right] P_{i,k}(u) du. \end{aligned} \tag{12}$$

Again comparing the constant terms in either side of the equation (11) and using the Bessel property $I_{-k}(.) = I_k(.)$, we will have

$$\begin{aligned} Q(t) + r_{i,k}(t) &= \int_0^t e^{-a(t-u)} \left\{ \left[-b\nu^{-1} I_1(\alpha(t-u)) - c\nu I_1(\alpha(t-u)) \right.\right. \\ &\quad \left.\left. + (a-\gamma) I_0(\alpha(t-u)) \right] [Q(u) + r_{i,k}(u)] + \gamma I_0(\alpha(t-u)) \right\} du \\ &\quad + \lambda\beta \int_0^t e^{-a(t-u)} \left[\nu^{-N+k} I_{N-k}(\alpha(t-u) \right. \\ &\quad \left. - \nu^{-N+k-1} I_{N-k+1}(\alpha(t-u)) \right] P_{i,N}(u) du + \sum_{m=o}^{N} P_{i,m} \nu^{-\tau(m)} e^{-at} I_{\tau(m)}(\alpha t) \\ &\quad + \lambda\beta \int_0^t e^{-a(t-u)} \left[\nu^{-1} I_1(\alpha(t-u)) - I_0(\alpha(t-u)) \right] P_{i,k}(u) du. \end{aligned} \tag{13}$$

Since left hand side of equation (11) doesn't contain any negative powers of z, then comparing the negative powers of z with zero and using the Bessel property $I_{-k}(.) = I_k(.)$, we will have

$$\begin{aligned} &\int_0^t e^{-a(t-u)} (a-\gamma)\nu^n I_n(\alpha(t-u)) [Q(u) + r_{i,k}(u)] du \\ &= \int_0^t e^{-a(t-u)} \left[b\nu^{n-1} I_{n+1}(\alpha(t-u)) + c\nu^{n+1} I_{n-1}(\alpha(t-u)) \right] [Q(u) + r_{i,k}(u)] du \\ &-\lambda\beta \int_0^t e^{-a(t-u)} \left[\nu^{n-N+k} I_{n+N-k}(\alpha(t-u) - \nu^{n-N+k-1} I_{n+N-k+1}(\alpha(t-u)) \right] P_{i,N}(u) du \\ &\quad -\lambda\beta \int_0^t e^{-a(t-u)} \left[\nu^{n-1} I_{n+1}(\alpha(t-u)) - \nu^n I_n(\alpha(t-u)) \right] P_{i,k}(u) du \\ &\quad - \int_0^t e^{-a(t-u)} \gamma\nu^n I_n(\alpha(t-u)) du - \sum_{m=o}^{N} P_{i,m} \nu^{n-\tau(m)} e^{-at} I_{n+\tau(m)}(\alpha t). \end{aligned} \tag{14}$$

Substituting the Equation (14) to the Equation (12) and after some algebra, we have

$$\begin{aligned} P_{i,k+n}(t) &= \lambda\beta\int_0^t e^{-a(t-u)}\Big\{\nu^{k+n-N}\left[I_{k+n-N}(\alpha(t-u))-I_{n+N-k}(\alpha(t-u))\right] \\ &\quad -\nu^{k+n-N-1}\left[I_{k+n-N-1}(\alpha(t-u))\ -I_{n+N-k+1}(\alpha(t-u))\right]\Big\}P_{i,N}(u)du \\ &\quad +n\nu^n\int_0^t\frac{e^{-a(t-u)}}{(t-u)}I_n(\alpha(t-u))P_{i,k}(u)du \\ &\quad +\sum_{m=o}^{N}P_{i,m}\nu^{n-\tau(m)}e^{-at}\left[I_{n-\tau(m)}(\alpha t)-I_{n+\tau(m)}(\alpha t)\right]. \end{aligned} \tag{15}$$

Let $P_{i,0}(0)=1$ and $\hat{f}(s)$ denotes the Laplace transform of $f(t)$. By taking the Laplace transform of the system of Equations (1)-(3) and applying boundary conditions, we have

$$\hat{Q}(s) = \frac{\gamma}{s(s+\eta+\gamma)}, \tag{16}$$

$$s\hat{P}_{i,0}(s)-1 = \mu\hat{P}_{i,1}(s)-(\lambda+\gamma)\hat{P}_{i,0}(s)+\eta\hat{Q}(s), \tag{17}$$

$$s\hat{P}_{i,n}(s) = \lambda\hat{P}_{i,n-1}(s)-(\lambda+\mu+\gamma)\hat{P}_{i,n}(s)+\mu\hat{P}_{i,n+1}(s). \tag{18}$$

3.2. Evaluation of $P_{i,n}(t)$. By the Equation (18), we will have

$$\frac{\hat{P}_{i,n}(s)}{\hat{P}_{i,n-1}(s)} = \frac{\lambda}{(s+\lambda+\mu+\gamma)-\phi(s)} \tag{19}$$

where

$$\phi(s) = \frac{\lambda\mu}{(s+\lambda+\mu+\gamma)-\frac{\lambda\mu}{(s+\lambda+\mu+\gamma)-\frac{\lambda\mu}{(s+\lambda+\mu+\gamma)-....}}},$$

$$\phi(s) = \frac{\lambda\mu}{(s+\lambda+\mu+\gamma)-\phi(s)}.$$

It is noted that $\phi(s)$ satisfies the quadratic equation, $\phi^2(s)-(s+\lambda+\mu+\gamma)\,\phi(s)+\lambda\mu=0$, the roots of which are $\theta(s),\vartheta(s)=\frac{w\pm\sqrt{w^2-4\lambda\mu}}{2}$ and $w=s+\lambda+\mu+\gamma$. It is clear that $\theta(s)=\frac{w-\sqrt{w^2-4\lambda\mu}}{2}$ is the unique real root within $[0,1)$ for $\gamma>0$ and $0\le s<1$.

Substituting $\theta(s)$ for the Equation (19) and after some algebra, we will have

$$\hat{P}_{i,n}(s) = \left[\frac{2\lambda}{w+\sqrt{w^2-4\lambda\mu}}\right]^n\hat{P}_{i,0}(s). \tag{20}$$

Taking the inversion of (20), we can obtain

$$P_{i,n}(t) = n\left(\frac{\lambda}{\mu}\right)^{\frac{n}{2}}\int_0^t P_{i,0}(u)e^{-(\lambda+\mu+\gamma)(t-u)}\frac{I_n\left(2\sqrt{\lambda\mu}(t-u)\right)}{(t-u)}du. \tag{21}$$

3.3. Evaluation of $P_{i,0}(t)$. By equation (17), we will have

$$\hat{P}_{i,0}(s) = \frac{1+\frac{\gamma\eta}{s(s+\eta+\gamma)}}{(s+\lambda+\gamma)-\phi(s)}. \tag{22}$$

Substituting $\theta(s)=\frac{w-\sqrt{w^2-4\lambda\mu}}{2}$ for the Equation (22) and after some algebra, we will have

$$\begin{aligned}\hat{P}_{i,0}(s) &= \frac{\lambda(\lambda-\eta+\gamma)}{(\lambda+\gamma)(\lambda-\eta)}\frac{1}{(s+\lambda+\gamma)}\sum_{n=0}^{\infty}\left[\frac{2\lambda\mu}{(s+\lambda+\gamma)(w+\sqrt{w^2-4\lambda\mu})}\right]^n \\ &+\frac{\gamma\eta}{(\eta+\gamma)(\lambda+\gamma)}\frac{1}{s}\sum_{n=0}^{\infty}\left[\frac{2\lambda\mu}{(s+\lambda+\gamma)(w+\sqrt{w^2-4\lambda\mu})}\right]^n \\ &+\frac{\gamma\eta}{(\eta+\gamma)(\eta-\lambda)}\frac{1}{(s+\eta+\gamma)}\times\sum_{n=0}^{\infty}\left[\frac{2\lambda\mu}{(s+\lambda+\gamma)(w+\sqrt{w^2-4\lambda\mu})}\right]^n \end{aligned}\tag{23}$$

Taking the inversion of the Equation (23), we will have

$$\begin{aligned}P_{i,0}(t) &= \frac{\lambda(\lambda-\eta+\gamma)}{(\lambda+\gamma)(\lambda-\eta)}\sum_{n=0}^{\infty}n(\lambda\mu)^{\frac{n}{2}}\int_0^t e^{-(\lambda+\gamma)(t-u)}\int_0^u e^{-(\lambda+\gamma)u}\frac{u^{n-1}}{(n-1)!} \\ &\times e^{-(\lambda+\mu+\gamma)(u-v)}\frac{I_n\left(2\sqrt{\lambda\mu}(u-v)\right)}{(u-v)}dudv \\ &+\frac{\gamma\eta}{(\eta+\gamma)(\lambda+\gamma)}\sum_{n=0}^{\infty}n(\lambda\mu)^{\frac{n}{2}}\int_0^t\int_0^u e^{-(\lambda+\gamma)u}\frac{u^{n-1}}{(n-1)!} \\ &\times e^{-(\lambda+\mu+\gamma)(u-v)}\frac{I_n\left(2\sqrt{\lambda\mu}(u-v)\right)}{(u-v)}dudv \\ &+\frac{\gamma\eta}{(\eta+\gamma)(\eta-\lambda)}\sum_{n=0}^{\infty}n(\lambda\mu)^{\frac{n}{2}}\int_0^t e^{-(\eta+\gamma)(t-u)}\int_0^u e^{-(\lambda+\gamma)u}\frac{u^{n-1}}{(n-1)!} \\ &\times e^{-(\lambda+\mu+\gamma)(u-v)}\frac{I_n\left(2\sqrt{\lambda\mu}(u-v)\right)}{(u-v)}dudv. \end{aligned}\tag{24}$$

3.4. Evaluation of $P_{i,N}(t)$. By taking the Laplace transform of the Equation (6) and after some algebra, we can obtain

$$\hat{P}_{i,N}(s) = \left[\frac{\lambda\beta}{(s+\mu+\gamma)}\right]^N\hat{P}_{i,0}(s). \tag{25}$$

Taking the inverse Laplace transform of the Equation (25), we will have

$$P_{i,N}(t) = (\lambda\beta)^N \int_0^t e^{-(\mu+\gamma)(t-u)} \frac{(t-u)^{N-1}}{(N-1)!} P_{i,0}(u)du. \quad (26)$$

3.5. Evaluation of $P_{i,k}(t)$. Taking the Laplace transform of the Equation (4) and after some mathematical calculation, we have

$$\frac{\hat{P}_{i,k}(s)}{\hat{P}_{i,k-1}(s)} = \frac{\lambda}{(s+\lambda+\mu+\gamma)-\psi(s)} \quad (27)$$

where

$$\psi(s) = \frac{\lambda\mu}{(s+\lambda\beta+\mu+\gamma)-\frac{\lambda\mu}{(s+\lambda\beta+\mu+\gamma)-\frac{\lambda\mu}{(s+\lambda\beta+\mu+\gamma)-....}}},$$

$$\psi(s) = \frac{\lambda\mu}{(s+\lambda\beta+\mu+\gamma)-\psi(s)}.$$

It is noted that $\psi(s)$ satisfies the quadratic equation, $\psi^2(s)-(s+\lambda\beta+\mu+\gamma)\psi(s)+\lambda\mu=0$, the roots of which are $\delta(s),\varepsilon(s)=\frac{\bar{w}\pm\sqrt{\bar{w}^2-4\lambda\mu}}{2}$ and $\bar{w}=s+\lambda\beta+\mu+\gamma$. It is clear that $\delta(s)=\frac{\bar{w}-\sqrt{\bar{w}^2-4\lambda\mu}}{2}$ is the unique real root within $[0,1)$ for $\gamma>0$ and $0\le s<1$.

Substituting $\delta(s)$ for the Equation (27) and after some algebra, we will have

$$\hat{P}_{i,k}(s) = \left[\frac{2\lambda}{\bar{w}+\sqrt{\bar{w}^2-4\lambda\mu}}\right]\hat{P}_{i,k-1}(s). \quad (28)$$

Substituting $n=k-1$ into the Equation (20), we will have

$$\hat{P}_{i,k-1}(s) = \left[\frac{2\lambda}{w+\sqrt{w^2-4\lambda\mu}}\right]^{k-1}\hat{P}_{i,0}(s). \quad (29)$$

By the Equations (28) and (29), we can obtain

$$\hat{P}_{i,k}(s) = \left[\frac{2\lambda}{\bar{w}+\sqrt{\bar{w}^2-4\lambda\mu}}\right]\left[\frac{2\lambda}{w+\sqrt{w^2-4\lambda\mu}}\right]^{k-1}\hat{P}_{i,0}(s). \quad (30)$$

Taking the inverse Laplace transform of the Equation (30), we can derive

$$P_{i,k}(t) = (k-1)\left(\frac{\lambda}{\mu}\right)^{\frac{k}{2}} \int_0^t e^{-(\lambda\beta+\mu+\gamma)(t-u)} \frac{I_1\left(2\sqrt{\lambda\mu}(t-u)\right)}{(t-u)} \times \int_0^u e^{-(\lambda+\mu+\gamma)(u-v)} \frac{I_{k-1}\left(2\sqrt{\lambda\mu}(u-v)\right)}{(u-v)} P_{i,0}(v)dudv. \quad (31)$$

3.6. Evaluation of $Q(t)$. Taking the inversion of Laplace transform of Equation (16) and using the convolution theorem, yields

$$Q(t) = \frac{\gamma}{\gamma+\eta}\left(1-e^{-(\gamma+\eta)t}\right). \quad (32)$$

4. Conclusion

An M/M/1/N queue with balking, catastrophes, server failures and repairs is considered and the explicit expression for the transient probabilities are obtained in terms of the modified Bessel function of first kind.

REFERENCES

1. Abou El-Ata, M. O. and Hariri, A. M. A. The M/M/C/N queue with balking and reneging Computers and //Operations Research. 1992. V. 19. P. 713–716
2. Haight, F. A. Queueing with balking //Biometrika. 1957. V. 44. P. 360–369
3. Kalidass, K., Gopinath, S., Gnanaraj, J. and Ramanath, K. Time dependent analysis of an M/M/1/N queue with catastrophes and a repairable server //Opsearch. 2012. V. 49. P. 39–61
4. Krishna Kumar, B. and Arivudainambi, D. Transient solution of an M/M/1 queue with catastrophes //Computers and Mathematics with Applications. 2000. V. 10. P. 1233–1240
5. Krishna Kumar, B. and Pavai Madheswari, S. Transient analysis of an M/M/1 queue subject to catastrophes and server failures //Stochastic Analysis and Applications. 2005. V. 23. P. 329–340
6. Krishna Kumar, B., Krishnamoorthy, A., Pavai Madheswari, S. and Sadiq Basha, S. Transient analysis of a single server queue with catastrophes, failures and repaires //Queueing Systems. 2007. V. 56. P. 133–141
7. Kumar, R. and Sharma, S. Transient analysis of an M/M/c queuing system with balking and retention of reneging customers //Communications in Statistics - Theory and Methods. 2018. V. 47(6). P. 1318–1327
8. Neuts, M. F. and Lucantoni, D. M. A Markovian queue with N servers subjects to breakdowns and repairs //Management Science. 1979. V. 25(9). P. 849–861
9. Tarabia, A. M. K. Transient and steady-state analysis of an M/M/1 queue with balking, catastrophes, server failures and repairs //Journal of Industrial and Management Optimization. 2011. V. 7. P. 811–823
10. Wang, K. H. and Chang, Y. C. Cost analysis of a finite M/M/R queueing system with balking, reneging and server breakdowns //Mathematical Methods of Operations Research. 2002. V. 56. P. 169–180

UDC: 519.248

On algorithms for effective speed-up simulation of reliability models

A. V. Borodina[1,2] and V. A. Tishenko [2]

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

[2]Petrozavodsk State University, Petrozavodsk, Russia

borodina@krc.karelia.ru, vitalik1tishenko@gmail.com

Abstract

We consider a regenerative degradation process composed of a sum of the successive phases, where preventative maintenance is used to prevent instantaneous failure. Calculation of the failure probability and other characteristics of the regeneration cycle is critical for the optimal control of such systems. When an instantaneous failure is a rare event, this model is a good reference for testing the variance reduction techniques that were proposed earlier and speed-up simulation algorithms. In more general cases, when we need to use simulation we propose two scenarios of the splitting method to evaluate the characteristics of the degradation process more effectively than naive Monte Carlo.

Keywords: failure probability, reliability analysis, degradation process, regenerative splitting, relative error, speed-up simulation

1. Introduction

Analysis and modeling of degradation and shock multi-stage processes is a key step in the development and implementation of modern highly reliable technologies. It is quite expected that the emergence of new models forces to modernize research methods for shock models with changing degradation rate, with soft and hard failures with a natural or predetermined threshold level, which is generally a random variable [1, 2]. In particular, effective methods for calculating the realistic model parameters and system reliability are in demand. Nevertheless, the *naive Monte Carlo method* is still popular in a large number of modern works, despite its well-known inefficiency, for example, see [3, 4, 5, 6]. For instance, as it proposed in [7], a standard solution is

The study was carried out under state order to the Karelian Research Centre of the Russian Academy of Sciences (Institute of Applied Mathematical Research KarRC RAS) and supported by the Russian Foundation for Basic Research, project 18-07-00187

based on the stationary distribution of the built-in Markov chain for the simplest cases. Then, for the generalized model, Monte Carlo simulation is used to approximate the cost of maintenance.

Nevertheless, the models become more complex, and analytical methods are less available. In particular, the Wiener process with independent and normally distributed increments is widely used for non-monotonic degradation. The Gamma process is useful in the stochastic modeling of monotonic and gradual degradation, characterized by the sequence of tiny increments, such as wear, fatigue, corrosion, crack growth, erosion, consumption, degrading health index. In addition, more complex models of two-stage or multi-stage degradation include Gamma-Gamma, Wiener-Wiener, Gamma-Wiener degradation models are also considered (for example, see [8, 9]). In this regard, we suggest looking for more effective alternatives for the naive simulation method that can be used to analyze the reliability of modern systems.

For a homogeneous case of exponential degradation stages of the system with gradual and instantaneous failures, analytical formulas were obtained and an advanced simulation technique was proposed in [10]. In [11] a heterogeneous degradation process was investigated analytically for exponential stages. Numerical experiments confirmed that using the standard Monte Carlo method impairs the accuracy of the probability estimates and other characteristics of the degradation process. In [12] the variance reduction technique has been extended to estimate the failure probability that a random sum exceeds a random variable V. In [13] a variance reduction technique using a special variant of *conditional Monte Carlo* approach proposed for heterogeneous dagradation process. It was shown by numerical examples that relative error is bounded and even is vanishing when the degradation stage has heavy-tailed distribution. On the other hand, our experiments show that this method has not an advantage for the light-tailed stages. A variance reduction technique based on the *Importance sampling* with an exponential change of measure for light-tailed degradation stages was introduced in [14].

All these techniques were tested on a model of the degradation process, which describes the thickness of the anti-corrosion coating and is described in [10]. Since it is possible to obtain analytical results in the simplest cases, this model is convenient for analyzing the effectiveness of accelerated simulation and variance reducing methods. In addition, the process has a regenerative structure, which is typical for degradation models. The proposed accelerated algorithm can also be extended for more complex models by replacing the procedure of simulation the time spent at the degradation stages.

2. Crude simulation of the degradation process

Following [10] consider the degradation process $X := \{X(t),\, t \geq 0\}$ with a finite state space $E = \{0, 1, \ldots, L, \ldots, M, \ldots, K; F\}$ describing the degradation stages of the system. Two-threshold policy (K, L) is considered, which means that the system is restored in the stage K and then proceeds to stage L (see Fig. 1. a)).

Let T_i be the transition time from i to $i+1$ stage. Random variables (r.v.) T_i are independent but not necessarily identically distributed. Starting in state $X(0) = 0$ the process successively passes $K-1$ intermediate degradation stages and reaches the state M. We denote by V a random time after which a failure can occur. Thus, after the stage M either event $\{S_{M,K} \geq V\}$ (*instantaneous failure*) may happen during a random period V, or event $\{S_{M,K} < V\}$ (*starting the preventive repair stage*) occurs during the time

$$S_{M,K} = \sum_{i=M}^{K-1} T_i.$$

The process X is strongly regenerative with moments

$$\tau_{n+1} = \inf\{Z_i > \tau_n :\, X(Z_i^+) = M\},\ n \geq 0,\ \tau_0 := 0,$$

where Z_k is the the hitting time of the stage $k \geq 1$, and cycle lengths $Y_k = \tau_{k+1} - \tau_k$, $k \geq 1$ are i.i.d. There are two types of regeneration cycles: with and without failure

$$Y = \begin{cases} Y_F = V + U_F + S_{0,M}, & \text{if } S_{M,K} \geq V \\ Y_{NF} = S_{M,K} + U_{K,L} + S_{L,M}, & \text{if } S_{M,K} < V, \end{cases} \tag{1}$$

where r.v. V, U_F, $S_{0,M} = \sum_{i=0}^{M-1} T_i$ with known distributions are independent as well as r.v. $S_{M,K}$, $U_{K,L}$, $S_{L,M} = \sum_{i=L}^{M-1} T_i$ After failure and repair, the process returns to the initial state 0. Thus (unconditional) regeneration cycle length Y can be written as

$$Y = Y_F \cdot I_{\{V \leq S_{M,K}\}} + Y_{NF} \cdot I_{\{S_{M,K} < V\}}. \tag{2}$$

where I_A denotes indicator function. The variable Y plays an important role in the analysis of the degradation process [10].

The main target is to find the probability of instantaneous failure within the regeneration cycle, that is

$$p_F = \mathbb{P}(S_{M,K} \geq V) = \mathbb{E}[F_V(S_{M,K})], \tag{3}$$

where F_V is the distribution function of the random variable V. But the simulation also necessary for other characteristics like the mean lifetime T

$$\mathbb{E}[T] = \mathbb{E}[Y_{NF}](\mathbb{E}[N] - 1) + \mathbb{E}[V|V \leq S_{M,K}],$$

where $\mathbb{E}[N] = 1/p_F$ is the mean number of cycles until complete failure; mean cycle length $\mathbb{E}[Y_F]$ with failure or $\mathbb{E}[Y_{NF}]$ without failure; reliability function

$$R(t) = \mathbb{P}[T > t|X(0) = 0],\ t \geq 0,$$

where T stands for the lifetime of the system.

Note that, for more complex shock models simulation allows us to estimate parameters of the model like the damage threshold, the intensity of random shock, critical shock inter-arrival time, scale, and shape parameters in Gamma models, etc.

If the failure is not a rare event and it is easy to simulate the r.v. V, U_F, $S_{0,M}$, $S_{M,K}, U_{K,L}$, $S_{L,M}$ on a computer and the performance function is computationally inexpensive to construct the regeneration cycles, then the p_F and other characteristics of degradation process can be approximated by *crude Monte Carlo* unbiased estimators. However, for highly reliable systems, such an assumption seems to be naive.

3. Splitting scenarios for the degradation process

The splitting procedure for a homogeneous degradation process was firstly proposed in [10], where the method showed the effectiveness of the estimation for exponential degradation stages. Taking into account the generalization of the method for other models, let's now compare *two possible splitting scenarios:* a) the process splits at each stage of degradation (Fig.1. a)); b) the levels of splitting l_i depend on the value of the accumulated amount of time $S_{M,K}$ (Fig.1. b)).

In both cases, the splitting of the trajectories occurs only in the area after stage M, when instantaneous failure becomes possible. Unlike processes with negative drift (which is typical for problems of evaluating rare events), it is impossible to perform optimal leveling due to the randomness of the threshold time to failure V for the degradation process. All methods for rare events probabilities simulation are designed to solve problems with a constant value of failure threshold [15, 16]. For standard splitting, an optimal distance between thresholds $\{l_i\}$ and splitting factors $\{R_i\}$ at each threshold are defined by pilot run [16]. The pilot run defines threshold partition in accordance with the requirement that conditional probabilities of transition between thresholds p_i is not so rare but gives the biased estimator. In addition, it is optimal if the number of branching trajectories does not grow exponentially on the one hand and the process is not damped on the other. Thus,

the optimal values are related by the ratio $R_i = 1/p_i$, but $p_i \approx 1$ for degradation process.

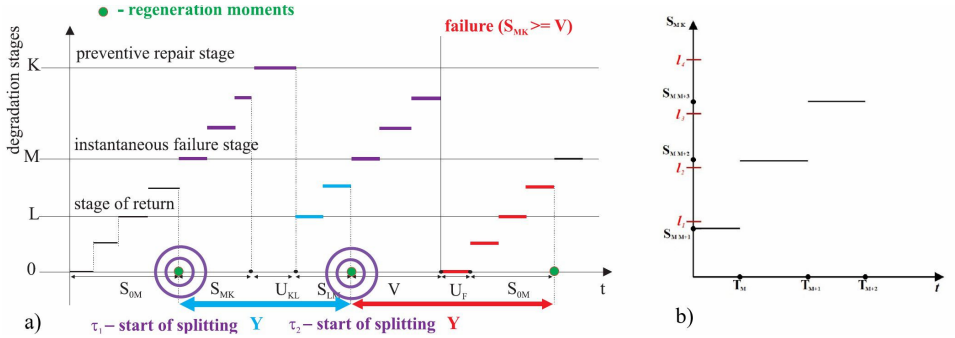


Fig. 1. Two splitting schemes: a) by stages b) by the value of $S_{M,K}$

Besides, given the artificial branching of the process, it is necessary to take into account the dependency between cycles. At each level we generate R_i copies of r.v. T_i, $M \le i \le K-1$, $R_{M+1} = 1$. So, each original path generates $D = R_M \cdots R_{K-1}$ (dependent) subpaths called *group of cycles*. The dependence is generated by the same pre-history of realizations of S_{MK} before the splitting point at each stage.

Each process trajectory started from the initial threshold l_0 gives the *group* of D dependable regeneration cycles. The cycles from different groups are independent by construction. For the degradation process, the groups started at the regeneration moment τ_i. The cycles from groups are independent. R_{M-1} is the total number of groups. The total number of the failures in the ith group is

$$A_i = \sum_{j=(i-1)\cdot D+1}^{i\cdot D} I^{(j)}, i = 1, \ldots, R_{M-1},$$

where $I^{(j)} = 1$ for the cycle with failure ($I^{(j)} = 0$, otherwise). Sequence $\{I^{(j)}, j \ge 1\}$ is discrete D-dependent regenerative with constant cycle length D and regeneration instants $i \cdot D$, $i \in [1, R_{M-1}]$.

The regenerative interpretation gives the strongly consistent estimator $\widehat{p}_F$ and the following $100(1-\delta)\%$ confidence interval for p_F based on the *regenerative variant of Central Limit Theorem*, that is well known from [17]

$$\widehat{p}_F = \frac{\sum_{j=1}^{R_{M-1}} A_j}{R_{M-1} \cdot D} \to \frac{\mathbb{E}A_1}{D} = p_F, \quad \left[\widehat{p}_F \pm \frac{z(\delta)\sqrt{v_n}}{\sqrt{n}}\right] \tag{4}$$

where quantile $z(\delta)$ satisfies $\mathbb{P}[N(0,1) \le z(\delta)] = 1 - \delta/2$, and

$$v_n = \frac{n^{-1}\sum_{i=1}^{n}[A_i - \widehat{p_F}D]^2}{D^2} \tag{5}$$

is a weakly consistent estimator of $\sigma^2 = \mathbb{E}[A_1 - p_F D]^2/D^2$ if $\mathbb{E}(A_1 - \gamma\alpha_1)^2 < \infty$. Under moment assumptions, $\mathbb{E}A_1^2 < \infty$, the estimate (5) is strongly consistent.

An experimental comparison of two splitting scenarios was made with respect to the following evaluation quality criteria: relative error RE and relative experimental error RER (if p_F is analytically available)

$$RE[\widehat{p}_F] = \frac{\sqrt{Var[\widehat{p}_F]}}{\mathbb{E}[\hat{p}_F]},\ RER[\widehat{p_F}] = |\hat{p_F} - p_F| \cdot 100/p_F.$$

Let's fix the parameters of the model $\nu = 0.5$, $\mu_F = 1.5$, $\mu = 2$, $L = 1$, $M = 5$, $K = 17$. To observe both RE and RER values we give an example of exponential degradation periods $T_j \sim Exp(\lambda_j)$ for the heterogeneous case where the analytical formulas are known from [11]. So, we will vary number of regeneration cycles n and sequence of values

$$\lambda_j = \lambda_{K-1} - (K - j - 1)s,\ j \in [0, K-2],$$

where λ_{K-1} will be initialized before starting the splitting procedure, and the other values will be shifted by step s, thus the condition

$$\lambda_0 < \cdots < \lambda_{K-1},\ \ \nu < \lambda_j,\ j = [0, K-1],$$

for increasing the degradation rate is guaranteed.

λ_{K-1}, s	n	p_F	t_{MC}	$t_{RS_{a)}}$	$t_{RS_{b)}}$
10^3, 50	10^4	$8.75 \cdot 10^{-3}$	0.318	0.019	0.228
10^4, $5 \cdot 10^2$	10^4	$8.79 \cdot 10^{-4}$	0.312	0.018	0.222
10^5, $5 \cdot 10^3$	10^5	$8.79 \cdot 10^{-5}$	3.06	0.180	2.16
10^6, $5 \cdot 10^4$	10^7	$8.79 \cdot 10^{-6}$	253	17.4	218

Table 1. Time estimator (sec.): MC vs. $RS_{a)}$ and $RS_{b)}$: $T_j \sim Exp(\lambda_j)$, $V \sim Exp(\nu)$

All numerical tests were executed on ultrabook HP ENVY Intel(R) Core(TM) i3 7100U 2.4GHz processor with 4GB of RAM, running Windows 10. Tables 1, 2, 3 show the results of running a programs in Python3 for the Monte Carlo method (MC) and the regenerative splitting method (RS) for a) and b) scenarios.

λ_{K-1}, s	n	p_F	RER_{MC}	$RER_{RS_{a)}}$	$RER_{RS_{b)}}$
10^3, 50	10^4	$8.75 \cdot 10^{-3}$	3.02	2.06	3.91
10^4, $5 \cdot 10^2$	10^4	$8.79 \cdot 10^{-4}$	4.59	3.46	4.17
10^5, $5 \cdot 10^3$	10^5	$8.79 \cdot 10^{-5}$	2.85	4.43	1.46
10^6, $5 \cdot 10^4$	10^7	$8.79 \cdot 10^{-6}$	0.79	1.23	4.79

Table 2. RER estimator: MC vs. $RS_{a)}$ and $RS_{b)}$: $T_j \sim Exp(\lambda_j)$, $V \sim Exp(\nu)$

λ_{K-1}, s	n	p_F	RE_{MC}	$RE_{RS_{a)}}$	$RE_{RS_{b)}}$
10^3, 50	10^4	$8.75 \cdot 10^{-3}$	0.18	3.02	1.5
10^4, $5 \cdot 10^2$	10^4	$8.79 \cdot 10^{-4}$	0.36	2.16	4.60
10^5, $5 \cdot 10^3$	10^5	$8.79 \cdot 10^{-5}$	0.31	3.23	4.65
10^6, $5 \cdot 10^4$	10^7	$8.79 \cdot 10^{-6}$	0.11	4.75	1.53

Table 3. RE estimator: MC vs. $RS_{a)}$ and $RS_{b)}$: $T_j \sim Exp(\lambda_j)$, $V \sim Exp(\nu)$

We compare all methods with the same number n of regeneration cycles and track the corresponding time t_{MC}, $t_{RS_{a)}}$, $t_{RS_{b)}}$ of each method in seconds. The number of cycles n is chosen for practical reasons so that the RER does not exceed 5% for all methods. For variance estimation, a sample of 100 values was constructed.

The Table 1 shows that the $RS_{a)}$ algorithm is significantly superior in time to the others MC and $RS_{b)}$. Despite the higher RE the a) splitting scenario provides a lower RER and is, therefore, closer to the analytical value than b).

It should be noted that for the scenario $RS_{b)}$, cases with a fixed number of splits R_i and a random one (exponential, uniform) were also studied, they are omitted here due to brevity, however, we did not see significant variations between them in terms of time, RE and RER.

4. Conclusion

We performed a series of experiments that demonstrated the effectiveness of the two splitting schemes for the accelerated simulation of regeneration cycles and estimation of the degradation process characteristics. Both a) and b) variant of splitting were compared with the Monte Carlo method in terms of running time, RE and RER estimates. In b) scenario the choice of splitting levels is performed in accordance with the value of the accumulated sum $S_{M,K}$, thus, this case exactly matches the standard splitting procedure, where the average value $\mathbb{E}V$ plays the role of a fixed failure threshold. Numerical results have shown that the a) scenario is more preferable for accelerated simulation of the degradation process with a random

threshold value than b), and besides, software implementation of the a) variant is much easier.

REFERENCES

1. H. Jensen, C. Papadimitriou, Reliability analysis of dynamical systems, Substructure Coupling for Dynamic Analysis (2019) 69–111.
2. A. Shahraki, O. P. Yadav, H. Liao, A review on degradation modelling and its engineering applications, International Journal of Performability Engineering 13 (3) (2017) 299–314.
3. Y. H. Lin, Y. F. Li, E. Zio, A comparison between monte carlo simulation and finite-volume scheme for reliability assessment of multi-state physics systems, Reliability Engineering & System Safety 174 (2018) 1–11.
4. H. Gao, L. Cui, D. Kong, Reliability analysis for a wiener degradation process model under changing failure thresholds, Reliability Engineering & System Safety 171 (2018) 1–8.
5. Q. Dong, L. Cui, A study on stochastic degradation process models under different types of failure thresholds, Reliability Engineering & System Safety 181 (2019) 202–212.
6. N. Yousefi, D. W. Coit, S. Song, Q. Feng, Optimization of on-condition thresholds for a system of degrading components with competing dependent failure processes, Reliability Engineering & System Safety 192 (2019) 106547.
7. Q. Sun, Z.-S. Ye, X. Zhu, Managing component degradation in series systems for balancing degradation through reallocation and maintenance, IISE Transactions 52 (7) (2020) 797–810.
8. X. Wang, P. Jiang, B. Guo, Z. Cheng, Real-time reliability evaluation for an individual product based on change-point gamma and wiener process, Quality and Reliability Engineering International 30 (4) (2014) 513–525.
9. D. Kong, N. Balakrishnan, L. Cui, Two-phase degradation process model with abrupt jump at change point governed by wiener process, IEEE Transactions on Reliability 66 (4) (2017) 1345–1360.
10. A. V. Borodina, D. V. Efrosinin, E. V. Morozov, Application of splitting to failure estimation in controllable degradation system, in: Communications in Computer and Information Science, Vol. 700, Springer International Publishing, 2017, pp. 217–230. doi:10.1007/978-3-319-66836-9.
    URL `http://www.springer.com/gb/book/9783319668352`
11. A. Borodina, V. Tishenko, Simulation of a heterogeneous degradation process in a system with gradual and sudden failures, Proceedings of the Karelian Research Center of the Russian Academy of Sciences 7 (2018. (in Russian)) 3–13. doi:https://doi.org/10.17076/mat836.

12. A. Borodina, O. Lukashenko, E. Morozov, On conditional monte carlo for the failure probability estimation, in: Proceedings of 2018 10th International Congress on Ultra Modern Telecommunications and Control Systems (ICUMT 2018), IEEE, 2018, pp. 202–207.
13. A. Borodina, O. Lukashenko, E. Morozov, A rare-event estimation of heterogeneous degradation process, in: 2019 11th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), IEEE, 2019, pp. 1–6.
14. A. Borodina, O. Lukashenko, E. Morozov, On estimation of rare event probability in degradation process with light-tailed stages, in: Proceedings the 22-th International Conference on Distributed Computer and Communication Networks: Control, Computation, Communications (DCCN'2019), 2019, pp. 477–483.
15. R. Y. Rubinstein, A. Ridder, R. Vaisman, Fast Sequential Monte Carlo Methods for Counting and Optimization, John Wiley & Sons, Inc., New Jersey, 2014.
16. R. Y. Rubinstein, D. P. Kroese, Simulation and the Monte Carlo method., John Wiley & Sons, Inc., New Jersey, 2017.
17. P. W. Glynn, D. L. Iglehart, Conditions for the applicability of the regenerative method, Management Science 39 (9) (1993) 1108–1111.

UDC: 004.733

Models and methods of usage of the heterogeneous gateways in the mesh LPWAN networks

V.A. Kulik[1], V.D. Pham[1], R.V. Kirichek[1,2]

[1]Bonch-Bruevich Saint-Petersburg State University of Telecommunications, 193232, Saint-Petersburg, Russia

[2]V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, 117997, Mosscow, Russia

vslav.kulik@gmail.com, fam.vd@spbgut.ru, kirichek@sut.ru

Abstract

This article considers a question for constructing LPWAN mesh networks. Authors in this article were investigated existing standards of the building gateways in the LPWAN networks, creating the structure of the mesh LPWAN with the special conversion structure - heterogeneous gateway. This structure was used to create a simulation model of this network. The simulation model was mapped with a model based on the more traditional star topology with a simple gateway and edge server. The results of the simulation can be used to design new mesh LPWAN networks.

Keywords: Heterogeneous gateways, Internet of Things, mesh networks, time analysis, simulation, LPWAN

1. Introduction

Modern society is undergoing rapid changes due to the occurrence and implementation of modern technologies in everyday life. These technologies have both positive and negative impacts on urban infrastructure and society. To minimize the impact of negative aspects, it is proposed to use special tools for monitoring and managing urban infrastructure based on the Internet of things (IoT) technologies. These technologies together with artificial intelligence, unmanned car/air transport management, environmental control, and emergency response systems are part of the concept of urban space organization — smart city (SC) [1, 2, 3].

When smart city technologies were implemented in the urban infrastructure, a number of problems related to ensuring connectivity of the system's endpoints

The publication has been prepared with the support of the grant from the President of the Russian Federation for state support of leading scientific schools of the Russian Federation according to the research project НШ-2604.2020.9.

arises. One of the main problems is providing a high data transmission distance in urban conditions. Modern mobile networks are not well suited for transmitting small amounts of data from low-performance computing devices with sensors/actuators connected to them, due to the high level of the energy consumption of cellular modules and a large amount of service traffic transmitted over these networks. To connect this type of device to the SC network infrastructure – low-power wide-area networks (LPWAN) are used [4, 5]. The range of data transmission in LPWAN networks varies from 1 to 10 km in urban conditions, depending on the power of the transceiver and the type of power supply of the end node (EN). According to the characteristics of the LPWAN switch, up to 50 thousand nodes can be working in each subnet, simultaneously transmitting data to the switch at a speed from 0.3 to 30 Kbit/s [6]. These technical characteristics allow us to implement a network with a high density of EN placement.

In most cases, LPWAN systems are based on the star network topology. However, in cases when it is not possible to provide a guaranteed high-quality connection EN with the network switch, the mesh network topology is used [7]. This topology allows organizing a network connection not only between the EN and the switch but also between the EN itself. This solution theoretically allows increasing the number of devices in the network per hub and increasing the reliability of data transmission in the network with the presence of multiple routes for data delivery to the destination.

In most cases, ENs transmit data to devices that are located outside the local LPWAN network. To solve these problems, devices called LPWAN gateways are used. These devices are necessary for ensuring interaction between EN and devices located in the external network. In most existing systems, the gateway and switch functions are combined into a single device, and the incoming data are processed on remote servers. This network structure is acceptable if the system doesn't have strict requirements for system response time, which are achieved by reducing the time of delivery and data processing. This problem is most often solved using two approaches: organizing a local edge or fog server in a local network, or by extending the functionality of the LPWAN network gateway. A gateway with extended functionality is called a heterogeneous gateway (HG) [8, 9]. This gateway allows us to dynamically add new software (SW) for data processing, depending on the requirements. Currently, the application of HG within LPWAN networks is a new unconventional task. In this paper, the authors conduct research on models and methods for ensuring interaction between LPWAN networks and other communication networks. Based on the research, a simulation model of the LPWAN mesh network with heterogeneous gateways is proposed, which was compared with a network with a more traditional star topology and without HG.

2. Structure of the LPWAN networks

LPWAN networks consist of the following elements [5]:

- end node – the low-performance device that is used to interact with sensors and/or actuators connected to them;
- switch – the device that acts as routers in the LPWAN networks;
- gateway – the device that receives and extracts useful data from packets of the LPWAN network format and then encapsulates them and sends it to the target network (usually the IP).

In its classic form, this network supports a star topology, where multiple nodes are connected to a single Central device that serves as a gateway and a switch. A group of such devices can be combined using external network technologies, such as IP, into a single network that allows them to interact with each other.

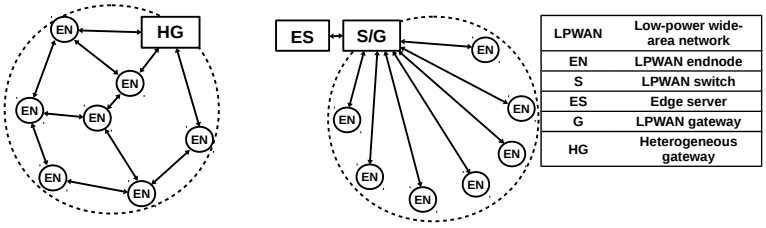
However, there are already existing such solutions that support the mesh network topology. In these networks, users can interact not only with the switch/gateway but also with each other. The function of dynamic mesh routing is experimental in WAN networks due to the high level of requirements for the service traffic bandwidth of the communication channel. LPWAN systems with pre-configured routing tables for each EN are a more common solution. Such tables can be generated manually by developers or network administrators, or by a single switch in the local network before the system running.

LPWAN systems are more often part of more complex network infrastructure, such as smart city systems [10]. Within these systems, LPWAN devices interact with both remote and local SC platforms, which are computer appliance (CA) for receiving, storing, and analyzing data coming from LPWAN systems. Based on the data analysis results, the CA either independently decides on the further functioning of the SC system, or if the situation is sufficiently critical, transmit information to the SC system operator.

LPWAN structure can be simplified and reduced the cost of implementing a number of devices, by using special devices that combine the functions of the switch, gateway, and SC edge platform that called the heterogeneous gateways which defined in ITU-T recommendations Q. 4060 "The structure of the testing of heterogeneous Internet of Things gateways in a laboratory environment" [11] and Q. 3055 "Signaling protocol for heterogeneous Internet of Things gateways" [12].

The HG can be used in the LPWAN networks only if there is a network interface that supports network information exchange technology and implements all the functions typical for the switch, gateway, and edge platform. All the described functions can be implemented using a semantic gateway, which works in the user's virtual workspace. Fig. 1a shows the structure of the LPWAN network, which includes a heterogeneous gateway for routing, edge processing, and further sending network

packets from LPWAN devices to the target network, and Fig. 1b shows the LPWAN network with the star topology and without using the HG.



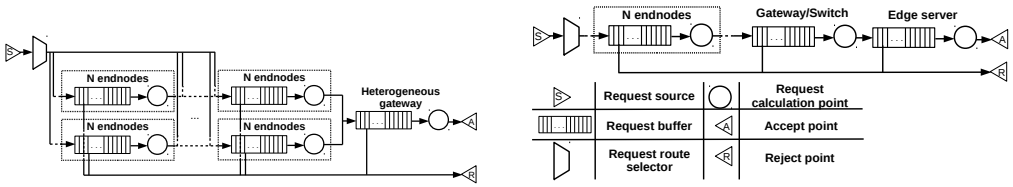
(a) (b)

Figure 1. LPWAN LAN Structure: a) using HG; b) without using HG

The advantages and disadvantages of the proposed LPWAN network topology using heterogeneous gateways should be determined. It is proposed to compare the operation of this network with the operation of a traditional LPWAN network with a star topology by the simulation models.

3. Analytical and simulation models

Based on the proposed structure, it is possible to describe the operation of a simulation model that can be used to study the properties of the proposed network, according to various parameters (for example, message service time, the amount of transmitted useful data per unit of time, the power consumption of devices in the network, etc.) and the model to which the model will be compared, including the heterogeneous gateway. These models are shown in Fig. 2a and 2b.

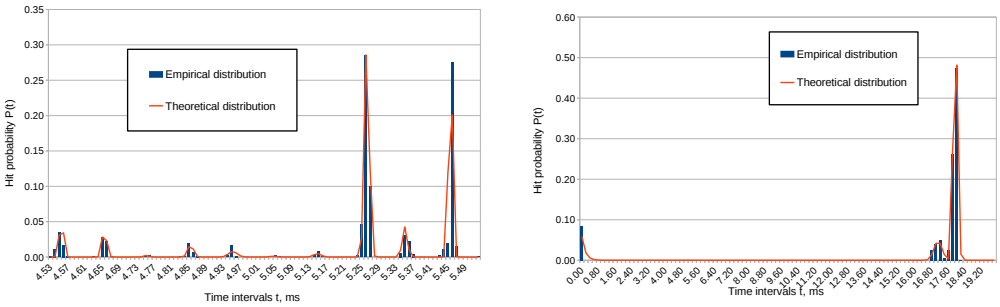


(a) (b)

Figure 2. Structure of the simulation model: a) using HG; b) without HG

It is necessary to take into account the features of LPWAN devices functioning to develop this model. The model network was developed, consisting of the LoRaWAN –

CubeCell Dev-Board end node and the Dragino LG02 gateway for this purpose. These devices were used to measure the intensity of message receipt without using software delay for the EN and the intensity of message service at the gateway. As a result, based on the method of least squares, the method of generalized reduced gradients, and the Kolmogorov-Smirnov test, analytical models describing the intensity of receiving and servicing messages for LoRaWAN devices were obtained. Graphs showing the ratio of the obtained analytical models to the original empirical distributions are shown in the Fig. 3a and 3b.



(a) (b)

Figure 3. The ratio of the empirical distribution to the theoretical distribution for: a) the network interface latency; b) the message service intensity

The intensity of the message service for the EN (2) and the server (4), and delay rates for the network interface (1) and for communication channel (3), can be described using the following probability density expressions:

$$f_a(x) = p_1\lambda_1 e^{\lambda_1 x} + p_2\frac{\lambda_2^{a_2}}{\Gamma(a_2)}xe^{-\lambda_2 x} + p_3\frac{\lambda_3^{a_3}}{\Gamma(a_3)}xe^{-\lambda_3 x} \quad (1)$$

$$f_c(x) = G_c(x, p_1, a_1, \lambda_1, c_1) + G_c(x, p_2, a_2, \lambda_2, c_2) + ... + G_c(x, p_{10}, a_{10}, \lambda_{10}, c_{10}) \quad (2)$$

$$f_{cd}(x) = p\frac{\lambda}{\Gamma(a)}xe^{-\lambda x} \quad (3)$$

$$f_s(x) = p\lambda e^{\lambda x} \quad (4)$$

$$G_c(x, p, a, \lambda, c) = p\frac{\lambda^a}{\Gamma(a)}(x-c)e^{-\lambda(x-c)} \quad (5)$$

where $f_a(x)$ – probability density of message arrival intensity, $f_c(x)$ – message service intensity, $G_c(x, p, a, \lambda, c)$ – bias Gamma probability density distribution, p_i - probability of falling into a given distribution, a_i – scale coefficient, λ_i – form coefficient, c_i – displacement coefficient.

These analytical models, together with the probabilistic distributions obtained during the experiment for the time intervals between message arriving, the latency indicators of the LoRaWAN network interface, the service time of messages on the EN, the latency on the communication channel and the service time of messages on the server were used to model a network of 1000 LoRaWAN end-nodes, according to Fig. 1a and 2a for the network using HG and Fig. 1b and 2b for the network without HG.

4. The results of the simulation

The obtained experimental network's probability distributions allow us to simulate the work of the previously described networks (Fig. 1a and Fig. 1b). The probability distributions used in the simulation using the Python library Ciw [13] are presented below:

1) Time intervals between arrival messages for the each end node: deterministic value, 1 message every 60 seconds.
2) Latency on network interface: $f_a(x)$ with the following parameters $p_1 = 0.116$, $\lambda_1 = 5988$, $p_2 = 0.115, a_2 = 8899.84, \lambda_2 = 518135, p_3 = 0.769, a_3 = 48643.63, \lambda_3 = 2695418$ (1).
3) Service time of messages on EN: $f_c(x)$ with the following parameters $p_{1..10} = [0.065, 0.052, 0.003, 0.028, 0.02, 0.002, 0.01, 0.434, 0.062, 0.434]$, $a_{1..10} = [$ 298, 1288, 2317, 3288, 2041, 5208, 6221, 24551, 27478, 93005], $\lambda_{1..10} = [9900990,$ 9900990, 1.0e7, 1.0e7, 4761905, 1.0e7, 1.0e7, 33333333, 33333333, $1.0e8]$, $c_{1..10} = 4.52e^{-3}$ (2).
4) Communication delay: $f_{cd}(x)$ with the following parameters $p = 1.00, a = 22.73, \lambda = 25253$.
5) Server service time: $f_s(x)$ with the following parameters $p = 1.00, \lambda = 100.47$.

These results are based on the study of the end nodes and processing devices for each type of network, which explains some differences in the probability distributions between nodes in these models. The simulation results with a confidence level of 95% are presented in Tab. 1.

According to the results of simulation with using HG, the service time is reduced for the model as a whole and both for the gateway and server. Thus, for a more complex assessment of the need of using a heterogeneous gateway, it is necessary to conduct experiments, both based on the energy consumption of the server, the gateway, and HG, and based on a real model network. Also, a study of emergency

Parameters	Distribution	Average value (ms)
Model without using HG		
Message Arrival Intensity	$Exponential(a = 16.57)$	61.850 ± 1.690
Message service time in the model	$Erlang(a = 3, \lambda = 44.06)$	84.310 ± 1.690
Message service time at the gateway	$Gamma(a = 600, \lambda = 104769.80$	5.710 ± 0.002
Message Service Time on ES	$Exponential(a = 82)$	10.730 ± 0.060
Model with using HG		
Message Arrival Intensity	$Exponential(a = 13.78)$	61.62 ± 0.55
Message service time in the model	$Erlang(a = 7, \lambda = 265.86)$	27.40 ± 0.07
Message service time at the HG	$Erlang(a = 4, \lambda = 307.61)$	15.52 ± 0.06

Table 1. Simulation Results

cases in which sending messages by each end node should be much more intensive than one message per second did not conduct in this paper. The authors assume that in this case, the network topology with a heterogeneous gateway will not have significant advantages, since both the traditional model (S/G) and the proposed model (HG) will have the narrowest and most critical place is the node providing access to the external network. But this issue may have unexpected results at the level of the radio communication channel and requires further researches.

5. Conclusion

In this paper, we have analyzed the model and methods for ensuring the interaction of LPWANs and other communication networks. Based on this analysis, we have proposed the structure and simulation model of the mesh LPWAN network using heterogeneous gateways. With the presented network structure, simulation series were carried out and the results of the mesh network using HG were compared with the network with the star topology and without using HG.

Looking into further study, we intend to develop a model mesh network based on the developed structure, including HG and dozens of ENs. The functional results of this model will be compared with the properties of the network with the star topology and without using HG. It is also proposed to conduct a study of the energy consumption of LPWAN network elements in this network: gateway, edge server, EN, and HG.

REFERENCES

1. Abu-Matar M., Mizouni R. Variability Modeling for Smart City Reference Architectures // 2018 IEEE International Smart Cities Conference (ISC2). 2018. P. 1-8. DOI: 10.1109/ISC2.2018.8656967.

2. Tolcha Y. K., Nguyen H. M., Byun J., Kwon K., Han J. et al. Oliot-OpenCity: Open Standard Interoperable Smart City Platform // 2018 IEEE International Smart Cities Conference (ISC2). 2018. P. 1-8. DOI: 10.1109/ISC2.2018.8656763.
3. Smart City Network Architecture Guide. Alcatel-Lucent. 2019. PP. 39. URL: https://www.al-enterprise.com/-/media/assets/internet/documents/smart-city-network-architecture-guide-en.pdf.
4. Sanchez-Iborra R., Cano M. D. State of the Art in LP-WAN Solutions for Industrial IoT Services // Sensors. V. 16(5), 708. P.1-14. DOI: 10.3390/s16050708.
5. LoRaWAN Specification V. 1.0.2. LoRa Alliance, Inc. PP. 70. URL: https://lora-alliance.org/sites/default/files/2018-05/lorawan1_0_2-20161012_1398_1.pdf.
6. SX1276/77/78/79 LoRa modules datasheet. Semtech Corporation. 2019. PP. 132. URL: https://semtech.my.salesforce.com/sfc/p/#E0000000JelG/a/2R0000001OKs/Bs97dmPXeatnbdoJNVMIDaKDlQz8q1N_gxDcgqi7g2o.
7. Pham V. D., Dinh T. D., Kirichek R. V. Method for organizing mesh topology based on LoRa technology // 2018 10th International congress on ultra modern tellecommunications and control systems and workshops (ICUMT). 2018. P. 1-6. DOI: 10.1109/ICUMT.2018.8631270.
8. Kulik V. A., Kirichek R. V. The heterogeneous gateways in the Industrial Internet of Things // 2018 10th International congress on ultra modern tellecommunications and control systems and workshops (ICUMT). 2018. P. 1-5. DOI: 10.1109/ICUMT.2018.8631232.
9. Kulik V., Kirichek R. Sotnikov A. Industrial Internet of Things classification and analysis performed on a model network // Internet of Things, smart spaces, and next generation networks and systems, Springer Verlag. 2019. P. 548-561. DOI: 10.1007/978-3-030-30859-9_48.
10. Ferreira C. M. S., Oliveira R. A. R., Silva J. S. Low-Energy Smart Cities Network with LoRa and Bluetooth // 2019 7th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud). 2019. P. 24-29. DOI: 10.1109/MobileCloud.2019.00011.
11. Q.4060 The structure of the testing of heterogeneous Internet of things gateways in a laboratory environment. ITU-T. 2018. URL: https://www.itu.int/rec/T-REC-Q.4060-201810-I.
12. Q.3055 Signalling protocol for heterogeneous Internet of things gateways. ITU-T. 2019. PP. 29. URL: https://www.itu.int/rec/T-REC-Q.3055-201912-I.
13. Palmer G., Knight V., Harper P., Hawa A. Ciw: An open-source discrete event simulation library // Journal of Simulation. 2018. P. 68-82. DOI: 10.1080/17477778.2018.1473909

УДК: 004.733

Применение гетерогенных шлюзов в ячеистых сетях LPWAN

В.А. Кулик[1], В.Д. Фам[1], Р.В. Киричек[1,2]

[1]Санкт-Петербургский Государственный Университет Телекоммуникаций им. проф. М.А. Бонч-Бруевича, 193232, Санкт-Петербург, Россия

[2]Институт Проблем Управления им. В.А. Трапезникова Российской Академии Наук, 117997, Москва, Россия

vslav.kulik@gmail.com, fam.vd@spbgut.ru, kirichek@sut.ru

Аннотация

В данной статье рассматриваются существующие стандарты построения гетерогенных шлюзов Интернета вещей, исследуются вопросы функционирования и взаимодействия гетерогенных шлюзов в рамках ячеистых сетей LPWAN. На основе проведённого анализа предлагается структура сети LPWAN, включающая в себя гетерогенные шлюзы Интернета вещей. На базе представленной структуры сети была описана структура имитационной модели, которую предлагается использовать для имитационного моделирования в будущих исследованиях.

Ключевые слова: Гетерогенные шлюзы, Интернет вещей, ячеистые сети, heterogeneous gateway, Internet of Things, mesh networks, LPWAN

1. Введение

Современное урбанизированное общество претерпевает стремительные изменения в связи с появлением и внедрением современных технологий в повседневную жизнь. Данные технологии оказывают как позитивное, так и негативное влияние на городскую инфраструктуру и общество. Для минимизации влияния негативных аспектов предлагается использовать специальные средства мониторинга и управления городской инфраструктурой, основанные на технологиях Интернета вещей (ИВ). Данные технологии в совокупности с системами искусственного интеллекта, управления беспилотным авто и авиа транспортом, контроля окружающей среды и реакции на чрезвычайные ситуации входят в

Исследование выполнено при финансовой поддержке гранта Президента Российской Федерации для государственной поддержки ведущих научных школ Российской Федерации в рамках научного проекта НШ-2604.2020.9.

концепцию организации городского пространства – умный город (УГ, Smart City – SC) [1, 2, 3].

При внедрении технологий умного города в состав городской инфраструктуры возникает ряд проблем связанных с обеспечением связности оконечных узлов данной системы. Одной из основных проблем является обеспечение высокой дальности передачи данных в городских условиях. Современные сети мобильной связи плохо подходят для задач передачи малых объёмов данных от низкопроизводительных вычислительных устройств, с подключёнными к ним датчиками/актуаторами, вследствие высокого уровня потребления модулей мобильной связи и большого объёма служебного трафика передаваемого по данным сетям. Для подключения такого типа устройств к сетевой инфраструктуре УГ и называемых оконечными узлами УГ, используются энергоэффективные сети с дальним радиусом действия – LPWAN [4]. Дальность передачи данных в сетях LPWAN варьируется от 1 до 10 км в городских условиях, в зависимости от мощности приемопередатчика и типа питания оконечного узла (ОУ). Согласно характеристикам концентратов узлов LPWAN, в каждой подсети может функционировать до 50 тыс. ОУ, одновременно передающих данные концентратору со скоростью от 0,3 до 37,5 Кбит/с [5, 6]. Приведённые технические характеристики позволяют реализовать на базе данных систем сеть с высокой плотностью размещения ОУ.

В большинстве случаев системы LPWAN реализуется на базе сетевой топологии – звёздная [5]. Тем не менее в случаях когда невозможно обеспечить гарантированно качественную связь ОУ с концентратором сети используются другая распространённая сетевая топология – ячеистая. Данная топология позволяет организовать сетевое соединение не только между ОУ и концентратором, но и ОУ между собой. Данное решение теоретически позволяет увеличить количество устройств в сети на один концентратор и увеличивает надёжность передачи данных в сети, вследствие наличия множества маршрутов доставки данных до цели назначения [7].

В большинстве случаев ОУ передают данные устройствам находящимся за пределами локальной сети LPWAN. Для решения данных задач используются устройства называемыми шлюзами LPWAN, необходимыми для обеспечения взаимодействия ОУ и устройств расположенных во внешней сети. В большинстве существующих систем функции шлюза и концентратора объединены в одно устройство, а обработка данных производится на удалённых серверах. Данная структура сети является приемлемой в том случае если к данной системе не предъявляются строгие требования к времени реакции системы, которые достигаются с помощью сокращения времени доставки и обработки данных. Данная проблема наиболее часто решается с помощью двух подходов: организации ло-

кального граничного или туманного сервера в локальной сети или с помощью расширения функциональности шлюза сети LPWAN. Шлюз с расширенной функциональностью называется гетерогенный шлюз (ГШ) [8, 9]. Данный шлюз позволяет динамически, в зависимости от требований, добавлять новое программное обеспечение (ПО) для обработки данных от ОУ в рамках локальной сети. В настоящее время применение ГШ в рамках сетей LPWAN является новой актуальной задачей. В рамках данной работы авторы проводят исследование моделей и методов обеспечения взаимодействия сетей LPWAN и других коммуникационных сетей. На основе проведённого исследования предлагается имитационная модель ячеистой сети LPWAN, с применением гетерогенных шлюзов, которая может быть использована для исследования свойств ячеистых сетей УГ.

2. Структура сетей LPWAN

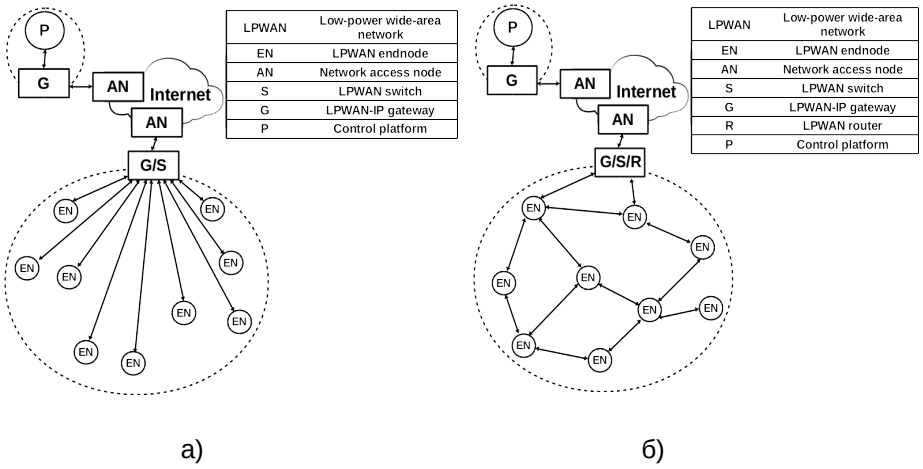
Сети LPWAN состоят из следующих элементов[6]:

- оконечные узлы (endnode – EN) – низкопроизводительные вычислительные устройства, используемые для взаимодействия с подключёнными к ним датчикам и/или актуаторам;
- концентраторы (switch – S) – устройства, выполняющие роль маршрутизатора в данных сетях;
- шлюз (gateway – G) – устройства выполняющие приём и извлечение полезных данных из пакетов формата сети LPWAN и их дальнейшую инкапсуляцию и отправку в целевую сеть (чаще всего сети Интернет).

В классическом виде данная сеть поддерживает звёздную топологию, где множество ОУ подключены к одному центральному устройству (ЦУ), выполняющему функции шлюза и концентратора. Группа ЦУ может быть объединена с помощью внешних сетевых технологий, например IP, в единую сеть, позволяющую им взаимодействовать друг с другом. Структура данной сети отображена на Рисунке 1а.

Тем не менее уже существуют решения поддерживающие ячеистую сетевую топологию, в рамках которой ОУ могут взаимодействовать не только с концентратором/шлюзом, но и между собой. Также в рамках данного решения одна сеть LPWAN может включать в себя несколько концентраторов, выполняющих задачи маршрутизации данных от ближайших к себе узлов и взаимодействующих между собой с помощью протокола, используемого в данной локальной сети LPWAN. Тем не менее функция динамической ячеистой маршрутизации в сетях LPWAN носит экспериментальный характер вследствие высокому уровню требований к пропускной способности канала связи из-за большого объёма передаваемого служебного трафика. Более часто встречаются системы LPWAN с заранее сконфигурируемыми таблицами маршрутизации для каждого из ОУ.

Данные таблицы могут быть сгенерированы как вручную разработчиками или администратором сети, так и единственным концентратором в локальной сети. Данная структура сети отображена на Рисунке 1б.



а) б)

Рис. 1. Структура сети LPWAN для: а) топологии звезда; б) ячеистой топологии

Системы LPWAN чаще всего являются частью более сложной сетевой инфраструктуры, например систем умного города [10]. В рамках данных систем устройства LPWAN взаимодействуют как с удалёнными, так и с граничными платформами УГ (SC Platform), фактически являющимися программно-аппаратными комплексами для приёма, хранения и анализа данных, поступающих от систем LPWAN. На основе проведённого анализа данных ПАК, либо самостоятельно принимает решение о дальнейшем функционировании системы УГ, либо если ситуация является достаточно критичной передаёт информацию и рекомендацию по решению данной проблемы оператору системы УГ. Для передачи данных через сети оператора связи (operator's network – ON), предоставляющего инфраструктуру для системы УГ, применяются специальные устройства выполняющие роль точек включения в сеть оператора связи. Структура функционирования сети LPWAN в рамках системы УГ изображена на Рисунке 2.

Для упрощения структуры сети LPWAN и сокращения издержек на внедрение целого ряда устройств предлагается реализовать устройство объединяющее функции концентратора, шлюза и граничной платформы УГ, с помощью ге-

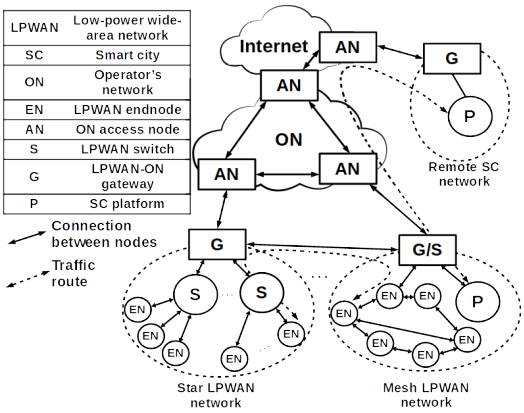


Рис. 2. Структура функционирования сети LPWAN в рамках системы УГ

терогенных шлюзов ИВ, определённых в рекомендациях МСЭ-Т Q.4060 «The structure of the testing of heterogeneous Internet of Things gateways in a laboratory environment»[11] и Q.3055 «Signalling protocol for heterogeneous Internet of things gateways»[12].

Гетерогенный шлюз ИВ – это программно-аппаратный комплекс, выполняющий преобразование, как протоколов физического, канального и сетевого уровня, с помощью существующих сетевых интерфейсов, так и протоколов прикладного уровня и форматов полезных данных, с помощью семантического шлюза ИВ [8, 9]. Семантический шлюз (СШ) – это программное обеспечение, функционирующее в рамках ГШ и выполняющее преобразование прикладных протоколов ИВ между собой и форматов полезных данных. ГШ включает в себя многозадачную ОС, которая имеет достаточно вычислительных ресурсов для работы систем виртуализации/эмуляции рабочего пространства пользователя (например, контейнеры в ОС основанных на Linux). Несмотря на некоторое падение производительности ПО используемого в средах виртуализации/эмуляции, в сравнении с запуском на основной ОС, данная структура позволяет динамически добавлять и проводить обновление новых приложений и услуг на данном устройстве. Данная система отображена на рисунках 3а и 3б.

ГШ может использоваться в сетях LPWAN только при условии наличия сетевого интерфейса, поддерживающего технологию сетевого обмена информацией и реализации всех функций характерных для концентратора, шлюза и гранич-

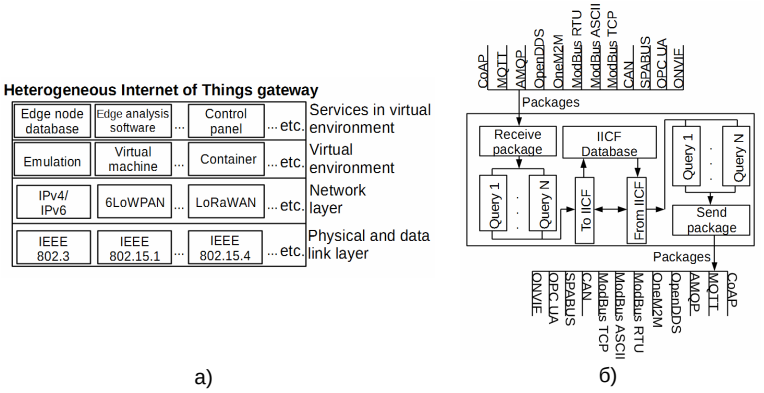


Рис. 3. Структура: а) гетерогенного шлюза ИВ; б) семантического шлюза ИВ

ной платформы. Все описанные функции могут быть реализованы, с помощью ПО, функционирующего в виртуальном рабочем пространстве пользователя. На Рисунке 4 отображена структура сети LPWAN, включающая в себя гетерогенный шлюз для маршрутизации, граничной обработки и дальнейшей отправки в целевую сеть сетевых пакетов от устройств LPWAN.

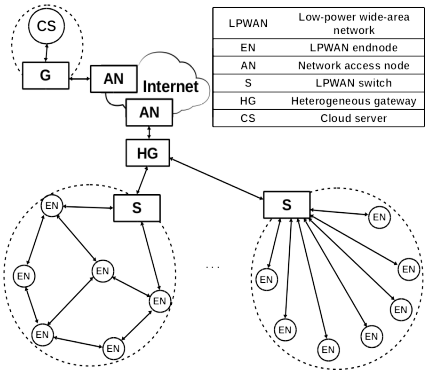


Рис. 4. Структура сети LPWAN с использованием ГШ

На основе предложенной структуры можно описать работу имитационной модели, которая может быть использована для исследования свойств работы

предложенной сети, согласно различным параметрам (например, время обслуживания сообщений, объем передаваемых полезных данных в единицу времени, энергопотребление устройств в сети и др.). Данная модель изображена на Рисунке 5.

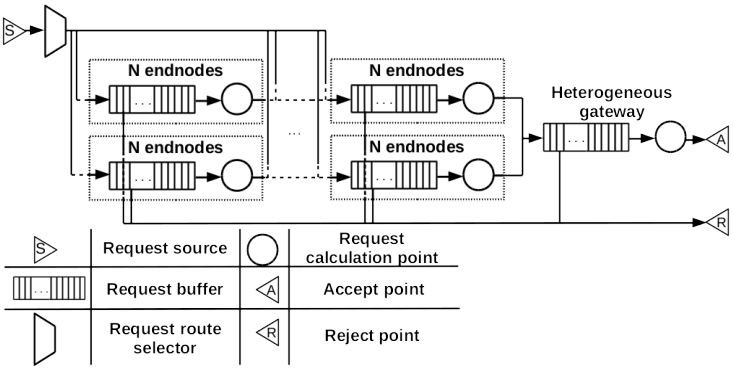


Рис. 5. Структура имитационной модели ячеистой сети LPWAN

3. Заключение

В рамках данной работы был проведён анализ моделей и методов обеспечения взаимодействия сетей LPWAN и других коммуникационных сетей. На основе данного анализа предлагается структура и имитационная модель ячеистой сети LPWAN, с применением гетерогенных шлюзов, которая может быть использована для исследования свойств ячеистых сетей УГ.

В будущих исследованиях на основе разработанной структуры будет разработана модельная сети и на её основе будет проведено измерение параметров оконечных устройств и шлюзов LPWAN. Полученные параметры предлагается использовать в рамках имитационной модели для исследования свойств гетерогенного шлюза, при его использовании в ячеистых сетях LPWAN, по сравнению с более традиционными шлюзами LPWAN.

Литература

1. Abu-Matar M., Mizouni R. Variability Modeling for Smart City Reference Architectures // 2018 IEEE International Smart Cities Conference (ISC2). 2018. P. 1-8. DOI: 10.1109/ISC2.2018.8656967.

2. Tolcha Y. K., Nguyen H. M., Byun J., Kwon K., Han J. et al. Oliot-OpenCity: Open Standard Interoperable Smart City Platform // 2018 IEEE International Smart Cities Conference (ISC2). 2018. P. 1-8. DOI: 10.1109/ISC2.2018.8656763.
3. Smart City Network Architecture Guide. Alcatel-Lucent. 2019. PP. 39. URL: https://www.al-enterprise.com/-/media/assets/internet/documents/smart-city-network-architecture-guide-en.pdf.
4. Sanchez-Iborra R., Cano M. D. State of the Art in LP-WAN Solutions for Industrial IoT Services // Sensors. V. 16(5), 708. P.1-14. DOI: 10.3390/s16050708.
5. LoRaWAN Specification V. 1.0.2. LoRa Alliance, Inc. PP. 70. URL: https://lora-alliance.org/sites/default/files/2018-05/lorawan1_0_2-20161012_1398_1.pdf.
6. SX1276/77/78/79 LoRa modules datasheet. Semtech Corporation. 2019. PP. 132. URL: https://semtech.my.salesforce.com/sfc/p/#E0000000JelG/a/2R0000001OKs/Bs97dmPXeatnbdoJNVMIDaKDlQz8q1N_gxDcgqi7g2o.
7. Pham V. D., Dinh T. D., Kirichek R. V. Method for organizing mesh topology based on LoRa technology // 2018 10th International congress on ultra modern tellecommunications and control systems and workshops (ICUMT). 2018. P. 1-6. DOI: 10.1109/ICUMT.2018.8631270.
8. Kulik V. A., Kirichek R. V. The heterogeneous gateways in the Industrial Internet of Things // 2018 10th International congress on ultra modern tellecommunications and control systems and workshops (ICUMT). 2018. P. 1-5. DOI: 10.1109/ICUMT.2018.8631232.
9. Kulik V., Kirichek R. Sotnikov A. Industrial Internet of Things classification and analysis performed on a model network // Internet of Things, smart spaces, and next generation networks and systems, Springer Verlag. 2019. P. 548-561. DOI: 10.1007/978-3-030-30859-9_48.
10. Ferreira C. M. S., Oliveira R. A. R., Silva J. S. Low-Energy Smart Cities Network with LoRa and Bluetooth // 2019 7th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud). 2019. P. 24-29. DOI: 10.1109/MobileCloud.2019.00011.
11. Q.4060 The structure of the testing of heterogeneous Internet of things gateways in a laboratory environment. ITU-T. 2018. URL: https://www.itu.int/rec/T-REC-Q.4060-201810-I.
12. Q.3055 Signalling protocol for heterogeneous Internet of things gateways. ITU-T. 2019. PP. 29. URL: https://www.itu.int/rec/T-REC-Q.3055-201912-I.

УДК: 004.72

Исследование протоколов маршрутизации для ячеистой сети дальнего радиуса действия

В.Д. Фам[1], Д.Т. Ле[2], Р.В. Киричек[1,3]

[1]Санкт-Петербургский Государственный Университет Телекоммуникаций им. проф. М.А. Бонч-Бруевича, 193232, Санкт-Петербург, Россия
[2]Университет Дананга – Университет Науки и Технологий, Дананг, Вьетнам
[3]Институт Проблем Управления им. В.А. Трапезникова Российской Академии Наук, 117997, Москва, Россия

fam.vd@spbgut.ru, letranduc@dut.udn.vn, kirichek@sut.ru

Аннотация

Данная статья рассматривает протоколы маршрутизации, используемые для организации ячеистой сети дальнего радиуса действия на основе технологии LoRa (Long-Range). Технология LoRa, которая появилась с преимуществами передачи данных на большие расстояния и малого энергопотребления, используется в приложениях Интернета вещей. На базе этой технологии было предложена модель ячеистой сети, которая предоставляет коммуникацию между фрагментами сенсорных сетей ближнего радиуса действия. При помощи разработанной имитационной модели, серии компьютерных экспериментов проведены с различным количеством узлов по различным размерам масштаба сети. По результатам моделирования проанализированы распределения задержек и коэффициенты доставки пакетов в данной сети.

Ключевые слова: Интернет вещей, LoRa, ячеистая сеть, AODV, DSDV, протокол маршрутизации, задержка, коэффициент доставки

1. Введение

В настоящее время, многие приложения были разработаны на основе коммуникационных технологий. Эти технологии играют важную роль в предоставлении связи между вещами, устройствами и людьми. В частности, Интернет вещей (ИВ) определяется для подключения вещей и их взаимодействия в единой сети. Парадигма ИВ привела появления других типов сетей связи, и также технологий

Исследование выполнено при финансовой поддержке Президента Российской Федерации для государственной поддержки ведущих научных школ Российской Федерации в рамках научного проекта НШ-2604.2020.9.

коммуникации [1, 2, 3]. В работе с большинством сенсорных устройств, приложения ИВ часто требуют малое энергопотребление и способность передачи данных на большие расстояния. Технология LoRa от корпорации Semtech считается одной из особенных технологии, предложенных для построения эффективной сети дальнего радиуса действия [4, 5]. С другой стороны некоторые технологии, такие как IEEE 802.11, IEEE 802.15.4 или BLE (Bluetooth Low Energy), также были разработаны для приложений ИВ. Однако эти технологии имеют невысокую дальность связи. С целями концентрации на способности передачи на большие расстояния и экономии энергии, мы рассматриваем технологию LoRa как один вариант для организации сенсорных сетей.

LoRa – одна из технологий входящих в группы энергоэффективной сетей дальнего радиуса действия (LPWAN). Таким образом, возникает возможность использование ячеистой сети дальнего радиуса действия как ячеистая сеть LoRa, которая организуют частную сеть для обеспечения связи между кластерами сетей малого радиуса действия или для устройств LoRa, которые находится далеко от базовой станции [6]. Однако технология LoRa обладает низкой скоростью передачи данных и другими характеристиками на физическом уровне. Параметры, такие как коэффициент распределения SF (Spreading Factor), ширина полосы пропускания BW (Bandwidth) и скорость кодирования CR (Coding Rate), сконфигурированы для радиомодулей LoRa в зависимости от выбора скорости передачи данных и чувствительности приема [5]. При разработке протоколов маршрутизации будет необходимо рассмотреть некоторые изменения для адаптации ячеистой сети LoRa. В данной работе, мы рассматриваем более известные протоколы маршрутизации для ячеистой сети LoRa на базе имитационного моделирования. На основе протоколов AODV и DSDV, узлы сети могут найти маршруты для переадрасации сообщений к узлу назначения.

2. Маршрутизация в ячеистой сети LoRa

Проектирование и разработка протоколов маршрутизации в беспроводных ячеистых сетях являются важной задачей, которая должна охватывать несколько показателей производительности, таких как минимальное количество переходов, предотвращение нарушения методов обслуживания в соответствия с концепцией устойчивости. Использование ячеистой инфраструктуры для максимального эффективного выполнения процесса маршрутизации и повышение масштабируемости протоколов маршрутизации для установки или поддержки маршрутов в ячеистой сети с большой емкостью.

Технология LoRa появилась с целями передачи данных на большое расстояние и экономии энергопотребления. При реализации беспроводных ячеистых сетей на базе технологии LoRa, зона покрытия может быть расширена для сенсорных

сетей, которые обычно работают в коротким расстоянии с многим количеством узлов. Чтобы реализовать идею построения модели ячеистой сети LoRa, нам нужно рассмотреть и выбрать соответствующий протокол маршрутизации. Хотя технология LoRa имеет преимущества по дальности передачи данных, она также имеет низкую скорость передачи данных. Поэтому время может заниматься больше для настройки или поиска маршрутов. Таким образом, исследование протоколов маршрутизации представлено в данной работе с помощью имитационного моделирования.

Типично, протоколы маршрутизации делятся на три категории в зависимости от информации о топологии сети, используемой для построения маршрутов: проактивный, реактивный и гибридный.

Проактивные протоколы маршрутизации, по которым информации об изменении топологии периодически обмениваются между всем узлами сети [7]. На основании этих служебной информации предварительная таблица маршрутизации определена для каждого узла. В памяти каждого узла сохраняется таблица маршрутизации, которая будет обновлена при изменении топологии сети. В качестве примеров можно привести некоторые протоколы, такие как DSDV (Destination Sequenced Distance Vector), OLSR (Optimized Link State Routing).

Реактивные протоколы маршрутизации, которые выбирают или ищут маршруты к другим узлам только тогда, когда они необходимы. Процесс поиска маршрута выполняется, когда узел хочет установить связь с другим узлом, для которого у него нет информации маршрутов в таблице маршрутизации [7, 8]. Такие протоколы отличаются от проактивных протоколов тем, что маршрут между двумя узлами определяется только по запросам передачи данных. Кроме того, время сохранения маршрутов ограничено в памяти, поэтому придется ожидание для поиска маршрута когда истекло время сохранения маршрутов. Примерами таких протоколов являются AODV (Ad-hoc On-Demand Distance Vector), DSR (Dynamic Source Routing).

Гибридные протоколы маршрутизации, которые объединяют проактивные и реактивные протоколы для снижения накладных расходов и задержек маршрутизации из-за процесса поиска маршрута. Преимуществами этих протоколов являются более высокая эффективность и масштабируемость. Однако недостатком является высокая задержка при поиске новых маршрутов.

3. Имитационная модель

В типичных приложениях ИВ, сети устройства могут делиться на несколько фрагментов от конечных устройств до удаленных облачных серверов. Коммуникационная технология обеспечивает сети связи между устройствами. В беспроводных сенсорных сетях можно найти несколько методов внедрения ячеистой

топологии между узлами. Ячеистые сети могут быть построены между конечным устройствами, между головными узлам кластеров, или между шлюзами. Применение технологии LoRa дает возможность передачи данных на большие расстояния, что мы можем использовать ее для обеспечения коммуникации фрагментов сетей малого радиусом действия к внешней сети, например к Интернету (Рис. 1).

На Рис. 1, ячеистая сеть LoRa организована для предоставления связи для других сетей, которые хотят отправить данные сенсоров к облачному серверу. Таким образом, сенсорные данные передаются через ячеистую сеть LoRa в сервер. В данной сети некоторые узлы LoRa имеют доступы к серверу, т.е. остальные узлы считаются транзитными узлами для переадресации пакетов до узла назначения.

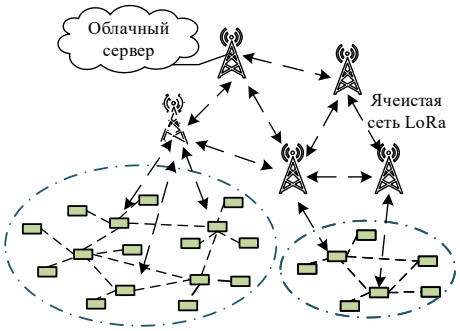


Рис. 1. Модель фрагмента сети

Для организации ячеистой сети протокол маршрутизации является важным элементом для настройки и поиска маршрутов между узлами. В данной работе, при помощи имитационного моделирования мы можем оценить эффективность ячеистой сети LoRa с протоколами маршрутизации AODV и DSDV.

На основании фреймворков OMNET++ и inet была разработана имитационная модель сети узлов LoRa. При этом на физическом уровне параметры, такие как $SF = 8$, $BW = 125$ кГц, $CR = 4/5$, сконфигурированы для радиомодуля LoRa. С данными параметрами узел может принимать радиосигнал до -127 дБм, что имеется высокая чувствительность на приеме [5]. Однако скорость передачи данных является низкой в этом случае. Поэтому интенсивность генерации сообщений рассматривается небольшой чтобы избежать перегрузки сети. Интервал времени между отправками сообщений подчиняется экспоненциальному закону распределения с среднем значением 120 с. Длина полезной нагрузки каждого

сообщения также по равномерному распределению генерируется в интервале от 20 до 150 байтов.

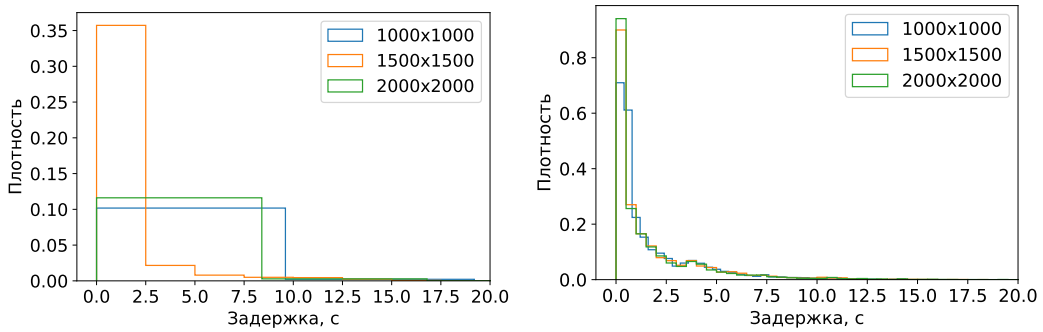
Исследуемая сеть рассматривается на квадратном поле, на котором равномерно распределены позиции узлов. Где один узел приемник является узлом, который находится в центре поля моделирования, а остальные узлы передают сообщения также с интервалом времени по случайному закону к узлу приемника. При рассмотрении протоколов маршрутизации для данной ячеистой сети, мы оцениваем параметры качества обслуживания, такие как распределение задержек и процент доставки пакетов к узлу приемника от остальных узлов. Анализ результатов моделирования сети проводится в двух случаях.

- Сеть с одинаковым количеством узлов развернута на поле различного размера. Компьютерные эксперименты проведены с сетью 25 узлов, распределенных на поле размерами 1000x1000, 1500x1500 и 2000x2000 m^2.
- Сеть имитируется с различным количеством узлов, распределенных на поле одинакового размера. Для этого, серия экспериментов проводится для сети с 16, 25 и 36 узлов на поле размером 2000x2000 m^2.

4. Результаты моделирования

4.1. Анализ по размерам масштаба сети. Рассматривается сеть 25 узлов, в которой один узел приемник принимает сообщения от остальных узлов. Распределения задержек доставки пакетов к узлу приемника изображены на Рис. 2 при использовании протоколов маршрутизации AODV (Рис. 2a), и DSDV (Рис. 2b). Можно видеть, что задержки в ячеистой сети LoRa распределены до нескольких секунд. Однако, размер поля сети не сильно влияет на задержку доставки при использовании протокола DSDV. В этом случае, распределение задержек похоже на экспоненциальное распределение, которое задается для интервала отправки сообщений на узлах источников. При использовании DSDV, таблица маршрутов между отравителями и получателем сохраняется после того, что маршруты найдены после первого запроса. После этого таблица обновлена если есть информации об изменении топологии сети. Поэтому время доставки не занимает много при поиске маршрутов в этом случае. Однако, при использовании AODV таблица маршрутов временно сохраняется на какой-нибудь продолжительности. Задержка доставки в этом случае может быть больше по сравнению с использованием DSDV (Рис. 3), так как время ожидания для поиска маршрутов после того, когда исчезло время сохранения таблицы маршрутизации в памяти.

Однако, вероятность успешной доставки пакетов при использовании AODV является больше, чем при использовании DSDV (Рис. 4a). Поскольку запрос для



(a) Использование AODV (b) Использование DSDV

Рис. 2. Распределение задержек по размерам масштаба сети

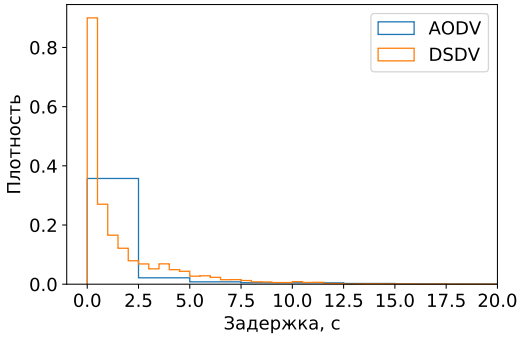
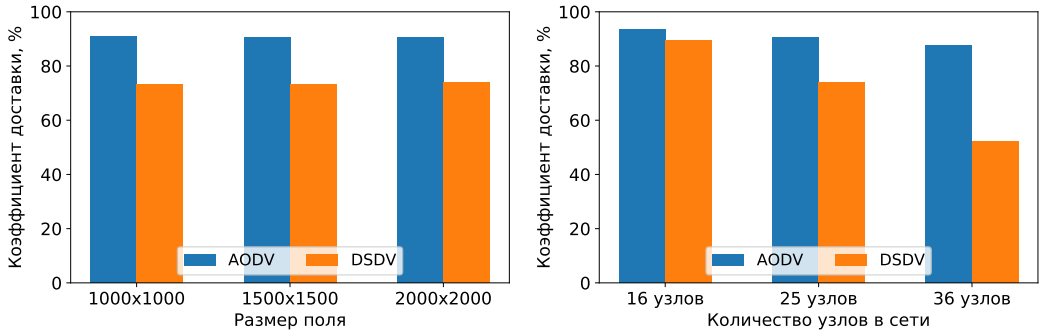


Рис. 3. Сравнение распределения задержек в сети 25 узлов в 1500x1500

поиска маршрутов повторяется при использовании AODV, сообщения передаются транзитным узлами после того когда маршруты найдены. Более того, что размер масштаба сети не влияет на коэффициенты доставки пакетов. Процент полученных пакетов в сети с масштабами 1000x1000 и 2000x2000 приблизительно является одинаковым. Таким образом, с фиксированным количеством узлов в сети мы можем развернуть сети по различным размерам поля. При этом гарантируется неизменение коэффициента доставки пакетов.

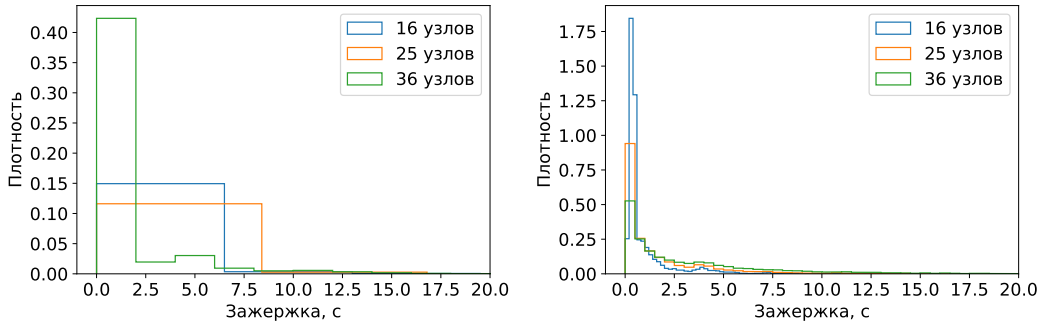
4.2. Анализ по количеству узлов в сети. С другой стороны, в поле сети размером 2000x2000 мы рассматриваем влияние изменения количества узлов на распределения задержек и коэффициент доставки пакетов. Сравнение распределения задержек при использовании протоколов AODV и DSDV в сети с различным количеством узлов представлено на Рис. 5. Очевидно, что увеличение количества узлов в сети изменяет задержки доставки. В том числе использовании



(a) Сети с различными размерами масштаба (b) Сети с различным количество узлов

Рис. 4. Коэффициент доставки пакетов

протокола DSDV, длительная задержка имеет вероятность больше при увеличении количества узлов в сети. Однако при использовании протокола AODV, сеть 36 узлов имеет задержки меньше, чем сети с 16 и 25 узлами. Но сеть 25 узлов имеет задержки больше чем сеть 16 узлов. Это может объясняется тем, что сеть 36 узлов имеет больше вариантов выбора маршрутов между узлами источников и узлом приемника. Поэтому время ожидания переадресации пакетов может быть меньше с сетью 25 узлов в данном размере масштабе.



(a) Использование AODV (b) Использование DSDV

Рис. 5. Распределение задержек с разным количеством узлов в сети

Более того, коэффициент доставки пакетов уменьшается при увеличении количества узлов в сети (Рис. 4b). Коэффициент потери больше увеличивается при использовании протокола DSDV, что в сети 16 узлов коэффициент доставки составляет больше 80 процентов, а в сети 36 узлов меньше 60 процентов сообщений

успешно было доставлено. В то время, коэффициент доставки при использовании AODV также уменьшается, но имеется небольшое значение изменения.

5. Заключение

В данной работе, мы рассмотрели возможности организации ячеистой сети на базе технологии LoRa, которая дает возможность передачи данных на большое расстояние и экономии энергии. При этом энергоэффективная сеть дальнего радиуса действия могут быть развернута в ячеистой топологии. Для получения этой цели, протоколы маршрутизации являются ключевыми элементами для организации сети. Проактивный протокол как DSDV показал малую задержку доставки сообщения и походит для фиксированной сети. С другой стороны, реактивный протокол как AODV показал эффективный коэффициент доставки и подходит для сети, в которой часто имеется изменения узлов. Таким образом для дальнейшего исследования, комбинированный протокол может быть использоваться в соответствии с типом узла в сети.

Литература

1. Кучерявый А. Е. Интернет вещей // Электросвязь. 2013. V. 1. P. 21–24.
2. Росляков А. В., Ваняшин С. В., Гребешков А. Ю., Самсонов М. Ю. Интернет вещей // под ред. А.В. Рослякова. Самара: ПГУТИ, ООО «Издательство Ас Гард». 2014. 340 p.
3. Кучерявый А. Е., Бородин А. С., Киричек Р. В. Сети связи 2030 // Электросвязь. 2018. V. 11. P. 52–56.
4. Kirichek R., Kulik V. Long-range data transmission on flying ubiquitous sensor networks (fusn) by using lpwan protocols // ser. Communications in Computer and Information Science. 2016. V. 678. P. 442–453.
5. Semtech Corporation. Datasheet SX1276/77/78/79 LoRa Transciever. 2019. 132 p.
6. Kirichek R., Vishenevsky V., Pham V. D., Koucheryavy A. Analytic Model of a Mesh Topology based on LoRa Technology // 22th International Conference on Advanced Communications Technology(ICACT). 2019. P. 251–255.
7. Haerri J., Filali F., Bonnet C. Performance comparison of aodv and olsr in vanets urban environments under realistic mobility patterns // in: Proceedings of the 5th IFIP mediterranean ad-hoc networking workshop. 2006. P. 14–17.
8. Kumar J. Comparative performance analysis of aodv, dsr, dymo, olsr and zrp routing protocols in manet using varying pause time // International Journal of Computer Communications and Networks (IJCCN). 2012. V. 3(1). P. 43–51.

УДК: 004.7

Метод определения координат узлов в беспроводной сенсорной сети с ячеистой топологией

В.Д. Фам[1], О.И. Ворожейкина[1], И.В. Гришин[1], Д.В. Окунева[1], Р.В. Киричек[1,2]

[1]Санкт-Петербургский Государственный Университет Телекоммуникаций им. проф. М.А. Бонч-Бруевича, 193232, Санкт-Петербург, Россия

[2]Институт Проблем Управления им. В.А. Трапезникова Российской Академии Наук, 117997, Москва, Россия

fam.vd@spbgut.ru, bonmot@yandex.ru, msp_sut@list.ru, darina_okuneva@mail.ru, kirichek@sut.ru

Аннотация

В статье рассматривается алгоритм определения пространственных координат датчиков в беспроводных сенсорных сетях с ячеистой топологией, состоящих из большого количества сенсорных узлов, на основе неполных данных о расстояниях между узлами.

Ключевые слова: Беспроводные сенсорные сети, сенсорный узел, многомерное шкалирование, мультилатерация, пространственное преобразование, опорные узлы

Назначение беспроводных сенсорных сетей (БСС) состоит в сборе информации о происходящих событиях или получения данных о ряде параметров окружающей среды. Сами сенсорные узлы (СУ) включают в свой состав: датчики, радиомодули и элементы питания, обеспечивающие работу устройства в автономном режиме на протяжении длительного периода времени [1].

В большинстве случаев необходимо точно знать, в какой точке пространства было зарегистрировано событие, или производились измерения, что оказывается возможным только при наличии данных о координатах сенсорных узлов в двухмерном или трехмерном пространстве [2]. В случае беспроводных сетей с ячеистой топологией и количеством узлов размещение узлов на заранее известных позициях оказывается маловероятным. Таким образом, возникает задача

Исследование выполнено при финансовой поддержке гранта Президента Российской Федерации для государственной поддержки ведущих научных школ Российской Федерации в рамках научного проекта НШ-2604.2020.9

определения координат узлов сенсорной сети. При этом системы спутниковой навигации в данном случае не могут быть задействованы, поскольку модуль GPS/ГЛОНАСС потребляет значительное количество энергии, что существенно сокращает срок работы устройства, и также не позволяет обнаруживать устройства, расположенные внутри зданий. Поэтому определение местоположения узлов сети в некоторой выбранной системе координат (СК) осуществляется на основе данных, получаемых из параметров принимаемых радиосигналов. Наиболее часто используются дальномерные методы, которые позволяют получить оценку расстояния между узлами:

$$\hat{d}_{m,n} = f(p) = \sqrt{(x_n - x_m + \epsilon_x)^2 + (y_n - y_m + \epsilon_y)^2 + (z_n - z_m + \epsilon_z)^2} = d_{m,n} + \epsilon_{m,n} \quad (1)$$

где $x_i, y_i, z_i, i = m, n$ – координаты узлов $\mathbf{x}_n = (x_n \quad y_n \quad x_n)^T$, $\epsilon_s, \epsilon_y, \epsilon_z$ – погрешности в оценке координат, $\epsilon_{m,n}$ – погрешности в оценке расстояния, – параметр радиосигнала.

В качестве такого параметра p могут выступать уровень по мощности принимаемого сигнала, зависящий от расстояния между излучающей и принимающей антеннами, или время распространения сигнала.

В случае оценки расстояния методом измерения уровня по мощности принимаемого сигнала RSSI величина $\epsilon_{m,n}$ может оказаться соизмеримой с величиной $d_{m,n}$, поскольку во многих случаях распространение радиосигналов осуществляется в условиях отсутствия прямой видимости (NLOS) и многолучёвости, а также многих других внешних факторах, влияющих на уровень сигнала по мощности.

В случае оценки расстояния как функции времени распространения радиосигнала основную погрешность вносит дрейф и сдвиг внутренних часов беспроводных узлов. В случае односторонней дальнометрии разница во временных сетках беспроводных узлов может приводить к недопустимо большим значениям $\epsilon_{m,n}$, однако методы двунаправленной дальнометрии позволяют минимизировать влияние сдвига и дрейфа часов и добиться высокой точности в оценке расстояния (TWR, SDS-TWR) [3].

Необходимо учитывать, что $d_{m,n} = d_{n,m}$, однако результаты косвенных измерений расстояний в прямом и обратном направлениях содержат погрешности, такие, что в общем случае $\epsilon_{m,n} \neq \epsilon_{n,m} \Rightarrow \hat{d}_{m,n} \neq \hat{d}_{n,m}$. Нарушение симметрии может быть устранено усреднением данных результатов. Тогда $\hat{d}_{m,n} = \hat{d}_{n,m} = d_{m,n} + 0.5(\epsilon_{m,n} + \epsilon_{n,m})$.

Очевидно, что косвенное измерение расстояния возможно только для беспроводных СУ, попадающих в зону радиопокрытия друг друга. Поскольку в большинстве случаев зона обслуживания БСС оказывается много больше зоны

радиопокрытия беспроводного СУ, то в данных о расстояниях между узлами сети будут иметься пропуски, и $\hat{d}_{m,n} \in \hat{D}'$, где $\hat{D}' \subseteq \hat{D}$, $\hat{D}$ - множество из оценок расстояний между различными парами узлов сети. Таким образом, задача сводится к определению координат узлов БСС, на основе неполных данных о расстояниях между узлами.

Расчет пространственных координат сенсорных узлов может осуществляться как в центральном блоке, так и распределенно, что подразумевает обработку данных непосредственно в самих узлах. Также в настоящее время рассматриваются совместные методы позиционирования, позволяющие объединить сильные и устранить слабые стороны централизованного и распределенного методов. В данном случае предполагается, что вычисления координат сенсорных узлов производятся централизованно. Следует также отметить, что рассматриваемой сети имеются опорные узлы с априорно известными координатами в исходной системе координат СК_0.

Решение данной задачи может быть осуществлено в несколько этапов.

1. На первом этапе осуществляется исключение некомплектных объектов, для чего беспроводная сенсорная сеть может быть рассмотрена как неориентированный взвешенный граф $G = \{V, E\}$ в трёхмерном пространстве, где $V = \{\text{У}_n\} \neq \varnothing$ – множество вершин, соответствующих СУ, $E \subset C_v^2$– множество ребер графа, соответствующих наличию или отсутствию соединений между узлами, $C_v^2 = \{(\text{У}_n, \text{У}_m) | \, \text{У}_n, \text{У}_m \in V; m \neq n\}$. Тогда в $G = \{V, E\}$ могут быть найдены клики $G_k = \{V_k, E_k\}$, $V_k \subseteq V$, $E_k \subseteq E$, содержащие полные данные о расстояниях между сети. Поиск клик может быть осуществлен с помощью алгоритма Брона-Кербоша [4].

Для каждой клики могут быть составлены симметрические матрицы весов $\mathbf{D} = \left(\hat{d}_{m,n}\right)$ и квадратов весов $\mathbf{R} = \left(\hat{d}^2_{m,n}\right)$. Задача нахождения пространственных координат вершин клики по M_k–мерной матрице $\mathbf{R}$ может быть решена с помощью метода метрического многомерного шкалирования (ММШ) [5, 6].

Согласно данному методу симметрическая матрица объект-признак $\mathbf{R}$ может быть представлена в виде произведения:

$$\mathbf{R} = \mathbf{U} \cdot \mathbf{U}^T \tag{2}$$

где $\mathbf{U} = (\mathbf{u}\ \mathbf{v}\ \mathbf{w})$–$M_k \times 3$ матрица координат вершин клики G_k в локальной правой прямоугольной системе координат ЛСК_k. Индекс номера клики k здесь и у ряда других матриц будет опущен.

Одновременно с этим матрица $\mathbf{R}$ может быть представлена спектральным разложением вида:

$$\mathbf{R} = \mathbf{V} \cdot \mathbf{\Lambda} \cdot \mathbf{V}^T \tag{3}$$

где $\mathbf{V}$ – ортогональная матрица, столбцы которой представлены собственными векторами $\mathbf{R}$, $\mathbf{\Lambda}$– диагональная матрица собственных значений матрицы $\mathbf{R}$, расположенные в порядке убывания $\lambda_1 \geq \lambda_2 ... \geq \lambda_M$. Откуда $\mathbf{R} = \mathbf{U} \cdot \mathbf{U}^T = \mathbf{V} \cdot \mathbf{\Lambda} \cdot \mathbf{V}^T$, и $\mathbf{U} = \mathbf{V} \cdot (\mathbf{\Lambda})^{1/2}$. Поскольку след матрицы $tr\mathbf{R} = 0$, то на главной диагонали матрицы собственных значений будет присутствовать хотя бы один отрицательный элемент, что будет приводить к комплексным значениям элементов матрицы координат вершин клики $\mathbf{U}$. Центрирование матрицы $\mathbf{R}$ по строкам и столбцам устраняет данную проблему:

$$\mathbf{C} \cdot \mathbf{R} \cdot \mathbf{C}^T = \left(\mathbf{I} - \frac{1}{N}\mathbf{E}\right) \cdot \mathbf{R} \cdot \left(\mathbf{I} - \frac{1}{N}\mathbf{E}\right) = \mathbf{R}_{\text{ц}} \tag{4}$$

где $\mathbf{I}$ – единичная матрица размерности M_k, $\mathbf{E}$ – матрица размерности M_k, все элементы которой равны 1.

Метод метрического многомерного шкалирования должен обеспечить нахождение координат вершин таких, чтобы матрица евклидовых расстояний между вершинами, максимально соответствовала матрице близостей $\mathbf{D}$, что соответствует минимуму стресса:

$$T = \sum_{m,n} \left(\hat{d}_{m,n} - \delta_{m,n}\right)^2 \tag{5}$$

где $\delta_{m,n} = \sqrt{(u_n - u_m)^2 + (v_n - v_m)^2 + (w_n - w_m)^2}$ - евклидова метрика. Поскольку $\hat{d}_{m,n}$ и $\delta_{m,n}$ являются оценками одной и той же величины расстояния между узлами в трёхмерном пространстве, полученные разными способами, то величина следа матрицы Λ будет преимущественно определяться суммой первых трех собственных значений, доля остальных собственных значений будет пренебрежимо мала. Таким образом, расположение вершин клики в пространстве относительно друг друга, рассчитанное методом ММШ, будет с высокой степенью точности соответствовать истинному расположению узлов.

$$\mathbf{U} = \mathbf{V} \cdot \mathbf{\Lambda}^{1/2} = (\mathbf{V}_1 \quad \mathbf{V}_2 \quad \mathbf{V}_3) \cdot diag\left(\sqrt{\lambda_1} \quad \sqrt{\lambda_2} \quad \sqrt{\lambda_3}\right) \tag{6}$$

2. На последующем шаге клики проверяются на наличие общих вершин, что дает возможность их объединения $G' = G_k \bigcup G_l$ при условии $|V_k \bigcap V_l| \geq 3$.

В том случае, когда условие выполнено, процесс объединения сопровождается пересчетом пространственных координат. Переход к новой прямоугольной системе координат осуществляется в несколько этапов, которые можно рассмотреть на примере одной из клик. Пусть количество общих для G_k и G_l вершин является минимально возможным и равно 3, данные вершины обозначены как $У_1, У_2, У_3$.

Координаты вершин в локальной системе координат подграфов G_k или G_l UVW могут быть представлены как $\mathbf{x'}_i = (u_i\ v_i\ w_i)^T,\ i$=1,2,3. Тогда:

1) осуществляется смещение системы координат ЛСК$_l$ на $\mathbf{s} = (u_1\ v_1\ w_1)^T = -\mathbf{x'}_1$;

2) осуществляется поворот ЛСК $U_1V_1W_1$ вокруг W_1 на угол γ_1, образуя новую ЛСК $U_2V_2W_2$;

3) осуществляется поворот ЛСК $U_2V_2W_2$ вокруг V_2 на угол β_2, образуя новую ЛСК $U_3V_3W_3$;

4) осуществляется поворот ЛСК $U_3V_3W_3$ вокруг W_3 на угол γ_3, образуя новую ЛСК $U'V'W'$.

Тогда вычисление координат узлов в новой локальной системе координат $U'V'W'$ $\mathbf{x''}_i = (u'_i\ v'_i\ w'_i)^T$ производится согласно выражению (5):

$$\mathbf{x''}_i = \mathbf{R}_{w3} \cdot \mathbf{R}_{v2} \cdot \mathbf{R}_{w1} \cdot \left(\mathbf{x'}_i - \mathbf{x'}_1\right) \tag{7}$$

где $\mathbf{R}_{wj} = \begin{pmatrix} \cos\left(\gamma_j\right) & \sin\left(-\gamma_j\right) & 0 \\ \sin\left(\gamma_j\right) & \cos\left(\gamma_j\right) & 0 \\ 0 & 0 & 1 \end{pmatrix}$, $\mathbf{R}_{vj} = \begin{pmatrix} \cos\left(\beta_2\right) & 0 & \sin\left(\beta_2\right) \\ 0 & 1 & 0 \\ \sin\left(-\beta_2\right) & 0 & \cos\left(\beta_2\right) \end{pmatrix}$, – матрицы вращения, $j = 1,3$, $\gamma_1 = -\text{arctg}\left(\frac{v_2 - v_1}{u_2 - u_1}\right)$, $\beta_2 = \frac{\pi}{2} - arctg\left(\frac{u_{2,2}}{v_{2,2}}\right)$, $\gamma_3 = \frac{\pi}{2}\left(1 - \frac{v_{2,3}}{|\mathbf{x}_{2,3}|}\right)$, $u_{2,2}, v_{2,2}$ – координаты У$_2$ в ЛСК $U_2V_2W_2$: такие что: $u_{2,2} = (u_2 - u_1)cos(\gamma_j) - (v_2 - v_1)sin(\gamma_j)$, $v_{2,2} = (u_2 - u_1)\sin\left(\gamma_j\right) - (v_2 - v_1)\cos\left(\gamma_j\right)$, $\mathbf{x}_{2,3} = (u_{2,3}\ v_{2,3}\ w_{2,3})^T$ – вектор координат У$_2$ в системе координат $U_3V_3W_3$.

Координаты У$_1$, У$_2$, У$_3$ в новой ЛСК $U'V'W'$ равны $(0\ 0\ 0)^T$, $(|\mathbf{x'}_2|\ 0\ 0)^T$ и $(u'_3\ v'_3\ w'_3)^T$ соответственно. Координаты вершин У$_1$, У$_2$, У$_3$ для G_k, G_l в новой системе координат должны совпасть.

3. Третий этап определения пространственных координат сенсорных узлов заключается в поиске вершин графа G, не принадлежащих G_k или $G' = G_k \bigcup G_l$, но смежных минимум 4-м вершинам, принадлежащих G_k или G', и определению их координат в локальных системах координат для G_k или G' методом мультилатерации, согласно которому искомая вершина графа определяется точкой пересечения 4-х и более сфер, центрами которых являются смежные вершины, и радиусы которых определяются по формуле (1). Пусть вершины смежные искомой обозначены как У$_1$, У$_2$, У$_3$, У$_4$, координаты которых определяются как $\mathbf{x'}_i = (u_i\ v_i\ w_i)^T,\ i$=1...4, координаты искомой вершины – $\mathbf{x'}_u = (u_u\ v_u\ w_u)^T$. Тогда для локальной системы координат подграфов G_k или $G' = G_k \bigcup G_l$ система из 3-х нелинейных уравнений квадратов расстояний принимает вида:

$$\hat{d}^2_{i,} = \left|\mathbf{x'}_i - \mathbf{x'}_u\right|^2 = (u_i - u_u)^2 + (v_i - v_u)^2 + (w_i - w_u)^2, i = 1,2,3 \tag{8}$$

Замена выражений в скобках на выражения вида: $(u_i - u_u) = (u_i - u_4 + u_4 - u_u) = (u_{i4} + u_{4u})$, $(v_i - v_u) = (v_i - v_4 + v_4 - v_{4u}) = (v_{i4} + v_4)$, $(v_i - v_u) = (v_i - v_4 + v_4 - v_u) = (v_{i4} + v_{4u})$ и дает:

$$\begin{cases} \delta^2_{1,u} = (u_{14} + u_{4u})^2 + (v_{14} + v_{4u})^2 + (w_{14} + w_{4u})^2 \\ \delta^2_{2,u} = (u_{24} + u_{4u})^2 + (v_{24} + v_{4u})^2 + (w_{24} + w_{4u})^2 \\ \delta^2_{3,u} = (u_{34} + u_{4u})^2 + (v_{34} + v_{4u})^2 + (w_{34} + w_{4u})^2 \end{cases} =$$

$$= \begin{cases} \delta^2_{1,u} = u^2_{14} + u^2_{4u} + v^2_{14} + v^2_{4u} + w^2_{14} + w^2_{4u} + 2(u_{14}u_{4u} + v_{14}u_{4u} + w_{14}w_{4u}) \\ \delta^2_{2,u} = u^2_{24} + u^2_{4u} + v^2_{24} + v^2_{4u} + w^2_{24} + w^2_{4u} + 2(u_{24}u_{4u} + v_{24}u_{4u} + w_{24}w_{4u}) \\ \delta^2_{3,u} = u^2_{34} + u^2_{4u} + v^2_{34} + v^2_{4u} + w^2_{34} + w^2_{4u} + 2(u_{34}u_{4u} + v_{34}u_{4u} + w_{34}w_{4u}) \end{cases} \quad (9)$$

Тогда:

$$\begin{cases} \delta^2_{1,u} - \delta^2_{2,u} = (u^2_{14} - u^2_{24} + v^2_{14} - v^2_{24} + w^2_{14} - w^2_{24} + ... \\ ... + 2(u_{4u}(u_{14} - u_{24}) + v_{4u}(v_{14} - v_{24}) + w_{4u}(w_{14} - w_{24})) \\ \delta^2_{2,u} - \delta^2_{3,u} = (u^2_{24} - u^2_{34} + v^2_{24} - v^2_{34} + w^2_{24} - w^2_{34} + ... \\ ... + 2(u_{4u}(u_{24} - u_{34}) + v_{4u}(v_{24} - v_{34}) + w_{4u}(w_{24} - w_{34})) \\ \delta^2_{3,u} - \delta^2_{1,u} = (u^2_{34} - u^2_{14} + v^2_{34} - v^2_{14} + w^2_{34} - w^2_{14} + ... \\ ... + 2(u_{4u}(u_{34} - u_{14}) + v_{4u}(v_{34} - v_{14}) + w_{4u}(w_{34} - w_{14})) \end{cases} \quad (10)$$

В выражениях (8, 9) неизвестными величинами являются u_{4u}, v_{4u}, w_{4u}. Тогда, обозначив слагаемые вида $\left(u^2_{i4} - u^2_{j4} + v^2_{i4} - v^2_{j4} + w^2_{i4} - w^2_{j4}\right)$ как $s_{i,j}$, можно получить систему линейных уравнений:

$$\begin{cases} \frac{\delta^2_{1,u} - \delta^2_{2,u} - s_{1,2}}{2} = u_{4u}(u_{14} - u_{24}) + v_{4u}(v_{14} - v_{24}) + w_{4u}(w_{14} - w_{24}) \\ \frac{\delta^2_{2,u} - \delta^2_{3,u} - s_{2,3}}{2} = u_{4u}(u_{24} - u_{34}) + v_{4u}(v_{24} - v_{34}) + w_{4u}(w_{24} - w_{34}) \\ \frac{\delta^2_{3,u} - \delta^2_{1,u} - s_{3,1}}{2} = u_{4u}(u_{34} - u_{14}) + v_{4u}(v_{34} - v_{14}) + w_{4u}(w_{34} - w_{14}) \end{cases} \quad (11)$$

которую можно решить одним из известных способов. Координаты искомой вершины будут равны $\mathbf{x'}_u = \begin{pmatrix} u_{4u} - u_4 & v_{4u} - v_4 & w_u - w_4 \end{pmatrix}^T$. Вершины с рассчитанными координатами добавляются в подграф. После их добавления производится поиск общих вершин в подграфах для возможного их объединения согласно пункту 2. Таким образом, последовательно осуществляется расчет координат для более удаленных узлов. Далее процедура повторяется для вершин, смежных 3-м вершинам подграфа G_k. В данном случае необходимо решить систему нелинейных уравнений, что дает неоднозначность решения для значения проекции на одну из главных осей локальной СК подграфа G_k. Расчеты с двумя решениями следует производить для проекции на такую главную ось, вдоль которой

наблюдается наибольший разброс значений. Наиболее вероятное значение координаты выбирается методом сравнения измеренных и рассчитанных расстояний с другими узлами. Случай, когда рассчитанное расстояние оказывается меньше порогового значения при отсутствии связи между узлами, свидетельствует о том, что выбранное значение координаты может быть ложным. При необходимости расчеты могут быть повторены для другой координаты.

4. Количество опорных узлов в сети должно быть достаточным для того, чтобы по завершении вычислений п. 1–3 в каждом полученном фрагменте сети было не менее 4-х опорных узлов, данных о которых позволяют перейти от локальной системы координат фрагментов сети к исходной $СК_0$. После перехода к исходной системе координат осуществляется уточнение координат сенсорных узлов.

5. Определяются координаты узлов, взаимодействующих с 2-мя узлами сети $У_1, У_2$ с известными координатами. Расчет координат может быть осуществлен методом многомерного шкалирования, описанным выше. Так как все известные координаты узлов даны в $СК_0$, то между данными узлами могут быть рассчитаны евклидовы расстояния. Это позволяет выбрать оптимальное количество узлов K, обеспечивающее точность вычислений при минимуме вычислений. Отсутствующие данные между данными узлами и искомым о расстояниях могут быть получены следующим образом.

1) пусть d_{max} – установленная максимальная дальность связи между узлами. Согласно правилу треугольника расстояние между узлом $У_m$ и искомым $d_{m,X}$ не превышает суммы расстояний $d_{m,X} \le d_{m,i} + d_{i,X},\ i = 1, 2$, где слагаемые в правой части неравенства известны. Тогда на первом этапе величина близости $\hat{d}_{m,X}$ берется равной:

$$\begin{aligned} \text{для 1-го узла: } \hat{d}_{m,X} &= d_{\max} + 0.5\left(\hat{d}_{m,1} + \hat{d}_{1,X} - d_{\max}\right) \\ \text{для 2-го узла: } \hat{d}_{m,X} &= d_{\max} + \min\left(0.5\left(\hat{d}_{m,i} + \hat{d}_{i,X} - d_{\max}\right)\right),\ i{=}1{,}2 \end{aligned} \tag{12}$$

3) методом многомерного шкалирования определяются координаты искомого узла , которые затем последовательно уточняются, исходя из минимума функции стресса (5), получаемой для K узлов с известными координатами.

Выводы: Разработанный в статье метод позволяет определять пространственные координаты узлов в больших беспроводных сенсорных сетях на основе неполных данных о расстояниях между узлами.

Литература

1. Кучерявый А. Е. Интернет вещей. Электросвязь 1 (2013). С. 21.
2. Kirichek R., Grishin I., Okuneva D., Falin M. Development of a node-positioning algorithm for wireless sensor networks in 3d space. 18th International Conference on Advanced Communication Technology (ICACT), IEEE, 2016, pp. 279-282.
3. «ГОСТ Р ИСО/МЭК 24730-5-2014 Информационные технологии (ИТ). Системы позиционирования в реальном времени (RTLS). Часть 5. Радиоинтерфейс расширения спектра методом линейной частотной модуляции (CSS) для связи на частоте 2,4 ГГц,» 01 01 2016. [В Интернете]. Available: http://docs.cntd.ru/document/1200115446. [Дата обращения: 25.05.2020].
4. Касьянов В. Н., Евстигнеев В. А. Графы в программировании: обработка, визуализация и применение. СПб.: БХВ-Петербург, 2003. 1104 с.
5. Ллойд Э., Ледерман У. Справочник по прикладной статистике. Том 2. М.: Финансы и статистика, 1990. 526 с.
6. Дэйвисон М. Л., Каменский В. С., Айвазян С. А. Многомерное шкалирование: методы наглядного представления данных. М.: Финансы и статистика, 1988. 254 с.
7. Роджерс Д., Адамс Дж. Математические основы машинной графики. М.: Мир, 2001. 604 с.

УДК: 004.7

Исследование использования протокола AODV в ячеистой сети LoRa

В.Д. Фам[1], Д.Т. Ле[2], Р.В. Киричек[1,3]

[1]Санкт-Петербургский Государственный Университет Телекоммуникаций им. проф. М.А. Бонч-Бруевича, 193232, Санкт-Петербург, Россия
[2]Университет Дананга – Университет науки и технологий, Дананг, Вьетнам
[3]Институт Проблем Управления им. В.А. Трапезникова Российской Академии Наук, 117997, Москва, Россия

fam.vd@spbgut.ru, letranduc@dut.udn.vn, kirichek@sut.ru

Аннотация

В этой статье, мы рассматриваем использование технологию LoRa для расширения покрытия сети сенсоров в процессе разработки умных устойчивых городов. Модель ячеистой сети LoRa предложена с использованием протокола AODV для маршрутизации потока данных между узлами. С имитационным моделированием на основе фреймворка OMNET++, серия компьютерных экспериментов была проведена с изменением различных параметров. В результате экспериментов, задержка и потерь пакетов были проанализированы в зависимости от количество узлов и размеров пакета в сети.

Ключевые слова: Интернет вещей, LoRa, ячеистая сеть, AODV, задержка, потерь пакетов

1. Введение

За последнее десятилетие, технология Интернет вещей (ИВ) уделяется большому вниманию не только в научных сферах, и в промышленности. Люди могут контролировать, мониторить и делать гораздо больше на удаленном расстоянии. Развитие ИВ создает потребность в новых беспроводных технологиях, способных поддерживать большое количество устройств в пространстве ИВ [1]. Эти системы требуют технологию, которая потребляет меньше энергии, а также покрывает большие расстояния. Однако, многие технологии, такие как ZigBee, WiFi, Bluetooth, широко используемые в настоящее время, потребляют большую

Исследование выполнено при финансовой поддержке гранта Президента Российской Федерации для государственной поддержки ведущих научных школ Российской Федерации в рамках научного проекта НШ-2604.2020.9.

мощность и не подходят для систем с батарейным питанием. Для удовлетворения условиям ИВ в области связи, нам нужна новая технология – глобальная сеть с низким энергопотреблением (LPWAN), которая обеспечивает радио-покрытие на большой площади посредством базовых станций и адаптирует скорости и мощность передачи данных, модуляцию, и т.д., так что конечные устройства потребляют очень малое энергопотребление для их подключения.

Согласно по рекомендации международной электросвязи в телекоммуникации МСЭ-Т У.4903/Л.1603 для развития умных устойчивых городов, необходимо обеспечить полный доступ к Интернету [2]. Одной из известных технологий, входящих в группу сети LPWAN, является технология LoRa (Long-Range), которая была разработана Semtech [3]. Большая дальность действия и маломощный характер делает LoRa интересным кандидатом для технологии интеллектуального сенсора в гражданских инфраструктурах (таких как мониторинг состояния здоровья, умные счетчики, мониторинг окружающей среды и т.д.), а также в промышленных приложениях.

Технология LoRa появилась с возможностью передачи данных на большие расстояния и при этом требуется низкое энергопотребление. LoRa заполняет технологический пробел сотовой сети связи и сетей WiFi/BLE, которые требуют либо высокой пропускной способности, либо высокой мощности, либо имеют ограниченный диапазон или неспособность проникать в глубокие внутренние помещения. В сущности, технология LoRa является гибкой для использования в сельских или внутренних помещениях, в умных устойчивых городах, интеллектуальной цепочке поставок и логистике. Поэтому ячеистая сеть LoRa может предоставить большое покрытие. Однако необходимо рассмотреть методы маршрутизации для организации такой сети. Как известно AODV – протокол динамической маршрутизации часто используется для ad-hoc и других беспроводных сетей. Таким образом, мы рассматриваем применение этого протокола для организации ячеистой сети LoRa.

2. Протоколы маршрутизации

Типично, протоколы маршрутизации делятся на три категории в зависимости от информации о топологии сети, используемой для построения маршрутов: проактивный, реактивный и гибридный.

Проактивные протоколы: этот вид протокола периодически обменивается информацией о топологии между всеми узлами сети. Следовательно, протокол упреждающей маршрутизации не позволяет обнаруживать маршруты, поскольку маршрут назначения сохраняется и поддерживается в таблице. Таблицы обычно необходимо обновлять. Эти протоколы используются там, где часто возникают требования к маршрутам. Однако недостатком этого протокола является то, что

он обеспечивает низкий уровень простоя для постоянного приложения [4]. Примерами являются DSDV (Destination Sequenced Distance Vector), OLSR (Opitmized Link State Routing).

Реактивной Протоколы: эти протоколы маршрутизации выбирают маршруты к другим узлам только тогда, когда они необходимы. Процесс обнаружения маршрута запускается, когда узел хочет связаться с другой станцией, для которой у него нет доступа к таблице маршрутов. Примерами являются AODV (Ad-hoc On-Demand distance Vector), DSR (Dynamic Source Routing).

Гибридной Протоколы: гибридная маршрутизация, объединяющая близлежащие проактивной протоколы и глобальные реактивной протоколы, чтобы уменьшить накладные расходы и задержку маршрутизации из-за процесса поиска маршрута. Преимущества этих протоколов – более высокая эффективность и масштабируемость. Однако недостатком является высокая задержка при обнаружении новых маршрутов [5]. Примером является ZRP (Zone Routing Protocol).

В каждой из этих трех категорий существуют различные протоколы, поэтому здесь невозможно дать исчерпывающий обзор. Вместе этого мы рассмотрим и оценим протоколы, которые обычно используются и упоминаются в литературе.

Существует множество исследований, посвященных сравнению упомянутых выше протоколов. В [6, 7] протокол AODV показал лучшую производительность (задержка, нагрузка маршрутизации, коэффициент доставки), чем другие протоколы маршрутизации при увеличении числа узлов. В [8] авторы показали, что хотя DSR хорошо работает при быстрой передаче, но имеет высокую потерю пакетов. Однако в [9] авторы доказали, что AODV имеет исполнение с точки зрения обычного джиттера и задержки по сравнению с DYMO, DSR, OSLR, ZRP.

В [10] авторы указали, что AODV рекомендуется для защищенной связи. Кроме того, этот протокол также дает хороее значение средней пропускной способности [11].

В дополнении к этим оценкам, мы может видеть, что OSLR хорошо подходит для больших и плотных сетей со случайным и нерегулярным трафиком. Однако нам нужно несколько главных узлов в сети LoRa для покрытия большой территории. Таким образом, не нужны дополнительные накладные расходы на выбор транзитных узлов и обновление информации о топологии в нашем случае [12].

Хотя DSDV имеет меньшие накладные расходы на управление, чем OLSR [13]. Тем не менее, непрерывные обновления не нужны для сетей со статическими узлами, как в типичном сценарии развертывания сети LoRa.

Поскольку для реактивных протоколов требуется совместное использование информации о топологии только при сбое маршрутов или необходимости создания нового маршрута, они позволяют снизить накладные расходы на управление и,

следовательно, затраты на энергию по сравнению с проактивными протоколами [14].

Протокол DSR разработан для сетей с потенциально высокой мобильностью. Поэтому это не совсем подходит для нашего случая [12].

Таким образом, хотя сравнения производятся в основном в сетях MANET и VANET, они также частично указывают на преимущества и недостатки протоколов. Судя по этим оценкам, возможно, подходящим протоколом маршрутизации для нашей ячеистой сети LoRa является AODV. Это причина, по которой мы решили использовать протокол AODV в ячеистой сети LoRa.

3. Сеть и имитационная модель

3.1. Модель сети. В настоящее время, ячеистые сети используются в многих приложениях ИВ. В различных случаях сенсорные сети развернуты далеко от точки доступа, подключенной к Интернету. В таких случаях предлагается модель на основе ячеистых сетей в двух сегментах, показанных на Рис. 1. Сенсорные сети связываются со шлюзами для соединения к облачному серверу. Шлюзы могут взаимодействовать между собой в ячеистой сети, в то время как некоторые шлюзы имеют доступ к Интернету. Более того, использование связи дальнего радиуса действия в качестве LoRa для сети шлюзов является основной идеей расширения зоны покрытия.

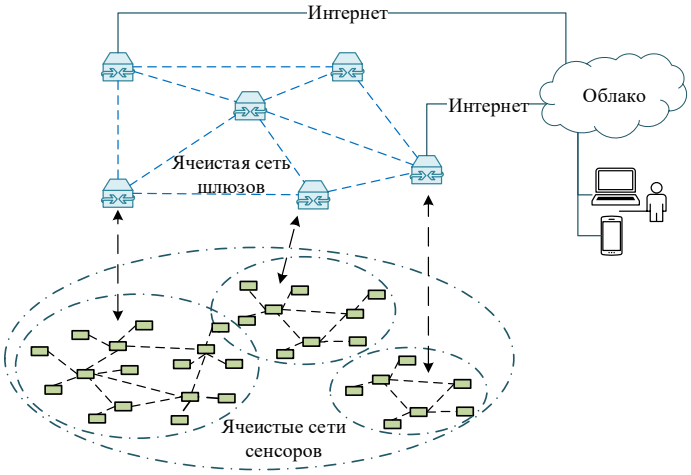


Рис. 1. Модель сети

В частности, ячеистая сеть может предоставить связь для устройств ИВ в приложениях умных устойчивых городов. Используя технология LoRa в качестве решения коммуникации на физическом уровне, устройства могут взаимодействовать с другими на расстоянии более ста метров. Как показано на Рис. 2, сеть устройств развернута в районе города Санкт-Петербурга. Каждый узел оснащен интерфейсом LoRa, настроенным с одинаковыми параметрами. Предполагается, что эти устройства используются для сбора данных с датчиков в зоне ближней связи, затем собранные данные передаются на узел приемника, который подключен к внешней сети к удаленному серверу. Протокол AODV предлагается для установления путей маршрутизации к узлу приемника.

Рис. 2. Пример фрагмента сети в городе

3.2. Методология и параметры моделирования. Для имитационного моделирования сети используются фреймворки OMNET++ и inet, которыми популярно воспользуются во многих областях моделирования проводных и беспроводных сетей. На основе этих фреймворков, был разработан модуль узла сети LoRa. Поскольку библиотеки и фреймворки построены на основе модульных и компонентно-ориентированных принципов, мы можем интегрировать модуль узла со встроенными модулями из фреймворков.

Узел LoRa включает модули, моделирующие протоколы радио и верхнего уровней. В радио-модуле LoRa мы можем задавать радиопараметры, соответствующие нашей аппаратной модели. В дополнение к конфигурации, параметры, такие как коэффициент распространения (SF) и скорость кодирования (CR),

ширина полосы пропускания (BW), настроены для узла LoRa. Согласно спецификации чипсета LoRa SX127x [3], выбранные настроенные параметры SF и BW влияют на чувствительность приема. При случае того, когда в сети сконфигурированы низкая пропускания и высокий коэффициент распространения, приемник имеет высокую чувствительность. Но в таком случае скорость передачи данных уменьшается, а дальность действия увеличивается. Анализируя результаты в работе [15], мы выбрали 125 кГц и 8, установленные для BW и SF соответственно.

Кроме того, в модуле среды LoRa модель распространения потерь на трассе сконфигурирована с учетом внедрения сети в городской среде. В данной работе параметры модели распространения сигналов были получены из серий измерений, выполненных в работе [16]. В частности, измерения в [16] соответствуют городской среде с зданиями, где устройства частично развернуты в помещениях.

Серия компьютерных экспериментов проводится с учетом двух случаев:

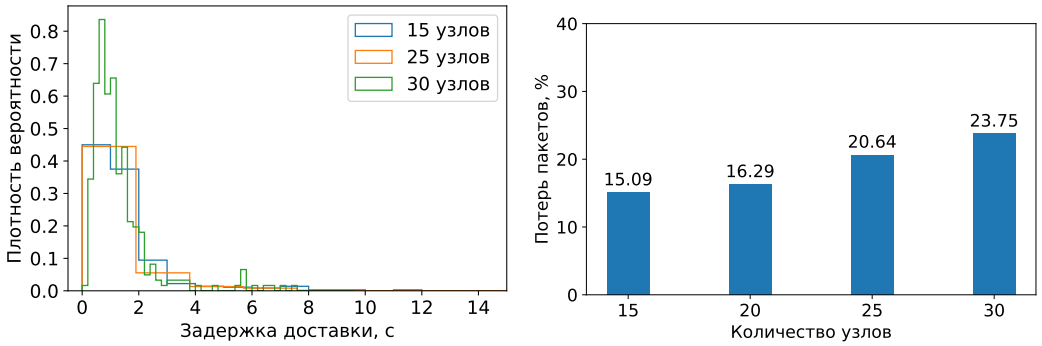
- Случай 1: количество узлов в сети изменяется, а пакеты данных генерируются со случайной длиной. Размер полезной нагрузки находится в интервале от 20 до 150 байт.
- Случай 2: хотя число узлов в сети постоянно, размер полезной нагрузки задается в байтах {20, 40, 60, 150}, соответствующих каждому эксперименту.

Интервалы между оправками сообщений генерируются случайным образом в соответствии с экспоненциальным распределением со среднем значением 120 с. На поле размером 2000x2000 m^2, координаты узла A и узла приемника зафиксированы. Узел A расположен в позиции (400, 400), а узел приемник расположен в (1500, 1500). Следовательно, необходимо иметь транзитные узлы для обеспечения связи между ними в рассматриваемой модели распространения.

В каждом эксперименте проанализированы сквозная задержка и процент потерь пакетов от узла A к узлу приемника. Полученные результаты приведены в следующем разделе.

4. Результаты моделирования

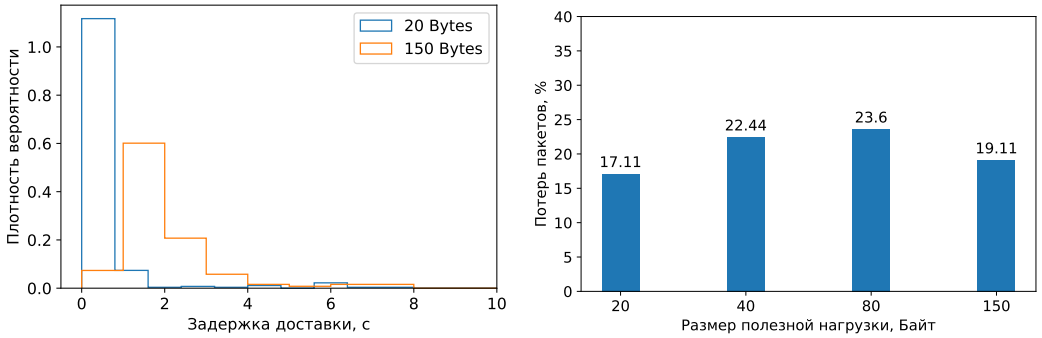
4.1. Анализ по количеству узлов. Эксперименты проводились с изменением количества узлов в сети.Координаты узлов расположены случайным образом по равномерном распределению в поле моделирования. Рис. 3a показывает распределения задержек, необходимой для доставки пакета данных из узла A в узел приемника. Задержка варьируется до нескольких секунд в такой сети. При этом интервале количества узлов задержка не сильно меняется. Как показан на рисунке, задержка составляет в интервале до 2 секунд с вероятностью равной $2 \cdot 0.45 = 0.9$ в сети 25 узлов. Однако, коэффициент потери пакетов увеличивается при увеличении количества узлов в сети (Рис. 3b). Добавление количества узлов существенно влияет на эффективность доставки пакетов.



(a) Распределение задержек (b) Процент потери пакетов

Рис. 3. Анализ по количеству узлов

4.2. Анализ по размеру полезной нагрузки. Кроме того, с фиксированным количеством узлов в сети рассматриваются задержки и потерь пакетов при изменении длины передаваемой полезной нагрузки. Результаты анализа изображены на Рис. 4. Задержка доставки может достичь 10 секунд через ретрансляционные узлы к приемнику. Однако, коэффициент потери пакетов также меняется при увеличении размера пакета. По сравнению результатов передачи пакетов размерам 20 и 150 байтов оказалось, что размер пакета не сильно влияет на вероятности потери в данной сети.



(a) Распределение задержек (b) Процент потери пакетов

Рис. 4. Анализ по размеру полезной нагрузки

5. Заключение

В данной работе, мы представили результаты исследовании ячеистой сети LoRa с использованием протокола AODV для поиска маршрута от узла-источника к узлу-приемнику. В большинстве, сети на основе LoRa развернуты в топологии "звезда". Однако, принимая во внимание преимущества связи дальнего радиуса действия и низкое энергопотребление, технология LoRa была рассмотрена для расширения покрытия сенсорных сетей с помощью ячеистой сети LoRa, которая предложена для передачи данных датчиков из различных кластеров в узел приемника, имеющий доступ к Интернету. Исследование было проведено в имитационной модели, разработанной на основе фреймворков omnet++ и inet. Результаты моделирования показали, что сквозная задержка в ячеистой сети LoRa довольно высока. Кроме того, изменение количества узлов в сети и размера полезной нагрузки повлияло на коэффициент потери пакетов.

На основании результатов исследования, протокол AODV может использоваться для ячеистой сети LoRa. Тем не менее, задержка должна рассматриваться в такой сети, что будет более эффективно сеть с небольшим количеством узлов. Заглядывая в будущее, мы намерены рассмотреть вопрос о разработке протокола, совместимого как с ячеистой сетью LoRa, так и с LoRaWAN.

Литература

1. A. Koucheryavy, Интернет вещей, Электросвязь (1) (2013) 21–24.
2. S. Ben Dhaou, N. Lopes, M. Meyerhoff Nielsen, Connecting cities and communities with the sustainable development goals (2017).
3. Semtech Corporation, Datasheet SX1276/77/78/79 LoRa Transciever (2019) 132.
4. R. Kumar, M. Dave, A comparative study of various routing protocols in vanet, arXiv preprint arXiv:1108.2094 (2011).
5. A. Chandra, S. Thakur, Qualitative analysis of hybrid routing protocols against network layer attacks in manet, Apoorva Chandra et al, International Journal of Computer Science and Mobile Computing 4 (6) (2015) 538–543.
6. J. Haerri, F. Filali, C. Bonnet, Performance comparison of aodv and olsr in vanets urban environments under realistic mobility patterns, in: Proceedings of the 5th IFIP mediterranean ad-hoc networking workshop, 2006, pp. 14–17.
7. J. A. Ferreiro-Lage, C. P. Gestoso, O. Rubiños, F. A. Agelet, Analysis of unicast routing protocols for vanets, in: 2009 Fifth International Conference on Networking and Services, IEEE, 2009, pp. 518–521.
8. P. K. Singh, K. Lego, T. Tuithung, Simulation based analysis of adhoc routing protocol in urban and highway scenario of vanet, International Journal of Computer Applications 12 (10) (2011) 42–49.

9. J. Kumar, Comparative performance analysis of aodv, dsr, dymo, olsr and zrp routing protocols in manet using varying pause time, International Journal of Computer Communications and Networks (IJCCN) 3 (1) (2012) 43–51.
10. B. Makodia, T. Patel, K. Parmar, S. Hadia, A. Shah, Implementing and analyzing routing protocols for self-organized vehicular adhoc network, in: 2013 Nirma University International Conference on Engineering (NUiCONE), IEEE, 2013, pp. 1–6.
11. V. N. Talooki, K. Ziarati, Performance comparison of routing protocols for mobile ad hoc networks, in: 2006 Asia-Pacific Conference on Communications, IEEE, 2006, pp. 1–5.
12. D. Lundell, A. Hedberg, C. Nyberg, E. Fitzgerald, A routing protocol for lora mesh networks, in: 2018 IEEE 19th International Symposium on"A World of Wireless, Mobile and Multimedia Networks"(WoWMoM), IEEE, 2018, pp. 14–19.
13. C. E. Perkins, P. Bhagwat, Highly dynamic destination-sequenced distance-vector routing (dsdv) for mobile computers, ACM SIGCOMM computer communication review 24 (4) (1994) 234–244.
14. K. Pandey, A. Swaroop, A comprehensive performance analysis of proactive, reactive and hybrid manets routing protocols, arXiv preprint arXiv:1112.5703 (2011).
15. R. Kirichek, V. Vishnevsky, V. D. Pham, A. Koucheryavy, Analytic model of a mesh topology based on lora technology, in: 2020 22nd International Conference on Advanced Communication Technology (ICACT), IEEE, 2020, pp. 251–255.
16. M. C. Bor, U. Roedig, T. Voigt, J. M. Alonso, Do lora low-power wide-area networks scale?, in: Proceedings of the 19th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, 2016, pp. 59–67.

UDC: 004.72

A study of using AODV protocol in LoRa mesh network

V.D. Pham[1], D.T. Le[2], R.V. Kirichek[1, 3]

[1]The Bonch-Bruevich Saint-Petersburg State University of Telecommunications, 193232, St. Petersburg, Russian Federation
[2]The University of Danang - University of Science and Technology, Danang, Vietnam
[3]V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, 117997, Moscow, Russian Federation

fam.vd@spbgut.ru, letranduc@dut.udn.vn, kirichek@sut.ru

Abstract

In this paper, we consider using the LoRa technology to expand sensor network coverage in the development of smart sustainable cities. A model of a LoRa mesh network is proposed using the AODV protocol to route data flow between nodes. With a simulation model developed based on OMNET++, a series of computer experiments was carried out with changing various parameters. In the results of the experiments, the end-to-end delay and packet loss ratio were analyzed in the dependence on the number of nodes and packet size in the network. The simulation results show that the latency is relatively high in the LoRa mesh network, but it might be accepted for some applications.

Keywords: IoT, LoRa, mesh network, LoRa mesh, AODV, delay, packet loss

1. Introduction

Over the past decade, the Internet of Things (IoT) has received significant attention in the scientific and industrial fields. People can control, monitor, and do a lot more from a remote distance. It is done by connecting various objects reducing physical distance. The IoT movement creates the need for new wireless technologies, capable of supporting the large numbers of devices in the IoT space. These systems require a technology that consumes less power and also covers long distances. However, many technologies such as ZigBee, WiFi, Bluetooth popularly used at present consumes high power and is not suitable for battery-operated systems. To fulfill the communication requirements of IoT, we need new technology. Low-Power

The publication has been prepared with the support of the grant from the President of the Russian Federation for state support of leading scientific schools of the Russian Federation according to the research project SS-2604.2020.9.

Wide Area Network (LPWAN) offers radio coverage over a large area by way of base stations and adapting transmission rates, transmission power, modulation, duty cycles, where end-devices incur a very low energy consumption due to their being connected.

According to the ITU-T Y.4903/L.1603 recommendation for the development of Smart Sustainable Cities, it is necessary to provide full Internet access covered in the city. One of the well-known technologies included in the LPWAN network group is LoRa (Long-Range) technology developed by Semtech. The long-range and low-power nature of LoRa makes it a promising candidate for smart sensing technology in civil infrastructures (such as health monitoring, smart metering, environment monitoring, etc.), as well as in industrial applications.

Many technologies are proposed to use in IoT applications. Every technology has its features, advantages, and disadvantages. However, no single technology can serve all IoT applications because different applications will have different requirements. Based on the requirement, we can only choose a technology that is best suited for the specific application from the existing technologies.

LoRa technology will revolutionize IoT by enabling data communication over a long-range distance while using very little power. LoRa fills the technology gap of Cellular and WiFi/BLE based networks that require either high bandwidth or high power or have a limited range or inability to penetrate deep indoor environments. In effect, LoRa Technology is flexible for rural or indoor use cases in smart cities, smart homes and buildings, smart agriculture, smart metering, and smart supply chain and logistics.

2. Routing in LoRa mesh networks

Designing and developing routing protocols in wireless mesh networks is a challenging issue that should cover multiple performance metrics such as minimum hop count, preventing disruption of the service methods according to robustness concepts. Using the mesh infrastructure to perform routing processes as efficiently as possible, and increasing the scalability of routing protocols to install or maintain routing paths in a mesh network with large capacity [1]. To implement a LoRa mesh network model with the LoRa gateways, we need to choose the appropriate routing protocol.

Typically, ad-hoc routing protocols are divided into three categories based on the network topology information used for route discovery: proactive, reactive, or hybrid.

• Proactive Routing Protocols: This kind of protocol periodically exchanges the topology information between all the network nodes. Therefore, proactive routing protocol has no route discovery since the destination route is saved and maintained within a table. The tables usually must be updated. These protocols are used where the route requirements are frequent. However, the drawback of this protocol is that it

gives low idleness to constant application [2]. DSDV (Destination Sequenced Distance Vector), OLSR (Optimized Link State Routing) are examples.

• Reactive Routing Protocols: These routing protocols choose routes to other nodes only when they are needed. A route discovery process is launched when a node wants to communicate with another station for which it does not possess any route table access. AODV (Ad-hoc On-Demand Distance Vector routing protocol), DSR (Dynamic Source Routing) are examples.

• Hybrid Routing Protocols: Hybrid Routing joining nearby proactive routing protocols and global reactive routing protocols to reduce routing overhead and delay due to route disclosure process. The advantages of these protocols are higher efficiency and scalability. However, the disadvantage is high latency for locating new routes [3]. Examples of these protocols are ZRP (Zone Routing Protocol).

There are many protocols within each of these three categories, and a comprehensive review cannot be provided here. We will instead examine and evaluate a representative selection of protocols commonly used and referenced in the literature.

There are many kinds of research focused on the comparison of the protocols mentioned above. In [4, 5], the authors found that AODV has shown better performance than other routing protocols (DSR, DSDV) with the increment of nodes number in all performance metrics (end-to-end delay, routing load, received packets, packet delivery ratio, dropped packets). In [6], Singh et al. showed that although DSR performs well in quick transmission, but it has high packet loss. Jogendra Kumar in [7] proved that AODV has the best execution as far as normal jitter and end-to-end delay in comparison with DYMO (Dynamic MANET On-Demand Routing Protocol), DSR, OSLR, ZRP.

In [8], Makodia et al. even pointed out that AODV is advised for secured communication. In addition, AODV also gives a good value of average throughput [9].

In addition to those evaluations, we can see that, due to its characteristics, OSLR is well-suited to large and dense networks with random and sporadic traffic. However, we need a few gateways to cover even a large area in the LoRa network. As such, the added overhead of choosing relays and updating topology information is unnecessary in our case [10].

Although DSDV has lower control overhead than OLSR [11], continuous updates are nonetheless unnecessary for networks with static nodes, as in a typical LoRa deployment scenario.

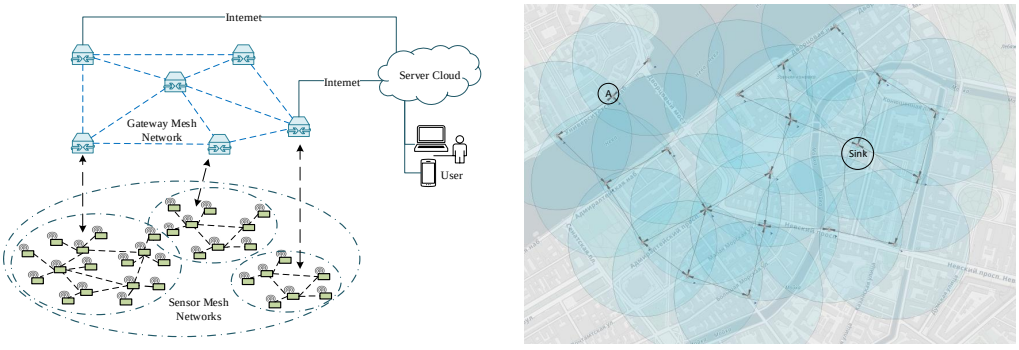
Since reactive protocols require sharing of topology information only when routes fail or a new route needs to be established, they allow for a reduced control overhead, and thus energy cost, in comparison to proactive protocols [12].

DSR protocol is designed for a network with potentially high mobility. Therefore it does not suit out case [10].

In summary, although the comparisons are made mainly in MANET and VANET networks, it also partly points out the advantages and disadvantages of the protocols. From those evaluations, perhaps, the suitable routing protocol for our LoRa mesh network is AODV. This is the reason why we choose to use the AODV protocol in the LoRa mesh network.

3. Network and simulation model

3.1. Network model. Currently, mesh networks are popularly used in various IoT applications. In many cases, sensor networks are deployed far away from the access point to the Internet. In these cases, a model is proposed based on mesh networks in two segments. As shown in Figure 1a, sensor networks communicate with the gateways to connect to the server cloud. The gateways are able to communicate with each other in the mesh network while some gateways have access to the Internet. Moreover, using long-range communication as LoRa for the gateway network is the main idea to expand the network coverage.



(a) Network model

(b) Deployment area

Figure 1. LoRa mesh network

In particular, the mesh network can provide connectivity for IoT devices in smart sustainable cities applications. Using the LoRa technology as a communication method at the physical layer, devices may communicate with the others over a hundred meters. As shown in Figure 1b, a network of devices is deployed in an area of the city Saint-Petersburg. Each node is equipped with the LoRa interface configured with the same parameters. We can consider these devices used to collect data from the short-range communication sensors, then collected data are transmitted

to the sink node connected to the external network to the remote server. The AODV protocol is proposed to establish routing paths to the sink node.

3.2. Simulation methodology and parameter. The frameworks OMNET++ and inet are used to carry out the network simulation. They are known well to be used in numerous domains for simulating wired and wireless networks. Based on these frameworks and the work in [13], we have developed a module of the LoRa node. Since the OMNET ++ library and frameworks are designed based on modular and component-oriented principles, the LoRa node can be integrated with the build-in modules from the inet framework.

The LoRa node consists of modules modeling radio and upper-layer protocols. In the LoRa radio module, we can set radio parameters that correspond to our hardware model. The other parameters, such as spreading factor and coding rate configured for the LoRa node in addition to the usual configurations. According to the datasheet of SX127x LoRa chipset [14], the bandwidth and spreading factor influence the reception sensitivity. The receiver has high sensitivity when the low bandwidth and the high spreading factor are used in the network. The data rate is also decreased while the communication range increases. Analyzing from work in [15], we have chosen to use 125 kHz and 8 configured for the bandwidth and the spreading factor, respectively.

Moreover, in the LoRa medium module, the path loss propagation model is configured in considering the wireless signal propagation in the urban environment. In this work, the propagation model parameters have been received from a series of measurements performed in [16]. In particular, the measurements in [16] correspond to the build-up urban environment, where devices are partially deployed indoors.

A series of experiments are carried out with considering two cases:

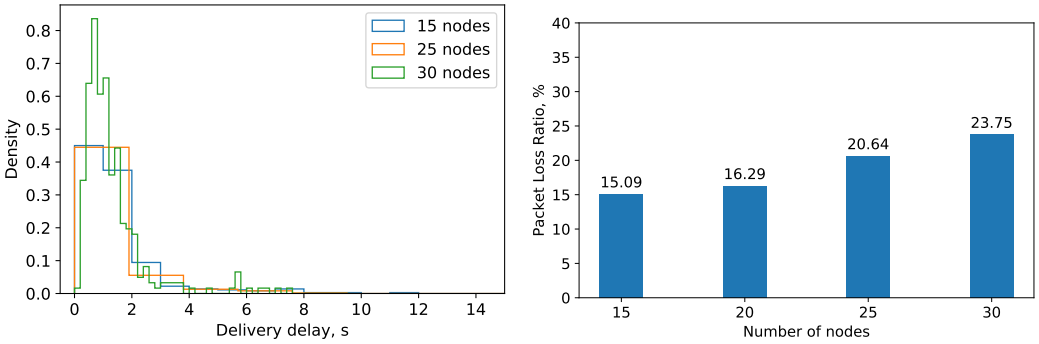
- Case 1: the number of nodes in the network is changed, while the data packets are generated with random length. The payload size is in the interval from 20 bytes to 150 bytes.
- Case 2: while the number of nodes is constant in the network, the payload size is set in {20, 40, 60, 150} bytes corresponding for each experiment.

The interval between messages is generated randomly according to the exponential distribution with a mean equal to 120 seconds. In a field with a size of 2000x2000 m^2, the coordinates of the node A and the sink node are fixed. The node A is located at the position (400, 400), and the sink node is located at (1500, 1500). Hence, it is required to have relay nodes to communicate between them in the considered propagation environment.

We have analyzed the end-to-end delay and packet loss ratio from the node A to the sink node in each experiment. The obtained results are shown in the next section.

4. Result analysis

4.1. Analysis by varying the number of nodes. The experiments were performed with a changing number of nodes in the network. The coordinates of nodes were generated randomly according to the uniform distribution in the interval (0, 2000). Fig. 2 shows a histogram of the delay required to deliver the data packet from the node A to the sink node. According to Fig. 2a, the delay does not change much when increasing the number of nodes in the network. Based on the probability density histogram, we can see that the delay varying in the interval to 2s has a probability equal to $2 \cdot 0.45 = 0.9$. However, the packet loss ratio increases when increasing the number of nodes in the network, as shown in Fig. 2b. Adding the number of nodes significantly affects the change of the packet loss ratio.



(a) Latency Distribution

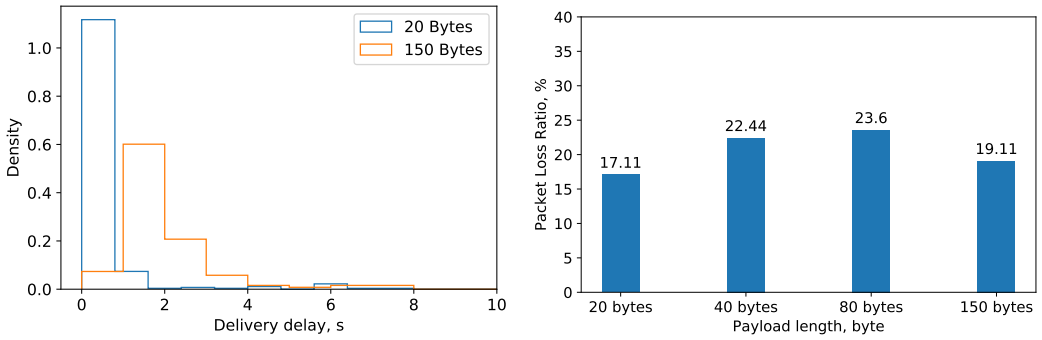
(b) Packet Loss Ratio

Figure 2. Analysis by varying the number of nodes

4.2. Analysis by varying payload length. In the second case, with the fixed number of nodes in the network, a series of experiments was performed with changing payload length generated randomly in the first case. In this case, the delay distribution and packet loss ratio are presented in Figure 3a and 3b, respectively. The delivery delay can reach up to several seconds via relay nodes to the destination. Besides, the packet loss ratio increases when increasing the payload length. However, the data packet size does not significantly affect the change of the packet loss ratio. Comparing the sending packets of size 20 and 150 bytes shows that the packet size does not greatly affect the probability of loss in this considered network.

5. Conclusion

This paper presented the results of studying a LoRa mesh network using the AODV protocol to find a route from a source node to a sink node. Mostly LoRa-based networks are deployed in the star topology. However, taking the advantages



(a) Latency Distribution

(b) Packet Loss Ratio

Figure 3. Analysis by varying payload length

of long-range communication and low power consumption, LoRa technology was considered to expand sensor network coverage. A LoRa mesh network was proposed to transmit sensor data from different clusters to the sink node connected to the Internet. The study was conducted in a simulation model developed based on the frameworks omnet++ and inet. The results of the experiments showed that the end-to-end delay is relatively high in the LoRa mesh network. Moreover, changing the number of nodes in the network and payload size affected the packet loss ratio.

Based on the study results, the AODV protocol might be used for the LoRa mesh network. However, the delay needs to be considered in such a network. Looking into the future, we intend to consider in developing the other protocol that has compatibility with both the LoRa mesh network and LoRaWAN.

REFERENCES

1. M. Eslami, O. Karimi, T. Khodadadi, A survey on wireless mesh networks: Architecture, specifications and challenges, in: 2014 IEEE 5th Control and System Graduate Research Colloquium, IEEE, 2014, pp. 219–222.
2. R. Kumar, M. Dave, A comparative study of various routing protocols in vanet, arXiv preprint arXiv:1108.2094 (2011).
3. A. Chandra, S. Thakur, Qualitative analysis of hybrid routing protocols against network layer attacks in manet, Apoorva Chandra et al, International Journal of Computer Science and Mobile Computing 4 (6) (2015) 538–543.
4. J. Haerri, F. Filali, C. Bonnet, Performance comparison of aodv and olsr in vanets urban environments under realistic mobility patterns, in: Proceedings of the 5th IFIP mediterranean ad-hoc networking workshop, 2006, pp. 14–17.

5. J. A. Ferreiro-Lage, C. P. Gestoso, O. Rubiños, F. A. Agelet, Analysis of unicast routing protocols for vanets, in: 2009 Fifth International Conference on Networking and Services, IEEE, 2009, pp. 518–521.
6. P. K. Singh, K. Lego, T. Tuithung, Simulation based analysis of adhoc routing protocol in urban and highway scenario of vanet, International Journal of Computer Applications 12 (10) (2011) 42–49.
7. J. Kumar, Comparative performance analysis of aodv, dsr, dymo, olsr and zrp routing protocols in manet using varying pause time, International Journal of Computer Communications and Networks (IJCCN) 3 (1) (2012) 43–51.
8. B. Makodia, T. Patel, K. Parmar, S. Hadia, A. Shah, Implementing and analyzing routing protocols for self-organized vehicular adhoc network, in: 2013 Nirma University International Conference on Engineering (NUiCONE), IEEE, 2013, pp. 1–6.
9. V. N. Talooki, K. Ziarati, Performance comparison of routing protocols for mobile ad hoc networks, in: 2006 Asia-Pacific Conference on Communications, IEEE, 2006, pp. 1–5.
10. D. Lundell, A. Hedberg, C. Nyberg, E. Fitzgerald, A routing protocol for lora mesh networks, in: 2018 IEEE 19th International Symposium on" A World of Wireless, Mobile and Multimedia Networks"(WoWMoM), IEEE, 2018, pp. 14–19.
11. C. E. Perkins, P. Bhagwat, Highly dynamic destination-sequenced distance-vector routing (dsdv) for mobile computers, ACM SIGCOMM computer communication review 24 (4) (1994) 234–244.
12. K. Pandey, A. Swaroop, A comprehensive performance analysis of proactive, reactive and hybrid manets routing protocols, arXiv preprint arXiv:1112.5703 (2011).
13. M. Slabicki, G. Premsankar, M. Di Francesco, Adaptive configuration of lora networks for dense iot deployments, in: NOMS 2018-2018 IEEE/IFIP Network Operations and Management Symposium, IEEE, 2018, pp. 1–9.
14. Semtech Corporation, Datasheet SX1276/77/78/79 LoRa Transciever (2019) 132.
15. R. Kirichek, V. Vishnevsky, V. D. Pham, A. Koucheryavy, Analytic model of a mesh topology based on lora technology, in: 2020 22nd International Conference on Advanced Communication Technology (ICACT), IEEE, 2020, pp. 251–255.
16. M. C. Bor, U. Roedig, T. Voigt, J. M. Alonso, Do lora low-power wide-area networks scale?, in: Proceedings of the 19th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, 2016, pp. 59–67.

UDC: 519.2

Approximate Sequencing of Virtual Reels with Genetic Algorithms

P.D. Petrov, G.B. Kostadinov, P.R. Zhivkov,

V.I. Velichkova, T.D. Balabanov[0000-0003-3139-069X]

Bulgarian Academy of Sciences
Institute of Information and Communication Technologies
acad. Georgi Bonchev Str., block 2, office 514
1113 Sofia, Bulgaria
http://iict.bas.bg/

p.petrov@iit.bas.bg g.kostadinov@iit.bas.bg pzhivkov@iit.bas.bg
vvelichkova@iit.bas.bg todorb@iinf.bas.bg

Abstract

Sequencing is a very popular mathematical problem in the field of genetics. DNA sequence information is organized as pairs of the four nucleotide bases - Cytosine, Guanine, Adenine, and Thymine. In some cases, only chunks are known but the full sequence is unknown. The problem of sequencing is a reconstruction of the full sequence from the known chunks. Sequencing is applied also in other fields as encoding and cryptography. This research proposes approximate sequencing of virtual reels used in gambling slot machine games. The optimization process is done with classical genetic algorithms, but optimality is estimated into chunks space instead of sequences space.

Keywords: Sequencing, Slot Machines, Genetic Algorithms

1. Introduction

Slot machines are gambling games organized as spinning reels with drawn symbols which stop on certain positions. In the beginning slot machines were mechanical with mechanical reels. Win of the player was appointed according to stop positions after reels spin. With the extensive evolvement of computers [1], integrated devices [2] and computerized gambling games in the last three decades of the 20th century, mechanical slot machines were replaced with computerized. Mechanical reels were replaced with virtual reels stored inside the computer's memory. Mechanical spin was replaced with animated virtual spin. The most spread electronic slot machines have

The publication has been prepared with the support of Velbazhd Software LLC and Bulgarian Ministry of Education and Science according to the research project No.D01–205/23.11.2018.

5 reels and only 3 symbols are visible on the screen. Such configuration gives a screen grid of 3x5 visible symbols. In most of the games, taken from left to right, different cells in each reel are checked as lines and if consequent symbols of a particular kind are met the win is awarded. Even that reels are virtual symbols are pre-ordered, using combinatorial approaches [3], by the mathematician who was responsible for the game design.

In almost all cases the gambling games are mathematically unfair. It means that in long term players are losing against the game operator. The rate of this loss is measured with return to player (RTP) percentage. In many countries where gambling is legalized the RTP is between 90% and 98%. Of course, there are exceptions like Nevada where RTP can be much lower. The RTP has a statistical meaning. If the player bets 100 dollars on a game with 95% RTP it means that statistically in a single run he/she will get back 95 dollars. The RTP of the game comes directly from the order of the symbols in the virtual reels. From a mathematical point of view, there is no reason the discrete distribution of the symbols on the reels to be unknown for the players. As it is very well known that gambling games are mathematically unfair and legal regulators are controlling strictly all gambling games, there will not be an advantage of the player if he/she knows what is the content of the virtual reels. The case is similar to the roulette where the order of the numbers and their colors are perfectly known to the players. However, slot machines are covered with additional mystery by the fact that virtual reels are not published in the game rules. If someone wants to estimate game's RTP without access to the original game source code there is no other way except reels sequences reconstruction from the observed chunks [4]. Such a sequence reconstruction task can be time-consuming because of its high-combinatorial nature [5].

Generally, in most of the sequencing problems, the final goal is an exact reconstruction of the analyzed sequence [6]. In the case of the virtual slot machines reels [7], the exact reconstruction is not mandatory. It is enough reels to be reconstructed in such a manner that the subjective feeling of the player is identical for the original and the reconstructed reels. This research proposes approximate slot machine reels reconstruction with genetic algorithms. The optimality of the solutions provided by the genetic algorithm is estimated by Euclidean distance between the chunks of the candidate solution and the chunks from the original sequences. The final goal of this sequencing is reconstructed virtual reels with identical properties as the original even that there will not be an exact match between the reconstructed and the original sequences.

After the introductory part this paper continues as follows: The second section describes the mechanism of genetic algorithm usage in sequencing problems. The

third section presents the experiments done and the results achieved. The last section concludes and some directions of further works are appointed.

2. Genetic Algorithms for Sequencing

Genetic algorithms are well known for the last three decades tool for global meta-heuristic optimization [8]. Genetic algorithms are applied to problems with higher dimensional solutions spaces with much greatest success than the exact numerical methods [9]. The optimization process relays on a set of candidate solutions [10]. This set is called population and the candidate solutions are called individuals or chromosomes. In most of the cases, the initial population is randomly generated [11], but starting with a set of known in advance suboptimal solutions is also possible. The optimization procedures in genetic algorithms are inspired from the ideas in natural evolution. At each epoch of artificial evolution, the population produces a new generation. The new generation is formed after the application of three basic recombination operators - selection, crossover, and mutation [12]. The selection operator is the base of genetic algorithm optimization convergence. The empirical expectation is that when better-fitted chromosomes are selected to produce offspring the offspring would be even better [13]. The crossover operator exchanges parts of two or more selected parents when the mutation operator does a random change in a single value of the chromosome produced after the crossover [14]. The best offspring individuals replace part of the population and this is the way in which the new generation is created. If elitism rule is applied a small amount of the best-found individuals can not be replaced and they do survive until the end of the optimization process.

Sequencing of slot machine virtual reels can be done with the exact reconstruction of the reels [15], but this is not needed because it is enough to achieve identical behavior of the game with the reconstructed reels in front of the players. When approximated reconstruction is applicable the process starts with collection of virtual reel chunks samples. In most cases, the length of the reel will not be known in advance. In such cases, statistical analysis should be done to estimate the number of samples that are needed for as better reconstruction as possible. Histogram of the chunks can reveal the appearance frequency of each observed chunk [16].

Chromosomes are encoded as candidate sequences formed from the set of the possible game symbols in the particular reel. From the candidate sequence, chunk samples are taken. The number of the samples taken is the same as the number taken from the original sequence. All estimations of the candidate solution quality are done according to these taken samples.

Estimation of the fitness value is done by calculation of average Euclidean distance for a sorted set of chunks in the candidate and the original as pairs. Sorting is done

in order for candidate chunks to correspond to original chunks when Euclidean distance is calculated between the pairs. Original chunks are sorted only once when they are collected as samples. The candidate chunks are sorted each time when the candidate sequence is changed (crossover and/or mutation). All distances between chunk pairs are summed and divided by chunks number in order for the average value to be offered as chromosome fitness. Average Euclidean distance is taken with a negative sign because as far is the candidate solution from the original as low is the chromosome fitness. By acceptance of this fitness value estimation, it means that fitness value of the original distanced with itself will be zero which is the highest possible fitness value.

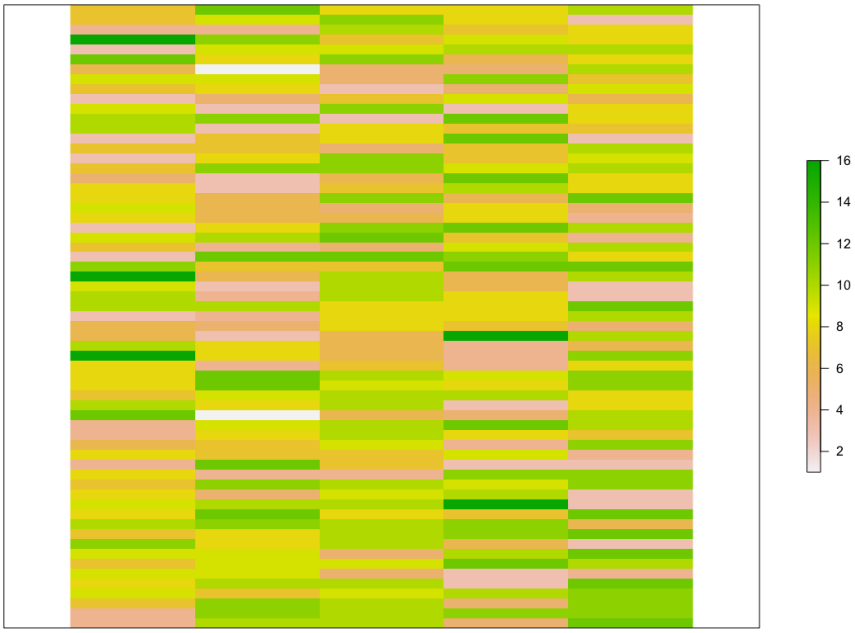


Fig. 1. Original sequences of five virtual reels

The fitness value estimation is done from the list of chunks, but crossover and mutation are done over candidate sequences. Such a change in the candidate sequences leads to an immediate recalculation of chunks samples. Modified uniform crossover is used with the application of normal distribution (mean 50% and a standard deviation of 20%) for the rate of participation of the two parents. It means that one of the parents has more influence in forming the offspring. Because the length of the original sequence is generally unknown the size of the offspring sequence length should be estimated during crossover process. As a lower bound of candidate sequence length,

the number of unique symbols in the virtual reels is taken. As an upper bound of candidate sequence length, the total length of all chunks taken together is taken. Such estimation of the candidate sequence length gives chance all symbols to participate (lower bound) and all chunks to be used in a single sequence (upper bound). Such an estimation of the length gives chance to the genetic algorithm to optimize this parameter too. Almost in all cases, parents are at different lengths. The offspring differs in length from parents too. When the offspring is longer than the parents, values are taken in a loop from the beginning. It is the natural way to produce longer offspring because virtual reels are used in looping during real-time game operation. The mutation is done by random replacement of a single value in the candidate sequence. The random value is taken from the chunks of the original sequence.

Fig. 2. Reconstructed sequences of five virtual reels

For the selection, three different chromosomes are selected randomly. Two of the three with better fitness are selected for parents. The third one is selected to be removed from the population and the newly generated offspring to take its place. This replacement is done only if the new offspring is better than the old one. With such a selection operator, the elitism rule is indirectly applied.

3. Experiments and Results

All experiments are done with a custom-created software project published in GitHub [16]. The source code is written in Java 8 as programming language. An empirical population of 137 individuals is chosen. The mutation rate is chosen to be 0.05% as a recommended value in the literature. Modified competitive selection is implemented where the crossover rate is almost 100%. Uniform crossover with a normally distributed participation threshold is chosen (mean of 50% and standard deviation of 20%). If the offspring is better than the worst of three individuals it takes its place and elitism rule is applied indirectly.

A set of 8 different virtual reels is used as input data of original sequences. Visual representation of one of them is presented in Fig. 1. Each of the five virtual reels is reconstructed separately. The result of the reconstructed reels is shown in Fig. 2. The genetic algorithm is stated for 100 generations. It is clearly visible that reconstructed sequences are pretty different in length than the original once. Fitness values for the five reconstructed reels are as follows: -3.521716216163723, -4.310895779885775, -2.3809034389618526, -3.8553857718048152, -3.6797186696978517.

In current experiments, only a single run of genetic algorithm is done, but many consecutive runs (simulated annealing as an analogy) will improve the achieved optimality because genetic algorithms can be stuck in local optimums.

4. Conclusion

The presented research is an approximate sequencing of slot machine gambling games virtual reels with genetic algorithms. The optimization process in reconstruction has as criteria of optimality the distance between the observed chunks from the candidate solution and the observed chunks from the original virtual reels. Results from the experiments clearly show that the proposed approach for approximate reconstruction can be efficiently used in the industrial production of slot machine gambling games. The main application is reverse engineering of the virtual reels. Such reverse engineering could be very useful in quality control and operational control of the games.

As for directions of further work, it will be an interesting discrete differential evolution to be used instead of genetic algorithms. In fitness estimation, Euclidean distance could be replaced with some other distance which can enforce faster optimization convergence. High-dimensional combinatorial optimization problems are usually time-consuming. In such situations, supercomputer systems as AVITOHOL [17] can be involved. A parallel implementation of population-based meta-heuristics is possible because of heir well known high degree of possible parallelism.

Acknowledgments

This research is funded by Velbazhd Software LLC and it is partially supported by the Bulgarian Ministry of Education and Science (contract D01–205/23.11.2018) under the National Scientific Program "Information and Communication Technologies for a Single Digital Market in Science, Education and Security (ICTinSES)", approved by DCM # 577/17.08.2018.

REFERENCES

1. Angelova, V.: Investigations in the Area of Soft Computing Targeted State of the Art Report, Cybernetics and Information Technologies, vol. 9, no. 1, 18-24, (2009).
2. Dineva, K., Atanasova, T.: Methodology for Data Processing in Modular IoT System, Distributed Computer and Communication Networks, Proceedings of 22-st International Conference, Springer Nature Switzerland, vol. 11965, 457-468, (2019).
3. Borissova, D., Mustakerov, I.: Open job shop scheduling via enumerative combinatorics, International Journal of Mathematical Models and Methods in Applied Sciences, vol. 9, 120-127, (2015).
4. Vaidyanathan, P.P., Phoong, S.M.: Reconstruction of Sequences from Nonuniform Samples, Proceedings of International Symposium on Circuits and Systems, vol. 1, 601-604, (1995).
5. Lewis, P.O.: A Genetic Algorithm for Maximum-Likelihood Phylogeny Inference Using Nucleotide Sequence Data, Molecular Biology and Evolution, vol. 15, no. 3, 277-283, (1998).
6. Parsons, R., Johnson, M.: A Case Study in Experimental Design Applied to Genetic Algorithms with Applications to DNA Sequence Assembly, American Journal of Mathematical and Management Sciences, vol. 17, no. 3-4, 369-396, (1997).
7. Tomov, P., Zankinski, I., Balabanov, T.: Slot Machine Reels Reconstruction with Monte-Carlo Search, Proceedings of International Scientific Conference UniTech, vol. 2, 384-387, (2017).
8. Balabanov, T, Ivanov, S. Ketipov, R.: Solving Combinatorial Puzzles with Parallel Evolutionary Algorithms, Proceedings of International Conference on Large-Scale Scientific Computing, Lecture Notes in Computer Science, vol. 11958, 493-500, (2020).
9. Tomov, P., Zankinski, I., Balabanov, T.: Genetic Algorithm Selection Operator Based on Recursion and Brute-Force, Abstracts of Annual Meeting of the Bulgarian Section of SIAM, Fastumprint, 93-93, (2019).

10. Balabanov, T., Sevova, J, Kolev, K.: Optimization of String Rewriting Operations for 3D Fractal Generation with Genetic Algorithms, Proceedings of International Conference on Numerical Methods and Applications, Lecture Notes in Computer Science, vol. 11189, 48-54, (2019).
11. Balabanov, T, Barova, M., Keremedchiev, D.: Image Construction with 2D Ellipses by Genetic Algorithms Optimization, Abstracts of Annual Meeting of the Bulgarian Section of SIAM, Fastumprint, 10-11, (2016).
12. Balabanov, T, Zankinski, I., Barova, M.: Strategy for Individuals Distribution by Incident Nodes Participation in Star Topology of Distributed Evolutionary Algorithms, Cybernetics and Information Technologies, vol. 16, no. 1, 80-88, (2016).
13. Balabanov, T, Zankinski, I., Dobrinkova, N.: Time Series Prediction by Artificial Neural Networks and Differential Evolution in Distributed Environment, Proceedings of International Conference on Large-Scale Scientific Computing, Lecture Notes in Computer Science, vol. 7116, 198-205, (2011).
14. Balabanov, T, Zankinski, I., Shumanov, B.: Slot Machine RTP Optimization and Symbols Wins Equalization with Discrete Differential Evolution, Proceedings of International Conference on Large-Scale Scientific Computing, Lecture Notes in Computer Science, vol. 9374, 210-217, (2015).
15. Tomov, P., Zankinski, I., Balabanov, T: Slot Machine Reels Reconstruction with Genetic Algorithms, Abstracts of Annual Meeting of the Bulgarian Section of SIAM, Fastumprint, 102-103, (2017).
16. Balabanov, T.: Approximated Sequences Reconstruction with Genetic Algorithms, https://github.com/TodorBalabanov/Approximated-Sequences-Reconstruction-with-Genetic-Algorithms
17. Tashev, T., Tasheva, R., Petrov, P.: Determination of the Computer Modelling Precision for Throughput of Switch Node with LPF-algorithm, Proceedings of International Conference on Computer Systems and Technologies, ACM, 141-145, (2019).

UDC: 519.248

On different approaches to study a double redundant renewable system under Marshall-Olkin failure model

Boyan Dimitrov[1], Vladimir Rykov[2], Sahib Esa[1,3]

[1]Department of Mathematics, Kettering University, MI USA
[2]Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation
[3]University of Kurdistan, Erbil, IQ

bdimitro@kettering.edu, rykov-vv@rudn.ru, sahib_esa@yahoo.com

Abstract

We present different approaches to analyze a dynamic Marshal-Olkin reliability model with dependent components functioning in parallel.

Keywords: Marshal-Olkin model in dynamics, Equilibrium balances, Kolmogorov equations, Probability interpretation of probability transformations.

1. Introduction

In the paper three approaches of the model, remind in the title, are considered. The first approach is based on detailed probability analysis of time dependent instant passages between the states of the process at a given moment.

The second approach uses the probability meaning of Laplace-Stieltjes transformation and of the probability generating functions to derive direct relationships between these find them explicitly.

The third approach uses that the input flow in each state must be equal to the respective output flow.

Materials of the paper are based on references [1]–[14].

2. The Marshal-Olkin dynamic model

In 1967 Marshall and Olkin [10] proposed a bivariate distribution, henceforth (MO), with dependent components, defined via three independent Poisson processes, which represent three types of shocks: individual to each component and commons to both.

The publication has been prepared with the support of the "RUDN University Program 5-100" and funded by RFBR according to the research project No.20-01-00575A.

Consider a heterogeneous two-component redundant hot standby renewable system, wherein components work in parallel and fail according to the original MO model, but are repaired according to description below. For lifetimes T_1 and T_2, the MO model is specified by the representation

$$(T_1, T_2) = (\min(A_1, A_3), \min(A_2, A_3)), \tag{1}$$

where non-negative continuous random variables A_1 and A_2 are the times to occurrence of independent "individual risk strikes" affecting individually each of the two devices working in parallel. The first risk strike affects only the first component, the second one affects only the second one, while the third type of risk strike represents the time to occurrence of the "common failure" A_3 that affects both components simultaneously, or just the working one, and leads to the failure of the entire system in any case. It is supposed that the risk strikes are governed by independent homogeneous Poisson processes, i.e., A_i's in (1) are exponentially distributed with parameters α_i $(i = 1, 2, 3)$.

About renovation it is assumed that after a partial failure (when only one component say i, fails) the repair of type i, with random duration B_i $(i = 1, 2)$ begins. This means that the system continues to function with the one working component. After a complete system failure a repair of the whole system (both components) begins, and lasts some random time, say B_3. It is assumed that the repair times B_k $(k = 1, 2, 3)$ have cumulative distribution functions (CDF) $B_k(x)$ $(k = 1, 2, 3)$ respectively. All repair times are assumed independent from the other random duration.

The system state space can be represented by $E = \{E_0, E_1, E_2, E_3\}$, where E_0 means that both components are working; E_i, $(i = 1, 2)$ shows that the i-th component is being repaired, and the other one is working; E_3 says that both components are in down states, the system has failed and is being repaired. To describe the system's behavior we introduce a random process $\{J(t),\ t \geq 0\}$ which takes values in the phase space E, such that

$$J(t) = j, \quad \text{if at time } t \text{ the system is in state } E_j \ (j = 0, 1, 2, 3).$$

Further, the following notations are used:

- $\alpha = \alpha_1 + \alpha_2 + \alpha_3$ is the summary risk intensity of the system failure;
- $b_k = \int_0^\infty x\, dB_k(x))$, $(k = 1, 2, 3)$ are the mean repair time of components and of the whole system;
- $\beta_k(s) = \int_0^\infty e^{-sx} dB_k(x)$ $(k = 1, 2, 3)$ is the LST of the repair time c.d.f. of components and the whole system;
- $T = \inf\{t :\ J(t) = 3\}$ is the system lifetime;

- W is the system life cycle which represents the portion of time interval when the system starts after a whole repair or both components being working, and ends with the complete repair of the whole system. Its LST is denoted by $\omega(s)$;

- $\tilde{b}_k(x) = (1 - B_k(x))^{-1} b_k(x)$ $(k = 1, 2, 3)$ is hazard rate function of components and the whole system given that elapsed repair time is x;

- $F(t) = \mathbf{P}\{T \leq t\}$ and $\tilde{f}(s)$ its LST;

- also to shorter some formulas the following notation is used

$$\phi_i(s) = \alpha_i \beta_i(s + \hat{\alpha}_{i^*}), \quad \text{with} \qquad \hat{\alpha}_i = \alpha_i + \alpha_3 \text{ and } i^* = (i+1)|_{mod 2}, \ (i = 1, 2)$$
$$\psi(s) = \phi_1(s) + \phi_2(s). \tag{2}$$

3. Kolmogorov equations in detailed time analysis

For the system behavior study the method of *additional variable* will be used. It consists of introducing an additional variables in order to describe the system's behavior via a Markov processes. In the case considered here, we use as such additional variable the time, spent by the state component in its J-th state subject to its last entry in it (the so-called elapsed time). We thus consider a two-dimensional Markov process $Z = \{Z(t), t \geq 0\}$, with $Z(t) = (J(t), X(t))$ where $J(t)$ is the system state at time t, and $X(t)$ represents the elapsed time of the process in the $J(t)$-th state after its last entering in it. The process phase space is given by $\mathcal{E} = \{0, (1, x), (2, x), (3, x)\}$. Corresponding probabilities (densities with respect to additional variables) are denoted by $\pi_0(t)$, $\pi_1(t; x)$, $\pi_2(t; x)$, $\pi_3(t; x)$ and we will refer to them as to the process (and the system) *micro-state* probabilities. The probabilities $\pi_i(t) = \mathbf{P}\{J(t) = j\}$ $(j = 0, 1, 2, 3)$ are called as *macro-state* process (and system) probabilities.

To calculate the time dependent system state probabilities during its life cycle the Markov process Z with absorbing state 3 should be used. Under the above assumptions, the following statement in [14] has been proved.

Theorem 1. The LT $\tilde{\pi}_i(s)$ of the time dependent system state probabilities $\pi_i(t)$, $(i = 0, 1, 2\}$ and LT $\tilde{R}(s)$ of the reliability function $R(t)$ for the considered

system are

$$\begin{aligned}
\tilde{\pi}_0(s) &= \frac{1}{s+\alpha_3+\psi(s)},\\
\tilde{\pi}_1(s) &= \frac{\phi_1(s)}{(s+\hat{\alpha}_2)(s+\alpha_3+\psi(s))},\ \ \tilde{\pi}_2(s) = \frac{\phi_2(s)}{(s+\hat{\alpha}_1)(s+\alpha_3+\psi(s))}.\\
\tilde{\pi}_3(s) &= \frac{\hat{\alpha}_1(s+\hat{\alpha}_2)\phi_2(s)+\hat{\alpha}_2(s+\hat{\alpha}_1)\phi_1(s)+\alpha_3(s+\hat{\alpha}_1)(s+\hat{\alpha}_2)}{s(s+\hat{\alpha}_1)(s+\hat{\alpha}_2)(s+\alpha_3+\psi(s))},\\
\tilde{R}(s) &= \frac{(s+\hat{\alpha}_1)(s+\hat{\alpha}_2)+(s+\hat{\alpha}_2)\phi_1(s)+(s+\hat{\alpha}_1)\phi_2(s)}{(s+\hat{\alpha}_1)(s+\hat{\alpha}_2)(s+\alpha_3+\psi(s))}, \qquad (3)
\end{aligned}$$

where notations above are used. □

By a substitution $s = 0$ one can find the mean time to the system failure.

Corollary 1. The mean system life time with the help of notations (2) can be represented as follows:

$$\mathbf{E}[T] = \tilde{R}(0) = \frac{\hat{\alpha}_1\hat{\alpha}_2+\alpha_1\hat{\alpha}_2(1-\beta_1(\alpha_2))+\alpha_2\hat{\alpha}_1(1-\beta_2(\alpha_1))}{\hat{\alpha}_1\hat{\alpha}_2(\alpha_3+\alpha_1(1-\beta_1(\alpha_2))+\alpha_2(1-\beta_2(\alpha_1))}. \qquad (4)$$

In [14] these results are used in sensitivity analysis of this system.

4. Probability interpretations of generating functions

The system-level characteristics in terms of its Laplace-Stieltjes transform (LST) for this model are derived, by use of probability meaning of the LSTs and avoiding cumbersome analytic mathematical details.

4.1. Life cycle and system life time. Since every life cycle W consists of a system work portion of time T and ends with next system repair time B_3, a repair type 3, it is true that $W = T + B_3$, and T and B_3 are independent. Using this fact it is possible to prove:

Lemma 1. The LST $\omega(s) = Ee\left[e^{-sW}\right]$ of life cycle W has a close form representation

$$\omega(s) = \frac{\alpha_3+[\alpha_1\frac{\hat{\alpha}_2}{\hat{\alpha}_2+s}(1-\beta_i(s+\alpha_2))+\alpha_2\frac{\hat{\alpha}_1}{\hat{\alpha}_1+s}(1-\beta_2(s+\alpha_1))]\beta_3(s)}{s+\alpha_3+\alpha_1(1-\beta_1(s+\hat{\alpha}_2))+\alpha_2(1-\beta_2(s+\hat{\alpha}_1))}.$$

From here the LST $\tau(s)$ of the system life time T simply can be found and as a result the mean work time $E(T)$ is represented in the following Corollary.

Corollary 2. The mean work time $E(T)$ of the system during a cycle is determined by the expression

$$E(T) = \frac{1+\frac{\alpha_1}{\alpha_2+\alpha_3}[1-\beta_1(\alpha_2+\alpha_3)]+\frac{\alpha_2}{\alpha_2+\alpha_3}[1-\beta_2(\alpha_1+\alpha_3]}{\alpha-\alpha_1\beta_1(\alpha_2+\alpha_3)-\alpha_2\beta_2(\alpha_1+\alpha_3)}.$$

The steady state system probabilities represented in the following theorem.

Theorem 2. If $\alpha_i > 0,\ (i = 1, 2, 3)$ and $0 < b_3 < \infty$, then the process is stable, and the macro state stationary probabilities

$$\lim_{t\to\infty} P(E_0 \cup E_1 \cup E_2, t) = \frac{E(T)}{E(T) + b_3},$$

and

$$\lim_{t\to\infty} P(E_3, t) = \frac{b_3}{E(T) + b_3}$$

do exist for any distributions of the repair times B_i $(i = 1, 2)$.

4.2. Number of passages between the states during a cycle. To study the number of changes between the states during a cycle of the system we use another probability interpretation of the probability generating functions together with the LST when changes occur. Again, we use the probability meaning of the PGF's combined with the LST, as referred above to the monograph of Gnedenko et al. [5].

Introduce the random variables (Symbol # means "counts in the set")

$$N_i = \#(passages\ into\ E_i\ during\ a\ cycle)$$

and denote by $\omega(\vec{z}, s)$ their joint with system life cycle W generating function,

$$\omega(\vec{z}, s) = E\left(z_0^{N_0} z_1^{N_1} z_2^{N_2} z_3 N_3 e^{-sW}\right)$$

Notice that

$$\omega(\vec{1}, s) = \omega(s) \ \text{ and } \ \omega(\vec{z}, 0) = \omega(z_0, z_1, z_2, z_3) \tag{5}$$

are the LST of the cycle duration, and the PGF of the number of passages a cycle correspondingly. It is true:

Lemma 2. The function $\omega(\vec{z}, s)$ is solution of the equation

$$\begin{aligned}
\omega(\vec{z}, s) &= \frac{\alpha_3}{\alpha + s} z_3 \beta_3(s) + \\
&+ \frac{\alpha_1}{\alpha + s} z_1 \beta_1(s + \hat{\alpha}_2) z_0 \omega(\vec{z}, s) + \\
&+ \frac{\alpha_2}{\alpha + s} z_2 \beta_2(s + \hat{\alpha}_1) z_0 \omega(\vec{z}, s) + \\
&+ \frac{\alpha_1}{\alpha + s} z_1 \frac{\alpha_2 z_2 + \alpha_3}{\hat{\alpha}_2 + s} [1 - \beta_1(s + \alpha_2 + \alpha_3)] z_3 \beta_3(s) \\
&+ \frac{\alpha_2}{\alpha + s} z_2 \frac{\alpha_1 z_1 + \alpha_3}{\hat{\alpha}_1 + s} [1 - \beta_2(s + \alpha_1 + \alpha_3)] z_3 \beta_3(s)
\end{aligned} \tag{6}$$

Corollary 3. The PGF of the number of passages in a cycle $\omega(z_0, z_1, z_2, z_3)$ is determined by the equation

$$\omega(\vec{z}) = \frac{\alpha_3 z_3 + \alpha_1 z_1 \frac{\alpha_2 z_2 + \alpha_3}{\hat{\alpha}_2}[1 - \beta_1(\hat{\alpha}_2)]z_3 + \alpha_2 z_2 \frac{\alpha_1 z_1 + \alpha_3}{\hat{\alpha}_1}[1 - \beta_2(\hat{\alpha}_1)]z_3}{\alpha_3 + \alpha_1[1 - z_1\beta_1(\hat{\alpha}_2)z_0] + \alpha_2 z_2[1 - \beta_2(\hat{\alpha}_1)z_0]}.$$

By partial derivations the average number of visits in each state during a cycle equals has been found.

4.3. Sojourn times during a life cycle. In this section the sojourn time in each state during the life cycle is calculated.

Theorem 3. (A0) The average sojourn time in state E_0 during a cycle equals

$$E(G_0) = \frac{\alpha_1\beta_1(\alpha_2 + \alpha_3) + \alpha_2\beta_2(\alpha_1 + \alpha_3)}{\alpha - \alpha_1\beta_1(\alpha_2 + \alpha_3) - \alpha_2\beta_2(\alpha_1 + \alpha_3)} \frac{1}{\alpha} \quad (i, j = 1, 2 \ i \neq j);$$

(Ai) The average sojourn time in state E_i during a cycle equals

$$\begin{aligned} E(G_i) &= \frac{\alpha_i + \alpha_j \frac{\alpha_i}{\alpha_i + \alpha_3}\beta_j(\alpha_i + \alpha_3)}{\alpha - \alpha_i\beta_i(\alpha_j + \alpha_3) - \alpha_j\beta_j(\alpha_i + \alpha_3)} \times \\ &\times \frac{1}{\alpha_j + \alpha_3}[1 - b_i(\alpha_j + \alpha_3)] \quad (i, j = 1, 2 \ i \neq j); \end{aligned}$$

(A3) The average sojourn time in state E_3 during a cycle equals

$$E(G_3) = E(B_3) = b_3.$$

An interesting dissection could be found if you compare the result of Corollary 2 and the last theorem. It must be true

$$E(T) = E(G_0) + E(G_1) + E(G_2),$$

since both expressions represent the work time on average during a life cycle.

4.4. Stationary probabilities. The transitions between the macro states E_i in the considered process form a Markov chain with finite number of states. According the theory (Feller, [4]). such chains always have stationary state and the stationary probabilities do exist. Namely, if $\pi_i(t)$ are the probabilities at the instant t the process

$$\pi_i = \lim_{t\to\infty} \pi_i(t), \quad (i = 0, 1, 2, 3)$$

are the stationary ones. We do not focus on the time dependent probabilities $\pi_i(t)$, but use the meaning of the stationary probabilities π_i. These are the portions of time in one unit of time, when the process spends in the state E_i, no matter how many times the process changes its states. Hence

Theorem 4. (P0) The Stationary probability to find the process in state E_0 when both components are functioning is

$$\pi_0 = \frac{E(G_0)}{E(T) + E(B_3)}$$

(Pi) The Stationary probability to find the process in state $E_i \quad i = 1, 2$ when only component i is functioning is

$$\pi_i = \frac{E(G_i)}{E(T) + E(B_3)}, \quad i = 1, 2;$$

(P3) The Stationary probability to find the process in state E_3 when both components 1 and 2 are not functioning, and the whole system is under repair is

$$\pi_3 = \frac{E(B_3)}{E(T) + E(B_3)}$$

where $E(G_i)$ and $E(T)$ are determined by the expressions in Theorem 3 and Corollary 2.

5. Stationary equilibrium equations approach

Using the input-output intensities for each of the four states, for our MO dynamic process we get the equilibrium equations, which should be completed by the normalizing equation.

Due to limitations in the size of the article, we leave details on this section for the full text that follows later.

6. Conclusion

The three discussed approaches produce equivalent results in regard the stationary probabilities. However, each approach offers different details about the behaviour of the progress and allow the use of these details for studying various process characteristics. If one is interested just in the steady state relationships, maybe then third approach is sufficient.

When non-stationary behaviour is important, especially in process's control, we would recommend first approach.

If one likes to find extra details within a process cycle (between two points of regeneration), then probably the Second approach should be preferred.

Our detailed discussion on MO dynamic model are new and we are glad to have the opportunity to present it here.

In our opinion, these approaches can be successfully applied in studying n-component systems with various modifications of the Marshal-Olkin type of maintenance models with renewals, as well as in modeling of k-out-of-n reliability systems under similar to our assumptions.

REFERENCES

1. Barlow, R.E., and Proshan, F. *Statistical theory of reliability and life testing: Probability models (To Begin With).* 1981 Silver Spring, MD.
2. Bocharov P.P., D'Apice C. and Pechinkin A.V. *Queueing Theory (Modern Probability and Statistics).* 2001. De Gruiter.
3. Dimitrov B. (1984) Asymptotic expansions of characteristics for queuing systems of the type M/G/1., in *Bulletin de l'Inst. de Math., Acad. Bulg. Sci.* 1984. v. XV. PP. 237-263 (in Bulgarian).
4. Feller W. *An Introduction to Probability Theory and its Applications,* Vol.II. 1966. John Wiley & Sons Inc. NY, London, Sydney.
5. Gnedenko B., Danielyan E., Klimov G., Matveev V. and Dimitrov B. (1973) *Prioritetnye Sistemy Obslujivania.* 1973. Moscow State University.
6. Kesten, H. and Runnenburg J T. Priority in Waiting Line Problems. 1957. vol. 60.*Koninklijke Netherlands Akademie van* . PP. 312–336.
7. Klimov G. P. *Stochastic Queuing Systems.* 1966. Nauka, Moscow (in Russian).
8. Li X., and Pellerey F. Generalized Marshall-Olkin distributions and related bivariate aging properties. *J. of Multivariate Analysis,* 2011. V.102. PP. 1399-1409.
9. Lin J., and Li X. Multivariate generalized Marshall-Olkin distributions and copulas. *Methodology and Computing in Applied Probability.* 2014. V.16. PP. 53-78.
10. Marshall A. , and Olkin I. (1967) A multivariate exponential distribution. *Journal of American Statistical Association.* 1967. V.62. PP. 30-44.
11. Omey E. and Willenkens E. (1989) Abelian and Tauberian Theorems for the Laplace Transform of Functions in Several Variables. *J.of Multivar.Annalysis,* V. 30. PP. 292-306.
12. Pakes A. G. (1969), Some Conditions for Ergodicity and Recurrence of Markov Chains. *Operations Research,* V. 17. PP.1048–1061.
13. I. G. Petrovsky. Lectures on the theory of ordinary differential equations. 1951. M.-L.: GITTL. 1952. 232p. (in Russian). .
14. Rykov V.V.and Dimitrov B. (2019) Renewal Redundant Systems Under the Marshall-Olkin Failure Model. Sensitivity Analysis. In: *Distributed Computer and Communication Networks,* Eds. V. Vishnevskiy, K. Samouylov and D. Kozyrev Proc. 22nd Intl Conf. DCCN 2019 Moscow, Russia, Sept. 23–27. 2019. Springer, LNCS 11965. PP. 234 - 248

UDC: 519.21

Ergodicity of generalized Markov modulated Poisson 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]Russian university of transport (MIIT), 9b9 Obrazcova Ulitsa, Moscow 127994, Russia

Abstract

A broader generalization of the Markov modulated Poisson process is considered – in comparison with the paper [4]. A method for proving the ergodicity of this process is described. Also, an algorithm for obtaining a strong upper bound for the rate of convergence in the total variation metric is given.

Keywords: generalization of Markov modulated Poisson processes, convergence rate, coupling method, generalized Lorden's inequality, successful coupling and it's modification

1. Introduction

In the paper [4], a simple generalization of Markov modulated Poisson process was considered. This was the situation where all intensities of the Markov Poisson processes involved in the composite process were bounded above and below by positive constants. Now, we consider more more difficult situation.

The main tool for studying such processes is a generalized Lorden's inequality.

Definition [Regenerative Process] *A random process is called regenerative if there exists an increasing sequence* $\{t_i\}_{i=0,1,2,\ldots}$*, such, that the random elements* $\Theta_i \stackrel{def}{=} \{X_t,\, t \in [t_{i-1}; t_i]\}$ *are i.i.d.* $\forall\, i = 1, 2, \ldots$.

Times t_i *are named regeneration times.*

Denote $\tau_i \stackrel{def}{=} t_{i+1} - t_i$*, and let* $\mathcal{P}_t$ *be a distribution of regenerative process at the time* t. ▷

It is well-known that:

1. If $\mathbf{E}\,\tau_i < C < \infty$, then the regenerative process is ergodic, i.e. there exists the probability distribution $\mathcal{P}$, such that $\mathcal{P}_t \Longrightarrow \mathcal{P}$.

2. If $\mathbf{E}\,(\tau_i)^k < \infty$, then $\exists\, K(\mathcal{P}_0):\quad \|\mathcal{P}_t - \mathcal{P}\|_{TV} \le \frac{K(\mathcal{P}_0)}{t^{k-1}}$.

3. If $\mathbf{E}\exp(\alpha\tau_i) < \infty$, then $\forall\ \beta < \alpha,\ \exists\ K(\mathcal{P}_0,\beta)$: $\|\mathcal{P}_t - \mathcal{P}\|_{TV} \le K(\mathcal{P}_0,\beta)\exp(-\beta t)$.

These results are the classic results, but they make it impossible to estimate the value K – see, e.g., [8, 12, 9] et all. The general method for obtain an upper bounds for the constant K for regenerative processes was invented in [3].

1.1. Some information of the coupling method. Consider two *independent Markov processes* with different initial states X_0 and $\widehat{X}_0$ and with the same transition function. Denote these processes by X_t and $\widehat{X}_t$ correspondingly. Let we can find the time τ where they are coincided. The time τ is called *coupling epoch* and it depends on X_0 and $\widehat{X}_0$. After the time $\tau(X_0,\widehat{X}_0)$, the distributions of the processes X_0 and $\widehat{X}_0$ are coincided – by Markov property. Thus, for all $t \ge \tau(X_0,\widehat{X}_0)$, and for all set $\mathcal{A}\in\sigma(\mathcal{X})$, $\mathbf{P}\{X_t\in\mathcal{A}\} = \mathbf{P}\{\widehat{X}_t\in\mathcal{A}\}$. It implies the basic coupling inequality:

$$\begin{aligned}|\mathbf{P}\{X_t\in\mathcal{A}\} - \mathbf{P}\{\widehat{X}_t\in\mathcal{A}\}| &= |\mathbf{P}\{X_t\in\mathcal{A}\ \&\ \tau>t\} - \mathbf{P}\{\widehat{X}_t\in\mathcal{A}\ \&\ \tau>t\} + \\ &+|\mathbf{P}\{X_t\in\mathcal{A}\ \&\ \tau\le t\} - \mathbf{P}\{\widehat{X}_t\in\mathcal{A}\ \&\ \tau\le t\}| = \\ &= |\mathbf{P}\{X_t\in\mathcal{A}\ \&\ \tau>t\} - \mathbf{P}\{\widehat{X}_t\in\mathcal{A}\ \&\ \tau>t\}| \le \mathbf{P}\{\tau>t\}.\end{aligned}$$

Then, if it possible to find the increasing positive function $\varphi(\tau)$ such that $\mathbf{E}\,\varphi(\tau(X_0,\widehat{X}_0)) < \infty$, then by Markov inequality, $\mathbf{P}\{\tau(X_0,\widehat{X}_0) \ge t\} = \mathbf{P}\{\varphi(\tau(X_0,\widehat{X}_0)) \ge \varphi(t)\} \le \frac{\mathbf{E}\varphi(\tau(X_0,\widehat{X}_0))}{\varphi(t)}$. From the last inequality the bounds for convergence of the distribution $\mathcal{P}_t$ can be obtained, namely.

If the process $\widehat{X}$ starts from the stationary distribution $\mathcal{P}$ of X_t, i.e. at any time, the distribution of $\widehat{X}_t$ is the same as the one of $\widehat{X}_0$, then

$|\mathbf{P}\{X_t\in\mathcal{A}\} - \mathcal{P}(A)\}| \le \frac{\int\varphi(\tau(X_0,\widehat{X}_0))\mathcal{P}(\,\mathrm{d}\widehat{X}_0)}{\varphi(t)}$.

This schema can be used for discrete Markov chain and for Markov chain in continuous time. But for the process X_t in continuous time, the "direct" coupling method is impossible, because for different values $X_0\ne\widehat{X}_0$, $\mathbf{P}\{\tau(X_0,\widehat{X}_0)<\infty\} = 0$. Thus, the modification of coupling method, or *successful coupling* will be used.

1.2. Successful coupling (see [10]). Let X_t and $\widehat{X}_t$ be two independent Markov processes with the same transition function, but with different initial states at time $t=0$.

Suppose that (*dependent*) processes Y_t and $\widehat{Y}_t$ are constructed on some probability space, in such a way that:

1. $Y_t \stackrel{\mathcal{D}}{=} X_t$ and $\widehat{Y}_t \stackrel{\mathcal{D}}{=} \widehat{X}_t$ for all *non-random* t;

2. $\mathbf{P}\{\tau(X_0,\widehat{X}_0)<\infty\} = 1$, where $\tau(X_0,\widehat{X}_0) = \tau(Y_0,\widehat{Y}_0) = \inf\{t>0:\ Y_t = \widehat{Y}_t\}$.

This pair of processes Y_t and $\widehat{Y}_t$ is called *successful coupling* for the processes X_t and $\widehat{Y}_t$, and $\tau(X_0,\widehat{X}_0)$ is called *coupling epoch.*

For successful coupling, the basic coupling inequality can be applied as:

$$\begin{aligned}|\mathbf{P}\{X_t \in \mathcal{A}\} - \mathbf{P}\{\widehat{X}_t \in \mathcal{A}\}| = |\mathbf{P}\{Y_t \in \mathcal{A}\} - \mathbf{P}\{\widehat{Y}_t \in \mathcal{A}\}| = \\ = |\mathbf{P}\{Y_t \in \mathcal{A}\ \&\ \tau > t\} - \mathbf{P}\{\widehat{Y}_t \in \mathcal{A}\ \&\ \tau > t\} + \\ + |\mathbf{P}\{Y_t \in \mathcal{A}\ \&\ \tau \le t\} - \mathbf{P}\{\widehat{Y}_t \in \mathcal{A}\ \&\ \tau \le t\}| = \\ = |\mathbf{P}\{Y_t \in \mathcal{A}\ \&\ \tau > t\} - \mathbf{P}\{\widehat{Y}_t \in \mathcal{A}\ \&\ \tau > t\}| \le \mathbf{P}\{\tau > t\}\end{aligned} \tag{1}$$

for any set $\mathcal{A} \in \sigma(\mathcal{X})$. Here, identical distribution of pairs $Y_t \stackrel{\mathcal{D}}{=} X_t$ and $\widehat{Y}_t \stackrel{\mathcal{D}}{=} \widehat{X}_t$ means only a coincidence of distributions in any time, but not the coincidence of finite-dimensional distributions of these processes.

Now, our goal is a construction of the successful coupling and an estimation of exponential moments of a random variable $\tau(X_0, \widehat{X}_0)$. For this construction the Basic Coupling Lemma is needed.

3. Basic Coupling Lemma (see, e.g., [11]). Here the simplest formulation of the Basic Coupling Lemma is given.

Lemma 1. If the random variable ϑ_1 and ϑ_2 have c.d.f. $\Phi_1(s)$ and $\Phi_2(s)$ correspondingly, and their common part $\kappa \stackrel{\text{def}}{=} \int\limits_{\mathbf{R}} \min\{\Phi_1'(s), \Phi_2'(s)\}\,\mathrm{d}\,s > 0$, then it can construct (on some probability space) the random variables $\widehat{\vartheta}_1$ and $\widehat{\vartheta}_2$ such, that

1. $\widehat{\vartheta}_1 \stackrel{\mathcal{D}}{=} \vartheta_1$, $\widehat{\vartheta}_2 \stackrel{\mathcal{D}}{=} \vartheta_2$; 2. $\mathbf{P}\{\widehat{\vartheta}_1 = \widehat{\vartheta}_2\} = \kappa$. ▷

The statement of Lemma 4 is naturally transferred to any finite number of random variables.

Lemma 2. Let ϑ_1, ϑ_2, ..., ϑ_n be the random variable with probability densities $\varphi_1(s)$, $\varphi_2(s)$, ..., $\varphi_n(s)$ correspondingly, and $\kappa \stackrel{\text{def}}{=} \int\limits_{\mathbf{R}} \min\limits_{i=1,\ldots,n} \{\varphi_i(s)\}\,\mathrm{d}\,s > 0$. Then on some probabilistic space it is possible to construct the random variables $\widehat{\vartheta}_1(s)$, $\widehat{\vartheta}_2(s)$, ..., $\widehat{\vartheta}_n(s)$ such that

1. $\widehat{\vartheta}_i \stackrel{\mathcal{D}}{=} \vartheta_i$, $i = 1, 2, \ldots n$; 2. $\mathbf{P}\{\widehat{\vartheta}_1 = \widehat{\vartheta}_2 = \ldots = \widehat{\vartheta}_n\} = \kappa$. ▷

Here we skip the proofs.

1.3. Lorden's inequality. Consider the renewal process $N_t \stackrel{\text{def}}{=} \sum\limits_{i=1}^{\infty} \mathbf{1}\{\sum\limits_{k=1}^{i} \xi_k \le t\}$, where $\{\xi_1, \xi_2, ...\}$ is the of i.i.d. positive random variables. N_t is a counting process which changes its value at the times $t_k = S_k \stackrel{\text{def}}{=} \sum\limits_{j=1}^{k} \xi_j$. The times t_k are the renewal times.

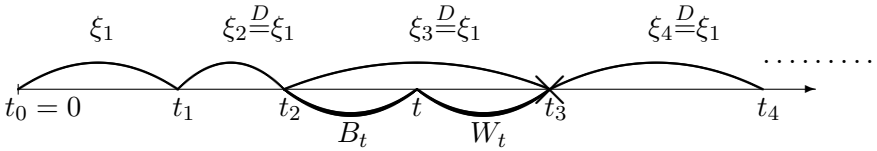


Fig. 1. B_t is a backward renewal time, and W_t is a forward renewal time at the fixed time t

In Fig.1, we see the *backward renewal time (or overshoot)* B_t, and the *forward renewal time (or undershot)* W_t at the **fixed** time t (See Fig.1):

$W_t \stackrel{\text{def}}{=} S_{N_t+1} - t = \sum\limits_{1}^{N_t+1} \xi_i - t; \qquad B_t \stackrel{\text{def}}{=} t - S_{N_t}.$

Theorem 1 (Lorden, G. (1970) [5]). Lorden's inequality states that the expectation of this overshoot is bounded as $\mathbf{E}\, B_t \le \frac{\mathbf{E}\,\xi^2}{\mathbf{E}\,\xi}$. ▷

These inequalities are the means for find the upper bounds for convergence rate in total variation metrics.

However, we need to use some generalization of the Lorden's inequality.

1.4. Generalized Lorden's inequality.

Some preliminary considerations. The random variables can be defined by distribution function, by its density, and by its *intensity*. Obviously, the intensity is a means for study the absolutely continuous distributions, and for d.f. $F(x)$ the intensity is equal $\lambda(x) = \frac{F'(x)}{1-F(x)}$.

This formula is correct:

$$F(x) = 1 - \exp\left(\int\limits_0^x -\lambda(s)\,\mathrm{d}s\right). \tag{2}$$

But we are interesting by mixed random variables, i.e. their d.f. can have jumps.

If $F(a-0) \ne F(a+0)$, the **we put** $\lambda(a) \stackrel{\text{def}}{=} -\ln\big(F(a+0) - F(a-0)\big)\delta(0)$, where $\delta(\cdot)$ is a "classic" δ-function. So, **we put**

$$f(s) = \begin{cases} F'(s), & \text{if } \exists\, F'(s); \\ 0, & \text{otherwise.} \end{cases}$$

Finally, (generalized) intensity is $\lambda(s) \stackrel{\text{def}}{=} \frac{f(s)}{1-F(s)} - \sum\limits_i \delta(s-a_i) \ln\big(F(a_i+0) - F(a_i-0)\big)$, where $\{a_i\}$ is a set of the discontinuous points of d.f. $F(s)$.

It is easy to see, that the formula (2) remains true for generalized intensity.

Let $Int_\xi(s)$ be an intensity of r.v. ξ.

Lemma 3. $Int_{\min\{\xi;\eta\}} = Int_\xi + Int_\eta$. ▷

Denotations and assumptions. Consider the sequence $\{\xi_1, \xi_2, ...\}$ of random variables.

Assumptions

1) $\xi_j = \min\{\zeta_j; \theta_j\}$, were $\{\zeta_j\}$ are i.i.d. r.v.'s defined by (generalized) intensities $\varphi_i(s)$, and $\zeta_i \perp\!\!\!\perp \theta_j$ for all i, j; θ_j are defined by (generalized) intensities $\mu_j(s)$;
2) There exists some (generalized) mesurable function $Q(s)$ such that for all $s \geq 0$, $\varphi(s) + \mu_j(s) = \lambda_i(s) \leq Q(s)$;
3) $\int\limits_0^\infty \varphi(s)\,\mathrm{d}s = \infty$, and $\int\limits_0^\infty \left(x^{k-1} \exp\left(-\int\limits_0^x \varphi(s)\,\mathrm{d}s \right) \right) \mathrm{d}x < \infty$ for some $k \geq 2$;
4) $Q(s)$ is bounded in some neighborhood of zero;
5) $\varphi(s) > 0$ a.s. for $s > T \geq 0$.

Definition. *If conditions 1–4 are satisfied, then the counting process*

$$N_t \stackrel{def}{=} \sum_{i=1}^{\infty} \mathbf{1}\{ \sum_{k=1}^{i} \xi_k \leq t\} \tag{3}$$

is named generalized renewal process. ▷

Remark 1. The condition 1 holds: the r.v.'s ξ_i and ξ_j are dependent, and this dependence is "weak" dependence in some sens. ▷

Remark 2. The conditions 3 and 4 hold: $\mathbf{E}\,\xi_i > 0, \qquad \mathrm{Var}\,\xi_i^2 > 0.$ ▷

Remark 3. The conditions 1 and 2 hold:

$$F_i(t) = 1 - \exp\left(\int\limits_0^t -\varphi_i(s)\,\mathrm{d}s \right) \geq 1 - \exp\left(\int\limits_0^t -Q(s)\,\mathrm{d}s \right) \;\Rightarrow\; \exists\, \mathbf{E}\,\xi_i^2 < \infty. \qquad \triangleright$$

Remark 4. The condition 5 reports that the renewal process under study is a delay renewal process, and a delay time does not exceed T. ▷

Theorem 2 (Generalized Lorden's inequality – see [7]). If the conditions 1–4 are satisfied, then for the process (3) the inequality

$$\mathbf{E}\,B_t \leq \mathbf{E}\,\eta + \frac{\mathbf{E}\,\eta^2}{2\mathbf{E}\,\zeta} = \Xi, \tag{4}$$

is thru, where $\mathbf{E}\,\eta^2 = \int\limits_0^\infty x^2\,\mathrm{d}\Phi(x)$; $\quad \mathbf{E}\,\zeta = \int\limits_0^\infty x^2\,\mathrm{d}G(x)$, and

$$G(x) = 1 - \exp\left(-\int\limits_0^x Q(t)\,\mathrm{d}t \right)\,\mathrm{d}s, \text{ and } \Phi(x) = 1 - \exp\left(-\int\limits_0^x \varphi(t)\,\mathrm{d}t \right)\,\mathrm{d}s. \triangleright$$

Remark 5. In the proof of this Theorem, there is an intermediate result:

If $\mathbf{E}\,\eta^k < \infty$, then for $\ell \in (0; k-1]$,

$$\mathbf{E}\,(B_t)^\ell = \int\limits_0^\infty \ell s^{\ell-1} \mathsf{P}\{B_t > s\}\,\mathrm{d}s = \int\limits_0^\infty \ell s^{\ell-1}(1-\Phi(s))\,\mathrm{d}s + \\ + \int\limits_0^\infty \ell s^{\ell-1} \left(\frac{1}{\mu} \int\limits_s^\infty 1-\Phi(u)\,\mathrm{d}u \right) \mathrm{d}s = \mathbf{E}\,\eta^\ell + \frac{\eta^{\ell+1}}{(\ell+1)\mathbf{E}\,\zeta}. \tag{5}$$

▷

2. Generalized Markov modulated Poisson processes (MMPP)

Consider n-dimensional Markov process on the state space $\mathbf{R}_+^n$. The state of this process $X_t = (x_t^{(1)}, x_t^{(2)}, x_t^{(3)}, \dots x_t^{(n)})$. If at the time t the component $x_t^{(i)}$ of X_t is equal to a: $x_t^{(i)} = a$, then:

1. with probability $1 - \lambda_t^{(i)}\Delta + o(\Delta)$, $x_{t+\Delta}^{(i)} = a + \Delta$;
2. with probability $\lambda_t^{(i)}\Delta + o(\Delta)$, $x_{t+\Delta}^{(i)} < \Delta$.

I.e. X_t is the multidimensional flow with intensity of i-th flow $\lambda_t^{(i)}$.

We suppose, that $\lambda_t^{(i)}$ depends of the full state $X_t = (x_t^{(1)}, x_t^{(2)}, x_t^{(3)}, \dots x_t^{(n)})$. So, these flows are dependent. But the arbitrary dependence does not allow to study the behaviour of the process X_t.

Hence, we suppose, that all intensities satisfy conditions 1–5.

In this case, we can prove the ergodicity of the process X_t and estimate its convergence rate.

3. Ergodicity of generalized MMPP

Consider some fixed positive number $\mathbf{T}$, and times $\mathbf{T}_i = i \times \mathbf{T}$. Suppose that the process X_t starts from the state $(0, 0, 0, \dots, 0)$.

At any times $\mathbf{T_i}$, the expectation of all parts of the process X_t does not exceed the constant Ξ. Let us fix some $\Theta > \Xi$. By Markov inequality, $\mathbf{P}\{$the backward renewal time of $x_{\mathbf{T_i}}^{(j)} < \Theta\} > 1 - \frac{\Xi}{\Theta}$ – this inequality is uniform by the numbers $1, 2, \dots, n$. Therefore, at any time $\mathbf{T_i}$, with probability greater then $(1 - \frac{\Xi}{\Theta})^n = \varkappa$, *all* backward renewal times less then Θ.

In this case, we use the method given in [3], and we can create the prolongations of the components of X_t by such a way that with probability greater then $\int\limits_0^\infty \inf\limits_{a_i<\Theta} \varphi(x + a_i)\,\mathrm{d}\,\mathrm{x} = \kappa > 0$ all components of X_t hit to zero at the same moment.

And this construction gives the process with the same marginal distribution as initial process X_t, similarly to the successful coupling method.

Therefore, we can create the "version" of X_t such that after any time $\mathbf{T_i}$ it hits to the state $(0, 0, \dots, 0)$ with probability greater then $\kappa\varkappa$, i.e. this "version" of X_t is regenerative. But original process X_t is not-regenerative! So, we call the process X_t quasi-regenerative process.

Hence, the process X_t is ergodic, and its stationary distribution can be found similarly to classic renewal theory (see [1]) – but by use distribution comparison:

$$\mathsf{P}\{\tilde{x}_t^{(i)} > x\} \le \Psi(x) \stackrel{\text{def}}{=} \frac{\int\limits_0^x (1 - \Phi(s))\,\mathrm{d}s}{\int\limits_0^\infty (1 - G(s))\,\mathrm{d}s}. \tag{6}$$

4. About convergence rate of generalized MMPP

For obtain an upper bounds for convergence rate of generalized MMPP, we consider two independent versions of studied process: X_t and Y_t.

For simplicity, put $X_0 = (0, 0, 0, \dots, 0)$. Put $Y_0 = (y_0^{(1)}, y_0^{(2)}, y_0^{(3)}, \dots, y_0^{(4)})$.

The times $\mathbf{T_i}$ we can use only after the time, when all components of Y_t will go to the point zero; this is $\max\{\text{residual times of} y_t^{(i)}\}$.

For simplicity, we can consider the sum $\mathcal{T}$ of these residual times. The distribution of residual times can be estimated by the initial states.

So, the first check of the backward renewal times we can do at the time $\mathbf{T}_i > \mathcal{T}$; for simplicity we begin the check at the times $\mathbf{T} + \mathcal{T}$, $2\mathbf{T} + \mathcal{T}$,..., $i \times \mathbf{T} + \mathcal{T}$,...

Here we use the coupling method similarly to [3] and method of successful coupling.

Then the result can be integrated by stationary distribution with known bounds (6).

5. Conclusion

The generalized MMPP describe the behaviour of complex reliability systems (see, e.g.,[6]) and great networks. So, knowing of its distribution and convergence rate is very useful.

Acknowledgments

This work is supported by a grant of the Russian Foundation for Basic Research project no. 20-01-00575_a.

REFERENCES

1. W. L. Smith, Renewal theory and its ramifications // J. Roy. Statist. Soc. Ser. B, 20:2 (1958), 243–302.
2. Zverkina G. On strong bounds of rate of convergence for regenerative processes // Communications in Computer and Information Science. 2016. 678, pp. 381–393.
3. Zverkina G. Lorden's inequality and coupling method for backward renewal process / Proceedings of XX International Conference on Distributed Computer and Communication Networks: Control, Computation, Communications (DCCN-2017, Moscow), 2017. pp. 484–491.
4. Zverkina G. On a generalization of Markov modulated Poisson flows (in russian) // XIII All-russian meeting on governance problems (VSPU-2019) Moscow June 17-20, 2019. P. 849–852
5. Lorden, G. (1970). "On Excess over the Boundary". The Annals of Mathematical Statistics. 41 (2): 520. doi:10.1214/aoms/1177697092. JSTOR 2239350
6. Zverkina G. A System with Warm Standby / Proceedings of the 26th International Conference "Computer Networks" (CN 2019, Kamień Śląski, Poland). Cham: Springer, 2019. P. 387–399.
7. Kalimulina E., Zverkina G. On some generalization of Lorden's inequality for renewal processes / arXiv.org. Cornell: Cornell university library, 2019. 1910.03381v1. P. 1–5.
8. Asmussen, S. Applied Probability and Queues. Second edition. New York: Springer-Verlag, 2003.
9. Gnedenko B. V., Kovalenko I.N., Introduction to Queuing Theory. Mathematical Modeling. Birkhaeuser Boston, Boston. 1989.
10. Griffeath, D. A maximal coupling for Markov chains // Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete — 1975 — Volume 31 — Issue 2, P. 95–106.
11. Kato, K. Coupling Lemma and Its Application to The Security Analysis of Quantum Key Distribution // Tamagawa University Quantum ICT Research Institute Bulletin Vol.4 No.1 : 23-30 (2014) P.23–30.
12. Thorisson, H. Coupling, Stationarity, and Regeneration. Springer, 2000.
13. Zverkina G. On strong bounds of rate of convergence for regenerative processes // Communications in Computer and Information Science, v.678, 2016, P.381–393.

UDC: 519.872

Asymptotic-Diffusion Analysis of Multiserver Retrial Queue with Two-Way Communication

A.A. Nazarov[1], T. Phung-Duc[2], S.V. Paul[1], O.D. Lizyura[1]

[1]Institute of Applied Mathematics and Computer Science, National Research Tomsk State University, 36 Lenina ave., Tomsk, 634050, Russia

[2]Faculty of Engineering Information and Systems, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan

nazarov.tsu@gmail.com, tuan@sk.tsukuba.ac.jp, paulsv82@mail.ru, oliztsu@mail.ru

Abstract

In this paper, we consider multiserver retrial queue with two-way communication. Incomimg calls arrive according to the Poisson process and reserve the server for an exponentially distributed time. If all of the servers are busy the incoming call joins the orbit and makes a delay for an exponentially distributed time before the next attempt to occupy the server. Idle servers also make outgoing calls following a distinct exponential distribution. Using the asymptotic-diffusion analysis method we derive the approximation for the stationary probability distribution of the number of calls in the orbit.

Keywords: multiserver retrial queue, two-way communication, incoming call, outgoing call, asymptotic-diffusion analysis, diffusion approximation

1. Introduction

Currently, more and more services are partially or fully working in a call-center mode. Banks use call centers to advise customers and to advertise their services; online stores utilize call centers to refine and confirm orders, and call centers also exist independently to conduct social surveys.

A lot of research papers are devoted to modeling call centers. Papers [1, 2, 3, 4] are devoted to the quality of customer service in telephone services. In [5], the authors present a study of incoming processes models in real call-centers. A statistical analysis of the work of call-centers taking into account various aspects of the functioning of the service is presented in the study [6].

Recently, retrial queues with two-way communication have been used as a model of a blended call-center. The most detailed studies on retrial queues can be found

The publication has been prepared with the support of RFBR according to the research project No.18-01-00277.

in [7, 8]. The main feature of model with two-way communication is that during idle time, the server makes outgoing calls and serves them along with incoming calls. As an outgoing call we refer to an operator's call to a client, or any other type of alternative operator's activities rather than serving incoming calls. Thus, retrial queueing models with two-way communication are flexible and allow you to simulate most modern telephone services. In [9] the model of multiserver retrial queue with two-way communication was was proposed and numerical analysis was presented.

In this paper, we consider the same model as in [9] but rather than numerical analysis of the stationary distribution, our focus is to obtain the diffusion limit of the underlying time-dependent Markov process [10]. As a byproduct, the limiting results are then used to approximate the probability distribution of the number of customers in the orbit in the stationary state.

The rest of our paper is organized as follows. Section 2 is devoted to the presentation of the model and preliminary analysis of the underlying Markov chain. Our main results are in Section 3 where we present the asymptotic-diffussion analysis to obtain the diffusion limit of the underlying Markov chain. In Section 4 we describe an algorithm for constructing a probability distribution approximation of the system state. Section 5 shows the diffusion approximation accuracy for several values of system parameters. Finally, Section 6 is devoted to some concluding remarks.

2. Mathematical model

We consider multiserver retrial queue with two-way communication. The input process is a stationary Poisson process with rate λ. Service times of incoming calls are exponentially distributed with rate μ_1. Calls that find servers fully occupied join the orbit and reattempt to access the server after an exponentially distributed delay with rate σ. When the server is idle it makes an outgoing calls with rate α and provides the service for an exponentially distributed time with rate μ_2. We denote N is the number of servers in the system.

Let $n_1(t), n_2(t)$ denote the number of servers busy serving incoming and outgoing calls at the time t, respectively. Also $i(t)$ is a number of calls in the orbit. Thus, we can see that three-dimensional random process $\{n_1(t),\ n_2(t),\ i(t)\}$ is a continuous time Markov chain.

Let $P(n_1, n_2, i, t) = P\{n_1(t) = n_1, n_2(t) = n_2, i(t) = i\}$ denotes the probability distribution of the process $\{n_1(t),\ n_2(t),\ i(t)\}$, which is the solution of Kolmogorov's system of equations

$$\frac{\partial P(0,0,i,t)}{\partial t} = -(\lambda + i\sigma + N\alpha)P(0,0,i,t) + \mu_1 P(1,0,i,t) + \mu_2 P(0,1,i,t),$$

$$\frac{\partial P(n_1,n_2,i,t)}{\partial t} = -(\lambda + i\sigma + (N-n_1-n_2)\alpha + n_1\mu_1 + n_2\mu_2)P(n_1,n_2,i,t) +$$

$$+\lambda P(n_1-1,n_2,i,t)+(i+1)\sigma P(n_1-1,n_2,i+1,t)+$$
$$+(N-n_1-n_2+1)\alpha P(n_1,n_2-1,i,t)+(n_1+1)\mu_1 P(n_1+1,n_2,i,t)+$$
$$+(n_2+1)\mu_2 P(n_1,n_2+1,i,t),\ 0<n_1+n_2<N,$$
$$\frac{\partial P(n_1,n_2,i,t)}{\partial t}=-(\lambda+n_1\mu_1+n_2\mu_2)P(n_1,n_2,i,t)+\lambda P(n_1-1,n_2,i,t)+$$
$$+\lambda P(n_1,n_2,i-1,t)+(i+1)\sigma P(n_1-1,n_2,i+1,t)+$$
$$+\alpha P(n_1,n_2-1,i,t),\ n_1+n_2=N. \quad (1)$$

Then we transform the system (1) into system for partial characteristic functions $H(n_1,n_2,u,t)=\sum\limits_{i=0}^{\infty}e^{jui}P(n_1,n_2,i,t)$, where $j=\sqrt{-1}$

$$\frac{\partial H(n_1,n_2,u,t)}{\partial t}=-(\lambda+N\alpha)H(n_1,n_2,u,t)+j\sigma\frac{\partial H(n_1,n_2,u,t)}{\partial u}+$$
$$+\mu_1 H(n_1+1,n_2,u,t)+\mu_2 H(n_1,n_2+1,u,t),\ n_1+n_2=0,$$
$$\frac{\partial H(n_1,n_2,u,t)}{\partial t}=-(\lambda+(N-n_1-n_2)\alpha+n_1\mu_1+n_2\mu_2)H(n_1,n_2,u,t)+$$
$$+j\sigma\frac{\partial H(n_1,n_2,u,t)}{\partial u}+\lambda H(n_1-1,n_2,u,t)-j\sigma e^{-ju}\frac{\partial H(n_1-1,n_2,u,t)}{\partial u}+$$
$$+(N-n_1-n_2+1)\alpha H(n_1,n_2-1,u,t)+(n_1+1)\mu_1 H(n_1+1,n_2,u,t)+$$
$$+(n_2+1)\mu_2 H(n_1,n_2+1,u,t),0<n_1+n_2<N,$$
$$\frac{\partial H(n_1,n_2,u,t)}{\partial t}=-(\lambda+n_1\mu_1+n_2\mu_2)H(n_1,n_2,u,t)+\lambda H(n_1-1,n_2,u,t)+$$
$$+\lambda e^{ju}H(n_1,n_2,u,t)-j\sigma e^{-ju}\frac{\partial H(n_1-1,n_2,u,t)}{\partial u}+$$
$$+\alpha H(n_1,n_2-1,u,t),\ n_1+n_2=N. \quad (2)$$

We denote $\mathbf{H}(u,t)$ is a matrix of functions $H(n_1,n_2,u,t)$ and rewrite the system (2) in form of

$$\frac{\partial \mathbf{H}(u,t)}{\partial t}=(\mathbf{A}+\lambda e^{ju}\mathbf{B})\mathbf{H}(u,t)+j\sigma(\mathbf{I}_0-e^{-ju}\mathbf{I}_1)\frac{\partial \mathbf{H}(u,t)}{\partial u}, \quad (3)$$

where $\mathbf{A},\ \mathbf{B},\ \mathbf{I}_0,\ \mathbf{I}_1$ are operators, which set in the following form

$$\mathbf{AH}(u,t) = \begin{cases} -(\lambda + N\alpha)H(n_1,n_2,u,t) + \mu_1 H(n_1+1,n_2,u,t) + \\ +\mu_2 H(n_1,n_2+1,u,t),\ n_1+n_2=0, \\ \\ -(\lambda + (N-n_1-n_2)\alpha + n_1\mu_1 + n_2\mu_2)H(n_1,n_2,u,t) + \\ +(N-n_1-(n_2-1))\alpha H(n_1,n_2-1,u,t) + \\ +\lambda H(n_1-1,n_2,u,t) + (n_1+1)\mu_1 H(n_1+1,n_2,u,t) + \\ +(n_2+1)\mu_2 H(n_1,n_2+1,u,t),\ 0<n_1+n_2<N, \\ \\ -(\lambda + n_1\mu_1 + n_2\mu_2)H(n_1,n_2,u,t) + \lambda H(n_1-1,n_2,u,t) + \\ +\alpha H(n_1,n_2-1,u,t),\ n_1+n_2=N, \end{cases} \tag{4}$$

$$\mathbf{BH}(u,t) = \begin{cases} 0, & n_1+n_2<N, \\ H(n_1,n_2,u,t), & n_1+n_2=N, \end{cases} \tag{5}$$

$$\mathbf{I}_0\mathbf{H}(u,t) = \begin{cases} H(n_1,n_2,u,t), & n_1+n_2<N, \\ 0, & n_1+n_2=N, \end{cases} \tag{6}$$

$$\mathbf{I}_1\mathbf{H}(u,t) = \begin{cases} 0, & n_1+n_2=0, \\ H(n_1-1,n_2,u,t), & n_1+n_2>0. \end{cases} \tag{7}$$

We denote $\mathbf{E}$ as an operator that summarizing over all available values of $n_1,\ n_2$ and present the additional equation that we need to provide analysis

$$\mathbf{E}\frac{\partial \mathbf{H}(u,t)}{\partial t} = \mathbf{E}(\mathbf{A} + \lambda e^{ju}\mathbf{B})\mathbf{H}(u,t) + j\sigma\mathbf{E}(\mathbf{I}_0 - e^{-ju}\mathbf{I}_1)\frac{\partial \mathbf{H}(u,t)}{\partial u}. \tag{8}$$

We also note that

$$\mathbf{E}(\mathbf{A}+\lambda\mathbf{B}) = 0,\ \mathbf{E}(\mathbf{I}_0 - \mathbf{I}_1) = 0. \tag{9}$$

3. Asymptotic-Diffusion Analysis

In operator equations (3) and (8) we introduce the following notations

$$\sigma = \varepsilon,\ u = \varepsilon w,\ \tau = \varepsilon t,\ \mathbf{H}(u,t) = \mathbf{F}(w,\tau,\varepsilon),$$

to obtain the equations

$$\varepsilon\frac{\partial \mathbf{F}(w,\tau,\varepsilon)}{\partial \tau} = (\mathbf{A} + \lambda e^{jw\varepsilon}\mathbf{B})\mathbf{F}(w,\tau,\varepsilon) + j(\mathbf{I}_0 - e^{-jw\varepsilon}\mathbf{I}_1)\frac{\partial \mathbf{F}(w,\tau,\varepsilon)}{\partial w}, \tag{10}$$

$$\varepsilon\mathbf{E}\frac{\partial\mathbf{F}(w,\tau,\varepsilon)}{\partial\tau}=\mathbf{E}(\mathbf{A}+\lambda e^{jw\varepsilon}\mathbf{B})\mathbf{F}(w,\tau,\varepsilon)+j\mathbf{E}(\mathbf{I}_0-e^{-jw\varepsilon}\mathbf{I}_1)\frac{\partial\mathbf{F}(w,\tau,\varepsilon)}{\partial w}. \quad (11)$$

We solve the system (11) taking the limit as $\varepsilon\to 0$ and present the result in following theorem.

Theorem 1. In considered retrial queue, the stationary probability distribution $R(n_1,n_2)$ of the two-dimensional process $\{n_1(t),\ n_2(t)\}$ is a solution of the system of operator equations

$$(\mathbf{A}+\lambda\mathbf{B}-x(\mathbf{I}_0-\mathbf{I}_1))\mathbf{R}=0,$$

$$\mathbf{E}\mathbf{R}=1, \quad (12)$$

where $\mathbf{R}$ is a matrix of probabilitites $R(n_1,n_2)$, function $x(\tau)$ is a solution of differential equation

$$x'(\tau)=\mathbf{E}\left[\lambda\mathbf{B}-x(\tau)\mathbf{I}_1\right]\mathbf{R}.$$

Denoting

$$a(x)=\mathbf{E}\left[\lambda\mathbf{B}-x\mathbf{I}_1\right]\mathbf{R} \quad (13)$$

and making the following replacements in the operator equations (3) and (8)

$$\mathbf{H}(u,t)=e^{j\frac{u}{\sigma}x(\sigma t)}\mathbf{H}^{(2)}(u,t), \quad (14)$$

we have

$$\frac{\partial\mathbf{H}^{(2)}(u,t)}{\partial t}+jux'(\sigma t)\mathbf{H}^{(2)}(u,t)=$$

$$=(\mathbf{A}+\lambda e^{ju}\mathbf{B}-x(\sigma t)(\mathbf{I}_0-e^{-ju}\mathbf{I}_1))\mathbf{H}^{(2)}(u,t)+j\sigma(\mathbf{I}_0-e^{-ju}\mathbf{I}_1)\frac{\partial\mathbf{H}^{(2)}(u,t)}{\partial u}, \quad (15)$$

$$\mathbf{E}\frac{\partial\mathbf{H}^{(2)}(u,t)}{\partial t}+jux'(\sigma t)\mathbf{E}\mathbf{H}^{(2)}(u,t)=$$

$$=\mathbf{E}(\mathbf{A}+\lambda e^{ju}\mathbf{B}-x(\sigma t)(\mathbf{I}_0-e^{-ju}\mathbf{I}_1))\mathbf{H}^{(2)}(u,t)+j\sigma\mathbf{E}(\mathbf{I}_0-e^{-ju}\mathbf{I}_1)\frac{\partial\mathbf{H}^{(2)}(u,t)}{\partial u}. \quad (16)$$

In operator equations (15) and (16) we introduce the following notations

$$\sigma=\varepsilon^2,\ \tau=\varepsilon^2 t,\ u=\varepsilon w,\ \mathbf{H}^{(2)}(u,t)=\mathbf{F}^{(2)}(w,\tau,\varepsilon), \quad (17)$$

to obtain the equations

$$\varepsilon^2\frac{\partial\mathbf{F}^{(2)}(w,\tau,\varepsilon)}{\partial\tau}+jw\varepsilon a(x)\mathbf{F}^{(2)}(w,\tau,\varepsilon)=$$

$$=(\mathbf{A}+\lambda e^{jw\varepsilon}\mathbf{B}-x(\mathbf{I}_0-e^{-jw\varepsilon}\mathbf{I}_1))\mathbf{F}^{(2)}(w,\tau,\varepsilon)+$$

$$+j\varepsilon(\mathbf{I}_0 - e^{-jw\varepsilon}\mathbf{I}_1)\frac{\partial \mathbf{F}^{(2)}(w,\tau,\varepsilon)}{\partial w}, \tag{18}$$

$$\varepsilon^2\mathbf{E}\frac{\partial \mathbf{F}^{(2)}(w,\tau,\varepsilon)}{\partial \tau} + jw\varepsilon a(x)\mathbf{E}\mathbf{F}^{(2)}(w,\tau,\varepsilon) =$$

$$= \mathbf{E}(\mathbf{A} + \lambda e^{jw\varepsilon}\mathbf{B} - x(\mathbf{I}_0 - e^{-jw\varepsilon}\mathbf{I}_1))\mathbf{F}^{(2)}(w,\tau,\varepsilon)+$$

$$+j\varepsilon\mathbf{E}(\mathbf{I}_0 - e^{-jw\varepsilon}\mathbf{I}_1)\frac{\partial \mathbf{F}^{(2)}(w,\tau,\varepsilon)}{\partial w}. \tag{19}$$

Solving the equations (18) and (19) taking the limit as $\varepsilon \to 0$ we present the following theorem.

Theorem 2. Probability density of diffusion limit $z(\tau)$ of the number of calls in the orbit given as follows

$$\Pi(z) = \frac{C}{b(z)}\exp\left\{\frac{2}{\sigma}\int\limits_0^z \frac{a(x)}{b(x)}dx\right\}, \tag{20}$$

where C is a normalization factor, function $a(x)$ is defined by (13), function $b(x)$ has the following form

$$b(x) = a(x) + 2[\mathbf{E}(\lambda\mathbf{B} - x\mathbf{I}_1)\mathbf{g} + x\mathbf{E}\mathbf{I}_1\mathbf{R}]. \tag{21}$$

Here $\mathbf{g}$ is the matrix of additional values. It has the same dimension as $\mathbf{R}$ and appears as the solution of the system of operator equations

$$(\mathbf{A} + \lambda\mathbf{B} - x(\mathbf{I}_0 - \mathbf{I}_1))\mathbf{g} = a(x)\mathbf{R} - (\lambda\mathbf{B} - x\mathbf{I}_1)\mathbf{R},$$

$$\mathbf{E}\mathbf{g} = 0. \tag{22}$$

From the obtained probability density $\Pi(z)$ we build the approximation of the probability distribution of the number of calls in the orbit using expression

$$PD(i) = \frac{\Pi(i\sigma)}{\sum\limits_{n=0}^{\infty}\Pi(n\sigma)}. \tag{23}$$

4. Algorithm of Calculating Drift and Diffusion Coefficients

To obtain the function $a(x)$ we need to calculate the elements of the matrix $\mathbf{R}$. We rewrite the system of operator equations (12) in scalar form

$$-(\lambda + N\alpha + x(\tau))R(n_1,n_2) + \mu_1 R(n_1+1,n_2) + \mu_2 R(n_1,n_2+1) = 0,\ n_1 + n_2 = 0,$$

$$-(\lambda+(N-n_1-n_2)\alpha+n_1\mu_1+n_2\mu_2+x(\tau))R(n_1,n_2)+$$
$$+(N-n_1-(n_2-1))\alpha R(n_1,n_2-1)+(\lambda+x(\tau))R(n_1-1,n_2)+$$
$$+(n_1+1)\mu_1 R(n_1+1,n_2)+(n_2+1)\mu_2 R(n_1,n_2+1)=0,\ 0<n_1+n_2<N,$$
$$-(n_1\mu_1+n_2\mu_2)R(n_1,n_2)+(\lambda+x(\tau))R(n_1-1,n_2)+\alpha R(n_1,n_2-1)=0,\ n_1+n_2=N.$$
$$\sum_{n_1=0}^{N}\sum_{n_2=0}^{N-n_1}R(n_1,n_2)=1. \tag{24}$$

We transform the system of equations (24) to the system of linear algebraic equations renumbering the elements of the matrix $\mathbf{R}$ in the following way

$$(n_1,n_2)\ \to 2n_1+n_2+\frac{(n_1+n_2)^2-(n_1+n_2)}{2}. \tag{25}$$

Then we obtain the system of equations

$$\widetilde{\mathbf{R}}\mathbf{D}(x)=0,\ \widetilde{\mathbf{R}}\mathbf{e}=1, \tag{26}$$

where $\widetilde{\mathbf{R}}$ is a vector of probabilities $R(n_1,n_2)$, $\mathbf{e}$ is a unit vector, the matrix $\mathbf{D}(x)$ is a matrix of the system with renumbered elements.

Solving the system of equations (26) we can express the function $a(x)$ as follows

$$a(x)=(\lambda+x)\widetilde{\mathbf{R}}\mathbf{e}_1-x, \tag{27}$$

where $\mathbf{e}_1$ is a vector, where the last $N+1$ elements are ones and the other elements are zeroes.

To derive the function $b(x)$ we need to calculate the elements of the matrix $\mathbf{g}$. We rewrite the system of operator equations (22) in scalar form

$$-(\lambda+N\alpha+x(\tau))g(n_1,n_2)+\mu_1 g(n_1+1,n_2)+\mu_2 g(n_1,n_2+1)=a(x)R(n_1,n_2),$$
$$n_1+n_2=0,$$
$$-(\lambda+(N-n_1-n_2)\alpha+n_1\mu_1+n_2\mu_2+x(\tau))g(n_1,n_2)+(N-n_1-(n_2-1))\alpha g(n_1,n_2-1)+$$
$$+(\lambda+x(\tau))g(n_1-1,n_2)+(n_1+1)\mu_1 g(n_1+1,n_2)+(n_2+1)\mu_2 g(n_1,n_2+1)=$$
$$=a(x)R(n_1,n_2)+x(\tau)R(n_1-1,n_2),\ 0<n_1+n_2<N,$$
$$-(n_1\mu_1+n_2\mu_2)g(n_1,n_2)+(\lambda+x(\tau))g(n_1-1,n_2)+\alpha g(n_1,n_2-1)=$$
$$=(a(x)-\lambda)R(n_1,n_2)+x(\tau)R(n_1-1,n_2),\ n_1+n_2=N.$$

$$\sum_{n_1=0}^{N}\sum_{n_2=0}^{N-n_1} g(n_1,n_2)=0. \tag{28}$$

We transform the system of the operator equations (22) to the system of linear algebraic equations renumbering the elements of the matrix $\mathbf{g}$ using (25). Thus, we obtain the system of equations

$$\widetilde{\mathbf{g}}\mathbf{D}(x)=\mathbf{d}(x),\ \widetilde{\mathbf{g}}\mathbf{e}=0, \tag{29}$$

where $\widetilde{\mathbf{g}}$ is a vector of elements of the matrix $\mathbf{g}$ with renumbered elements, vector $\mathbf{d}(x)$ is a vector of right parts of the system (28) with renumbered elements.

Solving the system of equations (29) we can express the function $b(x)$ as follows

$$b(x)=a(x)+2[(\lambda+x)\widetilde{\mathbf{g}}\mathbf{e}_1+x(1-\widetilde{\mathbf{R}}\mathbf{e}_1)]. \tag{30}$$

5. Numerical Examples

We fix the number of servers in the system $N=5$. We assume that the rates of service times are $\mu_1=1,\ \mu_2=2$. Outgoing calls rate is $\alpha=1$.

The accuracy of approximation we will determine using Kolmogorov range

$$\Delta=\max_{0\leqslant i\leqslant\infty}\left|\sum_{n=0}^{i}(P(n)-PD(n))\right|, \tag{31}$$

where $P(n)$ is the probability distribution of the number of calls in the orbit, obtained with simulation, $PD(n)$ is a diffusion approximation defined by (23). Calculation of Δ we provide while both $P(n)$ and $PD(n)$ are not equal to zero.

We assume that the approximation is acceptable when its accuracy $\Delta<0.05$. Table 1 depicts the accuracy of the diffusion approximation $PD(n)$ depending on parameters σ and ρ, where ρ characterizes the system load $\rho=\lambda/N\mu_1$.

Δ	$\sigma=5$	$\sigma=2$	$\sigma=1$	$\sigma=0.5$	$\sigma=0.2$	$\sigma=0.1$
$\rho=0.6$	**0,04**	**0,039**	**0,039**	**0,022**	**0,015**	**0,023**
$\rho=0.7$	0,07	0,063	0,051	**0,036**	**0,031**	**0,032**
$\rho=0.8$	0,09	0,078	0,064	**0,049**	**0,039**	**0,041**
$\rho=0.9$	0,081	0,069	0,065	0,053	**0,044**	**0,05**

Table 1. Kolmogorov range

6. Conclusion

We have considered multiserver retrial queue with two-way communication. Using asymptotic-diffusion analysis method we have built the diffusion process the distribution density of which we used to construct the approximation of the number of calls in the orbit.

We have presented numerical experiments where we have estimated the accuracy of the approximation by comparing to simulation.

REFERENCES

1. A. Gilmore, L. Moreland, Call centres: how can service quality be managed?, Irish Marketing Review 13 (1) (2000) 3.
2. S. Aguir, F. Karaesmen, O. Z. Akşin, F. Chauvet, The impact of retrials on call center performance, OR Spectrum 26 (3) (2004) 353–376.
3. H. G. Bernett, M. J. Fischer, D. M. B. Masi, Blended call center performance analysis, IT professional 4 (2) (2002) 33–38.
4. S. Bhulai, G. Koole, A queueing model for call blending in call centers, IEEE Transactions on Automatic Control 48 (8) (2003) 1434–1438.
5. A. Avramidis, A. Deslauriers, P. L'Ecuyer, Modeling daily arrivals to a telephone call center, Management Science 50 (7) (2004) 896–908.
6. L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao, Statistical analysis of a telephone call center: A queueing-science perspective, Journal of the American statistical association 100 (469) (2005) 36–50.
7. G. Falin, J. G. C. Templeton, Retrial queues, Vol. 75, CRC Press, 1997.
8. J. R. Artalejo, A. Gómez-Corral, Retrial Queueing Systems: A Computational Approach, Springer-Verlag Berlin Heidelberg, 2008.
9. T. Phung-Duc, K. Kawanishi, An efficient method for performance analysis of blended call centers with redial, Asia-Pacific Journal of Operational Research 31 (02) (2014) 1–35.
10. A. Moiseev, A. Nazarov, S. Paul, Asymptotic diffusion analysis of multi-server retrial queue with hyper-exponential service, Mathematics 8 (4) (2020) 531–546.

УДК: 519.872

Исследование циклической системы с повторными вызовами

А.А. Назаров, С.В. Пауль, П.Н. Ключникова

Томский государственный университет,пр. Ленина, 36, г. Томск, Россия

nazarov.tsu@gmail.com, paulsv82@mail.ru, polya.klyuch@gmail.com

Аннотация

В работе выполнено исследование математической модели циклической сети связи множественного доступа. В качестве модели такой сети рассматривается циклическая система с повторными вызовами, на вход которой поступает N простейших потоков заявок, продолжительности обслуживания которых имеют экспоненциальную функцию распределения. Заявки каждого потока формируют свою орбиту неограниченного объема. Пару поток и соответствующую ему орбиту назовем RQ-системой. Применяя метод систем с прогулками прибора, в данной работе найдено распределение вероятностей числа заявок на орбите в одной выделенной RQ-системе исходной циклической системы.

Ключевые слова: *циклическая система, система с повторными вызовами, RQ-система, система с прогулками прибора*

1. Введение

Специальные сети связи множественного доступа создаются для совместного выполнения миссий различными группировками. Естественной топологией для таких сетей связи может быть «звезда», центральный узел которой выполняет функции управления группировками и, в этом смысле, является общим ресурсом сети. Проблема разделения общего ресурса связи сети может в подобных случаях решаться выбором протокола множественного доступа абонентов сети к общему ресурсу.

Для эффективного разделения общего ресурса связи могут быть использованы циклические протоколы [1, 2] либо протоколы множественного в том числе случайного доступа. В данной работе предлагается рассмотреть математическую модель сети связи в виде циклической системы, в которой каждая подсистема представлена в виде системы с повторными вызовами (RQ-системы, Retrial Queueing System) [3, 4, 5, 6].

Работа выполнена при финансовой поддержке РФФИ, проект №18-01-00277 А

Ставится задача определения распределения вероятностей числа заявок на орбите в выделенной подсистеме с повторными вызовами циклической системы. Задача решается классическим методом «систем с прогулками прибора» [7, 8].

2. Математическая модель и постановка задачи

Рассмотрим циклическую систему с повторными вызовами (RQ-систему) с одним обслуживающим прибором, на вход которой поступает N простейших потоков с интенсивностью λ_n, $n = \overline{1, N}$. Заявки каждого потока формируют свою орбиту неограниченного объема. Будем называть пару n-го потока и соответствующую ему орбиту «n-й RQ-системой», $n = \overline{1, N}$.

Прибор посещает RQ-системы в циклическом порядке, начиная с первой и заканчивая N-ой, потом цикл повторяется. Время подключения прибора к каждой RQ-системе имеет экспоненциальную функцию распределения с параметром α_n, $n = \overline{1, N}$. В течение этого времени прибор обслуживает заявки, которые поступают из n-го входящего потока и соответствующей орбиты, с экспоненциальной функцией распределения с параметрами μ_n $n = \overline{1, N}$.

Если поступившая заявка входящего потока обнаруживает прибор занятым или не подключенным, она мгновенно уходит на соответствующую орбиту, где осуществляет случайную задержку в течение экспоненциального времени с параметром σ_n $n = \overline{1, N}$, после которой вновь обращается к прибору.

Если в течение времени подключения прибора к n-ой RQ-системе, в этой системе нет заявок, прибор все равно остается подключенным к системе, пока не истечет время подключения. Методом исследования циклической RQ-системы является метод систем с прогулками прибора.

3. Система с прогулками

Рассмотрим первую RQ-систему в циклической системе как систему с прогулками прибора. Имеем RQ-систему с одним обслуживающим прибором и орбитой. В систему поступает простейший поток заявок с интенсивностью λ. Прибор обслуживает заявки с экспоненциальной функцией распределения с параметром μ, которые поступают из входящего потока.

Если поступившая заявка из входящего потока обнаруживает прибор занятым, она мгновенно уходит на орбиту, где осуществляет случайную задержку в течение экспоненциального времени с параметром σ, после которой вновь обращается к прибору.

Продолжительность времени подключения прибора к потоку и орбите случайная и определяется экспоненциальной функцией распределения с параметром α_1.

От момента окончания этого интервала прибор уходит на «прогулку», продолжительность которой складывается из $N-1$ фаз. Каждая фаза имеет экспоненциальное распределение с параметрами α_v, $v=\overline{2,N}$. Каждая фаза прогулки соответствует времени подключения прибора в циклической системе к RQ-системам с номерами $n=\overline{2,N}$.

Во время прогулки, пришедшие в систему, заявки накапливаются на орбите и ждут, когда прибор вернется на обслуживание. Когда прибор уходит на прогулку, не закончив обслуживание заявки на приборе, то недообслуженная заявка уходит на орбиту.

Обозначим: $i(t)$ - число заявок на орбите в момент времени t; $k(t)$ - состояние прибора: 0 – прибор свободен, 1 – прибор обслуживает заявку, n – прибор на n-ой фазе прогулки, $n=\overline{2,N}$.

Найдем распределение вероятностей числа заявок на орбите в момент времени t в системе с прогулками прибора, тем самым определим распределение вероятностей числа заявок на орбите в выделенной подсистеме циклической RQ-системы.

Процесс $i(t)$ является немарковским. Рассмотрим двумерный марковский процесс $\{k(t),\ i(t)\}$, для распределения вероятностей $P_k(i,t)=P(k(t)=k,i(t)=i)$ которого составим систему дифференциальных уравнений Колмогорова, которую запишем в стационарном режиме:

$$\begin{gathered}
-(\lambda+i\sigma+\alpha_1)P_0(i)+\mu P_1(i)+\alpha_N P_N(i)=0,\\
-(\lambda+\mu+\alpha_1)P_1(i)+\lambda P_0(i)+(i+1)\sigma P_0(i+1)+\lambda P_1(i-1)=0,\\
-(\lambda+\alpha_2)P_2(i)+\alpha_1 P_1(i-1)+\alpha_1 P_0(i)+\lambda P_2(i-1)=0,\\
-(\lambda+\alpha_n)P_n(i)+\alpha_{n-1}P_{n-1}(i)+\lambda P_n(i-1)=0, n=\overline{3,N}.
\end{gathered}\tag{1}$$

Введем частичные характеристические функции:

$$H_n(u)=\sum_{i=0}^{\infty}e^{jui}P_n(i), n=\overline{0,N},$$

Систему уравнений Колмогорова (1) перепишем в виде:

$$\begin{gathered}
-(\lambda+\alpha_1)H_0(u)+j\sigma H_0'(u)+\mu H_1(u)+\alpha_N H_N(u)=0,\\
-(\lambda+\mu+\alpha_1)H_1(u)+\lambda H_0(u)-j\sigma e^{-ju}H_0'(u)+\lambda e^{ju}H_1(u)=0,\\
-(\lambda+\alpha_2)H_2(u)+\alpha_1 e^{ju}H_1(u)+\alpha_1 H_0(u)+\lambda e^{ju}H_2(u)=0,\\
-(\lambda+\alpha_n)H_n(u)+\alpha_{n-1}H_{n-1}(u)+\lambda e^{ju}H_n(u)=0, n=\overline{3,N}.
\end{gathered}\tag{2}$$

Характеристическая функция $H(u)$ числа заявок на орбите достаточно просто выражается через частичные характеристические функции $H_n(u)$ следующим равенством

$$H(u) = \sum_{n=0}^{N} H_n(u).$$

Теорема. *Характеристическая функция числа заявок на орбите в RQ-системе с прогулками прибора имеет вид:*

$$H(u) = \frac{H_0(u)}{f_0(u)}, \tag{3}$$

где функция $H_0(u)$ имеет вид

$$H_0(u) = r_0 exp\left\{\int_0^u f(x)\,dx\right\}. \tag{4}$$

Здесь величина

$$r_0 = \left(\sum_{n=1}^{N} \frac{\alpha_1}{\alpha_n}\right)^{-1} - \frac{\lambda}{\mu}. \tag{5}$$

Функция $f(u)$ имеет вид

$$f(u) = \\ = \frac{j}{\sigma}\left[\alpha_N \frac{f_N(u)}{f_0(u)} + \mu\frac{f_1(u)}{f_0(u)} - (\lambda + \alpha_1)\right]. \tag{6}$$

Здесь функции $f_0(u)$, $f_1(u)$ и $f_N(u)$ определяются равенствами

$$f_0(u) = \\ = \left[1 - \frac{\lambda}{\mu - (e^{ju}-1)\lambda} - e^{ju}\frac{\alpha_1}{\alpha_2 - \lambda(e^{ju}-1)} \cdot \frac{\lambda}{\mu + \lambda - e^{ju}\lambda} \times \right. \\ \left. \times \left[1 + \sum_{n=3}^{N}\prod_{v=3}^{n} \frac{\alpha_{v-1}}{\alpha_v - \lambda(e^{ju}-1)}\right]\right] \times \\ \times \left[1 + \frac{\alpha_1}{\alpha_2 - \lambda(e^{ju}-1)} \cdot \left[1 + \sum_{n=3}^{N}\prod_{v=3}^{n} \frac{\alpha_{v-1}}{\alpha_v - \lambda(e^{ju}-1)}\right]\right]^{-1}, \tag{7}$$

$$f_1(u) = \frac{\lambda}{\mu - (e^{ju} - 1)\lambda}, \tag{8}$$

$$f_N(u) =$$

$$= \frac{\alpha_1}{\alpha_2 - \lambda(e^{ju} - 1)} \left(f_0(u) + e^{ju} \frac{\lambda}{\mu + \lambda - e^{ju}\lambda} \right) \times$$

$$\times \prod_{v=3}^{N} \frac{\alpha_{v-1}}{\alpha_v - \lambda(e^{ju} - 1)}. \tag{9}$$

Пронумеруем входящие потоки так, чтобы исследуемая RQ-система имела номер один, положим равными $\lambda = \lambda_1$, $\mu = \mu_1$, тогда полученная допредельная характеристическая функция $H(u)$ определяет число заявок на орбите в первой RQ-системе в исходной циклической системе.

Дискретное распределение вероятностей числа заявок на орбите определяется обратным преобразованием Фурье по переменной u от характеристической функции и имеет вид

$$P(i) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-jui} H(u)\, du. \tag{10}$$

4. Численный пример

Обозначим загрузку системы ρ, тогда интенсивность λ входящего потока определяется следующей формулой

$$\lambda = \rho\mu \frac{\alpha_1^{-1}}{\sum\limits_{n=1}^{N} \alpha_n^{-1}}.$$

Пусть $N = 10$, $\alpha_n = n$, $n = \overline{1, N}$, $\mu_1 = 1$, $\rho = 0.8$, $\lambda = 0.27$.

Для предложенных параметров циклической RQ-системы получим графики дискретного распределения вероятностей для числа заявок на орбите в одной выделенной RQ-системе.

5. Заключение

В данной работе ставилась задача исследования математической модели циклической сети связи множественного доступа, на вход которой поступает N простейших потоков заявок. Было получено распределение вероятностей числа

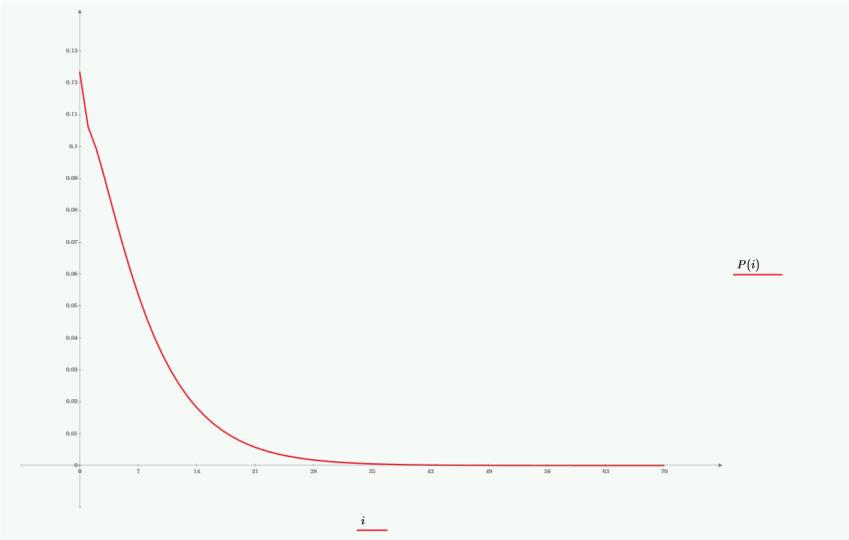


Рис. 1. Распределение вероятностей числа заявок на орбите при $\sigma = 5$

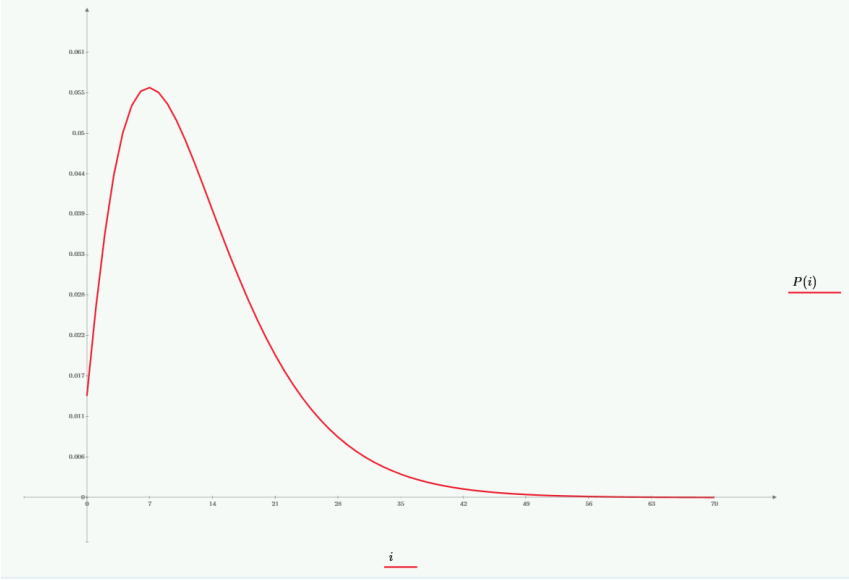


Рис. 2. Распределение вероятностей числа заявок на орбите при $\sigma = 1$

заявок на орбите в выделенной подсистеме с повторными вызовами в циклической системе.

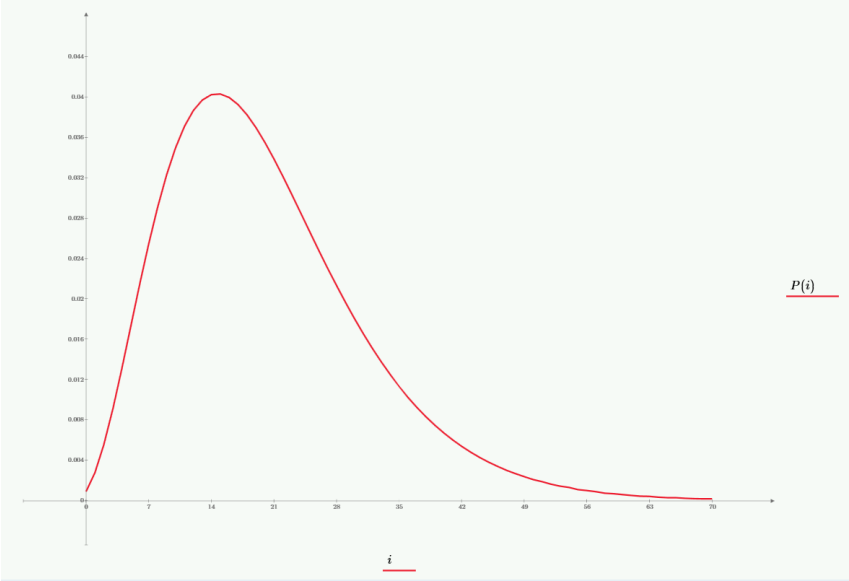


Рис. 3. Распределение вероятностей числа заявок на орбите при $\sigma = 0.5$

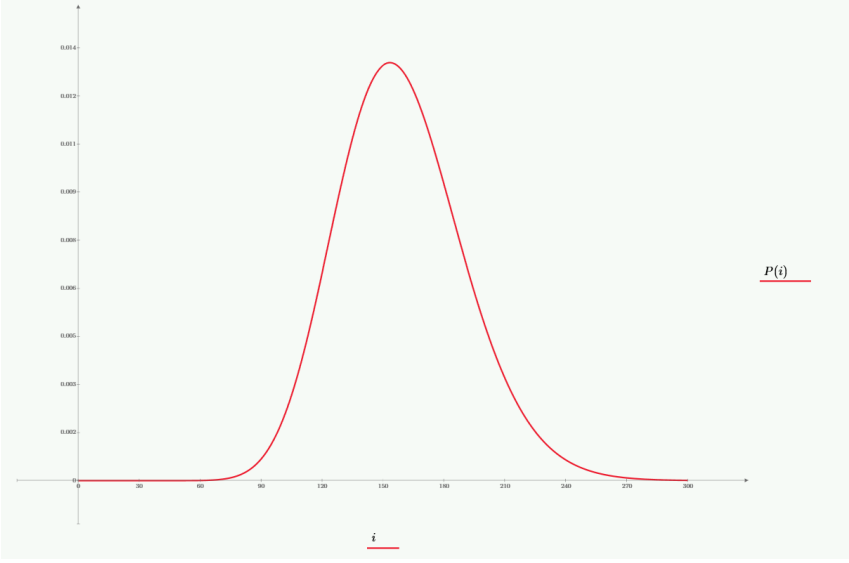


Рис. 4. Распределение вероятностей числа заявок на орбите при $\sigma = 0.05$

Задача решена классическим методом «систем с прогулками прибора».

ЛИТЕРАТУРА

1. В. М. Вишневский, О. В. Семенова, Системы поллинга: теория и применение в широкополосных беспроводных сетях. М.: Техносфера, 2007. 312 с., Информационные технологии и вычислительные системы (1) (2008) 98–99.
2. В. М. Вишневский, О. В. Семенова, Математические методы исследования систем поллинга, Автоматика и телемеханика (2) (2006) 3–56.
3. J. R. Artalejo, Accessible bibliography on retrial queues: progress in 2000–2009, Mathematical and computer modelling 51 (9-10) (2010) 1071–1081.
4. J. R. Artalejo, A classified bibliography of research on retrial queues: progress in 1990–1999, Top 7 (2) (1999) 187–211.
5. J. R. Artalejo, A. Gómez-Corral, Retrial queueing systems, Vol. 30, Springer, 1999.
6. J. R. Artalejo, Algorithmic Methods in Retrial Queues, Vol. 141, Springer, 2006.
7. A. Nazarov, S. Paul, A cyclic queueing system with priority customers and t-strategy of service, Communications in Computer and Information Science 678 (2016) 182–193.
8. A. Nazarov, S. Paul, A number of customers in the system with server vacations, Communications in Computer and Information Science 601 (2015) 334–343.

UDC: 004.057.4

An Algebraic Approach to Loop Free Routing

H.Khayou [1], M.A. Rudenkova[2], L.I. Abrosimov [3]

[1,2,3]National Research University "Moscow Power Engineering Institute",Krasnokazarmennaya 14, Moscow, Russia

hussein.khayou@gmail.com, RudenkovaMA@mpei.ru, AbrosimovLI@mpei.ru

Abstract

Validation models not only provide a better understanding of the system, but can also help in improving the reliability and robustness of the design. EIGRP metric can be modeled algebraically using semirings [1]. EIGRP uses DUAL algorithm which is the basis for loop free routing with non lexical metric. DUAL was validated using the classical shortest path problem [13, 22]. However, it was shown that DUAL does not perform as expected in the absence of monotonicity [25, 1]. This article approaches loop free routing from an algebraic perspective. Conditions for loop free routing and the relations between them were presented algebraically and proved correct. Then, we investigate loop free routing in the presence and the absence of monotoncity.

Keywords: Semirings, Routing Algebra, Loop Free Routing, Bellman Ford, Dijkstra, Diffusing Computation

1. Introduction

A routing loop is a common problem in computer networks. This happens when the path towards a particular destination contains a loop due to erroneous routing tables, thus packets destined to this destination will loop endlessly unless they are eventually dropped. This is especially true in early distance vector protocols such as routing information protocol (RIP). In link state protocols such as open shortest first (OSPF) and intermediate system to intermediate system (IS-IS), routing loops can still occur [14], however, they are short lived, as they disappear as soon as the information about the new topology is flooded across the network.

New distance vector protocols such as border gateway protocol (BGP) and enhanced interior gateway routing protocol (EIGRP) are equipped with loop prevention mechanisms. EIGRP protocol is a Cisco proprietary protocol based on the interior gateway protocol (IGRP). EIGRP was converted to an open standard in 2013 and

was published in RFC 7868 in 2016 [24]. EIGRP employs diffusing update algorithm (DUAL) which is a loop free routing algorithm. The convergence time with DUAL rivals that of any other existing routing protocol [15]. EIGRP employs SNC (Source Node Conditions), which is one of the DUAL's loop freedom sufficient conditions, as its feasibility condition. This condition is met when the neighbor's advertised distance for a particular destination is strictly less than the feasible distance for that destination [13, 22]. Other sufficient conditions includes DIC (Distance Increase Condition), and CSC (Current Successor Condition).
DUAL is proved to be loop free at every instance of time, and to converge in a finite time after the occurrence of link-cost changes [22]. However, EIGRP uses the same composite metric in IGRP which utilizes the available bandwidth, delay, load utilization, and link reliability for metric calculations. Composite metric is modeled in [1] using an algebraic construct called the functional product, where the authors showed that EIGRP's metric is non monotonic resulting that EIGRP solves a local optimal solution as opposed to the global optimal solution, which is solved by the classical shortest path problem. To the best of our knowledge, no formal proof of correctness of DUAL with a non monotonic metric has been given yet.
In this paper we investigate the concept of loop free routing with a generic metric using the matrix model with semirings. We provide an algebraic representation for the sufficient conditions of loop free routing in the semiring model. We also explore the relation between them and provide an algebraic proof for their correctness. We also introduce the concept of monotone routing. It was seen that if the routing was decreasing (or increasing) in one iteration, it will continue doing so in the next iterations as long as the topology is fixed. The generality of the model helps in showing some theory and guidelines on the design of new loop free routing protocols in the presence and absence of monotonicity.
This paper is organized as follows: next section surveys the literature in this field, an overview of the semiring model is presented in section 3, sufficient conditions for loop free routing are presented in section 4, monotone routing is introduced in section 5, and loop free routing in a non-monotone algebra is discussed in section 6, and section 7 concludes the paper.

2. Related Works

Recently, there have been some efforts to apply formal methods specifically algebraic specifications to existing routing protocols. In [5] and [2], Sobrinho developed an algebraic framework for investigating the convergence properties of distance vector and path vector protocols. It is shown that "monotonicity" (a property related to the inflationary property presented in Section 3) implies protocol convergence in every network but not necessarily to a "global optimal" (the notion of optimality

is defined in Section 3 also). However, "isotonicity" (which is a property related to distributivity, we refer to it as monotonicity in this article) assures convergence onto global optimal paths when the protocol converges. For link state protocols Sobrinho presented a more specific less general algebraic frame work [3]. It was seen that local optimality rather than global optimality is more appropriate for modeling inter-domain routing protocol such as BGP [5, 2, 6, 7].

Griffin and Sobrinho proposed metarouting as a means of defining routing protocols in a high-level and declarative manner [8]. Metarouting was based on Sobrinho's algebra. Sobrinho's model uses a partial order relation to construct the algebra. There is another approach to model path problems using algebraic structures called *semirings* ([9], and [10] contains modern surveys of this area). The two models can be related to each other [11]. Routing operations in the semiring model become matrix operations.
Diffusing Computation concept was first proposed by Dijkastra and Scholten [12]. After that, most of the work on loop free routing has been done by Garcia who presented DUAL [13, 14] and proved its loop freedom. DUAL was adopted later in EIGRP [15]. DIC was discussed in the literature prior to the work of Jaffe and Moss, however, they were first to prove that DIC is sufficient for loop freedom [18]. CSC and SNC were proposed and proved correct by Garcia [13, 14]. Gouda and Schneider [25] have provided a graph theoretical approach to show that IGRP and EIGRP protocol do not behave as expected because of the composite metric which is not nonmonotonic in general. Algebraic theory of routing was also used in more recent studies in [19, 17, 1, 26]. In [27] the author has developed algorithms to solve for optimal solutions of the shortest path problem for IGRP's like metrics.

3. Semirings and Graphs

We briefly describes in this section how semirings can be related to the shortest path problem. More details can be found in [23]. Semirings are structures of the form $(S, \oplus, \otimes, \overline{0}, \overline{1})$, where S is a non-empty and non-trivial set and the axioms in Table 1 hold. Semirings differ from rings in that the additive operation do not need to admit inverses. That is, $(S, \oplus)$ is only required to be a monoid, not a group [9]. This allows us to define a non-trivial "natural order" on S:

$$a \leq_{\oplus} b \equiv a = a \oplus b$$

We can also define the strict version of this order

$$a <_{\oplus} b \equiv a = a \oplus b \neq b$$

We need $\oplus$ to be idempotent ($\forall a \in S : a \oplus a = a$) so the relation $\leq_{\oplus}$ becomes reflexive. In this case $\leq_{\oplus}$ becomes a partial order. If $\oplus$ is also selective ($\forall a, b \in S :$

$a \oplus b \in \{a, b\}$) then $\leq_\oplus$ becomes total order. In addition, we define a "canonical order" on the semi-group $(S, \oplus)$

$$a \trianglelefteq_\oplus b \equiv \exists c \in S | a = b \oplus c$$

If $\oplus$ is commutative and idempotent then

$$\forall a, b \in S : a \trianglelefteq_\oplus b \Leftrightarrow a \leq_\oplus b$$

Table 1. Axioms for semirings

Axiom	Explanation
$\oplus$ Associative	$a \oplus (b \oplus c) = (a \oplus b) \oplus c$
$\oplus$ Commutative	$a \oplus b = b \oplus a$
$\oplus$ Identity	$a \oplus \overline{0} = \overline{0} \oplus a = a$
$\otimes$ Associative	$a \otimes (b \otimes c) = (a \otimes b) \otimes c$
$\otimes$ Identity	$a \otimes \overline{1} = \overline{1} \otimes a = a$
$\otimes$ Annihilator	$a \otimes \overline{0} = \overline{0} \otimes a = \overline{0}$
Left Distributivity	$a \otimes (b \oplus c) = (a \otimes b) \oplus (a \otimes c)$
Right Distributivity	$(a \oplus b) \otimes c = (a \otimes c) \oplus (b \otimes c)$

We define monotonicity from the left and the right respectively

$$a \leq_\oplus b \Rightarrow c \otimes a \leq_\oplus c \otimes b \tag{3.1}$$

$$a \leq_\oplus b \Rightarrow a \otimes c \leq_\oplus b \otimes c \tag{3.2}$$

Monotonicity holds in semirings due to distributivity. The semiring $(S, \oplus, \otimes, \overline{0}, \overline{1})$ can be used to define a semiring $(\mathbb{M}_n(S), \oplus, \otimes, J, I)$ of $n \times n$ matrices. The J and I matrices are:

$$J(i, j) = \overline{0} \tag{3.3}$$

$$I(i, j) = \begin{cases} \overline{1}, & \text{if } i = j; \\ \overline{0}, & \text{otherwise.} \end{cases} \tag{3.4}$$

The classical shortest path problem can be modeled with semirings [16]. Given a semiring $(S, \oplus, \otimes, \overline{0}, \overline{1})$ and a graph $G = (V, E)$, a *weight function* is a mapping $w : E \to S - \{\overline{0}\}$. The graph can be presented using what we call *weighted adjacency matrix*.

$$A(i, j) = \begin{cases} w(e), & \text{if } e = (i, j) \in E; \\ \overline{0}, & \text{otherwise.} \end{cases} \tag{3.5}$$

A path $p = v_1, v_2, \cdots, v_{k+1}$ of length k is a sequence of nodes such that $(v_m, v_{m+1}) \in E$ for each $m, 1 \leq m \leq k$. The weight of the path p is given by

$$w(p) = w(v_1, v_2) \otimes w(v_2, v_3) \otimes \cdots \otimes w(v_k, v_{k+1}) \tag{3.6}$$

Null paths are denoted by ϵ and are given the weight $\overline{1}$. Non-existing paths are given the weight $\overline{0}$.The power of a matrix A is defined inductively:

$$A^0 = I \tag{3.7}$$

$$A^{k+1} = A \otimes A^k \tag{3.8}$$

$$A^{(0)} = I \tag{3.9}$$

$$A^{(k+1)} = A^{k+1} \oplus A^{(k)} \tag{3.10}$$

Let $P^k(i,j)$ be the set of paths from node i to j which has exactly k arcs. $P^{(k)}(i,j)$ is the set of paths from node i to j with k arcs at most. $P(i,j)$ is the set of all possible paths from node i to j. We say that p is a simple path if it does not have loops. We denote by $SP(i,j)$ the set of all possible simple paths from node i to j. Additionally, $SP^{(k)}(i,j)$ is the set of all simple paths from i to j with at most k arcs of length.

Theorem 3.1.

$$A^k(i,j) = \bigoplus_{p \in P^k(i,j)} w(p) \tag{3.11}$$

$$A^{(k)} = \bigoplus_{p \in P^{(k)}(i,j)} w(p) \tag{3.12}$$

If there exists a $q \geq 0$ such that $A^{(q+1)} = A^{(q)}$ then $\forall k \geq q : A^{(k)} = A^{(q)}$. We say then that A is q-stable. We say also that $A^{(k)}$ **converges** to $A^* = A^{(q)}$, and we call A^* the **closure** matrix of A.

$$A^* = \bigoplus_{k \geq 0} A^{(k)} = \bigoplus_{p \in P(i,j)} w(p) \tag{3.13}$$

We interpret (3.13) as: A^* is the solution to the **global optimal** path problem. It remains to see when this solution exists. Theorem 3.2 gives that the inflationary property is sufficient for global optimality in semirings, and the solution is reached with at most $n-1$ iterations. Theorem 3.3 states that the solution will be the same when constructed using only simple paths. Since only single paths are used then we will reach the solution after $d \leq n-1$ iterations, where d is the diameter of the graph (the number of nodes in the longest path in the graph).

Theorem 3.2. Let $(S, \oplus, \otimes, \overline{0}, \overline{1})$ be an idempotent semiring, and let $\overline{1}$ be an annihilator for $\oplus$. Then, all weighted adjacency matrices in $(\mathbb{M}_n(S), \oplus, \otimes, J, I)$ are $(n-1)$-stable.

Theorem 3.3. Let $(S, \oplus, \otimes, \overline{0}, \overline{1})$ be an idempotent semiring, and let $\overline{1}$ be an annihilator for $\oplus$. Then,

$$A^{(k)}(i,j) = \bigoplus_{q \in SP^{(k)}(i,j)} w(q) \tag{3.14}$$

Bellman-Ford algorithm can be modeled using the following iteration [11, 19, 17]

$$A^{\langle 0 \rangle} = I \tag{3.15}$$

$$A^{\langle k+1 \rangle} = A \otimes A^{\langle k \rangle} \oplus I \tag{3.16}$$

We call this iteration *distance vector iteration*. If A is the weighted adjacency matrix of graph $G = (V, E)$, then (3.16) can be written as

$$A^{\langle k+1 \rangle}(i,j) = I(i,j) \oplus \bigoplus_{q \in V} A(i,q) \otimes A^{\langle k \rangle}(q,j) \tag{3.17}$$

If $N(i) \subseteq V$ is the set of node i neighbors, then we can restrict the sum in (3.17) to the set $N(i)$ because $A(i,q) = \overline{0}$ when $q \notin N(i)$

$$A^{\langle k+1 \rangle}(i,j) = I(i,j) \oplus \bigoplus_{q \in N(i)} A(i,q) \otimes A^{\langle k \rangle}(q,j) \tag{3.18}$$

We define the left and right strict inflationary property respectively

$$\forall a, b \in S, a \neq \overline{0}, b \neq \overline{1} : a <_{\oplus} b \otimes a \tag{3.19}$$

$$\forall a, b \in S, a \neq \overline{0}, b \neq \overline{1} : a <_{\oplus} a \otimes b \tag{3.20}$$

If we assume that the left strict inflationary property holds, and we do not allow links to have the weight $\overline{1}$, then we say the routing is **loop-free** for static topology i.e. only simple paths will be considered. This can be seen easily from Theorem 3.3. If $\otimes$ is cancellative (see (3.21) and (3.22)), and $\overline{1}$ is annihilator for $\oplus$ then the left and right strict inflationary properties hold.

$$a \neq \overline{0}, a \otimes b = a \otimes c \Rightarrow b = c \tag{3.21}$$

$$c \neq \overline{0}, a \otimes c = b \otimes c \Rightarrow a = b \tag{3.22}$$

We lose the loop free attribute in dynamic topology. If some of the links fail, then loops might occur causing the counting to convergence or counting to infinity problem. This can be explained easily algebraically [19]. Suppose that we start the distance vector iteration with an arbitrary matrix M rather than I. M represents the matrix of best path weights before the change in the topology, while A is the new weighted adjacency matrix after the change in the topology.

$$A_M^{\langle 0 \rangle} = M \tag{3.23}$$

$$A_M^{\langle k+1 \rangle} = A \otimes A_M^{\langle k \rangle} \oplus I, k \geq 1 \tag{3.24}$$

We find by induction that for $k \geq 1$

$$A_M^{\langle k \rangle} = A^k \otimes M \oplus A^{\langle k-1 \rangle} \tag{3.25}$$

If A is $(n-1)$-stable and for $k \geq n$ we have

$$A_M^{\langle k \rangle} = A^k \otimes M \oplus A^* \tag{3.26}$$

While the term A^* may be reached soon, but it could be that the term $A^k \otimes M$ is preferred. So, the iteration will continue until $A^* \leq_\oplus A^k \otimes M$. This might not happen if there is no longer a path that connects the nodes i and j [19] — counting to infinity. Solutions to this problem includes limiting the number of hops to 15 like in the RIP protocol [20]. This hop limit make networks connected with more than 15 routers unreachable. In BGP the whole path is advertised and the algorithm is modified to consider only simple paths [21]. Another solution is by using the DUAL algorithm which is proved to be loop free [13, 14].

4. Sufficient Conditions for Loop Freedom in Distance Vector Routing

We will assume in this section that $(S, \oplus, \otimes, \overline{0}, \overline{1})$ is a left strict inflationary, selective, and idempotent semiring. We call this a linear increasing semiring (LISR). In addition, $\overline{1}$ is not allowed as a link weight. We need selectivity to be able to define the successor node, which is the next hop node selected by a node i that corresponds to the best path toward a destination node j.

We will modify the model for distance vector routing in (3.15) and (3.16) to take account of topology changes. The change in the topology will be modeled as a change in the weighted adjacency matrix A. We will call the matrix computed by distance vector iteration the routing matrix. We say that a node $q \neq i$ is a downstream node for i in routing toward j if the path selected by i passes through q. And we say that q is an upstream node for i if i is a downstream node for q.

Equation (4.1) represents distance vector routing model in a dynamic topology. M_k represents the routing matrix (best paths weights) at stage k and A_{k+1} is the new weighted adjacency matrix that captures the new topology (it would be same as A_k if no topology changes occur at stage $k+1$ of routing)

$$M_{k+1} = A_{k+1} \otimes M_k \oplus I \tag{4.1}$$

The successor to i in routing toward j ($i \neq j$) at stage $k+1$ is then a node s selected by i such that

$$M_{k+1}(i,j) \quad = \quad \bigoplus_{q \in V} A_{k+1}(i,q) \ \otimes \ M_k(q,j) \quad = \quad A_{k+1}(i,s) \ \otimes \ M_k(s,j) \tag{4.2}$$

This is true because $\oplus$ is a selective operation. The importance of loop free routing is that it guarantees convergence in a dynamic topology as stated by Theorem 4.1.

Theorem 4.1. If we arrive in routing to the matrix M. And the topology is settled on a weighted adjacency matrix A (no further change in the topology). If the routing is loop free afterwards (only simple paths are inspected), then the routing will converge to A^*.

Proof. In (3.26), let $k \geq n$, then A^k will be equivalent to J as there is no simple path with a number of arcs greater than $n-1$ and the algorithm is assumed to consider only simple paths. Then, $A^{\langle k \rangle} = J \otimes M \oplus A^* = A^*$. ■

There are 3 sufficient conditions for loop free routing [14]. We express them algebraically in the following:

DIC *Distance Increase Condition* Node i is free to change its successor toward j to a node s that satisfies (4.2) if the distance is not increased i.e. $M_{k+1}(i,j) \leq_\oplus M_k(i,j)$. Otherwise node i must maintains its current successor.

CSC *Current Successor Condition* Node i is free to change its successor toward j to a node s that satisfies (4.2) if $M_k(s,j) \leq_\oplus M_{k-1}(s',j)$, where s' is the successor in the previous stage (step k). Otherwise node i must maintains its current successor.

SNC *Source Node Condition* Node i is free to change its successor toward j to a node s that satisfies (4.2) if $M_k(s,j) <_\oplus M_k(i,j)$. Otherwise node i must maintains its current successor.

In the rest of this section we will provide an algebraic proof for the correctness of these 3 conditions.

Theorem 4.2. The SNC is a sufficient condition for loop free routing.

Proof. In (4.1), suppose we arrive to a routing matrix M. Now, let i chooses a successor s_1 in the next step of routing toward j, so that the loop $(i, s_1, s_2, \cdots, s_k, i)$ will be formed. If all the nodes take into account the SNC when choosing their successors, then we have

$$\begin{aligned} M(s_1, j) &<_{\oplus} M(i, j) \\ M(s_2, j) &<_{\oplus} M(s_1, j) \\ &\vdots \\ M(s_k, j) &<_{\oplus} M(s_{k-1}, j) \\ M(i, j) &<_{\oplus} M(s_k, j) \end{aligned}$$

We have

$$M(i,j) <_{\oplus} M(s_k,j) <_{\oplus} M(s_{k-1},j) <_{\oplus} \cdots <_{\oplus} M(s_2,j) <_{\oplus} M(s_1,j) <_{\oplus} M(i,j)$$

leading to the contradiction $M(i, j) <_{\oplus} M(i, j)$. Then no loop can be formed if all the nodes respect SNC conditions while choosing their successors. ■

Theorem 4.3. DIC implies SNC.

Proof. In (4.1), suppose we arrive to a routing matrix M and the weighted adjacency matrix is A. Suppose that node i chooses node s as a successor in routing toward j. Then, the new distance to j is $A(i, s) \otimes M(s, j)$, if node i takes into account the DIC when selecting its successor. Then we have $A(i, s) \otimes M(s, j) \leq_{\oplus} M(i, j)$. We have also $M(s, j) <_{\oplus} A(i, s) \otimes M(s, j)$ since $A(i, s) \neq \overline{1}$. Therefore, $M(s, j) <_{\oplus} M(i, j)$ and the SNC holds. ■

Theorem 4.4. CSC implies SNC.

Proof. In (4.2), suppose that node i selects s as a successor in the $k + 1$ step of routing toward j. Let s' be the successor in the k step. Now if the CSC holds then $M_k(s, j) \leq_{\oplus} M_{k-1}(s', j)$. We have then $M_k(i, j) = A_k(i, s') \otimes M_{k-1}(s', j)$. Therefore, $M_{k-1}(s', j) <_{\oplus} M_k(i, j)$ and consequently $M_k(s, j) <_{\oplus} M_k(i, j)$. ■

Corollary 4.5. DIC and CSC are sufficient conditions for loop free routing.

5. Monotone Routing

Theorem 5.1 states that if the routing matrix increases (or decreases) in one distance vector iteration and the topology is fixed, then it will continue increasing (or decreasing) in the next iterations. Theorem 5.3 proves that the decreasing routing will eventually converge in a finite number of distance vector iterations.

Theorem 5.1. Let $(S, \oplus, \otimes, \overline{0}, \overline{1})$ be an idempotent semiring. In (3.23) and (3.24)

1) If $A_M^{\langle 1 \rangle} \leq_\oplus M$ then $\forall k \geq 0 : A_M^{\langle k+1 \rangle} \leq_\oplus A_M^{\langle k \rangle}$
2) If $M \leq_\oplus A_M^{\langle 1 \rangle}$ then $\forall k \geq 0 : A_M^{\langle k \rangle} \leq_\oplus A_M^{\langle k+1 \rangle}$

Proof. We will prove 1 by induction. 2 can be proven similarly. Let us assume the inequality in 1 holds for k then

$$A_M^{\langle k+1 \rangle} \oplus A_M^{\langle k \rangle} = A_M^{\langle k+1 \rangle}$$

We apply A on both sides. The distributivity of $\otimes$ on $\oplus$ implies

$$\left(A \otimes A_M^{\langle k+1 \rangle}\right) \oplus \left(A \otimes A_M^{\langle k \rangle}\right) = A \otimes A_M^{\langle k+1 \rangle}$$

Then we have

$$\left(\left(A \otimes A_M^{\langle k+1 \rangle}\right) \oplus I\right) \oplus \left(\left(A \otimes A_M^{\langle k \rangle}\right) \oplus I\right) = \left(A \otimes A_M^{\langle k+1 \rangle}\right) \oplus I$$

Because $I \oplus I = I$. Then

$$A_M^{\langle k+2 \rangle} \oplus A_M^{\langle k+1 \rangle} = A_M^{\langle k+2 \rangle}$$

So the inequality holds for $k+1$ then it holds $\forall k \geq 0$. ■

In general, the routing might be decreasing for some destination nodes and increasing for some others, nevertheless, routing toward a node is independent from routing toward other nodes as stated by Observation 5.2.

Observation 5.2. Routing toward a node j is independent from routing toward other nodes.

This is true because the values of column $M_{k+1}(-, j)$ are constructed using only the values in column $M_k(-, j)$. We can, therefore, restrict our focus on one destination node j.

Theorem 5.3. Let $(S, \oplus, \otimes, \overline{0}, \overline{1})$ be a left inflationary idempotent semiring. In (3.23) and (3.24) if $A_M^{\langle 1 \rangle} \leq_\oplus M$ then the routing will converge to $A^* \otimes (A^n \otimes M \oplus I)$.

Proof. From Theorem 3.2 we know that A will be $(n-1)$-stable. From (3.26) we have, for $k \geq 0$

$$A_M^{\langle k+n \rangle} = \left(A^{k+n} \otimes M \right) \oplus A^*$$

From Theorem 5.1 we have

$$A_M^{\langle k+n \rangle} \leq_\oplus A_M^{\langle k+n-1 \rangle} \leq_\oplus \cdots \leq_\oplus A_M^{\langle 1+n \rangle} \leq_\oplus A_M^{\langle n \rangle}$$

Then

$$\begin{aligned}
A_M^{\langle k+n \rangle} &= A_M^{\langle n \rangle} \oplus A_M^{\langle 1+n \rangle} \oplus \cdots \oplus A_M^{\langle k+n \rangle} \\
&= ((A^n \otimes M) \oplus A^*) \oplus \left(\left(A^{n+1} \otimes M \right) \oplus A^* \right) \oplus \cdots \oplus \left(\left(A^{n+k} \otimes M \right) \oplus A^* \right) \\
&= (A^n \otimes M) \oplus \left(A^{n+1} \otimes M \right) \oplus \cdots \oplus \left(A^{n+k} \otimes M \right) \oplus A^* \\
&= \left(\left(I \oplus A \oplus A^2 \oplus \cdots \oplus A^k \right) \otimes (A^n \otimes M) \right) \oplus A^* \\
&= \left(A^{\langle k \rangle} \otimes A^n \otimes M \right) \oplus A^*
\end{aligned}$$

Then for $k \geq n-1$ we have

$$\begin{aligned}
A_M^{\langle k+n \rangle} &= (A^* \otimes A^n \otimes M) \oplus A^* \\
&= A^* \otimes ((A^n \otimes M) \oplus I)
\end{aligned}$$

■

Corollary 5.4. Let $(S, \oplus, \otimes, \overline{0}, \overline{1})$ be a left strict inflationary, selective, and idempotent semiring, and $\overline{1}$ is not allowed as a link weight. In (3.23) and (3.24) if $A_M^{\langle 1 \rangle} \leq_\oplus M$ then the routing will converge to A^*.

This is true because the routing will be then loop free because of DIC. This gives the idea that in order to make the routing loop free after a topology change we have to manipulate the routing matrix M in a way so that the resulting routing will be decreasing.

6. Loop Free Routing in a Non-Monotone Algebra

Monotonicity is not needed in the proof that SNC is sufficient condition for loop free routing. We can also prove the correctness of DIC and CSC without monotonicity, where we assume the strict inflationary property instead.

Decreasing routing is the foundation stone for loop free routing. However, if the underlying algebra is not monotone, then routing will not be necessarily decreasing

even in a static topology. Let us consider the EIGRP metric as an example. We will use the default K values ($K_1 = K_3 = 1, K_2 = K_4 = K_5 = 0$). The metric then will take the form (bw, d). In the simple graph presented in Figure 1, Let the distance from node k to j be $(2 \times 10^6, 1)$.

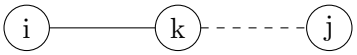


Fig. 1. Simple Graph

After applying the cost function on this distance we get

$$f(2 \times 10^6, 1) = 256 \times (5 + 1) = 1530$$

Let us assume also that the cost of the link from i to k is $(2 \times 10^6, 1)$. The computed distance in node i to node j, in this case, is

$$(\min(2 \times 10^6, 2 \times 10^6), 1 + 1) = (2 \times 10^6, 2)$$

Applying the cost function on the above distance we get

$$f(2 \times 10^6, 2) = 256 \times (5 + 2) = 1792$$

Now, let us assume that node k chooses another path towards j. Let the distance of the new path be $(5 \times 10^6, 3)$. This distance is preferred to the previous distance of k, because

$$f(5 \times 10^6, 3) = 256 \times (2 + 3) = 1280$$

The computed distance from i to j becomes

$$(\min(5 \times 10^6, 2 \times 10^6), 3 + 1) = (2 \times 10^6, 4)$$

Applying the cost function on the above distance we get

$$f((2 \times 10^6, 4)) = 256 \times (5 + 4) = 2304$$

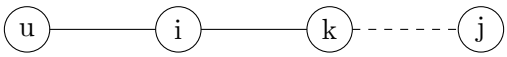


Fig. 2. Simple Graph 2

From the above discussion, the distance in node k decreased, however, the distance in i increased. It is clear, then, that the routing is not decreasing in this

case even when the topology is static. In loop free routing, this will cause a diffusing computation in node i if we are using the DIC. We can find similar scenarios for the other sufficient conditions for loop free routing. For example, if we use CSC, and there is a node u which uses node i as a successor in routing towards j (See Figure 2). The distance increase in node i will cause a diffusing computation in u. If we use SNC, and the distance in i after the increase becomes "greater" than the distance in u, then there will be a diffusing computation in u. Note that if we are using the original distributed Bellman-Ford algorithm in the latter case, node i will choose node u as a successor causing a routing loop. This means that DBF algorithm is not necessarily loop free in a static topology if the underlying algebra is not monotone, which explains why we are not bounded by $n-1$ iteration to reach convergence in non-monotone algebra.

In DUAL the affected node after topology changes starts a diffusing computations. So, that all upstream nodes change their distance to a proper value. As a result, routing will decrease to convergence when no other topology change occurs, and no future diffusing computation will occur [22]. This is true if monotonicity holds, as Garcia used original shortest paths problem in his discussion. However, in non monotone algebra, like (EIGRP metric), there is no guarantee that no diffusing computation will occur after the original diffusing computation terminates. However, the new diffusing computation if it occurs, and assuming the new topology did not change, it will affect only a subset of nodes from the original upstream nodes affected in the first diffusing computation. This means that each subsequent diffusing computation will affect a smaller set of nodes. Then, after a while no diffusing computation will ever happen, and the protocol will reach an equilibrium point and converge. The only difference is that the protocol will converge to a local optimal solution rather than a global optimal one due to the loss of monotonicity. The optimality is in the sense that a node can not change its successor to a better path considering neighboring nodes' choices. Algebraically, the local optimal solution is a fixed point for the equation $L = (A \otimes L) \oplus I$.

7. Conclusion

We have used the matrix model with semirings to investigate loop free routing. We have shown that, when the strict inflationary property holds, Bellman-Ford will calculate loop-free routing paths in a static topology. However, in a dynamic topology, routing loops may occur.

Loop free routing conditions have been presented algebraically and proved to be correct. We have also shown the relations between these conditions. In section 5 we have introduced the concept of monotone routing. We have shown that if the routing matrix decrease (or increase) in one distance vector iteration, it will continue

decreasing (or increasing) in the next iterations as long as the topology is fixed. Thus, we have to manipulate the routing matrix in order to make the routing decreasing so that DIC holds.

Finally, we have discussed the effect of loss of monotonicity. We have shown that DBF will no longer be loop free in a static topology, and we are no longer limited to $n-1$ iterations to achieve optimality (in this case, local optimality). We have also demonstrated that diffusion computations can occur even in a static topology when using DUAL, however, the algorithm will converge to a local optimum.

REFERENCES

1. Hussein Khayou & Bakr Sarakbi (2017). A validation model for non-lexical routing protocols. Journal of Network and Computer Applications.
2. Sobrinho, J. (2005). An Algebraic Theory of Dynamic Network Routing. IEEE/ACM Trans. Netw., 13.
3. Bob Albrightson, J.J. Garcia-Luna-Aceves, and Joanne Boyle. Eigrp - a fast routing protocol based on distance vectors. In *Proc. Networld/Interop 94*, 1994.
4. Sobrinho, J. (2002). Algebra and Algorithms for QoS Path Computation and Hop-by-hop Routing in the Internet. IEEE/ACM Trans. Netw., 10.
5. Sobrinho, J. (2003). Network Routing with Path Vector Protocols: Theory and Applications, in the proceedings of SIGCOMM '03.
6. Gurney, A. (2009). Construction and verification of routing algebras. (Doctoral dissertation, University of Cambridge)
7. Sobrinho, J., Griffin, T., & Term, M. (2010). Routing in Equilibrium. Mathematical Theory of Networks and System.
8. Griffin, T., & Sobrinho, J. (2005). Metarouting. In Proceedings of the 2005 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications. ACM.
9. Gondran, M., & Minoux, M. (2008). Graphs, Dioids and Semirings: New Models and Algorithms. Springer.
10. Baras, J., Theodorakopoulos, G., & Walrand, J. (2009). Path Problems in Networks. Morgan & Claypool.
11. Griffin, T., & Gurney, A. (2008). Increasing Bisemigroups and Algebraic Routing. (Vol. 4988) Springer Berlin Heidelberg.
12. Dijkstra, E., & Scholten, C. (1980). Termination Detection for Diffusing Computations. Information Processing Letters.
13. Garcia-Luna-Aceves, J. (1989). A Unified Approach to Loop-free Routing Using Distance Vectors or Link States. In Symposium Proceedings on Communications Architectures &Amp; Protocols. ACM.

14. Garcia-Lunes-Aceves, J. (1993). Loop-free routing using diffusing computations, IEEE/ACM Transactions on Networking.
15. Bob Albrightson, J.J. Garcia-Luna-Aceves, & Joanne Boyle (1994). Eigrp - A Fast Routing Protocol Based On Distance Vectors. In Proc. Networld/Interop 94.
16. Jaffe, J., & Moss, F. (1982). A Responsive Distributed Routing Algorithm for Computer Networks, IEEE Transactions on Communications.
17. Dynerowicz, S., & Griffin, T. (2013). On the forwarding paths produced by Internet routing algorithms. In 2013 21st IEEE International Conference on Network Protocols (ICNP).
18. Mohri, M. (2002). Semiring Frameworks and Algorithms for Shortest-distance Problems. J. Autom. Lang. Comb., 7(3).
19. Alim, M., & Griffin, T. (2011). On the Interaction of Multiple Routing Algorithms. In Proceedings of the Seventh COnference on Emerging Networking EXperiments and Technologies. ACM.
20. Malkin, G. (1998). RFC 2453: RIP Version 2
21. Rekhter, Y., Li, T., & Hares, S. (2006). RFC 4271: A Border Gateway Protocol 4 (BGP-4).
22. J. J. Garcia-Lunes-Aceves (1993). Loop-free routing using diffusing computations. IEEE/ACM Transactions on Networking, 1(1).
23. Timothy Griffin. (2010). Lecture notes in An Algebraic Approach to Internet Routing.
24. D. Savage , et al (2016). RFC 7868: Cisco's Enhanced Interior Gateway Routing Protocol (EIGRP).
25. M. G. Gouda and M. Schneider, "Maximizable routing metrics," in IEEE/ACM Transactions on Networking, vol. 11, no. 4, pp. 663-675, Aug. 2003, doi: 10.1109/TNET.2003.815294.
26. J. L. Sobrinho, "Correctness of Routing Vector Protocols as a Property of Network Cycles," in IEEE/ACM Transactions on Networking, vol. 25, no. 1, pp. 150-163, Feb. 2017, doi: 10.1109/TNET.2016.2567600.
27. Mohamed Saad, "Non-isotonic routing metrics solvable to optimality via shortest path" in journal of Computer Networks, vol. 145, 2018.

UDC: 654.09, 519.21

The mathematical model of Front-End calculating in DPI system

B.S. Goldstein[1] and V.V. Fitsov[2]

[1,2]The Bonch-Bruevich Saint-Petersburg State University of Telecommunications, St.Petersburg, Russia

bgold@niits.ru, noldi@iks.sut.ru

Abstract

This article describes the specialized servers that build up the DPI system architecture. Some initial data for calculating DPI system based on traffic statistics have been formalized. A mathematical model for calculating Front-End server in the DPI system, based on the model by Ventcel-Ovcharov, is provided. The DPI simulation model in GPSS World is briefly described. The results of the mathematical and simulation modeling are compared.

Keywords: DPI, QoS, queuing system (QS), queuing network, math model.

1. Introduction

Many telecom operators use deep packet inspection (DPI) systems to manage and offload their networks, analyze user interests, behavioral targeting, implement personal tariffs, protect copyrighted content, provide lawful interception according to the laws of their country, additional network protection against hacker attacks.

However, DPI requires a significant investment in hardware resources. Meanwhile the efficiency of using hardware resources in DPI systems remains understudied due to the complexity and novelty of the issue. There is a lack of necessary mathematical and simulation models for determining the parameters of a DPI architecture.

2. Related Work

There is a fairly large number of works dedicated to the questions of mathematical description, analysis and classification of network traffic. In this article, when describing and formalizing the process of transferring flows for analysis, the studies [1, 2, 3] were used. The closest is the work [2], in which a mathematical model of DPI interaction with additional servers is presented and the traffic flow is taken as an example (which will be discussed below). In Russia, the mathematical description

of packet traffic is developed by Stepanov S.N., Samuilov K.E., GaidamakaYu.V. [4], Levakov A.K., Sokolov N.A., Zaitsev V.S. [5], and others.

There are various mathematical models, but when applying, one should take into account the peculiarities of a packet traffic. Modern western research suggests that network traffic is similar to itself or fractal in structure. This kind of traffic is most successfully described by the Pareto and Weibull distributions or as fractal Brownian motion (FBM)[6, 7].

When there are several interacting queuing systems (QS), they make up a queuing network (QN). In a QN, the interest is the parameters of the output after processing in QS1, which determine the models that can be used to describe the subsequent QS (QS2). For the mathematical description of the QS as a part of the QN which receives packet traffic, the models G/M/1, G/G/1 described in [6, 8, 9] and others can be used. However, they restrict the model to one device. The G/G/V model should be used to overcome this limitation and to describe modern systems, but it cannot be calculated [6]. Most of the known models suggest an arriving Poisson flow of requests. For example, M/G/1, which is suitable for calculating the QS with one device. M/M/V and M/G/V with an infinite queue, which do not take into account the possibility of simultaneous processing of a request by several devices. As well as processor sharing models [10] and so-called Ventcel-Ovcharov model with an equal mutual assistance (where several devices work to serve one request) [11, 12, 13, 14].

According to Burke's theorem for QS1 (M/M/V and M/M/1), the distribution of time intervals between outgoing requests, as well as the time intervals between incoming requests, are distributed exponentially with the same parameter. Which was proved mathematically by Burke, based on the following: when QS1 becomes empty after it's done with the query, the time interval when the next request leaves the QS1 will be the sum of the time until the next request arrives at the QS1 and the service time of the next request; when there is a next request in the QS1 queue, after the previous one is finished, then the time interval when it leaves QS1 is distributed in the same way as the service time.

The mathematical description of the output flow of requests after processing in QS1 (of M/G/1 type, with Poisson input flow of requests) is described in [8]. At the same time, note that the value of the coefficient of the interval duration of requests moving between systems cannot be considered as a sufficient condition for the correspondence of the distribution function of the output flow to the exponential distribution law. But it is a necessary condition for the exponential approximation. The coefficient of variation of the output flow is close to 1 in two cases: when the primary server load is very low or when the variation coefficient of the request service time is very close to 1. It was shown for the M/M/1 model in [8] that the output flow is also a Poisson flow. In [9], it is said that for a primary server G/G/1 with

an unlimited capacity storage unit operating without overloads, the intensity of the outgoing flow of requests is equal to the intensity of the incoming flow (since the mathematical expectations of the intervals between successive requests at the exit and the entrance coincide). In addition, it is said in [9] that for the M/M/1 model, the variation coefficient of the outgoing flow is equal to one.

According to the study [15], the value of the variation coefficient of the resulting flow of requests (packet traffic) on the input of QS1 of a large number of sources with similar distributions for certain cases approaches one. The subsequent verification carried out in [15] using the Kolmogorov-Smirnov criterion showed that the addition of a large number (more than 100) of flows with the Weibull distribution gives the resulting flow a manner similar to Poisson flow. In this case, one should take into account the magnitude of errors in such a simplification.

Let us consider the cases in which a forced assumption will be made about the exponential distribution of the flow exiting QS1 and entering the QS2, in order to compare the results of the QS2 calculations, which can be given by the so-called Ventcel-Ovcharov model with the results obtained in the DPI simulation model described in [16]. Proceeding from the requirement to avoid packet loss, the QN (consisting of QS1 and QS2) is designed in such a way that it can be represented as a system with an infinite queue, which is important for the cases considered below.

In the first case, QS1 receives aggregated packet traffic received from more than 100 sources, where each traffic from each source can be approximately described by the Weibull distribution. Then, according to [15], we assume that the distribution of the flow of requests entering the QS1 may be close to the Poisson distribution, but it should be borne in mind that there are some errors associated with this assumption. Then, according to [9], let us assume for one QS1 service device that the output flow entering QS2 is also close to the Poisson distribution.

In the second case, if we assume that QS1 processes requests according to an exponential distribution, and is in a mode when the average intensity of requests is equal to the average intensity of request processing, but the system remains in a stable state and the waiting time in the queue does not become infinite. Then, according to Burke's theorem, the intensity of the output flow of requests from QS1 to QS2 will be distributed in the same way as the service time in QS1. This means that if we assume that requests in QS1 are treated exponentially, then the output flow will have an exponential distribution.

The third case is less interesting, since imposes even greater restrictions on the QS1. First, according to the conditions and assumptions of the first case, it is assumed that a flow close to Poisson arrives at the QS1. And secondly, QS1 must process requests according to an exponential distribution. Then, according to Burke's theorem, the distribution of time intervals between outgoing requests, as well as the

time intervals between incoming requests, are distributed exponentially. In this case, there is no limitation per 1 device as in the first case, and there is no limitation of the mode necessarily operating in a limited but stable state, as in the second case.

Further mathematical calculation of the QS2 is carried out based on these three cases.

3. DPI

The basis of the DPI system is the Bypass server, the hardware filter (HWF), Front-End (FE), PCRF (Policy and Charging Rules Function) and Back-End. Each of the DPI servers performs its own tasks and actively interacts with the rest.

Front-End - is the main element of the system, as it analyzes data from the traffic flow previously identified by the hardware filter. It uses signature analysis, statistical methods, behavioral analysis, and other approaches [17]. Having recognized the application that generated the traffic flow, Front-End asks the PCRF server for a decision on what to do with this traffic. Further, based on this decision, it receives more detailed instructions on filtering from the Back-End server. Then it gives the flow and instructions for execution on the hardware filter.

Thus, the DPI system can be represented as a QN, consisting of several QS. Where the hardware filter acts as QS1, which receives the packet data, and the output flow from it goes to QS2 (FE). Subsequent DPI servers are not taken into account in this work, due to the relatively smaller number of request sentering them.

4. The mathematical model

4.1. Formalization DPI input parameters. For the practical use of mathematical models, it is necessary to determine methods for obtaining quantitative characteristics of the operating conditions of the DPI system. This article describes the formalization of some common initial parameters and the number of requests processed on the Front-End server (1).

To calculate, you need to know or set the intensity of incoming requests for the DPI system. Therefore, it is rational to use the statistics of the transmitted traffic on the network where you plan to install the DPI system. A peculiarity of DPI is that QS1 (hardware filter) processes all incoming packets and identifies traffic flows from them, and QS2 (Front-End) receives a request to analyze a specific flow. For analysis, a certain number of packets of the flow (n_f) is transmitted, about which studies have been carried out [1, 18].

From the traffic statistics collected by wireshark or cisco NetFlow, you can get the number of packets and the number of flows during statistical analysis, and then the average number of packets in the flow (n_{af}) and the rate of arrival of the packets (λ_0). In this case, you can get the number and frequency of occurrence of new

unknown flows, the probability that the flow was previously known (P_{kn}), the time of occurrence of the flow and the average duration of the flow.

The rate of arrival of unknown packets will be determined by the product of the probability of an unknown flow $(1 - P_{kn})$ and λ_0. Based on the above description of the interaction of DPI servers, only a part of the packets of the flow as a request is sent for analysis, which is set by the ratio of the number of analyzed packets (n_f) to the n_{af}. For each analyzed flow, FE sends a response to the hardware filter, which is set by the ratio of 1 to the n_{af}. From these components, the number of requests processed at the Front-End is obtained, determined by the formula (1).

$$\lambda_{fe} = (1 - P_{kn}) \times (\frac{n_f + 1}{n_{af}}) \times \lambda_0 \quad (1)$$

The parameters defined here, as well as other parameters of traffic characteristics and DPI system features, are presented in table 1.

4.2. Ventcel-Ovcharov model and Front-End. Previously, a forced assumption was made in the cases considered in which the exponential distribution of the flow leaving the hardware filter and entering the Front-End, in order to compare the results of Front-End calculations, which can provide the so-called Ventcel-Ovcharov model with the results obtained in the DPI simulation model.

Ventcel and Ovcharov distinguish two types of mutual assistance: full and partial (equal). This article discusses equal mutual assistance. Since the DPI system must process all requests, to simplify the calculations of the Front-End mathematical model, it is advisable to take a model with an infinite queue and with equal mutual assistance [11] mentioned, but not completely described, in the works of Ventcel.

The concept of the model with equal mutual assistance is to combine channels into groups for the joint service of requests. In this case, a system with equal mutual assistance will have 3 modes of operation : I - the number of requests is less than the maximum number of groups (like a classical QS), II - the number of requests is greater than the maximum number of groups, but less than the number of channels (transient mode), III - the number of requests is greater than the number of channels (like a classical QS). One of the advantages of the considered mathematical model is the use of all possible resources of the system before the number of requests equals the number of channels. The Ventcel-Ovcharov model with equal mutual assistance describes the operation of servers, when many service processors can distribute computing power to simultaneously work on a single request in the system, or evenly to work on several requests received in the system.

Let us assume that the intensity of servicing one request by a group of channels will be directly proportional to the number of involved channels. Taking into account the above information about the system, you can get the relevant formulas. Let us

denote V - the number of devices in the system, l - the number of devices in one group, h - the maximum possible number of groups. For this model, the probability of system downtime (P_{0fe}), the ratio of the intensity of incoming requests to the intensity of processing by one group (α), the ratio of the intensity of incoming requests to the intensity of processing by all devices of the FE (β), the formula for the average time spent in the queue ($\bar{t_q}$), the formula for the average service time ($\bar{t_w}$) is defined in [11]. The main indicator of the performance of the DPI system is the average time spent by a request in the system (2). We get it by adding the average service time ($\bar{t_q}$) and the average waiting time ($\bar{t_w}$) (with substituting P_{0fe}):

$$\bar{T_{fe}} = \frac{\left(\frac{1}{\sum_{i=1}^{h} i\times l\times\mu+\sum_{j=h+1}^{V} j\times\mu} + \frac{\beta}{V\times\mu} \times \frac{\alpha^h}{h!} \times \beta \times \frac{1}{(1-\beta)^2}\right)}{\left(\sum_{i=0}^{h} \frac{\alpha^i}{i!} + \frac{\alpha^h}{h!} \times \frac{\beta^{h+1}}{1-\beta}\right)} \tag{2}$$

The mathematical model for the Front-End DPI server was shown. Part of the initial data for calculating the DPI system has been determined. The formula of the final processing time of requests (2) of the Front-End DPI server is presented.

4.3. Data sets and calculation. In this section, we will briefly present the calculated average time spent by a request on the Front-End server of the DPI system obtained for a given set of initial data. All values are summarized in table 1.

Name	λ_0	P_{kn}	n_{af}	n_f	V_{fe}	μ_{fe}	A_{fe}	λ_{fe}	$\bar{T_{fe}}$
Value	409875	0.78	1093	10	1	917	0.99	908	0.10847

Table 1. Initial and calculated data

To obtain the initial data for the calculations, it was necessary to study the statistics of network traffic that is supposed to be passed through the DPI system. One must find the total number of flows, n_{af} and $1 - P_{kn}$. To determine $1 - P_{kn}$, it is necessary to divide the number of new flows by the total number of flows received during a given period of time. For the calculations presented, we used the traffic collected in the dormitories of SPbSUT using Cisco NetFlow equipment. FE performance was taken to comply with system stability factors. For simplicity, the calculations were performed for one FE server (which reduces the visibility of the Ventcel-Ovcharov model). The result of the calculation showed that equipment with a given performance successfully copes with processing the load with a stability coefficient of 0.99. However, a temporary increase in the number of requests to the Front-End can lead to a significant increase in the time spent by a request on the FE server. With the obtained and given in table 1 value of the average time of finding a request on the FE server, there is no need to change the server performance.

5. Simulation

For the DPI system, a simulation model (SM) was created in GPSS [16]. In the SM, the initial parameters are set, presented in table 1. However, to describe the traffic arrival, the Weibull distribution (for the HWF) is used, and for the processing law, the exponential distribution (for the HWF and FE). It is possible to apply a SM in GPSS to obtain the probabilistic-temporal characteristics of the DPI system and compare with the results of the calculation using the mathematical model.

Model type	$\bar{T_{dpi}}$, **ms**	**HWF req., items**	$\bar{T_{hwf}}$, **ms**	**FE req., items**	$\bar{T_{fe}}$, **ms**
SM	108.9	884128	24.278	908	84.593
MM	142.8	847983	33.306	908	108.47

Table 2. Time characteristics of hardware filter and Front-End

The simulation modeling results showed that DPI system hardware can handle the load with a stability coefficient of 0.99 in this case. Changes in the characteristics of the distribution of incoming traffic to the DPI system, which was described in the simulation model by the Weibull distribution, significantly affects the size of the queue, and through it, the processing time of requests in the hardware filter. Comparison of the calculation results based on the mathematical and simulation models given in table 2 indicates the possibility of their use.

6. Conclusion

The aforementioned work formalized the initial data for calculating the DPI system based on traffic statistical data. The mathematical model of Ventcel-Ovcharov is presented. The possibility of practical application of this model for calculating specialized DPI servers is shown.

This paper describes the need for a mathematical model to determine the parameters of the Front-End server in the DPI architecture. The mathematical model is used to calculate the average analysis time for a given number of processors in a dedicated DPI server. The use of the DPI mathematical model will help reduce the purchase cost of DPI equipment and avoid overloads on communication networks early. The developed formalization of the ratio of flows and packets can be useful in calculating DPI systems and other similar systems.

REFERENCES

1. A. Dainotti, A. Pescape, C. Sansone, Early classification of network traffic through multi-classification, in Proc. Traffic Monitoring and Analysis III (2011) 122–135.

2. B. Niang, Bandwidth management - a deep packet inspection mathematical model, in Proc. ICUMT-2014 (2014) 169–175.
3. Y. Zeng, S. Guo, Deep packet inspection with delayed signature matching in network auditing, in Proc. ICICS (2018) 75–91.
4. G. Basharin, Y. Gaidamaka, K. Samuilov, N. Yarkina, Models for analyzing the quality of service in next generation communication networks, RUDN, M., 2008.
5. K. Samuilov, A. Levakov, , N. Sokolov, V. Zaitcev, Method of arrival process description for packet switch, in Proc. of the Workshop on ITTMM (2020) 47–56.
6. C. Grimm, G. Schluchtermann, IP Traffic Theory and Performance, Springer, 2008.
7. A. Lozhkovsky, V. Kaptur, , O. Verbanov, Mathematical model of packet traffic, Bulletin of the National Polytechnic University KhPI (9) (2011) 113–119.
8. N. Sokolov, Telecommunication networks planning tasks, BHV, Spb, 2012.
9. T. Aliev, Basics of modeling discrete systems, SPbGU ITMO, Spb, 2009.
10. L. Kleinrock, Time-shared system: a theoretical treatment, J. Assoc. Comput. 14 (2) (1967) 242–251.
11. A. Novikov, V. Fitsov, Application of the wentzel-ovcharov mathematical model with uniform mutual assistance for modern nfv systems, in Proc. Actual problems of information and telecommunications in science and education VIII Conf. 1 (2019) 705–709.
12. E. Wentzel, Operations research, Soviet Radio, M., 1972.
13. L. Ovcharov, Applied problems of the theory of queuing, Mashinostroenie, M., 1969.
14. E. Wentzel, Probability theory, Nauka, M., 1969.
15. V. Zaitcev, Characteristics of the total flow of ip packets at the input of the switching node of a multiservice network, in Proc. XII conference Technologies of the Information Society (2018) 178–182.
16. V. Fitsov, A simulation model of the dpi system based on the gpss world software, in Proc. Actual problems of information and telecommunications in science and education V Conf. (2016) 539–545.
17. W. Song, M. Beshley, K. Przystupa, H. Beshley, O. Kochan, A. Pryslupskyi, D. Pieniak, J. Su, A software deep packet inspection system for network traffic analysis and anomaly detection, Sensors 20(6): 1637 (2020).
18. Z. Wang, S. Zhu, Y. Cao, Z. Qian, C. Song, S. Krishnamurthy, K. Chan, T. Braun, Symtcp: Eluding stateful deep packet inspection with automated discrepancy discovery, in Proc. NDSS-2020 Symposium (2020).

UDC: 519.6

Simulation-based Analysis of Mobility Models for Wireless UAV-to-X Networks

V. Khalina[1], V. Prosvirov[1], Yu. Gaidamaka[1,2], J. Pokorny[3], J. Hosek[3], K. Samouylov[1,2,3]

[1]Peoples' Friendship University of Russia (RUDN Universiry), Moscow, Russian Federation
[2]Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russian Federation
[3]Brno University of Technology, Brno, Czech Republic

viktoriya.khalina@gmail.com, gnarwhal18@gmail.com, gaydamaka-yuv@rudn.ru, jiri.pokorny@vutbr.cz, hosek@feec.vutbr.cz, samuylov-ke@rudn.ru

Abstract

Recently, the use of air base stations located on unmanned aerial vehicles (UAVs) has attracted great attention. Static deployment of a sufficient number of such UAVs allows for uniform wireless coverage in the demanded areas, where the existing cellular infrastructure has white spots or insufficient capacity. However, UAVs mobility may be required for applications, where UAVs are used to provide communications for mobile groups of users (e.g., massive sport or community events like marathon or music festival) or for patrolling tasks with soft requirements for data transmission delays (for example, when collecting information from a large number of mMTC sensors). In such tasks, the movement of UAVs can significantly increase the efficiency of the system, since in this case the coverage of the area can be provided by a smaller number of UAVs following the dynamics of ground users. Nowadays, more and more often the question arises about the mobile communications availability in a remote area, for example, during events or search operations. The high-quality communication lack on demand in such areas is unacceptable in modern conditions. Therefore, the study of the behavior of a dynamic UAV network is necessary for decision-making in such scenarios. The main contribution to this work is making the user behavior more human-alike according to the real scenarios: users set and their mobility model. The paper considers two models of UAVs movement, the effectiveness of which is estimated from the point of view of the coverage probability and average fade duration of the signal.

Keywords: UAV, user, coverage probability, mobility models, wireless on-demand connectivity.

1. Introduction

Currently, unmanned aerial vehicles (UAVs) are easily deployable air devices that can be used as base stations and repeaters to provide additional network coverage at various public events, in emergency situations, as well as to collect data and perform environmental monitoring [3, 9, 15]. UAVs can be classified into devices, whose working height is measured in tens of kilometers, and on devices operating at heights of several tens of meters, as well as on drones with a fixed or a rotating wing [12].

To support the upcoming 5G technology, simply replacing the antennas at all base stations is not enough. In addition, more and more often the question arises about the availability of mobile communications in a remote area, for example, during search and rescue operations. The lack of communication in such areas is unacceptable in modern conditions.

Therefore, it is necessary, with a high degree of accuracy, to analyze existing network modeling mechanisms using auxiliary devices/retranslators in order to meet the 5G requirements. In the well-known literature [2, 4, 14], the main results are concentrated on modeling a dynamic network of Unmanned Aerial Vehicles (UAVs) over a stationary user. The interest in this work is due to the consideration of a scenario closer to reality - when a certain number of users move inside the area according to the movement models close to the considered system scenarios.

The need to increase communication coverage and/or capacity is determined by the presence of recent trends: a rapidly growing number of network users, quality of service requirements for such a large number of subscribers, and technology requirements (for example, high sensitivity and the absence of critical delay) [11]. Therefore, by deploying a UAV network, you can fulfill the above requirements. Therefore, the task of the characteristics of mobile users' coverage evaluation is stated.

In this paper, the ring patrol model [4] and the random patrol model [14] are taken as a basis. The analysis of the user coverage probability is carried out, which is defined as the proportion of subscribers for whom the signal power at the user's receiver from the closest UAV to the receiver exceeds a predetermined threshold necessary for communication, as well as the average fade duration.

2. System Model

We consider a network model of M UAVs deployed to serve N users. Let the service area be a parallelepiped, at the base of which is a square with side $S_A = 2R$ and height H.

All UAVs move inside the simulation area, and at each moment, the nearest UAV is selected for user service. The distance from the user to the nearest UAV is indicated by ρ. The user is intercepted by another UAV when it becomes the closest

to the user. All other airborne base stations are considered as interference sources for the users.

2.1. User mobility. In this paper, as a model of user movement, we consider the Random Direction Mobility model (RDM) [10]. In the RDM model, at each moment of time, the i-th user ($i = 1, \ldots, N$) selects a random direction uniformly distributed over $(0, 2\pi)$, and moves in this direction at a constant speed v_i over a period of time that has an exponential distribution with the parameter $1/E(\tau_i)$, where $E(\tau_i)$ is the average movement time. Note that the speed and direction of the movement of one user are independent of the speed and direction of other users.

2.2. UAVs mobility. In this paper, we use two models for UAVs movement: the random patrol model and the ring patrol model. These patrol models, as the most accessible for analysis and deployment, were chosen to assess the quality of user service when UAVs are not aware of their behavior.

Random patrol model (based on [14]). The model of random patrolling involves alternating the vertical and horizontal movement of the i-th ($i = 1, \ldots, M$) UAV (Figure 1), changing it's height $h_i(t)$ and the horizontal position $z_i(t)$ in accordance with the Random Waypoint Mobility (RWP) model. Initially, the UAV is launched at a random height H_1 uniformly distributed in the interval $[H_{min}; H_{max}]$. Then the UAV selects a random point at a height of H_2, also uniformly distributed in the interval $[H_{min}; H_{max}]$, and moves toward it at a speed v (the same for all UAVs) for several time intervals, without changing direction, until it reaches its destination. Upon reaching a point at a height of H_2, the UAV remains at this point for the delay time T_s, uniformly distributed in the interval $[\tau_{min}; \tau_{max}]$, where τ_{min} is the minimum, and τ_{max} is the maximum delay time. During this time, the UAV moves in the horizontal plane, choosing a point uniformly distributed on the segment $[z_i(t)–R, z_i(t) + R]$ and moves toward it with a speed v.

Note that a change in the horizontal position of the UAV at the current height may occur several times during the delay, and the UAV may not have time to reach a point, in this case, its movement is interrupted and it remains in the current horizontal position. After the time T_s, the UAV selects the vertical point again and flies to it, further performing the same iterations.

Ring patrol model (based on [4]). This model is based on a deterministic ring trajectory (see Figure 2). All UAVs fly at the same height H_D. To achieve uniform coverage of the patrol area with a network of M UAVs in the case of a ring path, the following conditions must be met:

- The i-th UAV moves in a circle with a radius:

$$R_i = \sqrt{\frac{i}{M+1}}, \; i = \overline{1, M}; \tag{1}$$

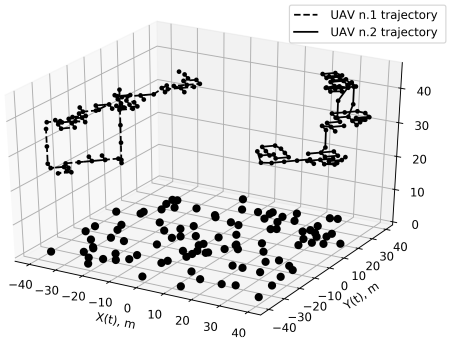


Fig. 1. UAVs motion paths for the random patrol model

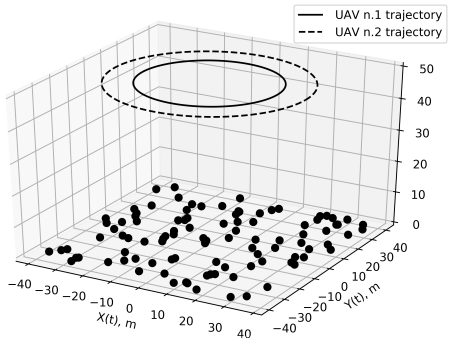


Fig. 2. UAV motion paths for the ring patrol model

- The i-th UAV makes a full turn in τ seconds with a constant speed:

$$v_i = 2\pi \frac{R_i}{\tau},\ i = \overline{1, M}; \tag{2}$$

- All UAVs fly at the same angular velocities, so the angle between adjacent UAVs is $\frac{2\pi}{M}$ and is kept at any time.

However, in the random patrol model, all UAVs have the same speed v. Therefore, in order to get results that are fair for comparison, we will calculate the speed of each UAV in its orbit for the ring patrol model based on the average speed v_{av}, which should be equal to the speed of each UAV in the random spatial patrol model ($v_{av} = v$):

$$\tau_{av} = \frac{2\pi \sum_{i=1}^{n} R_i}{v_{av} M}, \quad v_i = 2\pi \frac{R_i}{\tau_{av}}, i = \overline{1, M}. \tag{3}$$

3. Metrics of Interest and Numerical Results

To assess the applicability of the mobility model in the proposed scenario, the coverage probability indicator is estimated, which is defined as the proportion of users, for whom the signal power at the receiver, from the j-th user to the nearest UAV receiver (SNR_j), exceeds a predetermined threshold (γ) necessary for the communication:

$$P_c = \frac{1}{N} \sum_{j=1}^{N} 1\{SNR_j \geq \gamma\}. \tag{4}$$

The SNR_j (dB) indicator between the user and the UAV is calculated as:

$$SNR_j = 10 \log_{10} \frac{P_{rx_j}}{N_0}, \tag{5}$$

where P_{rx_j} is power of the received signal [W] at the receiving antenna (device of the j-th user); N_0 is the noise power [W]. The power of the signal received by the user device is modeled using the Friis transfer formula:

$$P_{rx_j} = P_{tx} + G_{tx} + G_{rx} + 20\log_{10}\left(\frac{c}{4\pi r_{jk} f_c}\right). \quad (6)$$

Here, P_{tx} denotes the transmit power (antenna power) [dBm], G_{tx} and G_{rx} are the transmit and receive antenna gains [dBi], c is the speed of light [m/s], f_c is the frequency [GHz], r_j is the distance between user j and the nearest UAV [m]. Noise power $N_0 = -174 + 10\log_{10}(B)$ [dBm], where B is the bandwidth [Hz], -174 is the noise power [dBm], emitted for 1 Hz. The conversion from dBm to W is performed as $W = \frac{10^{\frac{dBm}{10}}}{1000}$.

If the value of SNR_j exceeds a predetermined threshold value ($\gamma = 20$ dBm) [7], then the connection is available and the user is considered covered.

Let us consider the stochastic process (SP) $S_j(t)$, which characterizes the quality of the connection for the j-th user at time t, so that the value of the SP will coincide with the SNR at the receiver of the j-th user. Let the threshold γ exist for the process under consideration. The characteristics of SP τ_{ij}^- is the duration of the i-th signal absence period and τ_{ij}^+ is the duration of the i-th communication period between the devices [5]. Note that such an indicator of the quality of service (QoS) as the average fade duration (AFD), studied in [4, 13, 6], corresponds to the mean of random variable τ_{ij}^-.

Let $N_j(T)$ be a counting SP, whose value corresponds to the number of positive intersections of the threshold γ by the $S_j(t)$ during the simulation time T. Thus, $N_j(T)$ coincides with the number of periods of lack of communication between the interacting receiver and transmitter [5], i.e., periods during which the level of the signal received by the user was insufficient to provide a service with the required quality.

Here, τ_{ij}^- is a random value of i-th lack period duration of communication for the j-th user on the modeling interval with length T, $i = 1, \ldots, N_j(T)$. Then, for $N_j(T) > 0$, the average fade duration $\overline{\tau_j}$ for the j-th user during the simulation time T can be found by the formula:

$$\overline{\tau_j} = \frac{1}{N_j(T)} \sum_{i=1}^{N_j(T)} \tau_{ij}^-. \quad (7)$$

To calculate the AFD δ for a user group, we use the formula:

$$\delta = \frac{1}{M} \sum_{j=1}^{M} \overline{\tau_j}, \quad (8)$$

where M is the number of users.

For numerical analysis, the simulator was developed, consisting of two models of user mobility and two models of UAV movement. As the initial data for modeling, we used the values given in Table 1, which were chosen as the most consistent with the real scenarios of UAV user coverage [8].

Symbol	Value	Description
P_{tx}	24 dBm	Transmitting power
G_{tx}	3 dBm	Transmitting antenna gain
G_{rx}	3 dBm	Receiving antenna gain
N_0	27.434 dBm	Noise power
B	0.56 GHz	Bandwidth
γ	20 dB	SNR threshold
M	3 or 5	Number of UAVs
N	100	Number of users
S_A	100*100 m^2	Area of interest
H_D	20	UAV altitude for the ring patrol model
$[H_{min}, H_{max}]$	[15; 20] m	UAV flight altitude range for the random patrol model
v_i	1.4 m/s	Users speed
v_{av}	5 m/s	UAVs average speed

Table 1. Simulation parameters

In Figure 3, showing the coverage probabilities, we can observe that the curves corresponding to the random patrol model have the most stable character. Due to the features of this model, UAVs can "hover" over a group of users or a large number of them for a long time, and therefore the smallest deviation from the coverage probability averaged over launches is ensured. In addition, this result is facilitated by the fact that in the random patrol model (RPM), UAVs can change their height, and with UAVs height decreasing, the radius of users' coverage is increasing. At the same time, the coverage probability indicator for the ring patrol model is spasmodic in nature with the largest discrepancy interval compared to the random patrol model. Since the model involves constant movement, at certain points in time, UAVs can be located above the smallest congestion of users.

However, in Figure 3 we can see, that for the random patrol model, the curves characterized the AFD behaves worse than for the ring patrol model. As mentioned above, UAVs can be located over one cluster of users or a group of users for a long time, while not serving another cluster or group of users for a longer time.

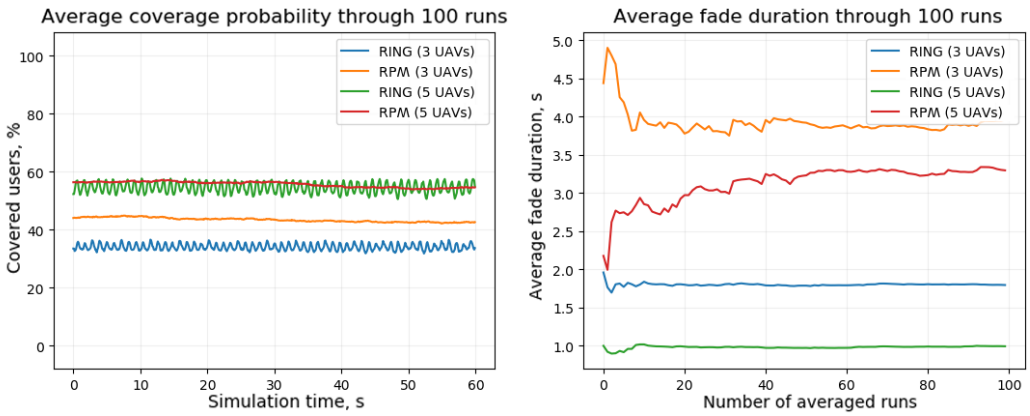


Fig. 3. Metrics of interest

4. Conclusion

The results obtained in the article showed that none of the models is universal from the point of view of the coverage probability. The advantage of the ring patrol model is the uniform coverage of the region at a low average fade duration. Also, such a model gives good results for non-feedback scenarios, when the drones have no information about the movement of users. In the latter case, the random patrol model yields results that are superior to the ring patrol. It is planned to continue the investigation of the random patrol model for use in the 3GPP 5G Integrated Access and Backhaul (IAB) [1].

5. Acknowledgments

The publication has been prepared with the support of the "RUDN University Program 5-100". The reported study was partially funded by RFBR, projects 18-07-00576 and 20-07-01064. This article is based upon the support of international mobility project MeMoV, No. CZ.02.2.69/0.0/0.0/16 027/00083710 funded by European Union, Ministry of Education, Youth and Sports, Czech Republic, and Brno, University of Technology.

REFERENCES

1. 3GPP RP-17148. New SI: Study on Integrated Access and Backhaul for NR. Rel.15. Dubrovnik, Croatia. March 2017.
2. Chetlur V.V., Dhillon H.S. "Downlink coverage analysis for a finite 3-D wireless network of unmanned aerial vehicles"". In IEEE Transactions on Communications 2017. Vol. 65, No. 10. P. 4543-4558.

3. Cheng F., Gui G., Zhao N., Chen Y., Tang J., Sari H. "UAV Relaying Assisted Secure Transmission With Caching". In IEEE Transactions on Communications 2019. Vol. 67, No. 5. P. 3140-3153.
4. Enayati S., Saeedi H., Pishro-Nik H., Yanikomeroglu H. "Moving Aerial Base Station Networks: A Stochastic". In IEEE Transactions on Communications 2019. Vol. 18, No. 6. P. 2977-2988.
5. Orlov Yu., Fedorov S., Samuylov A., Gaidamaka Yu., Moltchanov D. "Simulation of devices mobility to estimate wireless channel quality metrics in 5G networks". In AIP Conference Proceedings 2017. Vol. 1863, No. 090005.
6. Hadzi-Velkov Z. "Level Crossing Rate and Average Fade Duration of EGC Systems With Cochannel Interference in Rayleigh Fading". In IEEE Transactions on Communications 2007. Vol. 55, No. 11. P. 2104-2113.
7. How to: Define Minimum SNR Values for Signal Coverage. Available at: http://www.wireless-nets.com/resources/tutorials/define_SNR_values.html (accessed 7 February 2020)
8. Mezzavilla M. et al. "End-to-End Simulation of 5G mmWave Networks". In IEEE Communications Surveys & Tutorials 2018. Vol. 20, No. 3. P. 2237-2263.
9. Mozaffari M., Saad W., Bennis M., Debbah M. "Communications and Control for Wireless Drone-Based Antenna Array". In IEEE Transactions on Communications 2019. Vol. 67, No. 1. P. 820-834.
10. Mozaffari M., Saad W., Bennis M., Debbah M. "Unmanned aerial vehicle with underlaid device-to-device communications: Performance and tradeoffs". In IEEE Transactions on Communications 2016. Vol. 15, No. 6. P. 3949-3963.
11. Muthanna A., Masek P., Hosek J., Fujdiak R., Hussein O., Paramonov A., Koucheryavy A. "Analytical Evaluation of D2D Connectivity Potential in 5G Wireless Systems". In International Conference on Next Generation Wired/Wireless Networking Conference on Internet of Things and Smart Spaces 2016. Vol. 9870. P. 395-403.
12. Nain P., Towsley D., Liu B., Liu Z. "Properties of random direction models". In proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies 2005. Vol. 3. P. 1897-1907.
13. Shankar P.M. "Fading and Shadowing in Wireless Systems". Springer Science+Business Media, LLC 2012.
14. Sharma P.K., Kim D.I. "Random 3D Mobile UAV Networks: Mobility Modeling and Coverage Probability". In IEEE Transactions on Wireless Communications 2019. Vol. 18, No. 5. P. 2527-2538.
15. Zhang S., Zhang H., Di B., Song L. "Cellular UAV-to-X Communications: Design and Optimization for Multi-UAV Networks". In IEEE Transactions on Communications 2019. Vol. 18, No. 2. P. 1346-1359.

UDC: 530.145

On the quantum teleportation of Bell states performed on 5-qubit IBM Q computers

V.P. Gerdt[1,2,3] and E.A. Kotkova[1,3]

[1]Joint Institute for Nuclear Research, Dubna 141980, Russian Federation
[2]Peoples' Friendship University of Russia, Moscow 117198, Russian Federation
[3]Dubna State University, Dubna 141982, Russian Federation

gerdt@jinr.ru, ekaterina.a.kotkova@gmail.com

Abstract

Keywords: quantum teleportation, IBM Quantum Experience

Noisy Intermediate-Scale Quantum (NISQ) [1] technology is rapidly developing over last years. The near-term quantum computers with 50-100 qubits are able to perform tasks which surpass the capabilities of today's classical digital computers. However, noisy qubits and quantum gates lead to limitation of the size of quantum circuits that can be executed reliably. In the given talk we present our results on implementation on the 5-qubit IBM computers accessible via the cloud platform IBM Quantum Experience [2] and Qiskit framework [3], the protocol [4] of quantum teleportation of Bell (EPR) states. We adapt the original version of this protocol to devices connection graphs and the set of gates built-in IBM Q, which includes the measurement gate and the 2-qubit control-$\oplus$ (CNOT) gate and the following one-qubit parametric gates

$$U2(\phi,\lambda) = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{\exp i\lambda}{\sqrt{2}} \\ \frac{\exp i\phi}{\sqrt{2}} & \frac{\exp i(\lambda+\phi)}{\sqrt{2}} \end{pmatrix}, \tag{1}$$

$$U3(\theta,\phi,\lambda) = \begin{pmatrix} \cos\left(\frac{\theta}{2}\right) & -\sin\left(\frac{\theta}{2}\right) e^{i\lambda} \\ \sin\left(\frac{\theta}{2}\right) e^{i\phi} & \cos\left(\frac{\theta}{2}\right) e^{i(\lambda+\phi)} \end{pmatrix}, \quad \theta,\lambda,\phi \in [0,2\pi]. \tag{2}$$

The gates used in the teleportation protocol,

$$\oplus = U3(\pi,0,\pi) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad H = U2(0,\pi), \tag{3}$$

are transpiled by IBM to the previously mentioned parametric ones.

The Bell states are the maximally entangled 2-qubit states.

$$|q_0\rangle = |x\rangle\,, |q_1\rangle = |y\rangle \longrightarrow |\beta_{xy}\rangle = \frac{1}{\sqrt{2}}\left(|0, y\rangle + (-1)^x\,|1, 1-y\rangle\right), \quad x, y \in \{0, 1\}. \quad (4)$$

Since any gate introduces an error, it is worthwhile to match the qubits in the quantum algorithm with computer qubits in a way that minimizes the number of auxiliary gates. We made the first implementations of the protocol on the IBM Q Yorktown (Fig. 1) when all connections between the qubits of the device were unidirectional, and the measurement error of qubit q1 equal to 0.30275 was much higher than errors of other qubits (0.01325–0.044). Due to these limitations, in our realization we chose the qubit matchings shown in Table 1. It means that the Bell states are prepared on qubits q0, q1 for the implemented Circuits 1 and 2 and on q3, q4 for Circuit 3. The states are transported to qubits q2, q4 for Circuit 1, to q2, q3 for Circuit 2, and to q2, q0 for Circuit 3.

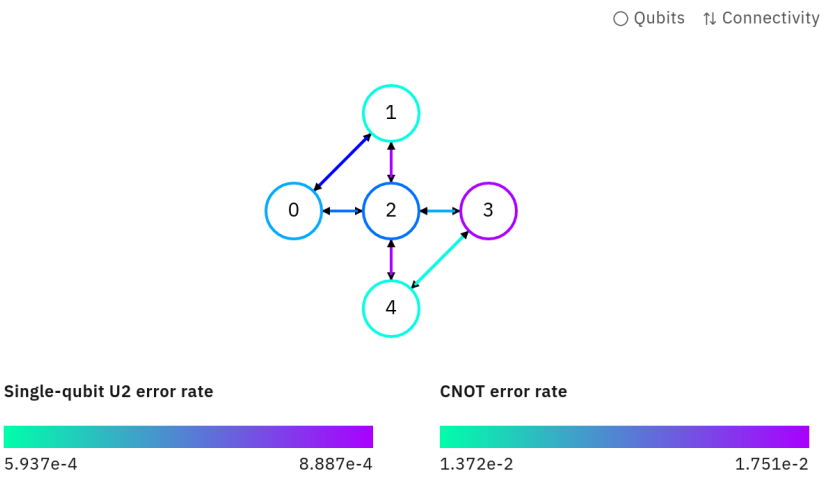


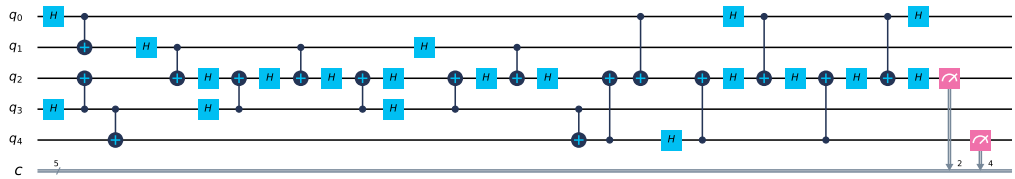
Fig. 1. Quantum computer IBM Q Yorktown

After certain modification of IBM Q Yorktown when all connections between qubits had become bidirectional, a reduction of implemented circuits depths became possible, since a programmatic change of some CNOT gates directions was no longer necessary. The initial and reduced versions of Circuit 1 are shown in Fig. 2.

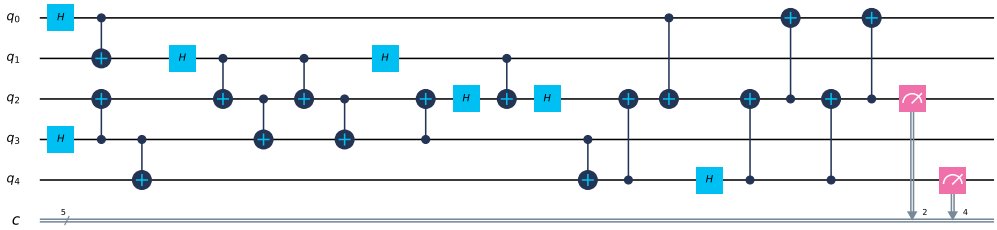
You can see the results for the initial and reduced versions of Circuit 1 in Fig. 3. As you may notice, the changes in error rates of the device play more crucial role in the quality of the implementation rather than the choice of the reduced or initial version of the circuit.

Qubit of the protocol	Implementation 1	Implementation 2	Implementation 3
0	0	0	3
1	1	1	4
2	3	4	1
3	2	2	2
4	4	3	0

Table 1. Matchings of the quantum algorithm qubits to the qubits of the IBM Q Yorktown



a)



b)

Fig. 2. Quantum Circuit 1 on IBM Q Yorktown a) initial version, b) reduced after the device modification

We have also implemented the teleportation protocol on IBM Q Rome and have compared the results for its implementation on IBM Q Burlington, IBM Q Essex, IBM Q London, IBM Q Ourense, and IBM Q Vigo. Despite the technical improvement of the IBM quantum hardware advent of quantum computers such as IBM Q Vigo, IBM Q Essex, IBM Q Ourense, and IBM Q Rome on which the teleportation protocol was implemented successfully, in many cases a further decrease of the hardware errors is needed.

REFERENCES

1. R.S.Sutor. Dancing with Qubits: How quantum computing works and how it can change the world. Packt, 2019.
2. IBM Quantum Experience, `https://www.ibm.com/quantum-computing/technology/experience/`
3. Qiskit, `https://qiskit.org/`
4. V.N.Gorbachev and A.I.Trubilko. Quantum teleportation of an Einstein-Podolsky-Rosen pair using an entanglement three-particle state // Journal of Experimental and Theoretical Physics. 2000. V. 118, no. 5. P. 1036–1040. arXiv:9906110 [quant-ph].

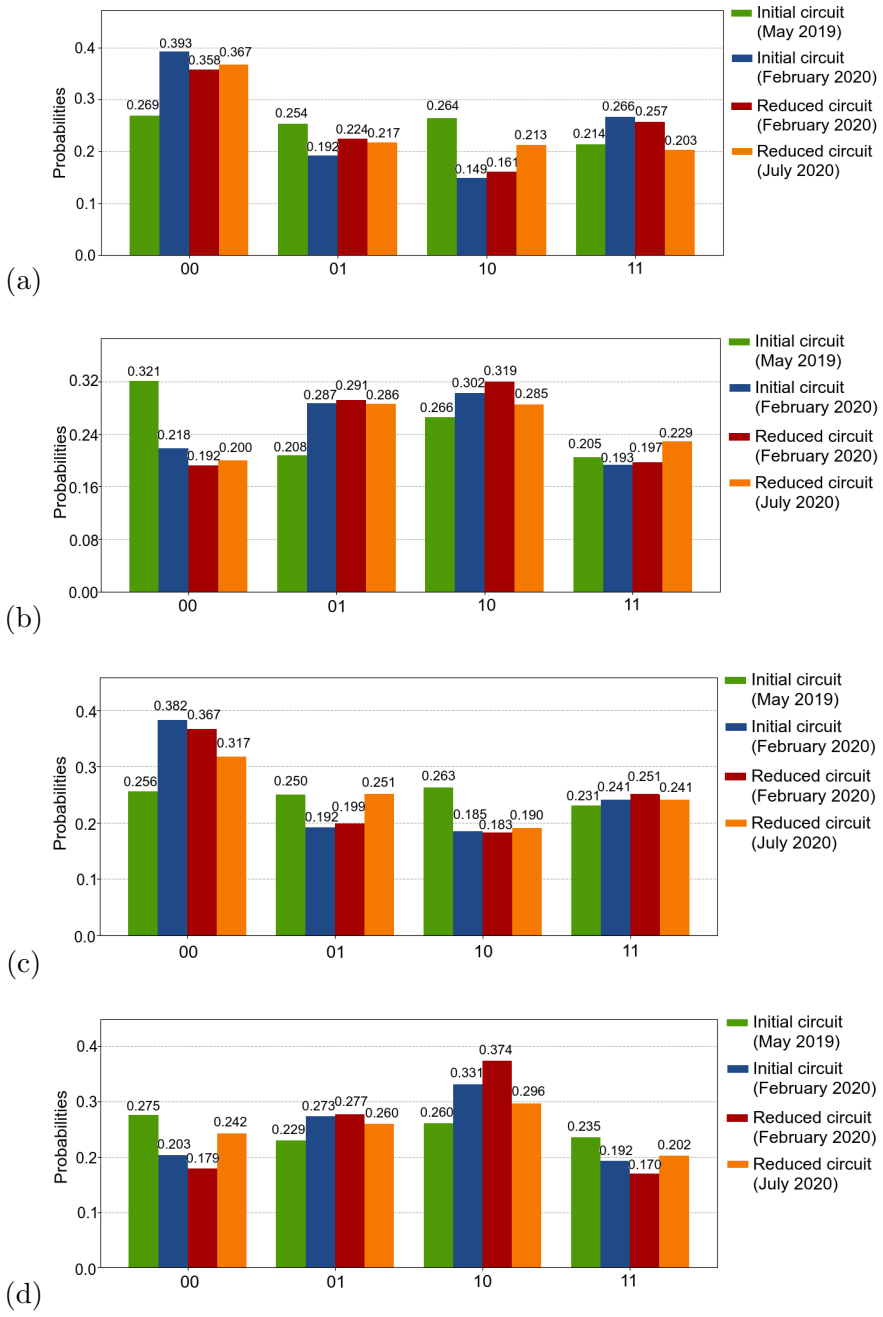


Fig. 3. Results of teleportation of the Bell states on the IBM Q Yorktown with readout in the classical basis. a) state $|\beta_{00}\rangle$, b) state $|\beta_{01}\rangle$, c) state $|\beta_{10}\rangle$, d) state $|\beta_{11}\rangle$

УДК: 004.738

Подходы к определению приоритетов обслуживания сетевого трафика для гетерогенных шлюзов промышленного Интернета вещей

В.А. Кулик[1], Д.А. Галлямов[1], Р.В. Киричек[1]

[1]ФГБОУВО «Санкт-Петербургский государственный университет телекоммуникаций им. проф. М.А. Бонч-Бруевича», пр. Большевиков д.22, корп.1, Санкт-Петербург, Россия

vslav.kulik@gmail.com, gallyamovda@yandex.ru, kirichek@sut.ru

Аннотация

В данной статье рассматриваются подходы к заданию приоритетов обслуживания сетевого трафика приложений промышленного Интернета вещей, согласно ранее определенной классификации. Была определена общая структура модельной сети для исследования приоритизации сетевого трафика промышленного Интернета вещей, затем для данной модельной сети были заданы основные методы приоритизации трафика – на основе VLAN, на основе используемых протоколов, портов и адресов, на основе систем DPI и систем машинного обучения.

Ключевые слова: промышленный Интернет вещей, качество обслуживания, фильтрация трафика, модельная сеть, IIoT, QoS, traffic filtration, model network

1. Введение

В настоящее время в рамках современных сетей связи появляется множество новых видов сетевого трафика и приложений, связанных с концепцией Интернета вещей (ИВ). Данные виды приложений могут включать в себя как критичные к высоким уровням задержек системы (например, пожарная сигнализация, системы мониторинга дорожного движения, системы контроля работы промышленного оборудования и т.д.), так и толерантные к ним (например, системы мониторинга состояния окружающей среды, системы позиционирования объектов в сельском хозяйстве и др.) [1]. В технологии ИВ в настоящее время принято включать: медицинские сети, Интернет нановещей, системы «умный город» и «умный дом»,

Исследование выполнено при финансовой поддержке гранта Президента Российской Федерации для государственной поддержки молодых российских ученых - докторов наук МД-2454.2020.9.

тактильный Интернет, всепроникающие сенсорные сети и др. [2, 3] Одной из самых важнейших концепций, развивающихся в рамках ИВ является промышленный Интернет вещей (ПИВ), который подразумевает использование концепции и технологий Интернета вещей в рамках задач промышленной автоматизации [4]. В рамках ПИВ особенно важной составляющей является обеспечение низких показателей задержек и джиттера, так множество производственных процессов имеют высокие требования к точности выполнения (например, аддитивная печать, калибровка авиационных и автомобильных деталей и т.д.), а для этого необходимо вовремя, с помощью датчиков, определить потенциально возможное нарушение технического процесса и прервать выполнение операции в максимально короткие сроки. Таким образом, в рамках ПИВ, приоритизация трафика является одним из самых важных процессов [5, 6].

Для исполнения задач приоритизации трафика в рамках мультисервисных сетей связи традиционно используются локальные роутеры клиента, поставляемые провайдером доступа к услугам Triple Play. В более ранних работах авторами был предложен новый класс устройств для сетей ИВ и ПИВ – гетерогенный шлюз ПИВ, который объединяет в себе традиционный шлюз, выполняющий обеспечение доступа к внешним сетям для различных локальных сетей, функционирующих на основе различных технологий канального и сетевого уровня и семантических шлюзов обеспечивающих совместимость технологий на прикладном и семантическом уровнях [7, 8, 9]. Ранее авторами не рассматривались вопросы обеспечения должного качества обслуживания для данного типа устройств. В данной работе будет определена структура модельной сети для исследования вопросов приоритизации сетевого трафика, с помощью гетерогенных шлюзов ПИВ, будут исследованы возможные подходы к определению приоритетов обслуживания сетевого трафика. Данные подходы к приоритизации трафика выставляются согласно основным параметрам качества обслуживания, указанным в ITU Y.1564, в зависимости от требований к виду трафика.

2. Модельная сеть

На рисунке 1 изображена модельная сеть для исследования вопросов приоритизации сетевого трафика, с помощью гетерогенных шлюзов ПИВ [10, 11]. Данная модельная сеть состоит из следующих элементов:

- Системы мониторинга окружающей среды – системы, включающие в себя множество датчиков, собирающих данные о состоянии окружающей среды и чаще всего использующие для передачи данных технологии беспроводных сенсорных сетей.
- Системы позиционирования – системы, которые используются для позиционирования объектов, как в локальном, так и глобальном пространстве. Для

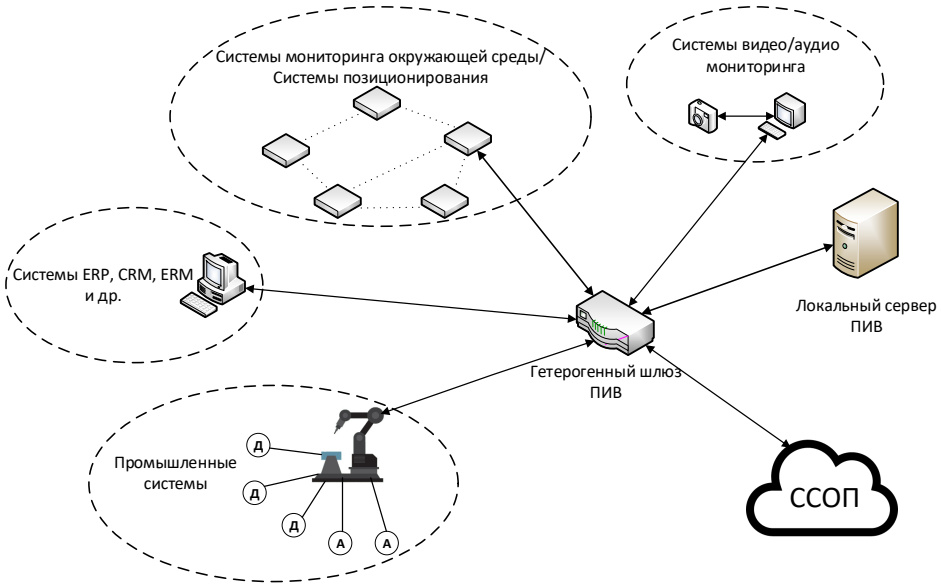


Рис. 1. Модельная сеть для исследования приоритизации трафика

локального позиционирования, наиболее часто используются технологии беспроводных сенсорных сетей, а для глобального – системы спутниковой навигации – GPS, ГЛОНАСС и др.

- Системы аудио и видео мониторинга – мультимедийные системы, применяющиеся для аудиовизуального контроля безопасности исполнения технического процесса на предприятии.
- Промышленные системы – промышленное оборудование, включающее в себя датчики и исполнительные устройства – актуаторы (на рисунке «Д» и «А»), а также вычислительные устройства для мониторинга и управления их состоянием (например, станки ЧПУ). Данные вычислительные устройства представляют собой специализированный компьютер или микроконтроллер, подключенный к сети связи и позволяющий контролировать работу оборудования удаленно.
- Системы управления ресурсами предприятия – системы, отвечающие за экономический и логистический учет ресурсов предприятия – состояния обо-

рудования, количества и стоимости товаров, заработной платы персонала и т.д.

- Сеть связи общего пользования (ССОП) – система, представляющая собой комплекс взаимодействующих сетей электросвязи, в том числе сети связи для трансляции телеканалов и (или) радиоканалов. В данном случае под ССОП подразумевается сеть связи Интернет.
- Локальный сервер ПИВ – локальный сервер, используемый для контроля работы всего предприятия. В данном случае данный сервер выполняет роль пункта назначения для сетевого трафика от включенных в данную модель подсистем.
- Гетерогенный шлюз ПИВ (ГШ) – вычислительное устройство, отвечающее как за обеспечение взаимодействия различных технологий канального, сетевого и транспортного уровней между собой, так и за преобразование форматов полезных данных между собой, т.е. за преобразование на семантическом уровне.

В рамках данной модельной сети приоритизация трафика должна проходить при поступлении трафика от систем ПИВ на ГШ.

3. Подходы к приоритизации трафика

Для вышеопределенной модельной сети (рис.1) могут быть выделены следующие подходы к приоритизации трафика ПИВ:

- Приоритет обслуживания по VLAN. Для приоритизации сетевого трафика, поступающего от различных систем ПИВ может быть использован подход к организации виртуальных локальных вычислительных сетей (VLAN), весь трафик от которых будет помечаться специальной меткой класса обслуживания на канальном уровне (CoS) и меткой DSCP (кодовая точка дифференцированных услуг), в зависимости от требований к качеству обслуживания для каждого из типов трафика.
- Приоритет обслуживания по протоколам, портам и сетевым адресам источника/назначения. Приоритет обслуживания может выставляться по протоколам, портам и сетевым адресам.
- Приоритет обслуживания по системам глубокой инспекции пакетов (DPI). Вид трафика может определяться с помощью характерных для отдельных видов трафика ПИВ свойств. Например, в качестве свойств может использоваться интенсивность поступления пакетов, размеры сетевых пакетов, зашифрован ли пакет или нет, порт назначения/источника и т.д.
- Приоритет обслуживания по системам машинного обучения. Наименее используемый на реальных сетях, экспериментальный метод, заключающейся

в использовании систем машинного обучения для классификации трафика ПИВ и установки необходимых флагов (CoS и DSCP).

Наиболее простым, легко реализуемым и наиболее часто используемым на реальных сетях методом является приоритизация трафика по VLAN. Тем не менее не все устройства ПИВ поддерживают технологию VLAN, таким образом использовать только данный метод для определения приоритетов обслуживания невозможно.

Метод определения приоритетов обслуживания, с помощью протоколов, портов и сетевых адресов также является удобным методом для реализации на уже существующих сетях, но часто бывает, что определить используемый протокол прикладного уровня невозможно вследствие ряда факторов, например, использования протоколов шифрования, что позволяет скрыть прикладной протокол от оператора сети, использование случайных портов источника/назначения и т.д.

Таким образом наиболее гибкими и подходящими для реализации системы приоритизации трафика ПИВ, но при этом самыми сложными, являются методы, основанные на системах DPI и машинного обучения.

На основе данных подходов и описанной модельной сети предлагается подготовить и произвести реальный эксперимент, а также имитационное моделирование по определению свойств качества обслуживания сетевого трафика от различных видов источников, для всех вышеописанных протоколов.

4. Заключение

В данной статье была рассмотрена структура модельной сети для приоритизации сетевого трафика на основе гетерогенных шлюзов ПИВ. Для данной модельной сети были определены подходы к приоритизации сетевого трафика. На основе данных методов предлагается собрать модельную сеть и провести реальный эксперимент по определению свойств качества обслуживания для каждого из видов трафика. В дальнейшем на основе полученных результатов предлагается провести имитационное моделирование при большем количестве устройств в данной системе.

Литература

1. Махмуд О.А. Моделирование влияния трафика Интернета вещей на качество обслуживания / О.А. Махмуд, А.И. Парамонов // Электросвязь. - 2018. - № 9. - С. 39-44.
2. Кучерявый, А.Е. Перспективы научных исследований в области сетей связи на 2017-2020 годы / А.Е. Кучерявый, А.Г. Владыко, Р.В. Киричек и др. // Информационные технологии и телекоммуникации. - 2016. - Т. 4. - № 3. - С. 1-14. URL: https://www.sut.ru/doci/nauka/review/20163/1-14.pdf.

3. Smart City Network Architecture Guide. Alcatel-Lucent. 2019. PP. 39. URL: https://www.al-enterprise.com/-/media/assets/internet/documents/smart-city-network-architecture-guide-en.pdf.
4. Y.4003 : Overview of smart manufacturing in the context of the industrial Internet of things. ITU-T. 2018. PP. 16. URL: https://www.itu.int/ITU-T/recommendations/rec.aspx?rec=13634&lang=en.
5. Singh M., Baranwal G. Quality of Service (QoS) in Internet of Things. 2018 3rd International Conference On Internet of Things: Smart Innovation and Usages (IoT-SIU). 2018. P. 1-6. DOI: 10.1109/iot-siu.2018.8519862.
6. Al-Masri E. QoS-Aware IIoT Microservices Architecture. 2018 IEEE International Conference on Industrial Internet (ICII). 2018. P. 171-172. DOI: 10.1109/ICII.2018.00030.
7. Кулик, В.А. Модель семантического преобразования пакетов для гетерогенного шлюза промышленного Интернета вещей / В.А. Кулик, С.А. Вахитов, Р.В. Киричек // Электросвязь. - 2020. - № 3. - С. 49-54. DOI: 10.34832/ELSV.2020.4.3.007.
8. Q.4060 The structure of the testing of heterogeneous Internet of things gateways in a laboratory environment. ITU-T. 2018. URL: https://www.itu.int/rec/T-REC-Q.4060-201810-I.
9. Q.3055 Signalling protocol for heterogeneous Internet of things gateways. ITU-T. 2019. PP. 29. URL: https://www.itu.int/rec/T-REC-Q.3055-201912-I.
10. Кулик, В.А. Исследование и генерация трафика промышленного Интернета вещей / Р.В. Киричек, В.А. Кулик // Труды учебных заведений связи. - 2019. - Т. 5. - № 3. - С. 27-36. DOI: 10.31854/1813-324X-2019-5-3-27-36.
11. Кулик, В.А. Классификация и исследование трафика промышленного Интернета вещей на модельной сети / В.А. Кулик, А.И. Парамонов, Р.В. Киричек // Электросвязь. - 2019. - № 8. - С. 22-28.

UDC: 004.94

Delivering Multicast Traffic in mmWave Systems: Challenges and Performance Analysis

Nadezhda Chukhno[1,2], Olga Chukhno[1,3], Giuseppe Araniti[1], Antonio Iera[4], Antonella Molinaro[1,5], Sara Pizzi[1]

[1]University Mediterranea of Reggio Calabria, Reggio Calabria, Italy
[2]Universitat Jaume I, Castelló de la Plana, Spain
[3]Tampere University, Tampere, Finland
[4]University of Calabria, Rende, Italy
[5]Université Paris-Saclay, Gif-sur-Yvette, France

{olga.chukhno, nadezda.chukhno, araniti, antonella.molinaro, sara.pizzi}@unirc.it, antonio.iera@dimes.unical.it

Abstract

Millimeter wave (mmWave) radio technology is considered as a comprehensive foundation for fifth-generation (5G) networks, which are claimed to efficiently and effectively support both multicast and unicast traffic. One of the main features of mmWave communications is the exploitation of highly directional antennas, at both user and access point sides, that allows providing very-high speed and ultra-low latency services. However, the delivery of multicast traffic in mmWave systems requires special attention because of the nature of the group-oriented services in which receivers simultaneously fed by a single transmission can be located at different positions. The aim of this paper is to discuss the main challenges that must be faced to take advantage of mmWave communication for multicast data delivery. A performance analysis is carried out with the aim to provide a comparison among unicast, sequential multicast, and multicast transmission modes.

Keywords: 5G, mmWave, 802.11ad/ay, Unicast, Multicast

1. Introduction

Recently, millimeter wave (mmWave) wireless networks have become increasingly popular thanks to their capability to cope with the escalation of mobile data demands caused by the unprecedented proliferation of smart devices in upcoming fifth-generation (5G) communication systems [1]. Furthermore, according to both academic and industrial communities, mmWave technology is expected to play a fundamental role even in beyond-5G networks to ensure efficient massive data transmissions [2]. In this regard, mmWaves have been envisaged for a wide range of emerging applications that mainly require the dissemination of a large amount of

data traffic with low latency, such as autonomous driving, mobile video streaming, virtual/augmented/mixed reality (VR/AR/XR) applications, public/road safety, road infotainment, among others. The 3GPP New Radio (NR) technology will exploit the mmWave spectrum to achieve wider bandwidths and higher data rates [3]; a similar approach is used by IEEE 802.11ad/ay families, which will be the focus of this paper.

In this scenario, multicast is a beneficial technique for the improvement of the system bandwidth efficiency. In an IEEE 802.11-based multicast system, a device, which acts as a personal basic service set (PBSS) central point (PCP) or access point (AP), may transmit the same packet to a group of receivers simultaneously, by utilizing the same frequency and modulation and coding scheme (MCS). MmWave transmissions use highly directional antennas to guarantee the gigabit capability and overcome the short propagation range, but they make multicasting more complex to implement in comparison with microwave networks where omnidirectional antennas are typically applied. MmWaves complicate the multicast deployment by posing additional challenges [4], such as beam steering and proper selection of beamwidth. These significant challenges stem from the tiny wavelength and high-frequency band properties of mmWave systems, where the signal is sensitive to rapid channel variations, atmospheric absorption, and severe attenuation. In order to clarify these issues, we investigate the challenges and advantages of mmWave multicast communication in this paper.

The rest of the paper is organized as follows. In Section 2, background on IEEE 802.11ad/ay is presented. Section 3 discusses the design challenges of multicast and unicast modes with mmWaves. In Section 4, we describe the system model. Simulation results are given in Section 5, followed by concluding remarks.

2. IEEE 802.11ad/ay specifications

The IEEE 802.11ad/ay standards of the Wi-Fi family operate in the 60 GHz band. The former was ratified in 2012, offering real gigabit data rates. Its successor, IEEE 802.11ay, exploits the same band and provides ultra-high-speed and super low-latency services by introducing advanced physical layer (PHY) features. Furthermore, 802.11ay improved power-saving features make it ideal for wearable devices [5].

2.1. IEEE 802.11ad. The beacon interval structure in IEEE 802.11ad is illustrated in Fig. 1. The Medium Access Control (MAC) design for 802.11ad may utilize both carrier sensing multiple access with collision avoidance (CSMA/CA) and scheduled service periods (SPs) channel access schemes depending on the type of application. In the case of SPs, 802.11ad uses Time Division Multiple Access (TDMA), where PBSS PCP utilizes the polling mechanism by asking devices and receiving their feedback. Alternatively, CSMA/CA is used for contention-based

periods (CBP), where devices are allowed to use the same radio channel without pre-coordination with the help of *listen-before-talk* operating procedure. In this work, we consider SPs only.

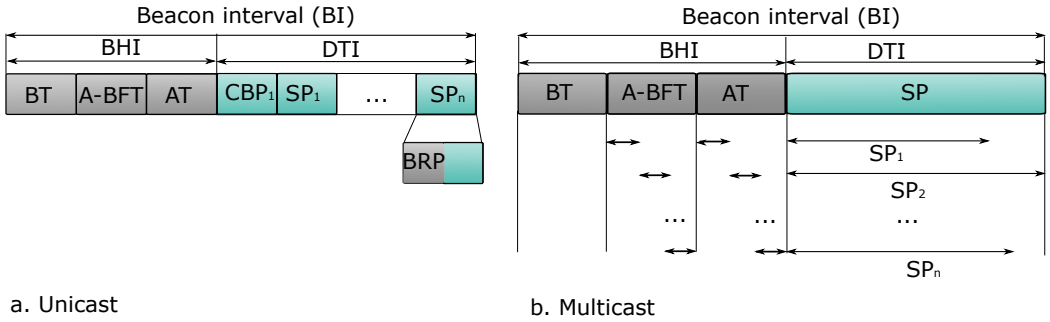


Fig. 1. IEEE 802.11.ad beacon structure.

The time is divided into beacon intervals (BIs) of total length T; each BI incorporates: (i) a beacon header interval (BHI), where devices perform initial beamforming which generally involves sectored antennas (that is, sector-level sweep, SLS), and (ii) the data transmission interval (DTI), including SPs of different connected clients, while containing a beam refinement protocol (BRP) to improve the resulting instantaneous data rate.

More specifically, each BI starts with a beacon time (BT) interval during which the *initiator* transmits sector sweep (I-TXSS) beacons across all M_{SLS} sectors with half-power beamwidth (HPBW) $\theta_{\text{SLS,Tx}} = 2\pi/M_{\text{SLS}}$. The receive sector sweep (RSS) process and feedback during the association beamforming training (A-BFT) announced by the initiator are performed after I-TXSS by the *receiver*. Practically, in A-BFT interval, if more than one client selects the same transmission opportunity (up to eight slots for 802.11ad), the signals collide and devices cannot establish a connection in the current BI.

The receive antenna operates in an omnidirectional mode during SLS and, after measuring the receive signal strength (RSS) across all N_{SLS} sectors, with $\theta_{\text{SLS,Rx}} = 2\pi/N_{\text{SLS}}$, it provides the SLS feedback to the transmitter identifying the sector with maximum RSS value. Based on RSS indicator as well as by using an angle of arrival (AoA), or time difference of arrival (TDoA), the AP can determine the user location information. The training packets are transmitted with the low-power low-rate MCS 0, which provides the reliable communication required to establish the initial beamformed link.

In the announcement time interval (ATI), management information is exchanged between the PCP/AP and the receivers. Once the best sector pair is identified, the

beam refinement phase (BRP) iteratively trains the transmit and receive antenna beams found during the SLS to select a beam pattern pair with finer beamwidths, which are determined by the beam refinement factor b, $b > 1$. Therefore, for the transmit antenna training, both devices sweep through exactly b narrower beams (within the initial transmit sector), while during the receive training, all $M = bM_{\text{SLS}}$ or $N = bN_{\text{SLS}}$ directions should be covered.

To adjust for channel changes, the optional *beam tracking phase* is used during data transmission (DT). Beam tracking is accomplished by appending training (TRN) fields to data packets [6]. More details can be found in [7].

2.2. IEEE 802.11ay. IEEE 802.11ay is an amendment of great interest for applications ranging from high-speed short-range links to wireless backhaul that enable 100 Gbps communications in the unlicensed 60 GHz mmWave band. IEEE 802.11ay incorporates a variety of technical advancements at the PHY over IEEE 802.11ad standard, such as channel bonding and aggregation, single-user (SU) and downlink (DL) multi-user (MU) Multiple-Input Multiple-Output (MIMO) transmissions, and nonuniform modulation constellation, as well as improved channel access and enhanced beamforming training. For an overview of the IEEE 802.11ay amendment, the reader is referred to [8, 9].

3. Design Challenges of mmWave Multicast and Unicast Modes

3.1. Unicast in Directional Networks. A considerable amount of literature has been published on mmWave communications by focusing on unicast data transmission optimization [10], whereby every user is served independently of the others. The AP sweeps many different beams with minimum beamwidth (to provide high data rate) in the TDMA fashion. The reliability of the transmission is very high since the beamwidth is equal to the resolution and provides the maximum available Signal-to-Noise Ratio (SNR). However, the AP requires a long time to serve all users, as it generates a separate beam for each user to transmit the data sequentially.

3.2. Multicast in Traditional Networks. The multicast concept consists in feeding a group of users by transmitting data packets only once [11]. It consequently improves the bandwidth efficiency compared to unicast mode since all users are served simultaneously by using a single wide transmission beam (omnidirectional), which generally provides very short transmission duration. However, pure multicast schemes are almost infeasible in mmWave systems due to the propagation properties of extremely high frequency (EHF) bands.

3.3. Multicast in Directional Networks. A significant difference between mmWave and traditional networks (e.g., LTE) consists in the use of highly directional transmission beams to cope with high attenuation at EHF bands. We emphasize that mmWave multicast is performed in a sequential manner. Moreover, differently

from conventional omnidirectional networks, the beam orientation and the beam resolution (beamwidth) need to be adjusted in addition to the beam radius.

According to the Friis transmission equation, the received power is directly proportional to the transmitter channel gain, which in turn strongly depends on beam resolution and its orientation. When delivering multicast services in mmWave systems, the following challenges have to be considered [4]:

1) Wide beams are more likely to reach all multicast receivers since they can cover a larger angle range and, thus, serve more users simultaneously. However, due to the lower antenna gain that wide beams provide, the supported transmission rate is limited.
2) Narrow beams provide higher antenna gain and thus can support higher transmission rates. However, they are limited in coverage in terms of the aperture angle, and may not serve a number of users simultaneously. As a consequence, multiple unicast transmissions are required to reach all multicast users.
3) The presence of moving users is more challenging for multicast transmission. In fact, in the case of unicast, the AP is beamformed toward the only receiver, and small movements of the receiver still allow to guarantee a good reception. Differently, in the case of multicast, beams are steered in between users. Hence, some receivers may be close to the edge of the coverage area of a beam.

4. System Model

In this paper, we consider a general public scenario where owners of *wearable devices* are interested in receiving the same content, i.e., the AP transmits data to multiple users thought multicast mmWave links. We assume analog beamforming only to analyze the performance of sequential multicast in the TDMA fashion. This means that the AP can transmit through a single beam at a time to serve the users.

4.1. Antenna and Channel Models. In what follows, we assume that devices transmit directionally with the same antenna beam pattern, which is symmetrical w.r.t. the boresight [12]. By this symmetrical assumption, we mean that antennas have a unique beam shape in both elevation and azimuth planes, i.e., their antenna pattern is akin to a conical shape.

In terms of channel model, when the HPBW θ is used, the received signal power at the receiver i is calculated by the Friis equation:

$$P_{\mathrm{rx},i} = \frac{P_{\mathrm{tx}} D_0 \rho(\alpha_i) \lambda^2}{(4\pi)^2 r_i^{\kappa}}, \tag{1}$$

where P_{tx} is the transmit power, α_i is the current angular deviation of the transmit/receive direction from the antenna boresight for receiver i, $\rho(\alpha_i) \in [0; 1]$ is

a piece-wise linear function that scales the antenna directivity D_0 [12]*, λ is the wavelenght, r_i is the separation distance between the transmitter (Tx) and receiver (Rx) i, and κ is the path loss exponent.

We assume the Line of Sight (LoS) path only. Hence, the maximum achievable rate D for the multicast group depends on the user with the worst channel conditions and could be estimated according to Shannon's channel capacity as:

$$D = W \log_2 \left(1 + \min_i \left(\frac{P_{\text{rx},i}}{P_{\text{noise}}}, 0 | P_{\text{rx},i} < P_{\text{thr}}\right)\right), \tag{2}$$

where $P_{\text{rx},i}$ incorporates both transmit and receive antenna gains after the BRP phase for link Tx-Rx$_i$, P_{thr} guarantees the minimum required sensitivity threshold for data transmission, W is the bandwidth, P_{noise} is noise power in the channel, which corresponds to $P_{\text{noise}} = W\, N_0\, \text{NF}$, N_0 is the power spectral density of noise per 1 Hz, and NF is the noise figure.

Then, the total duration of data transmission can be found as [13]:

$$T = T_{\text{SLS}} + U(T_{\text{BRP}} + T_{\text{DT}}) + T_0, \tag{3}$$

where $T_{\text{SLS}} + U(T_{\text{BRP}})$ is the overhead on beam training, $T_{\text{DT}} = B/D$ is the data transmission duration, B is the packet size, U is the average number of clients per AP, and T_0 is the total signaling overhead independent on the number of beams.

5. Performance Analysis

In this section, we assess the performance of the multicast transmission in directional mmWave networks via simulations. As a representative scenario, we focus on a group of people in a museum equipped with high-end wearable devices that communicate via the IEEE 802.11ad protocol at 60GHz. We consider resource-hungry applications in a scenario with low or no mobility. We also analyze two patterns of users' distribution: in a line (Fig. 2) and within a sector (Fig. 3). The transmit power is fixed at the level of $P_{\text{tx}} = 23$ dBm, whereas $P_{\text{thr}} = -68$ dBm (MCS 1) [7].

In Fig. 2a, we show results in terms of achievable data rate for multicast, unicast, and sequential multicast schemes when the group of *ten users* is located within a line of length d. We investigate the maximum possible distance d for three considered schemes. As one may observe, the distance d does not affect unicast transmission as each user is served by a separate aligned beam. Regarding the pure multicast, we can see that there is a threshold that determines the maximum angle coverage

*$\rho(\alpha_i)=1-\frac{\alpha_i}{\theta}$, if $\alpha_i \leq \theta$, otherwise $\rho(\alpha_i) = 0$; $\rho(\alpha_i) = 1$ corresponds to the antenna boresight in the case of perfect alignment (e.g., unicast transmission after the beamforming procedure). In the case of multicast transmission, each user deviates on angle α from the boresight of the transmitter.

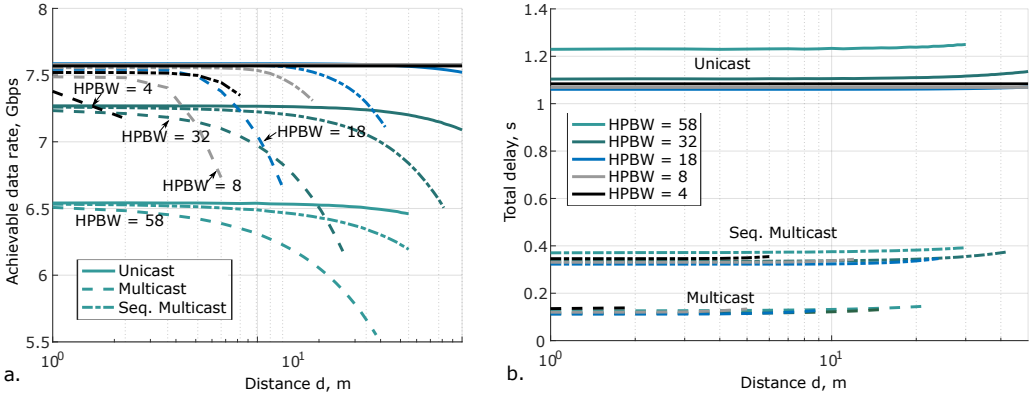


Fig. 2. (a) Data rate and (b) total delay vs. distance d for unicast, multicast, and sequential multicast (with 3 beams).

of the beam. For example, HPBW= 58° provides the larger distance d, whereas narrow beam HPBW= 4° shows the lowest distance d. In Fig. 2b, we present the total data transmission duration T for all thee transmission modes. Due to the sequential nature of unicast in mmWave, pure multicast guarantees the shortest data transmission duration. We also emphasize that the beamforming overhead for narrow beams is greater than for wide beams and affects the total transmission delay.

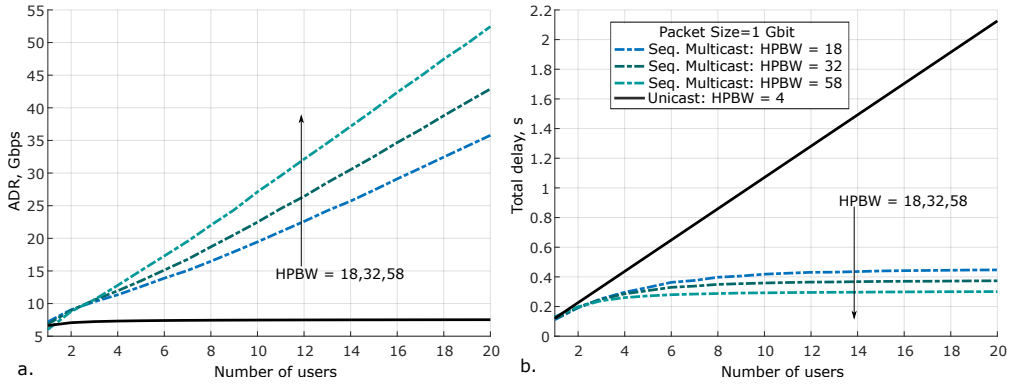


Fig. 3. (a) Aggregated data rate and (b) total transmission duration vs. number of users.

In Fig. 3, we analyze the system performance according to the aggregated data rate (ADR) and the time T required for serving all users for sequential multicast and unicast schemes and evaluate them using a Montecarlo approach (10^6 simulations). For this purpose, we uniformly distribute users within a sector of 90° of radius

Rd = 40m. The choice of the service area can be explained by the fact that sequential multicast with width $\theta = 58°$ has the smallest number of sequential beams, whereas it can cover the smallest Rd in comparison with more narrow beams. As one may notice, sequential multicast with $\theta = 58°$ guarantees the shortest data transmission duration, hence, the lower delay. However, it provides lower SNR value as well as higher outage probability due to the lower antenna directionality. Using the widest possible beam at EHF bands severely limits the data rate and transmission range. In contrast, narrow beams require a longer data transmission duration. Therefore, we may conclude that the resource management algorithms are of the crucial importance, which dynamically make decisions on the number and resolution of beams in directional multicast by taking into consideration (i) multicast group size, (ii) the shape and size of the service area, (iii) users' locations and density, as well as (iv) QoS requirements. This is of special interest for our future work.

Acknowledgement

The authors gratefully acknowledge funding from European Union's Horizon 2020 Research and Innovation programme under the Marie Skłodowska Curie grant agreement No. 813278 (A-WEAR: A network for dynamic wearable applications with privacy constraints, `http://www.a-wear.eu/`).

REFERENCES

1. T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!," *IEEE access*, vol. 1, pp. 335–349, 2013.
2. X. Lu, V. Petrov, D. Moltchanov, S. Andreev, T. Mahmoodi, and M. Dohler, "5G-U: Conceptualizing Integrated Utilization of Licensed and Unlicensed Spectrum for Future IoT," *IEEE Communications Magazine*, vol. 57, no. 7, pp. 92–98, 2019.
3. S. Ahmadi, *5G NR: Architecture, Technology, Implementation, and Operation of 3GPP New Radio Standards*. Academic Press, 2019.
4. A. Biason and M. Zorzi, "Multicast via point to multipoint transmissions in directional 5G mmWave communications," *IEEE Communications Magazine*, vol. 57, no. 2, pp. 88–94, 2019.
5. C. R. da Silva, A. Lomayev, C. Chen, and C. Cordeiro, "Analysis and Simulation of the IEEE 802.11 ay Single-Carrier PHY," in *2018 IEEE International Conference on Communications (ICC)*, pp. 1–6, IEEE, 2018.
6. S. Kutty and D. Sen, "Beamforming for millimeter wave communications: An inclusive survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 949–973, 2015.

7. IEEE 802.11 Working Group, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band," 2012.
8. IEEE 802.11 Working Group, "Enhanced throughput for operation in license-exempt bands above 45 GHz," tech. rep., IEEE 802.11ay/D0.3, Mar. 2017.
9. Y. Ghasempour, C. R. da Silva, C. Cordeiro, and E. W. Knightly, "IEEE 802.11 ay: Next-generation 60 GHz communication for 100 Gb/s Wi-Fi," *IEEE Communications Magazine*, vol. 55, no. 12, pp. 186–192, 2017.
10. A. M. Al-samman, M. H. Azmi, and T. A. Rahman, "A survey of millimeter wave (mm-Wave) communications for 5G: Channel measurement below and above 6 GHz," in *International Conference of Reliable Information and Communication Technology*, pp. 451–463, Springer, 2018.
11. F. Rinaldi, S. Pizzi, A. Orsino, A. Iera, A. Molinaro, and G. Araniti, "A Novel Approach for MBSFN Area Formation Aided by D2D Communications for eMBB Service Delivery in 5G NR Systems," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 2, pp. 2058–2070, 2019.
12. O. Chukhno, N. Chukhno, O. Galinina, Y. Gaidamaka, S. Andreev, and K. Samouylov, "Analysis of 3D Deafness Effects in Highly Directional mmWave Communications," in *2019 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, IEEE, 2019.
13. N. Chukhno, O. Chukhno, S. Shorgin, K. Samouylov, O. Galinina, and Y. Gaidamaka, "Maximizing achievable data rate in unlicensed mmwave networks with mobile clients," in *Internet of Things, Smart Spaces, and Next Generation Networks and Systems*, pp. 282–294, Springer, 2019.

UDC: 123.456

Comparison of Different Levels of Local Purchase Quantities in a Geo/Geo/1 Production Inventory System

Anilkumar M. P. [1] and K. P. Jose [2]

[1]Department of Mathematics, T. M. Govt. College, Tirur-676 502, Kerala, India
[2]Department of Mathematics, St.Peter's College, Kolenchery - 682 311,Kerala, India.

[1]anilkumarmp77@gmail.com, [2]kpjspc@gmail.com

Abstract

This paper considers a Geo/Geo/1 production inventory system in which demand occurs according to a Bernoulli process and service time follows a geometric distribution. The maximum inventory that can be accommodated in the system is S. When the on-hand inventory is reduced to a preassigned level of s due to service completion (and consequent purchase of exactly one item by each customer), production is started. The production time for each item (inter-production time) follows a geometric distribution. When the inventory level becomes zero, an instantaneous local purchase of one/s/S units is made to meet the demand. These three types of local purchases are discussed as three separate models. Using the closed-form solution obtained for the steady-state probability vector and by constructing an appropriate cost function, compare these models with the help of a few numerical work.

Keywords: Discrete-time production inventory, Bernoulli process, Geometric distribution, Matrix-Analytic Method

1. Introduction

The first reported work on product form solution in queueing inventory system is by Schwarz et al. [1] in which the authors assumed the different policies. The product form solution to this model is obtained only with the assumption that no customers are allowed to enter the system when the inventory level is zero. Later, there are few papers with product form solutions in queueing inventory system with positive service time and zero lead time. These works are mentioned in the survey paper by Krishnamoorthy et al. [2]. Schwarz and Daduna [3] developed approximation for performance measures in the M/M/1 queueing system in which the issue of inventory is considered as service with the assumption that customers

can join in the system even when the inventory level is zero. To overcome the loss of customers, due to the lack of inventory in the product form solutions, Schwarz et al. [4] considered a queueing inventory model in which product form solution is obtained with the assumption that the demand that occurs during the stock out period is re-routed to other service stations. Instantaneous replenishment during the stock out the period with high replenishment cost is considered by Saffari and Haji [5] to obtain the product form solution. M/M/1 queueing inventory system under (r, Q) policy analyzed by Saffari et al. [6] in which, demand during the stock out period was assumed to be lost. An explicit expression for long-run performance measures was obtained and carried out cost optimization. The investigation of stochastic decomposition of production (s, S) inventory system in continuous time received much attention from researchers since the paper by Krishnamoorthy and Viswanath [7]. To get the explicit expression for the steady-state distribution, the authors restricted the entry of customers according to the inventory level. They obtained an explicit expression for the production cycle and optimized the cost function associated with the model with respect to maximum storage S. Deepthi [8] extended this model to discrete-time. Krishnamoorthy et al. [9] optimized the N in (s, Q) inventory system in continuous time so that local purchase of $N + Q$ items is done when the on-hand inventory level is $s - N$. Recently Krishnamoorthy et al. [10] also considered continuous-time (s, S) production inventory system with positive service time and obtained stochastic decomposition of the system by introducing one unit of local purchase during the stock out period. The analogue work on discrete-time inventory is not known to the best of our knowledge. This article analyses the quantity of instantaneous replenishment, instead of one unit which is mentioned in Krishnamoorthy et al. [10], during the stock out period in discrete-time set-up.

Notable work on the discrete-time queue is done by Meisling [11]. Dafermos and Neuts [12] approximated a continuous-time model by a discrete-time single server queueing model as a limiting case. Lian et al. [13] introduced inventory in a discrete-time inventory system having common life. We use the discrete version of the Matrix-Analytic Method (MAM), explained in Alfa [14, 15], to analyse the model. For the elementary details of MAM, one can refer to Neuts [16]. The present paper is an attempt to avoid the loss of customers in Krishnamoorthy and Viswanath [7] by introducing local purchases in a discrete-time set up with a closed-form solution. During the stock out period, one unit, s units or S unit is locally purchased with the high cost and these are studied in three different models. In the first two cases, the replenishment order is not canceled whereas in the third it is canceled. These three models are compared based on a suitable cost function and obtained the best profitable model.

2. Mathematical Modelling

We consider a single sever production (s, S) production inventory system in which the arrival of customers follows a Bernoulli process with parameter p, service time follows a geometric distribution with parameter q. Each customer receives one inventory after completing the service. When the inventory level depletes s due to demands, production starts. The production time of the individual item in the inventory follows a geometric distribution with parameter r. The production is stopped when the inventory is reached to the maximum level of S. We assume that arrival and service completion occurs at the beginning of the slot boundary and production of the individual item takes place at the end of the slot boundary. In any epoch, if the inventory level becomes zero due to service and production lag, an instantaneous local purchase of one/s/S units with high purchasing cost is made to meet the demand. These three types of local purchases are discussed as three separate models-Model 1, Model 2 and Model 3 respectively.

Notations

$N(n)$: Number of customers in queue at an epoch n.

$I(n)$: Inventory level at the epoch n.

$C(n)$: The production status, which is $\begin{cases} 0, \text{when production is off} \\ 1, \text{when the production is on} \end{cases}$

$\bar{x} : 1 - x, \text{ for } 0 \leq x \leq 1.$

Then $\{(N(n), I(n)), c(n); n = 0, 1, 2, 3, ..\}$ is a Quasi Birth Death process (QBD) with state space

$$\{(i, j); 1 \leq j \leq s\} \cup \{(i, j, k); s + 1 \leq j \leq S - 1, k = 0, 1\} \cup \{(i, S)\}, \text{for } \ i \geq 0$$

2.1. Stability . Using the stability of level independent QBD discussed in Neuts [16], we can coclude that the above QBD is stable if and oly if $p\bar{q} < \bar{p}q$, which leads to $p < q$.

2.2. Steady-State Analysis.

Theorem 1. The steady-state probability vector $\Pi^{(i)} = (\pi_0^{(i)}, \pi_1^{(i)}, \pi_2^{(i)}, \dots)$ of the Model i, i=1, 2, 3 is given by

$$\pi_n^{(i)} = \begin{cases} (\frac{q-p}{q})\widehat{\Pi}^{(i)} \text{ for } n = 0 \\ (\frac{q-p}{q})\frac{p}{\bar{p}q}\rho^{i-1}\widehat{\Pi}^{(i)} \text{ for } n \geq 1 \end{cases} \tag{1}$$

where $\rho = \dfrac{p\bar{q}}{\bar{p}q}$. and $\widehat{\Pi}^{(i)} = (\widehat{\pi}_1^{(i)}, \ldots, \widehat{\pi}_s^{(i)}, \widehat{\pi}_{s+1,0}^{(i)}, \widehat{\pi}_{s+1,1}^{(i)}, \ldots, \widehat{\pi}_{S-1,0}^{(i)}, \widehat{\pi}_{S-1,1}^{(i)}, \pi_S^{(i)})$ in which,

$$\widehat{\pi}_j^{(1)} = \frac{p}{(r-p)}(1-k^{S-s})k^{s-j}\widehat{\pi}_S^{(1)} \text{for } 1 \le j \le s,$$

$$\widehat{\pi}_S^{(1)} = \frac{(1-k)(r-p)}{p(i-k-k^{s+1}+k^{S+1})+r(1-k)(S-s)},$$

$$\widehat{\pi}_{j,1}^{(1)} = \frac{p}{r-p}(1-k^{S-j})\widehat{\pi}_S^{(1)} \text{ for } s+1 \le j \le S-1,$$

$$\widehat{\pi}_j^{(2)} = \frac{p(k^s-k^S)(1-k^j)}{(r-p)(1-k^s)k^s}\widehat{\pi}_S^{(2)} \text{ for } 1 \le j \le s,$$

$$\widehat{\pi}_S^{(2)} = \frac{(1-k^s)(r-p)}{r(1-k^s)(S-s)-ps(k^S-k^s)},$$

$$\widehat{\pi}_{j,1}^{(2)} = \frac{p}{r-p}(1-k^{S-j})\widehat{\pi}_S^{(2)} \text{ for } s+1 \le j \le S-1,$$

$$\widehat{\pi}_j^{(3)} = \frac{pk^{(s-j)}(1-k^{(S-s)})(1-k^j)}{(r-p)(1-k^S)}\widehat{\pi}_S^{(3)} \text{ for } 1 \le j \le s,$$

$$\widehat{\pi}_S^{(3)} = \frac{(1-k^S)(r-p)}{r(1-k^s)(S-s)-ps(k^s-k^S)},$$

$$\widehat{\pi}_{j,1}^{(3)} = \frac{p(1-k^s)}{(r-p)(1-k^S)}(1-k^{S-j})\widehat{\pi}_S^{(2)} \text{ for } s+1 \le j \le S-1,$$

$$\widehat{\pi}_{s+1,0}^{(i)} = \widehat{\pi}_{s+2,0}^{(i)} = \cdots = \widehat{\pi}_{S-1,0}^{(i)} = \widehat{\pi}_S^{(i)} \text{ for i=1,2,3}$$

and $k = \dfrac{p\bar{r}}{\bar{p}r}$.

3. Performance Measures

Let $\Pi^{(k)} = (\pi_0^{(k)}, \pi_1^{(k)}, \pi_2^{(k)}, \ldots)$ for the model $k = 1, 2, 3$. Then, the corresponding important system performance measures considered are given below.

i) Expected queue length, $EQ = \sum_{i=0}^{\infty}\sum_{j=1}^{S} i\pi_{ij}^{(k)} = \dfrac{p\bar{p}}{(q-p)}$

ii) Expected inventory level, $EIL = \sum_{i=0}^{\infty}\sum_{j=1}^{S} j\pi_{ij}^{(k)}$

iii) Expected rate of production,EPR, is

$EPR = r\sum_{j=1}^{s}\widehat{\pi}_i^{(k)} + r\sum_{i=s+1}^{S}\widehat{\pi}_{i,1}^{(k)} = (1-(S-s)r\widehat{\pi}_S^{(k)})$

iv) Expected local purchase rate, $ELP = q(1-r)\sum_{i=1}^{\infty}\pi_i^{(k)} = \widehat{\pi}_1^{(k)}p(1-r)$

v) Expected production switching on rate, $E_{ON} = q\sum_{i=1}^{\infty}(\frac{q-p}{q})\frac{p}{\bar{p}q}\rho^{i-1}\widehat{\pi}_{s+1,0}^{(k)} = p\widehat{\pi}_S^{(k)}$

4. Distribution of Number of Local Purchase in Specified Time

In order to calculate the number of local purchase in a given time duration, first, we truncate the size of the queue. For this, choose $\epsilon > 0$ and N large enough so that

$$\sum_{i=N+1}^{\infty}\rho^{i-1} < \frac{\epsilon}{(\frac{q-p}{q})\frac{p}{\bar{p}q}}$$

On simplification, it reduces to $\rho^N < \frac{\epsilon}{q}$.

Let $L(n)$ denote the number of local purchases during the time $[o,n]$. $N(n)$, $I(n)$ and $c(n)$ respectively denote the number of customers, inventory level and server-status at an epoch n. Consider the Markov chain $\{\big(L(n),N(n),I(n),c(n)\big);n\geq 0\}$ with state space $\{\Delta\}\cup\{0,1,2\dots\}\times\{0,1\dots,N\}\times\{1,2,\dots,s,(s+1,0),(s+1,1),\dots,(S-1,0),(S-1,1),S\}$, where Δ represents the absorbing state on the realization of the random clock with some probability δ and N is the truncation level mentioned above. The transition probability matrix of the process P_L, is of the form

$$P_L = \begin{bmatrix} 1 & \mathbf{0} \\ \boldsymbol{\delta} & \boldsymbol{U} \end{bmatrix}$$

in which $\boldsymbol{U} = \begin{bmatrix} N_1 & N_0 & & \\ & N_1 & N_0 & \\ & & \ddots & \ddots \end{bmatrix}$, and $\boldsymbol{\delta} = \begin{bmatrix} \delta\boldsymbol{e} \\ \delta\boldsymbol{e} \\ \vdots \end{bmatrix}$.

Let x_k denote the probability that k local purchase is done before the realization of the random clock. In order to calculate x_k, we consider the top left submatrix $\boldsymbol{U}^*$of $\boldsymbol{U}$ having order $(k+1)(2S-s)$.

Then the probability that absorption will take place at k^{th} level is $\boldsymbol{\beta}N'\delta\boldsymbol{e}$, where N'

is the $(k+1)^{th}$ block of the first row of $(\boldsymbol{I}-\boldsymbol{U}^*)^{-1}$ and $\boldsymbol{\beta}$ is the initial probability vector. Hence,

$$x_0 = -\delta\boldsymbol{\beta}(I-N_1)^{-1}\boldsymbol{e},$$
$$x_k = (-1)^{k-1}\delta\boldsymbol{\beta}((I-N_1)^{-1}N_0)^k(I-N_1)^{-1}\boldsymbol{e}$$

The initial probability vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_N)$, where β_i's are given by $\beta_0 = \dfrac{1-\frac{p}{q}}{1-\frac{p}{q}\rho^N}\widehat{\Pi}$ and for $i \geq 1$, $\beta_i = \dfrac{(1-\frac{p}{q})\frac{p}{\bar{p}q}\rho^{i-1}}{1-\frac{p}{q}\rho^N}\widehat{\Pi}$, where $\widehat{\Pi} = \widehat{\Pi}^{(i)}$ for $i = 1, 2, 3$ depending on the model.

5. Numerical Experiments

5.1. Cost Function. Based on the above performance measures, we define a suitable cost function. For this, we consider the individual costs c_0, c_1, c_2, c_3 and c_4 as

c_0 : switching cost for the production
c_1 : production per unit inventory per unit time
c_2 : holding of inventory per unit per unit time
c_3 : cost due to local purchase unit items
c_4 :cost per local purchase
c_5 :cancellation cost of a production process due to local purchase ($c_5 = 0$ for the model I and II)

Define expected total cost (ETC) per unit time as

$$ETC = c_0E_{ON} + c_1EPR + c_2EIL + (c_3Q + c_4 + c_5)ELP$$

where Q is the number of items locally purchased.

5.2. Graphical illustrations. To determine the profitable model, the comparison of three models on the basis of the expected total cost is made. For this, first, fix all parameters and individual costs associated with the model except one. Then compare the cost associated with these models from the graph.

Fig. 1 illustrates the variation ETC with p corresponding to the other parameters in the figure. From this figure, when $p = 0.34$, ETC for all the three models are approximately the same. When p increases from there, ETC first decreases for these models. For $0.34 < p < 0.5$, the Model 1 is the most profitable and Model 2 is

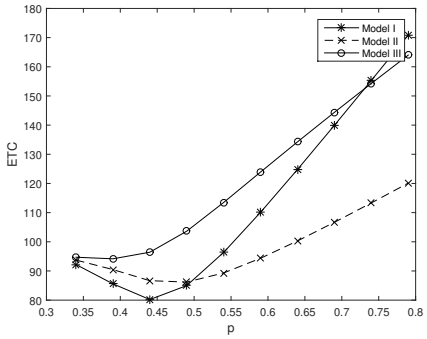


Fig. 1. ETC vs. p
$c_0 = 250; c_1 = 100; c_2 = 5; c_3 = 120; c_4 = 200 : c_5 = 600 : q = 0.9; r = 0.4; s = 8; S = 20$

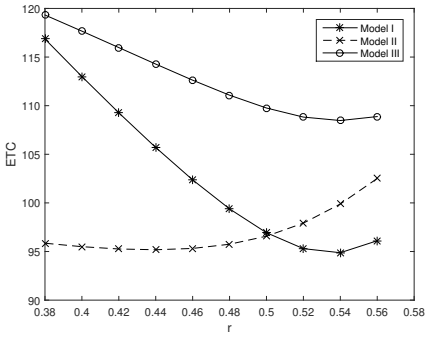


Fig. 2. ETC vs. r
$c_0 = 250;\ c_1 = 100;\ c_2 = 5;\ c_3 = 120;\ c_4 = 200 : c_5 = 100 : q = 0.8; p = 0.6; s = 8; S = 20$

profitable than Model 3. As p increases further, Model 2 is the most profitable and model 1 is more profitable than model 2. For $p > 0.74$ the model 3 is profitable than model 1. The greater ETC for Model 3 in the interval $0.34 \leq p \leq 0.74$ is due to the cancellation cost c_5 of the production associated with that model. If we increase c_5 further, the Model 3 will be the least profitable model for all the values of p

Fig. 2 illustrates the variation of ETC with r. For lower values of r, that is for $r \leq 0.5$, model 2 is the best model and for $r > 0.5$, model 1 will be the suitable model. This indicates that when the replenishment rate is high, the minimum unit of local purchase makes the firm more profitable.

Since all performance measures considered for cost function are independent of q, variations in q will not affect ETC. From Fig. 1 and 2, one can observe that the Model 2 is the most acceptable model.

6. Conclusion

This article analyzes a $Geo/Geo/1$ production inventory system with a local purchase during the stock-out period. We analyzed three models based on the number of items locally purchased. A closed-form solution is obtained for all these three models. Based on suitable cost function we compared the models by varying the parameters. The distribution of the number of times locally purchased during a specified period is also calculated. A simple extension of this model can be done by taking the variations in the local purchase quantity. Further extensions are also possible by considering a discrete-time MAP for the arrival process or discrete-time phase-type distributions for service time and lead-time or both.

REFERENCES

1. M. Schwarz, C. Sauer, H. Daduna, R. Kulik, R. Szekli, M/M/1 Queueing systems with inventory, Queueing Systems 54 (1) (2006) 55–78.
2. A. Krishnamoorthy, B. Lakshmy, R. Manikandan, A survey on inventory models with positive service time, Opsearch 48 (2) (2011) 153–169.
3. M. Schwarz, H. Daduna, Queueing systems with inventory management with random lead times and with backordering, Mathematical Methods of Operations Research 64 (3) (2006) 383–414.
4. M. Schwarz, C. Wichelhaus, H. Daduna, Product form models for queueing networks with an inventory, Stochastic Models 23 (4) (2007) 627–663.
5. M. Saffari, R. Haji, Queueing system with inventory for two-echelon supply chain, in: 2009 International Conference on Computers & Industrial Engineering, IEEE, 2009, pp. 835–838.
6. M. Saffari, S. Asmussen, R. Haji, The M/M/1 queue with inventory, lost sale, and general lead times, Queueing Systems 75 (1) (2013) 65–77.
7. A. Krishnamoorthy, N. C. Viswanath, Stochastic decomposition in production inventory with service time, European Journal of Operational Research 228 (2) (2013) 358–366.
8. C. Deepthi, Discrete time inventory models with/without positive service time, Ph.D. thesis, Cochin University of Science and Technology (2013).
9. A. Krishnamoorthy, R. Varghese, B. Lakshmy, An (s; Q) Inventory System with Positive Lead Time and Service Time Under N-Policy, Calcutta Statistical Association Bulletin 66 (3-4) (2014) 241–260.
10. A. Krishnamoorthy, R. Varghese, B. Lakshmy, Production inventory system with positive service time under local purchase, in: International Conference on Information Technologies and Mathematical Modelling, Springer, 2019, pp. 243–256.
11. T. Meisling, Discrete-time queuing theory, Operations Research 6 (1) (1958) 96–105.
12. S. C. Dafermos, M. F. Neuts, A single server queue in discrete time., Tech. rep., Purdue Univ Lafayette Ind Dept of Statistics (1969).
13. Z. Lian, L. Liu, M. F. Neuts, A discrete-time model for common life time inventory systems, Mathematics of Operations Research 30 (3) (2005) 718–732.
14. A. S. Alfa, Discrete time queues and matrix-analytic methods, Top 10 (2) (2002) 147–185.
15. A. S. Alfa, Applied discrete-time queues, Springer, 2016.
16. M. F. Neuts, Matrix-geometric solutions in stochastic models: an algorithmic approach, Courier Corporation, 1994.

UDC: 004.7

Architecture and functionality of the collective operations subnet of the Angara interconnect

A.S. Simonov[1] and O.M. Brekhov[2]

[1]JSC "NICEVT", Varshavskoe shosse 125, Moscow, Russia
[2]National research university for aeronautical engineering, Volokolamkoe shosse 4, Moscow, Russia

alexey.s.simonov@rambler.ru, obrekhov@mail.ru

Abstract

The Angara interconnect developed by JSC ≪NICEVT≫ is designed to connect the nodes of supercomputers and computing clusters. The paper describes the main architectural solutions, algorithms and functionality of the collective operations subnet of the Angara interconnect and presents the forecast of its characteristics based on the simulation modeling and actual operation. The proposed solutions allow bringing the time complexity of the collective operations execution to the theoretical limit for the kD-torus topology network.

Keywords: Angara interconnect, multiprocessor computing system, supercomputer

1. Introduction

Multiprocessor computing systems (hereinafter referred to as MCS) are important for solving application tasks aimed at increasing the scientific and technical potential of the economy and strengthening the country's defense capability. After NEC SX-6 Earth Simulator [1] appeared, it became clear how the characteristics and capabilities of the interconnect are important for ensuring high scalability of the MCS performance in solving computationally complex problems, primarily computer simulation, processing of large data arrays and forecasting.

JSC "NICEVT" started the development of the first-generation Angara high-speed interconnect (hereinafter referred to as Angara interconnect) in 2006. The operation principles and technical appearance of the Angara interconnect were formed basing on the analysis of the world experience in creating custom-made interconnect solutions for the highest performance range supercomputers, primarily the IBM BlueGene series [2-4] and CRAY SeaStar/Gemini [5-7], as well as on the results of a number of studies conducted at JSC "NICEVT" using simulation modeling tools.

Angara interconnect is a Direct Network that supports topologies from 1D-mesh to 4D-torus and allows to create MCS of up to 32K nodes. Its first production samples were presented in 2013.

During the development of the Angara interconnect the emphasis was placed on ensuring high scalability of the MCS performance in solving computationally complex problems. Algorithms for solving these problems, as a rule, are based on numerical methods with spatial decomposition of the computational domain, which require the execution of collective operations and boundary condition exchange operations after every iteration between the computational nodes of MCS that are involved in solving the problem.

Collective operations are among the main primitives of the interaction of computational processes in most parallel programming standards oriented to distributed memory computer systems (MPI, SHMEM, PGAS-languages — UPC, X10), and they can make up a significant part of communication exchanges. Such operations primarily include:

- broadcasting a packet from one node of MCS to other nodes allocated to the application (broadcast);

- collecting information from the nodes allocated to the application into one node and performing reduction (reduce).

This paper presents the main architectural solutions, algorithms and functionality of the collective operations subnet of the Angara interconnect and forecast of its characteristics based on the simulation modelling and actual operation.

2. Architecture of the collective operations subnet of the Angara interconnect

A trivial way to implement collective operations at the hardware level is to adequately replace them with many point-to-point operations, in which each broadcast is replaced with many writes to the memory coming from the root node to all other nodes of MCS allocated to the application, and each operation of collection is replaced by many operations of reading data from the MCS nodes memory to the root node.

This approach has undoubted advantages - simplicity of implementation and predictability of the result, and it can be used to build small MCS up to 100-200 nodes, for which, due to the small number of nodes, broadcast traffic will not have a significant impact on the load on the interconnect, and due to the small diameter of the network and the number of hops, the communication delay will be negligible.

For the medium and large size MCS the situation is different. In this case a large share of duplicate traffic can significantly affect the load of the interconnect, which, together with a large diameter, can negatively affect the latency and bandwidth of the interconnect and will lead to a significant deterioration in MCS performance [8].

That is why the hardware support of collective operations will positively affect the MCS performance scalability [9].

Obviously, a significant increase in the duration of collective operations due to an increase in the number of MCS nodes allocated to the application is associated with two factors - increased traffic due to an increase in the number of packets in the network and an increase in the number of hops due to an increase in the diameter of the network. In this regard, the problem of reducing the number of packets when performing collective operations for medium and large size MCS is very urgent.

The analysis of the trajectories of packets on the network in the case of implementation of the collective operations as a set of simple read and write operations in the memory of a remote node showed that in most cases the trajectories overlap each other. Considering the fact that the proposed solution requires its own routing algorithm that allows duplication and multiplication of packets when passing through transit nodes, it is advisable to implement hardware support for collective operations within a dedicated collective operations virtual subnet based on two virtual channels. A virtual subnet has the topology of a tree built in a torus (developers of the SMPO-10G interconnect came to a similar solution). There are two directions of movement in the tree: from root to leaves and from leaves to root. Each direction has its own virtual channel, VcDown – from root to leaves, and VcUp – from leaves to root. There may be transit nodes in the system; neither injection nor ejection of traffic occurs in these nodes.

When the broadcast operation is performed, each node, when receiving a packet from a node located higher in the tree, sends it to all nodes located lower in the tree. When a packet is injected into the network not from the root node, first a broadcast request is generated, which is sent to the root via the VcUp virtual channel, after which the broadcast operation itself is performed from the root via the VcDown virtual channels.

When the reduce (or all reduce) operation is performed, the node waits for packets from the node's processor if the node isn't transit, and from all the nodes located lower in the tree, performs the commutative associative binary operation indicated in the packet and sends the finished result up to the root. The current implementation supports the operations of maximum, minimum and sum of integers. The reduce operation ends with a hardware sending (without ejection) of the result to the given node using point-to-point operation (broadcast for all reduce). In order to determine direction to and from the root a routing table for collective operations subnet is set at each node. The table form is shown in Table 1.

The routing table in addition to dir-bits indicating the direction of the packet distribution along the subnet down the tree, has isRoot field that determines whether the node is root, toRoot field that indicates the direction to the root node, PE field

TreeId	isRoot	toRoot	PE	dirSum	dir-bits			
					+X	-X	+Y	-Y
0x00	0	+Y	1	3	1	1	0	1
0x01	0	+X	0	2	0	0	1	1
...								
0x0F								

Table 1. Form of collective operation subnet routing table

that determines whether this node is transit or not and dirSum field which stores the number of directions to the leaves of the tree.

To set the correct tree, the following criteria must be met:

a) there is only one root;

b) if at some node the direction is set down the tree, then in this direction there should be a node that belongs to the tree in which the direction up the tree is set opposite to the given;

c) directions to nodes down the tree can only be:

- directions next to the direction, specified by the toRoot field, according to the direction order;

- the direction opposite to the direction, specified by the toRoot field.

Since many tasks can run simultaneously on the MCS, the collective operations virtual subnet allows to build up to 16 intersecting trees. At the same time, one task can use several trees. Each tree has its own TreeId identifier. For each TreeId identifier there is a corresponding routing table entry that determines routing direction. A subnet supports up to 16 different reduce packets for each TreeId. Each reduce performing on this tree has its own reduceId identifier (from 0 to 15).

The generalized algorithms of the virtual channels VcDown and VcUp of the collective operation subnet are shown in Fig. 1 and 2. The algorithm presented in Fig. 1 works as follows. From the header fleet received for routing, the TreeId field is selected, according to which the corresponding line is searched in the routing table of the collective operations subnet. If the line is found, the PE field in the routing table is checked. If it is 1, i.e. the node is not transit; the packet is ejected into the node.

Next, the reading and execution of the bitwise conjunction operation of dir-bits of the packet header fleet with dir-bits from the routing table is performed. The resulting bit vector is used to perform a loop search over directions (+X, -X, +Y, -Y...) and duplicate the packet into those for which the corresponding bit of the resulting vector is in state 1.

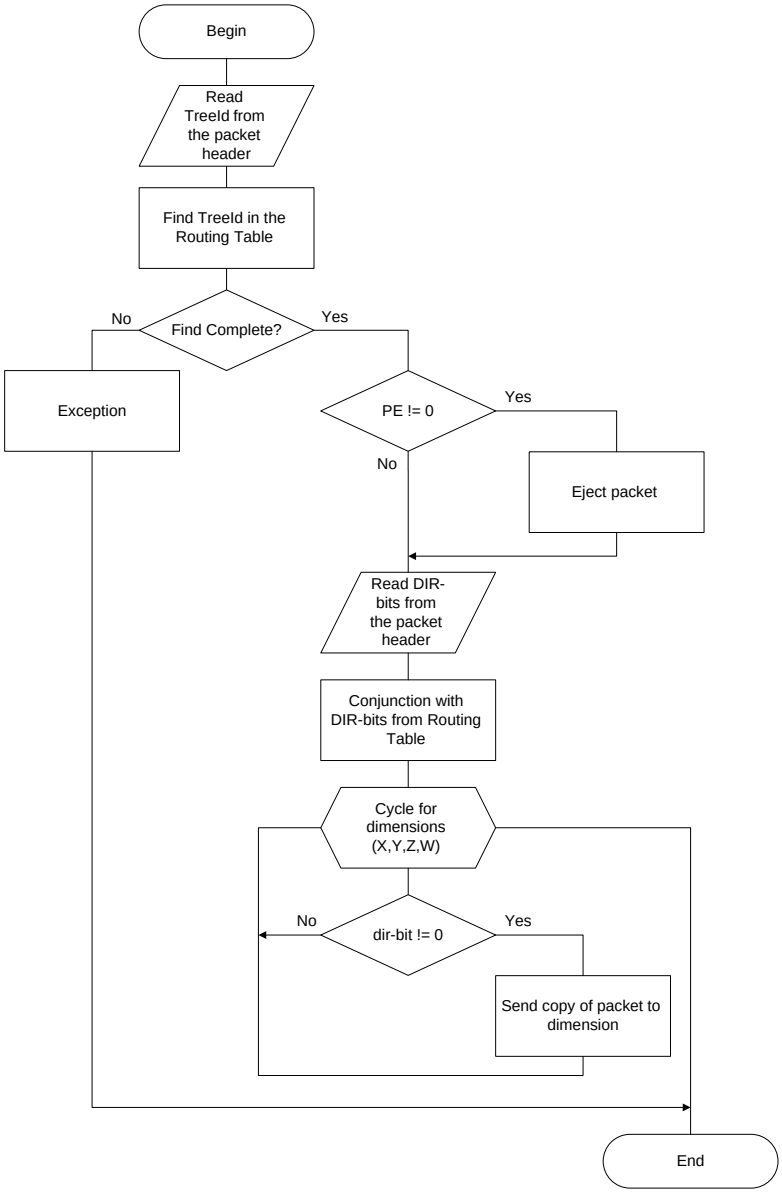


Fig. 1. The algorithm of the virtual channel VcDown of the collective operations subnet of the network with kD-torus topology

The algorithm presented in Fig. 2 works as follows. The packet type is checked after checking the TreeId identifier. If this is a broadcast packet the field isRoot

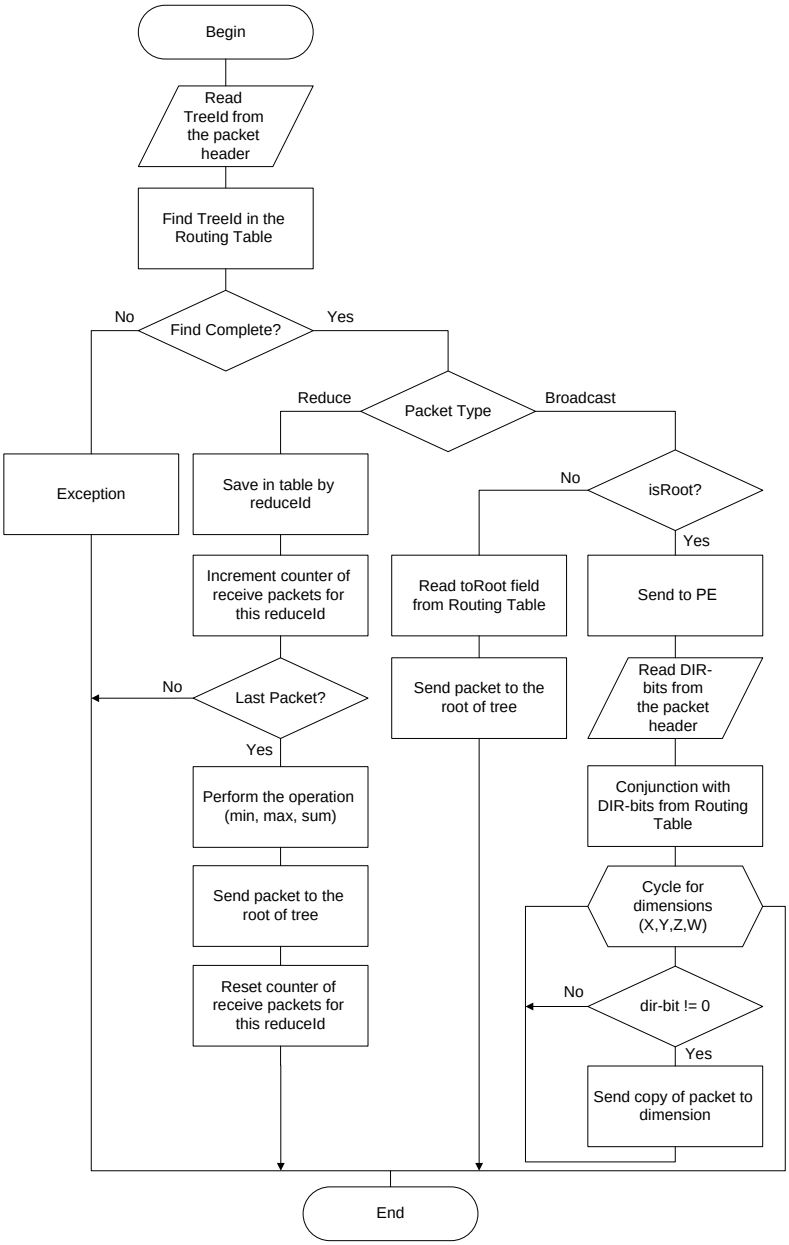


Fig. 2. The algorithm of the virtual channel VcUp of the collective operations subnet of the network with kD-torus topology

of the routing table is checked to see if this node is root. If this is a root node the package is ejected into the node and distributed similar to previous algorithm. Otherwise, the toRoot field of the routing table is read and the packet is sent in the direction to the root node of the tree.

Reduce type packets that have got into the virtual channel VcUp are processed as follows. Data is extracted from the package and stored in a special table in the line corresponding to the value of the reduceId field in the package, after which the received packets counter allocated for this reduceId is increased by 1. If packets came from all directions, i.e. the value of received packets counter has become equal to the dirSum value of the routing table, the operation is performed, the packet is sent up the tree to the root node and the received packets counter is reset.

From the point of view of the application programmer, the basic versions of collective operations are one-way asynchronous operations. The processor is not blocked after sending the message, and the result is written into memory without the active participation of the receiving party, which allows overlapping computation and communication. Synchronization mechanisms based on collective operations are implemented in order to determine if the collective operation is completed and the result is available to the computational nodes.

Considering the specificity of toroidal networks, the distribution of nodes by tasks is usually carried out on the principle of minimizing the distance between them. As a result, the nodes allocated to the task in the structure of the kD-torus are generally limited to a rectangular region. Since it is advisable to use the principle of minimizing the distances to the most distant nodes of the tree when choosing the root node, it is obvious that it is advisable to choose the node located in the geometric center of the rectangular area of the MCS nodes allocated to the task as the root node.

Two mechanisms are used in collective operations virtual subnet to prevent deadlocks caused by different tasks trees overlapping:

- bubble routing;
- various trees are constructed according to the X, Y, Z, W direction order.

3. Characteristics of the collective operations subnet of the Angara interconnect

To confirm the efficiency of proposed algorithms time complexity should be estimated. Diameter of a network is the shortest distance between the two most distant nodes in it. In general, for an equilateral kD-torus with even number of nodes diameter equals

$$r = \frac{k}{2}\sqrt[k]{\omega} \tag{1}$$

where r - is the network diameter;
k - is a number of dimensions in torus;
ω – is the total number of MCS nodes.
For an arbitrary 4D-torus networks r gives the lower-bound estimate.

When using VCT routing method, network message delivery time is determined by the communication delay and the time, required to inject the message of a given size into the communication channel (link) with a certain bandwidth, that is

$$T_d = t^0 + (l-1)t^1 + \frac{Q}{V_L} \tag{2}$$

where T_d - is the network message delivery time;
t^0 – is the communication latency between two adjacent nodes;
l – is the route length (number of hops);
t^1 – is the communication latency of a hop;
Q – is the message size;
V_L – is the communication channel bandwidth.

If collectives are implemented using point-to-point operations, the time complexity estimation for the broadcast and reduce algorithms will heavily depend on the host interface bandwidth of the root node

$$T(A^1_\omega) = T(A^\omega_1) = (\omega - 1)(t^1 + \frac{Q}{V_P}) + (\frac{k}{2}\sqrt[k]{\omega} - 1)t^1 + t^0 \tag{3}$$

where A^1_ω – is a broadcast algorithm;
$T(A^1_\omega)$ – is an estimation of the time complexity of the broadcast algorithm;
A^ω_1 – reduce algorithm;
$T(A^\omega_1)$ - is an estimation of the time complexity of the reduce algorithm;
V_P – host processor interface bandwidth.

Considering that in the VcDown and VcUp virtual channels algorithms the multiplication of packets during broadcast execution and reduction during reduce execution are performed at all nodes in the tree, the proposed method can significantly reduce both the number of packets and the total execution time. As a result, the estimate of time complexity will be close to the theoretical limit as it is determined by the distance to the most distant nodes. Applied to kD-torus network

$$T(A^1_\omega) = T(A^\omega_1) = t^0 + (\frac{k}{2}\sqrt[k]{\omega} - 1)t^1 + \frac{Q}{V_L} \tag{4}$$

Collective operations execution time were estimated to verify the above relation using simulation model. The tests consisted of sending one packet of 32 flits (256 bytes) using the broadcast operation for a different number of nodes of the simulated network.

Tables 2, 3 present the results obtained under the following initial conditions: $t^0 = 700$ ns; $t^1 = 80$ ns; $Q = 256$ bytes; $V_L = 6$ GB/s; $V_P = 8$ GB/s.

Parameter	Number of nodes of MCS			
	8	64	512	4096
3D-torus, point-to-point collective operations				
Analytical calculation, µs	1,64	8,16	58,81	461,18
Simulation modelling results, µs	2,18	6,07	40,16	312,64
Divergence, %	24,5%	25,6%	31,7%	32,2%
3D-torus, proposed method				
Analytical calculation, µs	0,9	1,14	1,62	2,58
Simulation modelling results, µs	0,96	1,20	1,76	2,80
Divergence, %	5,9%	4,8%	7,8%	7,8%

Table 2. Comparison of the estimated execution time of the broadcast operation depending on the number of computational nodes for the 3D-torus network

Parameter	Number of nodes of MCS		
	16	256	4096
4D-torus, point-to-point collective operations			
Analytical calculation, µs	2,62	29,82	460,54
Simulation modelling results, µs	2,90	21,78	312,65
Divergence, %	9,6%	26,9%	32,1%
4D-torus, proposed method			
Analytical calculation, µs	0,98	1,3	1,94
Simulation modelling results, µs	1,04	1,44	2,16
Divergence, %	5,5%	9,5%	10,0%

Table 3. Comparison of the estimated execution time of the broadcast operation depending on the number of computational nodes for the 4D-torus network

The above results indicate an acceptable value of the inaccuracy in estimating the time complexity of collective operations execution using expression 4. Expression 3, unfortunately, gives a higher inaccuracy, which, according to the author, is caused by incomplete consideration of certain aspects of the network.

The results obtained using the simulation model made it possible to preliminarily confirm the hypothesis that the proposed method for making a collective operations subnet gives a significant gain in comparison with the use point-to-point collective operations. At the same time, its hardware implementation in comparison with the software implementation also provides a significant gain.

4. Angara interconnect design versions

Currently there are two version of the Angara interconnect:

- switchless version – full-height full-length PCI Express card (see. Fig. 3) that allows to connect up to 32K computing nodes with an 8x16x16x16 4D-torus topology;
- switch version – 19" 24-port switch and low-profile PCI Express card (half-height full-length) (see. Fig. 4). This version allows to connect up to 2048 computing nodes with 2D-torus topology by connecting up to 256 switches.

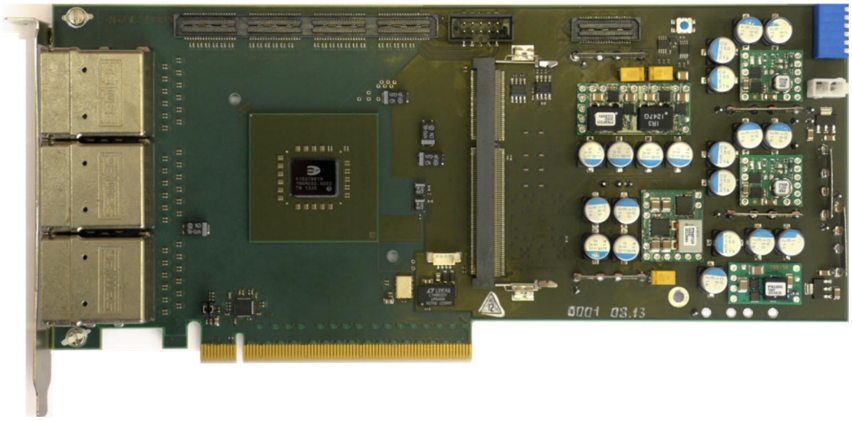


Fig. 3. Switchless version of the Angara interconnect

Fig. 4. The Angara interconnect 24-port switch (a) and low-profile PCI Express card (b)

There are several MCS with high performance scalability based on the Angara interconnect in the Russian Academy of Sciences institutions, research institutes and industrial enterprises.

The Angara interconnect allows achieving good performance scalability of MCS both on evaluation tests and applied computer simulation tasks (see. Fig. 5) [10-16].

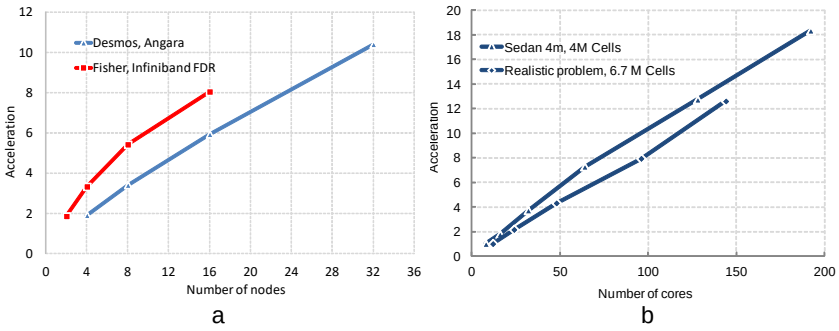


Fig. 5. MCS performance scalability (a – VASP - software package for quantum-chemical calculations, Desmos supercomputer; b –ANSYS FLUENT, Angara-K1 cluster)

5. Conclusion

1. The paper proposes architectural and algorithmic solutions for the collective operations subnet for a kD-torus topology network.

2. The analytical assessment of the developed algorithms time complexity is presented in the paper.

3. A comparison of the analytical assessment with the simulation modeling results was performed, which showed a 10% divergence (for proposed method).

4. The analysis of the developed algorithms was performed and the preliminary confirmation of the hypothesis that the proposed architectural and algorithmic solutions allows bringing the time complexity of the collective operations execution to the theoretical limit for the kD-torus topology network was obtained using the simulation model.

The proposed architectural and algorithmic solutions have a positive effect on the scalability of MVS performance in solving computationally complex problems, primarily computer simulation, processing large data arrays, planning and forecasting.

REFERENCES

1. Habata S. et al. The Earth Simulator system // NEC Research and Development. – 2003. – Т. 44. – №. 1. – С. 21–26.

2. Gara A. Overview of the Blue Gene/L system architecture // IBM Journal of research and development. – Vol. 49. – 2005. – Pp. 195–212. – http://rsim.cs.illinois.edu/arch/qual_papers/systems/19.pdf.
3. Overview of the IBM Blue Gene/P project / Gheorghe Almasi, Sameh Asaad, Ralph E Bellofatto et al. // IBM Journal of Research and Development. – 2008. – Vol. 52, no. 1–2. –Pp. 199–220. – http://scc.acad.bg/ncsa/documentation/team.pdf.
4. The IBM BlueGene/Q interconnection fabric / Dong Chen, Noel Eisley,Philip Heidelberger et al. // IEEE Micro. – 2012. – Vol. 32, no. 1. – Pp. 32–43. – https://www.researchgate.net/publication/220290398_The_IBM_blue_geneQ_interconnection_fabric.
5. Abts D, Storm C R. The Cray XT4 and Seastar 3-D torus interconnect. – 2010. https://static.googleusercontent.com/media/research.google.com/ru//pubs/archive/36896.pdf.
6. Alam S R. Cray XT4: an early evaluation for petascale scientific simulation // Proceedings of the 2007 ACM/IEEE Conference on / IEEE. – 2007. – Pp. 1–12. – https://doi.org/10.1145/1362622.1362675.
7. Alverson R, Roweth D, Kaplan L. The gemini system interconnect // High Performance Interconnects (HOTI), 2010 IEEE 18th Annual Symposium / IEEE. – 2010. – Pp. 83–87. – https://doi.org/10.1109/HOTI.2010.23.
8. Bala V., Bruck J., Cypher R. et al. CCL: A Portable and Tunable Collective Communication Library for Scalable Parallel Computers // Parallel Processing Symposium. – c. 835-844. – Proceedings, ISBN 0-8186-5620-6. – 1994.
9. Almasi G., Dozsa G., Erway C., Steinmacher-Burow B.. Effcient Implementation of All reduce on BlueGene/L Collective Network, Recent Advances in Parallel Virtual Machine and Message Passing Interface. – c.57-66. – Springer Berlin, Heidelberg. – 2005.
10. Mukosey A., Simonov A., Semenov A. Extended Routing Table Generation Algorithm for the Angara Interconnect // Russian Supercomputing Days. – Springer, Cham, 2019. – C. 573-583.
11. Stegailov, V., Dlinnova, E., Ismagilov, T., Khalilov, M., Kondratyuk, N., Makagon, D., Semenov, A., Simonov, A., Smirnov, G., Timofeev, A.: Angara interconnect makes GPU-based Desmos supercomputer an efficient tool for molecular dynamics calculations. The International Journal of High Performance Computing Applications, 2019.
12. Stegailov, V., Smirnov, G., Vecher, V.: VASP hits the memory wall: Processors efficiency comparison. Concurrency and Computation: Practice and Experience p. e5136. – 2019, https://doi.org/10.1002/cpe.5136.

13. Polyakov, S., Podryga, V., Puzyrkov, D.: High performance computing in multiscale problems of gas dynamics. Lobachevskii Journal of Mathematics 39(9). – P. 1239–1250. – 2018.
14. Ostroumova, G., Orekhov, N., Stegailov, V.: Reactive molecular-dynamics study of onion-like carbon nanoparticle formation. Diamond and Related Materials 94, P. 14–20. – 2019.
15. M. Tolstykh, G. Goyman, R. Fadeev, and V. Shashkin. Structure and algorithms of slav atmosphere model parallel program complex. Lobachevskii Journal of Mathematics, 39(4):587–595, 2018.
16. V. Akimov, D. Silaev, A. Aksenov, S. Zhluktov, D. Savitskiy, and A. Simonov. Flowvision scalability on supercomputers with angara interconnect. Lobachevskii Journal of Mathematics, 39(9):1159–1169, 2018.

UDC: 004

Mathematical model for horizontal on-demand vEPC scalability in SDN-based environment

A. Tsarev[1] and P. Abaev [1]

[1]Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation

atsarev@sci.pfu.edu.ru, abaev_po@rudn.university

Abstract

Next generation networks such as 5G provide a new approach for building highly scalable network infrastructure owing to NFV/VNF and SDN technologies. Infrastructure scalability allows a network operator to offer fine-graded on-demand services deployment. We consider virtualized Evolved Packet Core which functions are represented in form of several Virtual Network Functions. Computational capacity of every function could be extended via scaling mechanism with non-instantaneous activation/deactivation. In this paper we build a mathematical model of horizontal scaling for vEPC function in terms of queuing model with finite buffer size, several group of servers and queue thresholds. As a result, analytical formulas were derived and QoS metrics were proposed. Analytical evaluation shows that correct thresholds selection allows to significantly improve system performance.

Keywords: virtualized EPC, Evolved Packet Core, Queuing Model, Horizontal Scalability, Non-Instantaneous Activation/Deactivation, VNF, NFV, SDN, 5G

1. Introduction

Rapid development and evolution of modern computer networks is a network operator's response to increasing data consumption [1]. Cisco analytics predicted that two-thirds of humanity will have access to the Internet by 2023, and, consequently, generate data traffic. Further, the number of mobile users will reach 5.7 billion and the number of mobile connections will reach 13.1 billion [2]. Beyond that, one should take into the account increased user mobility and network traffic heterogeneity that implies different QoS/QoE requirements.

In order to cope with the mentioned challenges, telecommunication operators need properly update network infrastructure. One of the emerging trends in next

The publication has been prepared with the support of "RUDN University Program 5-100" program.

generation networks became 5G technology standardized by ITU. According to forecasts, more than 50 operators from 38 countries plan to launch 5G between 2018 and 2023 [3].

Since 5G includes concepts such as Software Defined Networking and Network Functions Virtualization [4], telecommunication operators could build more flexible and scalable infrastructure. The transition of Evolved Packet Core to the virtual one with help of Virtual Network Functions allows to increase the use of network resources and reduce OPEX [5]. At the same time one of key features of virtualization is possibility of resource scaling. What is more, the network load by user data usually has peaks during the day [6]. Therefore, one of the optimal strategies for mobile operators could be resource scaling in order to cope bottlenecks.

One of approaches for modelling scale in/out process and overcoming either overloads or resource allocation issues alongside with the hysteresis technique successfully used in [7]-[10], [14]-[16] could be non-instantaneous activation and deactivation of additional resources studied in this work.

The rest of the paper is organized as follows. Section II provides a brief overview of scalability researches for virtualized EPC. Section III describes a system model and approach for scale in/out process. Section IV deals with a mathematical model in terms of queueing model with finite buffer, finite servers number and queue thresholds. Section V provides results of analytical evaluation. Section VI contains conclusions.

2. Background

In the literature researchers focus on different aspects of virtual EPC scalability. Abaev et al. study in [9] case of hybrid evolved packet core with assumption there are legacy and virtualized parts of EPC. The system has two thresholds: one for scaling in and another for scaling out. Once scaling process initiated, VNF orchestrator deploys new VNF consecutively one by one. Proposed model is analyzed in a custom simulator.

Tobar et al. develop in [10] a scaling mechanism for evolved packet core based on network functions virtualization taking into the account both horizontal and vertical scalability. Proposed framework allows to find optimal workload between horizontal and vertical scaling and avoid waste of resources. Also, extensive scalability evaluation was performed in AWS on the base of vEPC from Indian Institute of Technology Bombay.

Liebsch et al. propose in [11] extension for traditional VNF scaling. The main idea of developed approach consists in scaling VNF from vEPC Data Plane that allows to avoid failure of overloaded Data Plane functions. In order to validate the proposal, the authors perform testing in real environment. Evaluated results demonstrate efficiency of runtime offload.

Kempf at al. consider in [12] the possibility to move EPC to the cloud. The main contribution of this work consists in describing how SDN could be applied Evolved Packet Core. The authors describe enhancements for Open Flow protocol and Open Flow switches necessary for EPC virtualization. Neither analytical nor numerical evaluation carried out.

Arteaga et al. investigate in [13] adaptive scaling mechanism in application to NFV-based EPC. The authors propose adaptive scaling algorithm based on Q-Learning and Gaussian Processes. Simulation shows better accuracy then algorithms with static thresholds.

Overall, all the mentioned papers have no analytical results. Alternatively, if we ignore SDN/VNF topic we can find such interesting works regarding either search of optimal hysteresis strategy in multi-server system [14] and analysis of multi-server threshold systems with hysteresis in [15]-[16]. In this paper we study single VNF deployment case and propose analytical model for VNF deployment in form of multi-server queuing model with thresholds and non-instantaneous activation and deactivation.

3. System Model

Evolved Packet Core is an important element of any modern network. Furthermore, EPC is responsible for authentication, authorization, channel bandwidth calculation, user mobility processing, etc. The main functions of the EPC include Access and Mobility Management Function (AMF), Session Management Function (SMF), User Plane Function (UPF), Unified Data Management (UDM) and Policy Control Function (PCF). SDN support in 5G networks allows network operators to virtualize mentioned functions via software and hardware decoupling.

EPC virtualization involves the transformation of EPC functions into Virtual Network Function (VNF), whose lifecycle is managed using VNF Manager. The decision whether to deploy a new function or disable the old one is realized by NFV Orchestrator. Thus, if a bottleneck occurs at the level of any of the EPC functions, network functions management and orchestration will allow you to react to the situation promptly and provide additional computing power to the problem node. An example of a virtualized EPC architecture is shown in the Fig. 1.

We consider a VNF group of the same type. The minimum number of deployed VNF is enough for the normal functioning of the network, we denote this group as c_0. When bottleneck occurs, it is feasible to allocate additional computational resources in order to provide required QoS level, particularly for delay critical applications. Since new VNF deploy is not instant and requires some time, it is worthwhile to use batch deploy approach. We consider the same assumption for undeploy procedure.

We consider reaching the threshold H_i in the queue of waiting requests as a trigger for the VNF group $c_i, 1 \leq i < k$ deploy.

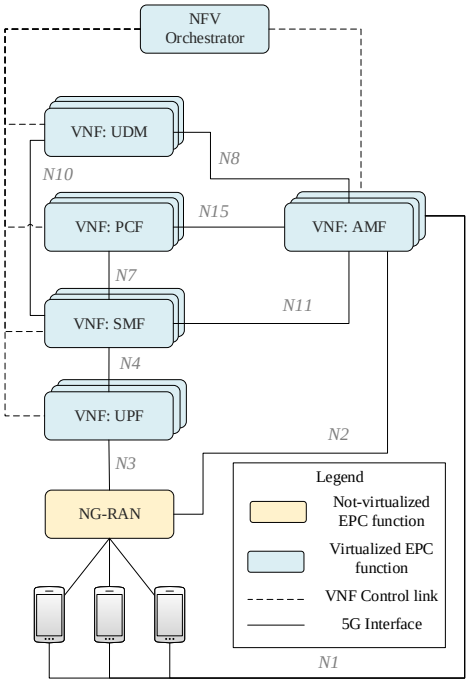


Fig. 1. Virtualized EPC architecture

4. Mathematical Model

The model is represented by the finite buffer of capacity r and C servers, as shown in the Fig. 2(a).

The Poisson requests flow reaches the system with rate $\lambda, 0 < \lambda < \infty$. We consider FCFS behavior for all incoming requests. Any server accepts one request at a time. All the servers are divided into $k + 1$ groups with exponentially distributed service rate $\mu_i, 0 < \mu_i < \infty, 0 \leq i \leq k + 1$. The system's buffer has a set of thresholds $H = H_i, 0 < H_i < r, 1 \leq r \leq k$.

Threshold H_i correlates to server group i. Initially, only servers from group 0 handle incoming requests. As soon as the buffer size reaches the threshold H_i, all the servers from group i become available for requests handling after scale in procedure that follows exponential distribution with rate $\theta, 0 < \theta < \infty$; and vice

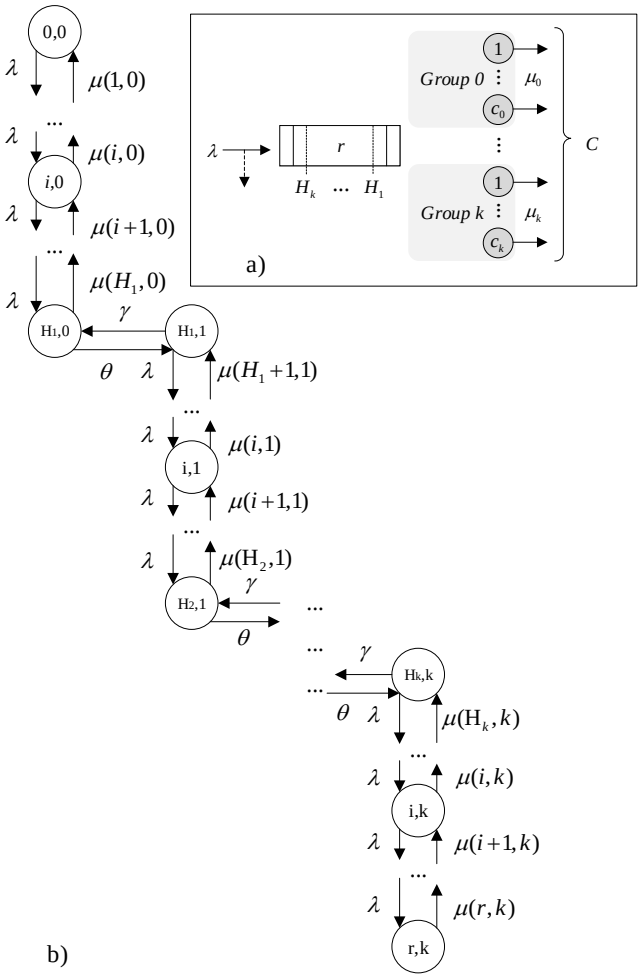


Fig. 2. (a) Queuing model with hysteric control via buffer thresholds
(b) Transition rate graph

versa in case buffer size drops below the threshold H_i, all the servers from group i become unavailable after scale out procedure that follows exponential distribution with rate $\gamma, 0 < \gamma < \infty$.

Let us use the following assumption: during activation or deactivation of a servers group the number of requests in the buffer cannot change. Another assumption is that the server, that provides service to a request, can not be changed if requests servicing is in process.

Assume $X(t) = (n(t), m(t))$ is a two-dimensional Poisson process over the set of states $\chi(t) = (n, m) : (0 \leq n \leq r, 0 \leq m \leq k)$ where n defines the number of requests in the system and m defines the number of working server groups. It is evident $X(t)$ is ergodic and its stationary distribution exists. Let us denote $p(i, j) = \lim_{t\to\infty} P\{n(t) = i, m(t) = j\}$ as stationary distribution. Transition rate graph for described system is presented in the Fig.2(b)

Assume $g(m)$ as a set of states for group m taking into the account the following assumption: $H_0 = 0$ and $H_{k+1} = r$:

$$g(m) = \{(H_m, m), ..., (H_{m+1}, m)\}, 0 \leq m \leq k \tag{1}$$

Note, that:

$$\sum_{m=0}^{k} \sum_{(i,j)\in g(m)} p(i, j) = 1 \tag{2}$$

Consider for brevity $\mu(n, m)$ as serving rate for n requests in the system and m deployed server groups:

$$\mu(n, m) = \begin{cases} min(n, c_0) \cdot \mu_0, m = 0 \\ \sum_{j=0}^{m-1} c_j \cdot \mu_j + min(n - \sum_{u=0}^{m-1} c_u, c_k) \cdot \mu_k, m > 0 \end{cases} \tag{3}$$

The transition rate graph of the process is shown in the Fig. **??**.

For case $m = 0$ stationary distribution could be expressed as:

$$p(n, 0) = \frac{\lambda^n}{\prod_{i=1}^{n} \mu(i, 0)} \cdot p(0, 0), 1 \leq n \leq H_1 \tag{4}$$

Note, that partial balance equations exist for neighboring groups:

$$p(n, m) = \frac{\theta}{\gamma} \cdot p(n, m - 1), 1 \leq m \leq k, n = H_m \tag{5}$$

Taking into the consideration (5), for case $m = 1$ we obtain:

$$p(n, 1) = \frac{\theta}{\gamma} \cdot \frac{\lambda^{n-H_1}}{\prod_{i=1}^{n-H_1} \mu(i, 1)} \cdot \frac{\lambda^{H_1}}{\prod_{i=1}^{H_1} \mu(i, 0)} \cdot p(0, 0), H_1 \leq n < H_2 \tag{6}$$

Correspondingly, in general case we have:

$$p(n, m) = \lambda^n \cdot \left(\frac{\theta}{\gamma}\right)^m \cdot \prod_{v=0}^{m} \prod_{(i,j)\in g(v)} \frac{1}{\mu(i, j)} \cdot p(0, 0) \tag{7}$$

Since all $p(n,m)$ are expressed through $p(0,0)$ we have:

$$p(n,m) = q(n,m) \cdot p(0,0) \tag{8}$$

Hence, from (2), (7) and (8) we obtain $p(0,0)$:

$$p(0,0) = \frac{1}{\sum_{v=0}^{m} \sum_{(n,m)\in g(v)} q(n,m)} \tag{9}$$

Thus, stationary distribution could be calculated analytically using (8) and (9). Main performance metrics could be calculated by the formulas (10), (11), (12), (13) and (14). Let B denote the blocking probability:

$$B = p(r,k) \tag{10}$$

Let Q denote mean queue size:

$$Q = \sum_{v=0}^{k} \sum_{(n,m)\in g(v)} \left(n - min(n, \sum_{j=0}^{m} c_j) \right) \cdot p(n,m) \tag{11}$$

Let N denote mean number of requests in the system:

$$N = \sum_{v=0}^{k} \sum_{(n,m)\in g(v)} n \cdot p(n,m) \tag{12}$$

According to Little's Law, let W_Q denote mean time that a request spends in the queue and W_N denote mean time that a request spends in the system:

$$W_Q = \frac{Q}{\lambda \cdot (1-B)} \tag{13}$$

$$W_N = \frac{N}{\lambda \cdot (1-B)} \tag{14}$$

5. Analytical Evaluation

We provide analytical evaluation applied to four difference scenarios. Input values for evaluation are provided in table 1. We consider that μ^{-1} equals 1 as a time unit. Server groups number equals 3 and includes zero group, hence we have only 2 thresholds.

Parameter	Comment	Value
λ	Incoming requests rate	1-20
Q	Queue length	50
θ	Servers group deployment rate	0.005
γ	Servers group undeployment rate	0.00005
c_0	Group 0 servers count	10
c_1	Group 1-2 servers count	5
μ_0	Service rate in group 0-2	1
H_1	Threshold for group 1	12
H_2	Threshold for group 2	25

Table 1. Considered input for evaluation

5.1. Varying queuing size. In this scenario we investigate the influence of buffer size on the system performance. Evaluation demonstrates that buffer capacity growth leads to packet loss reduction illustrated on the Fig. 3(a) and mean queue size rise illustrated on the Fig. 4(a). Although, there is no effect of hysteretic control on blocking probability, anyway we can see serious benefit in terms of mean waiting time and slight impact on queue size.

5.2. Varying the biggest group number. In this scenario we analyze resource allocation among fixed number of server groups. Fig. 3(b) shows that there is no difference in terms of blocking probability for different group setups. However, it is worth noting that allocating more resources to the last servers group improves QoS: mean queue size drastically falls on the Fig. 4(b) as well as mean waiting time on the Fig. 5(b).

5.3. Varying groups count. In this scenario we study optimal number of groups for case of equal resource allocation. We consider 2 groups of 10 servers, 3 groups of 6 and 7 servers and, finally, 4 groups of 5 servers. Numerical evaluation clearly demonstrates on the Fig. 3(c) that small number of groups with more resources is better in terms of blocking probability than large number of groups with fewer resources. Despite the fact that 2 groups of 10 servers shows higher mean queue size on the Fig. 4(c) and waiting time on the Fig. 5 at low values of λ , with growth of λ this setup achieves better performance.

5.4. Varying group service rate. In the last scenario we evaluate impact of different service rates on the system. We use $\mu = 1$ for all the groups as a reference result.Fig. 3(d) demonstrates that service rate boost by 10%, 20% and 30% considerable decreases blocking probability. Nevertheless, Fig. 4(d) and Fig. 5(d) show slightly decline for mean queue size and waiting time.

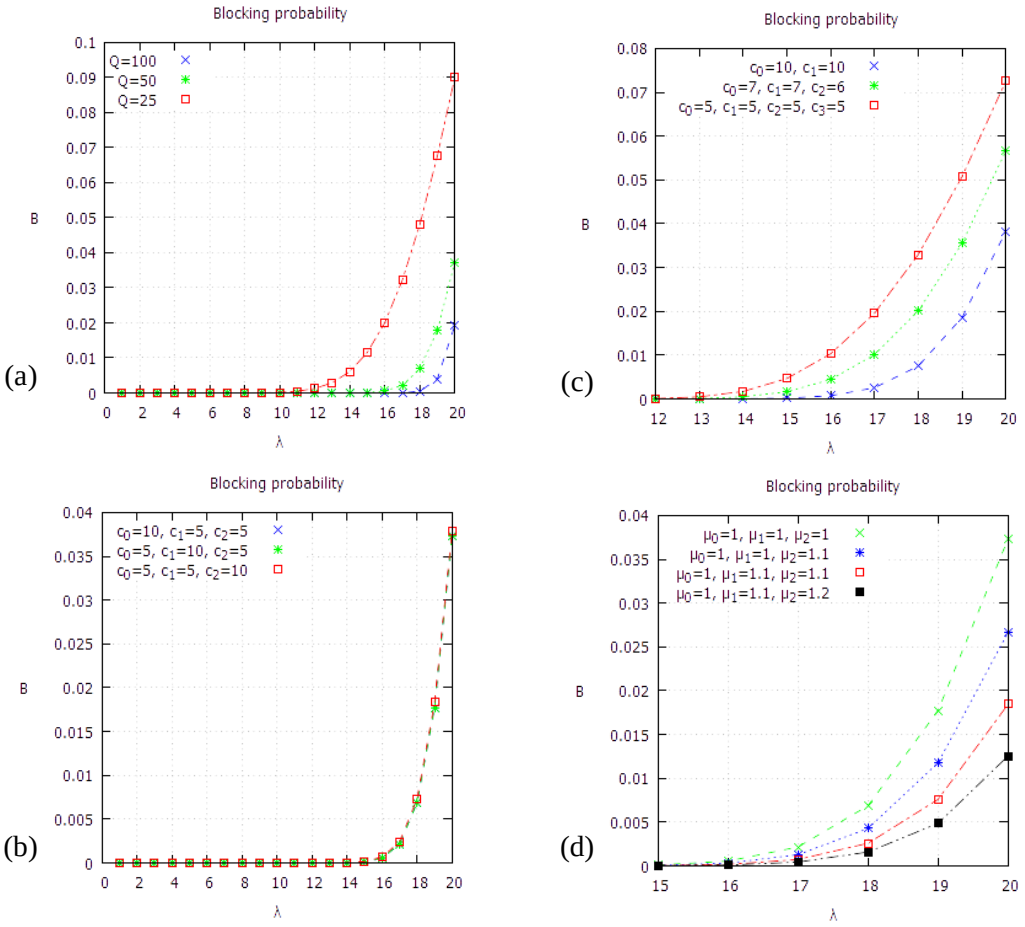


Fig. 3. Probability to block incoming request

6. Conclusions

In this paper we give an overview of virtualized Evolved Packet Core which throughput could be extended by allocation of additional computational resources in result of scaling in procedure. The hysteretic control approach was applied in order to handle horizontal scaling. Mathematical model was proposed and analytically evaluated. It was shown that horizontal resource scaling could significantly improve the system performance. To find optimal server groups partition and thresholds position optimization problem should be formulated and solved.

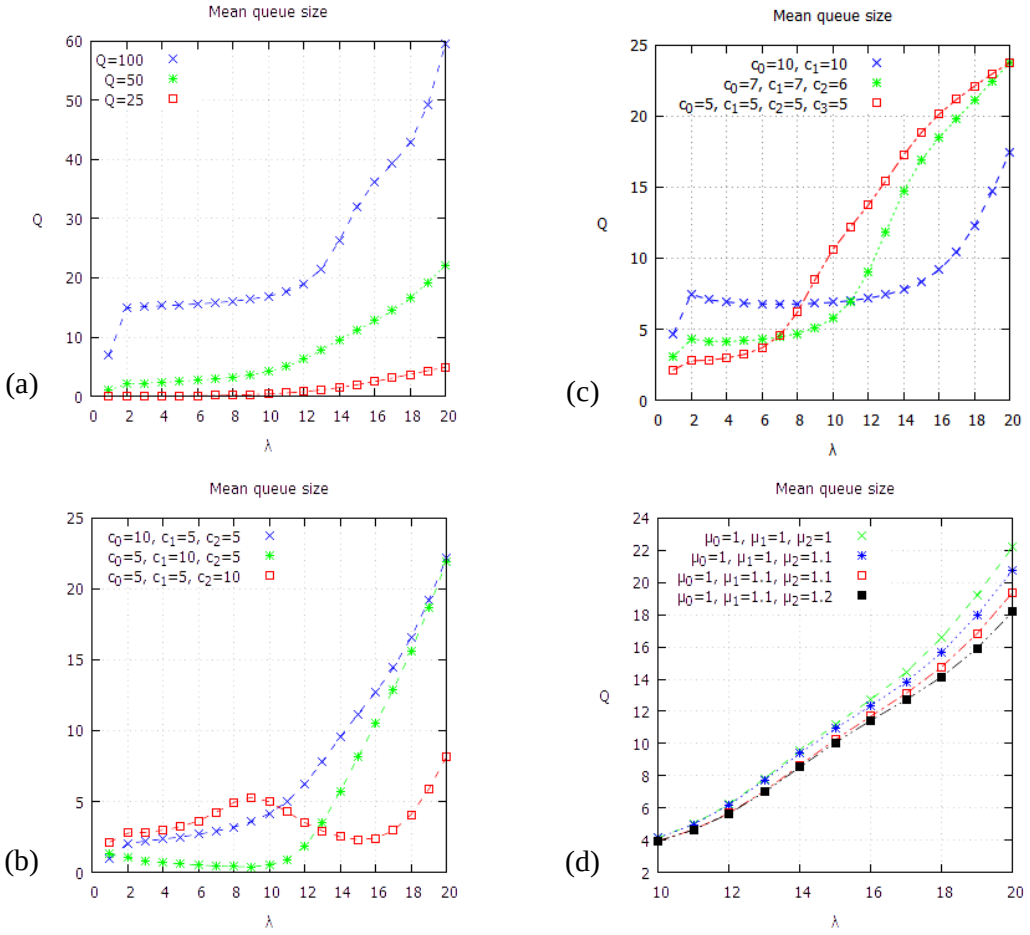


Fig. 4. Mean length of system's queue

REFERENCES

1. S. Soliman and B. Song. "Fifth Generation (5G) Cellular and the Network for Tomorrow: Cognitive and Cooperative Approach for Energy Savings" // Journal of Network and Computer Applications, vol. 85., pp. 84-93, May 2017.
2. Cisco Annual Internet Report (2018–2023) // White Paper, March, 2020.
3. Global Progress to 5G - Trials, Deployments and Launches // GSA Report, July 2018.

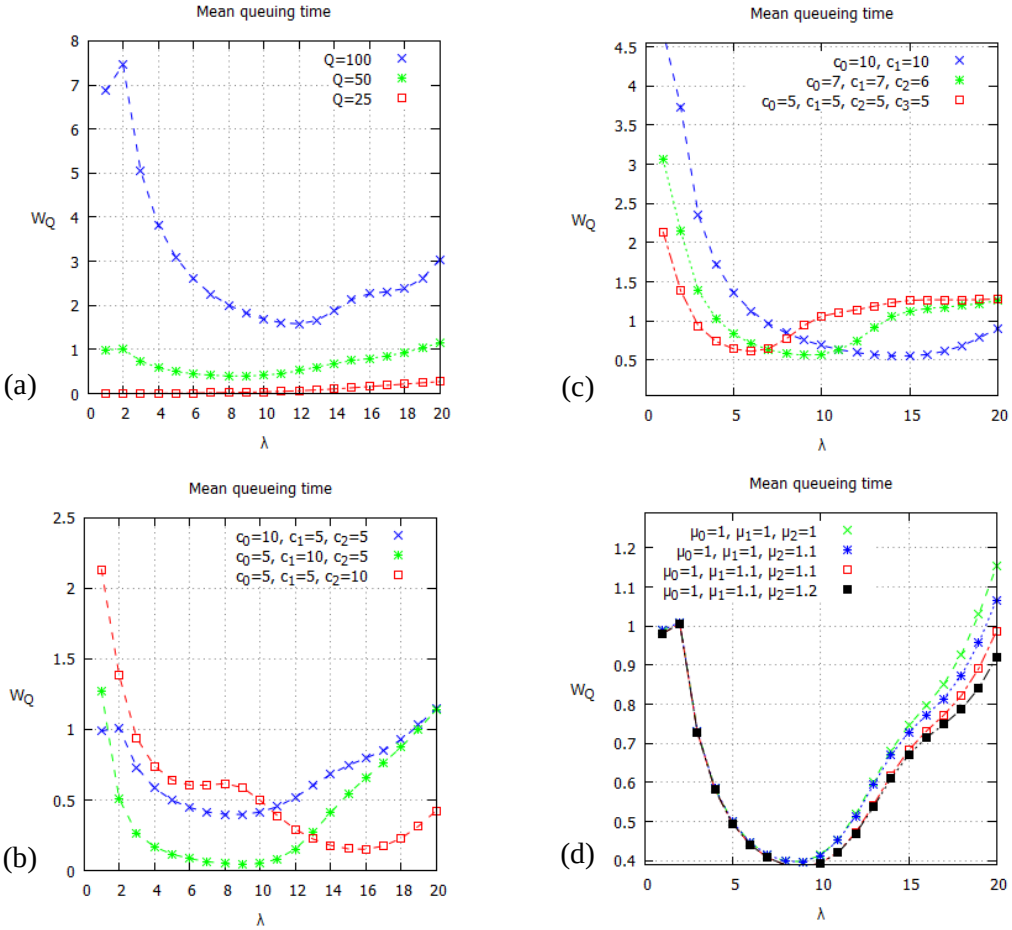


Fig. 5. Mean serve waiting time

4. M. Agiwal, A. Roy and N. Saxena. “Next Generation 5G Wireless Networks: A Comprehensive Survey” // IEEE Communications Surveys & Tutorials, vol. 18(3), pp. 1617–1655, 2016.
5. Economic Benefits of Virtualized Evolved Packet Core // White Paper, IDC, June 2016
6. P. Abaev P., R. Razumchik and I. Uglov. “Statistical Analysis and Modeling of SIP Traffic for Parameter Estimation of Server Hysteretic Overload Control” // Journal of Telecommunications and Information Technology, vol. 4, pp. 22-31, 2013.

7. P. Abaev, Y. Gaidamaka, K. Samouylov, A. Pechinkin, R. Razumchik and S. Shorgin. “Hysteretic control technique for overload problem solution in network of SIP servers” // Computing and Informatics, vol. 33, pp. 218-236, 2014.
8. Y. Gaidamaka, P. Abaev and K. Samouylov, “Modeling of Hysteretic Signaling Load Control in Next Generation Networks” in Lecture Notes in Computer Science, vol. 7469 // Springer, Berlin, Heidelberg, 2012, ch. Internet of Things, Smart Spaces, and Next Generation Networking, pp.440-452.
9. P. Abaev, A. Tsarev, “Hysteretic Mechanism for 5G Hybrid Evolved Packet Core Resource Management” // Proc. 10th International Congress on Ultra-Modern Telecommunications and Control Systems and Workshops (ICUMT), 2018, pp. 1-6.
10. C. H. Tobar Arteaga, F. B. Anacona, K. T. Tobar Ortega and O. M. Caicedo Rendon (December 2019). “A Scaling Mechanism for an Evolved Packet Core based on Network Functions Virtualization” // IEEE Transactions on Network and Service Management. pp.1-14
11. M. Liebsch and F. Faqir. “Virtualized EPC - Runtime Offload for Fast Data-Plane Scaling” // Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), 2016, pp. 1-6.
12. J. Kempf, B. Johansson, S. Petterson, H. Luening and T. Nilsson, “Moving the mobile Evolved Packet Core to the Cloud” // Proc. 8th International Conference on Wireless and Mobile Computing, Networking and Communications, pp. 784-791, 2012.
13. C. Tobar, F. Risso, and O. Caicedo, “An adaptive scaling mechanism for managing performance variations in network functions virtualization: A case study in an nfv-based epc” // Proc. International Conference on Network and Service Management (CNSM), Nov 2017, pp. 1–7.
14. C.S.Kim, A.Dudin, S.Dudin, O. Dudina ”Hysteresis Control by the Number of Active Servers in Queueing System with Priority Service” // Performance Evaluation, 2016, v. 101, pp. 20-33
15. J. C. S. Lui, L. Golubchik ”Stochastic complement analysis of multi-server threshold queues with hysteresis” // Performance Evaluation, v. 35(1-2), 1999, pp. 19–48.
16. C. Chou, L. Golubchik, J. Lui ”Multiclass Multiserver Threshold-Based Systems: A Study of Noninstantaneous Server Activation” // IEEE Transactions on Parallel and Distributed Systems, v. 18(1), 2007, pp. 96–110.

UDC: 519.2

Estimating the overflow probability in single-server retrial system with two classes of customers

Ksenia Zhukova[1]

[1]Karelian Research Centre RAS, Institute of Applied Mathematical Research,
Pushkinskaya st., 11, Petrozavodsk, Russia

Abstract

We discuss the asymptotic of the large deviation probability in single-server retrial queue with two classes of service time of customers. Constant retrial rate policy is considered. The input is assumed to be a general renewal process and the retrial attempts follow an exponential distribution. The systems are described with a regenerative process. We are interested in the large deviation probability that the orbit size of the system reaches a level N within regeneration cycle. We apply the idea to interpret the original retrial system with two classes of customers as a classic buffered system to estimate the upper and lower bounds of the large deviation probability.

Keywords: retrial queue, large deviations, overflow probability, simulation, constant retrial rate, orbit.

1. Introduction

Large deviation analysis is directly related to the evaluation of the quality of service parameters of the telecommunication and computer systems. One of such an important parameters is the overflow probability (the probability that the buffer content exceeds a given large threshold N). The problem of calculating and estimating overflow probability in classical buffered system was well-studied previously in [1].

In this paper, we apply the approach suggested in [2] to approximate a single-server retrial system with two classes of customers as an equivalent buffered single-server system. It allows to construct the upper and lower bounds for the large deviation (overflow) probability that the orbit size of the system reaches a level N within regeneration cycle, as $N \to \infty$. Retrial models are motivated by practical applications in the modern telecommunications systems. The problem of calculating and estimating overflow probability in retrial systems was considered previously in [3], [4] for queue size process. The lower and upper bounds for the decay rate of a large deviation asymptotics of the overflow probability during a regeneration cycle

in retrial system with constant and classical retrial policy are shown in [2]. These results were extended for the case of overflow probability during a full busy cycle in multiserver retrial system, see [5].

2. Overflow probability estimation

We consider a single-server retrial system with a renewal input of customers arriving at the instants $\{t_n\}$, with independent identically distributed (iid) interarrival times $\tau_n := t_{n+1} - t_n$, $t_1 = 0$, and with the iid service times of two classes

$$S_n = \begin{cases} S^{(1)}, \text{with probability } \ p \\ S^{(2)}, \text{with probability } \ 1-p, \end{cases} \tag{1}$$

where $S^{(1)}$ (for the first class), $S^{(2)}$ (for the second class) – independent random variables (r. v.), $n \geqslant 1$. We denote input rate $\lambda = 1/\mathsf{E}\tau$ and rate of the different service time classes $\mu_1 = 1/\mathsf{E}S^{(1)}$, $\mu_2 = 1/\mathsf{E}S^{(2)}$. We consider constant retrial policy and assume that if a new customer finds the server busy, it joins an infinite-capacity virtual orbit and attempts to occupy server after an exponentially distributed time with rate γ regardless of the class of a customer.

We are interested in the large deviation probability P_N, that the number of customers in the retrial system reaches a level N during a regeneration cycle, as $N \to \infty$. Regenerations of the system occur when an arrival meets completely empty system (for details, see [2]). To estimate upper and lower bounds for the asymptotics of P_N, we apply the key idea from [2] and interpret the original retrial system as a system, where each new customer joins the "end" of the orbit regardless of the state of server. In the case of busy server, the behaviour of an arrival in both systems is identical. If the server in the new system is idle at the arrival instant, then the oldest orbital customer of the the class jumps to server instead of the new arrival to start service immediately. After service completion, the server remains idle for a time until the next arrival or the attempt of the orbital customer happens. Due to stochastic equivalence, this replacement does not change distributional properties of the the number of customers. Moreover, the possible idle time before the service can be interpret as a part of service time of the next customer in the buffered system with the same input. Such an interpretation of the original system allow us to compare it to classical buffered systems with the same input and minoring/majoring values of service time.

Now we consider the interpretation of the original system as an buffered system described above. The service time $\{\hat{S}_n, n \leq 1\}$ of customers in this system can be written as

$$\hat{S}_n =_{st} S^{(1)} \cdot \mathsf{I}_n + S^{(2)} \cdot (1 - \mathsf{I}_n) + \xi_n, \tag{2}$$

where I_n is an indicator function for the class of service time

$$\mathsf{I}_n = \begin{cases} 1, \text{if } n\text{-th customer on the server is first class customer} \\ 0, \text{otherwise,} \end{cases}$$

and ξ_n is a possible idle time of the server before the n-th customer gets service (can equals zero). We assign the service times in the order in which service initiations occur, in other words, S_n is the service time used for the n-th such initiation. Assume that

$$S^{(1)} \leqslant_{st} S^{(2)}. \tag{3}$$

Then the minoring (S_l) and majoring (S_u) values for service time $\hat{S}$ can be easily constructed respectively as

$$S_l =_{st} S^{(1)} \leqslant_{st} \hat{S}, \tag{4}$$

$$\hat{S} \leqslant_{st} S_u =_{st} S^{(2)} + \xi, \tag{5}$$

where ξ is exponentially distributed r. v. with parameter γ.

Consider two classic buffered systems with the same inter-arrival times τ as in original retrial system and service times S_l and S_u. Because of stochastic monotonicity [2], these systems are minorant and majorant for the original one respectively, and it allows us to construct the lower and upper bounds for the overflow probability P_N.

For a random variable X, we remind the logarithmic (log) moment generating function, $\Lambda_X(\theta) = \log \mathsf{E}e^{\theta X}$, $\theta > 0$, and define

$$\hat{\theta} = \sup(\theta > 0 : \mathsf{E}e^{\theta S} < \infty) > 0. \tag{6}$$

The lower and upper bounds are presented in the following statement:

Lemma 1. Assume that condition (3) and

$$\frac{1}{\mu_2} + \frac{1}{\gamma} < \frac{1}{\lambda} \tag{7}$$

hold. Then the overflow probability P_N satisfies

$$\Lambda_\tau(-\theta_*) \leq \limsup_{N\to\infty} \frac{1}{N} \log \mathsf{P}(K_N < K_0) \leq \Lambda_\tau(-\theta^*),$$

where θ_* and θ^* are defined as

$$\theta_* = \sup\Big(\theta \in (0,\, \hat{\theta}) : \Lambda_\tau(-\theta) + \Lambda_{S^{(1)}}(\theta) \leqslant 0\Big) \tag{8}$$

and

$$\theta^* = \sup\Big(\theta \in (0,\, \min(\hat{\theta}, \gamma)) : \Lambda_\tau(-\theta) + \Lambda_{S^{(2)}}(\theta) + \log\frac{\gamma}{\gamma - \theta} \leqslant 0\Big). \tag{9}$$

The similar statement holds with all limsups replaced by liminf.

Condition (7) provide the positive recurrence of all the systems described above and it necessary for the stationarity of the systems (for details see [1], [2]).

3. Example

In this section we present an example that shows the estimation of the overflow probability in a retrial system and the constructed theoretical bounds (8).

First we simulate $M/GI/1$ retrial system with input rate $\lambda = 2$ and exponentially distributed r. v. $S^{(1)}$ and $S^{(2)}$ with parameters $\mu_1 = 1.5$, $\mu_2 = 1.25$ and $p = 0.5$. Let retrial rate $\gamma = 10$. Note that the chosen parameters satisfy condition (7), and thus the statements of Lemma 1 hold true. We estimate the overflow probability that number of customers in the retrial system exceeds given level N during a regeneration cycle. The estimation is constructed on $k = 10000$ cycles.

Second, we calculate the upper and lower bounds of the overflow probability in the described system for several values N (see Fig. 1) using the theoretical result (8). In our setting $\hat{\theta} = 1.25$. We need to derive

$$\Lambda_\tau(\theta) = \log \mathsf{E} e^{\theta\tau} = \log \frac{\lambda}{\lambda - \theta}. \tag{10}$$

and

$$\Lambda_{S^{(1)}}(\theta) = \log \frac{\mu_1}{\mu_1 - \theta}. \tag{11}$$

Using (10) and (11) we can calculate

$$\begin{aligned} \theta_* &= \sup\left(\theta \in (0, \hat{\theta}) : \Lambda_\tau(-\theta) + \Lambda_{S^{(1)}}(\theta) \le 0\right) \\ &= \sup\left(\theta \in (0, \hat{\theta}) : \frac{\lambda}{\lambda + \theta} \cdot \frac{\mu_2}{\mu_2 - \theta} \le 1\right). \end{aligned} \tag{12}$$

In our settings, the value $\theta_* = 0.5$, so $\Lambda_\tau(-\theta_*) = -0.41$.

Now we calculate θ^*:

$$\begin{aligned} \theta^* &= \sup\left(\theta \in (0, \min(\hat{\theta}, \gamma)) : \Lambda_\tau(-\theta) + \Lambda_{S^{(2)}}(\theta) + \log \frac{\gamma}{\gamma - \theta} \le 0\right) \\ &= \sup\left(\theta \in (0, \hat{\theta}) : \frac{\gamma}{\gamma - \theta} \cdot \frac{\lambda}{\lambda + \theta} \cdot \frac{\mu_2}{\mu_2 - \theta} \le 1\right). \end{aligned} \tag{13}$$

In our settings, the value $\theta^* = 0.13$, so $\Lambda_\tau(-\theta^*) = -0.126$.

We compare the estimation of the overflow probability on the regeneration cycle with the theoretical bounds (8). The simulation result presented in Fig. 1 shows that

estimate of the overflow probability is indeed between the lower and upper bounds, but the bounds become quite rough as overflow level N increases. The reason is that the majoring (5) and minoring (4) values have to be independent so that it is possible to use the result of [1]. On the other hand, the service time $\hat{S}_n$ (2) are dependent. This leads us to construct a quite rough bounds but we suppose that there are some opportunities to improve the bounds with the help of empirical approach.

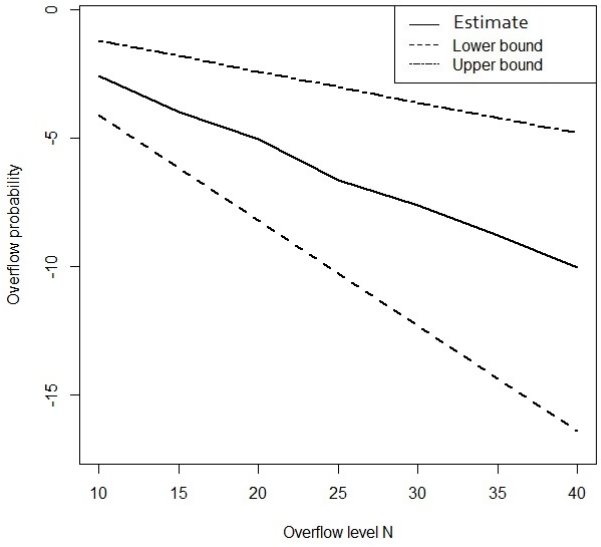


Fig. 1. Estimates of the overflow probability in retrial system and theoretical asymptotics vs. overflow level N; logarithmic scale.

4. Conclusion

Single-server retrial system with constant retrial rate and two classes of service time is considered. The probability that the orbit size of the system reaches a high level N, within regeneration cycle, is studied. It is shown that the original retrial system can be treated as an equivalent buffered system with service times of a special type. The lower and upper bounds for the asymptotic of large deviation probability are constructed.

5. Acknowledgements

The research was carried out under state order to the Karelian Research Centre of the Russian Academy of Sciences (Institute of Applied Mathematical Research KarRC RAS) and supported by the Russian Foundation for Basic Research, projects 18-07-00147 and 19-57-45022.

REFERENCES

1. Sadowsky J. S. Large deviations theory and efficient simulation of excessive backlogs in a GI/GI/m queue // IEEE Trans. Autom. Control. 1991. V. 36(12). P. 1383–1394.
2. Morozov E., Zhukova K. A large deviation analysis of retrial models with constant and classic retrial rates // Performance Evaluation. 2019. V. 135.
3. Kim J., Kim B. Tail asymptotics for the queue size distribution in the MAP/G/1 retrial queue // Queueing System. 2010. V. 66. P. 79–94.
4. Kim J., Kim B., Ko S.-S. Tail asymptotics for the queue size distribution in an M/G/1 retrial queue // J. Appl. Probab. 2007. V. 44. P. 1111-–1118.
5. Zhukova K., Morozov E. An Upper Bound of the Large Deviation Probability in Multi-Server Constant Retrial Rate System // Vishnevskiy V., Samouylov K., Kozyrev D. (eds) Distributed Computer and Communication Networks. DCCN 2019. Communications in Computer and Information Science. Springer, Cham. 2019. V. 1141. P. 325-337.
6. Artalejo J. R., Resing J. A. C. Mean value analysis of single server retrial queues // Asia-Pacific Journal of Operational Research. 2010. V. 27(3). P. 335-–345.

УДК: 519.872

Исследование бесконечнолинейной СМО с интенсивностью входящего потока, зависящей от состояния системы

Е.П. Полин[1], С.П. Моисеева[1], А.Н. Моисеев[1]

[1]НИ ТГУ, пр. Ленина 36, Томск, Россия

polin_evgeny@mail.ru, smoiseeva@mail.ru, moiseev.tsu@gmail.com

Аннотация

Рассматривается математическая модель бесконечнолинейной системы массового обслуживания с входящим пуассоновским потоком с интенсивностью, зависящей от числа занятых приборов. Дисциплина обслуживания определяется тем, что заявка занимает любой из свободных приборов в системе, на котором выполняется ее обслуживание в течение случайного времени, распределенного по экспоненциальному закону. Методом производящих функций определены выражения для вероятностных характеристик числа занятых приборов в системе в стационарном режиме. Получена производящая функция рассматриваемого случайного процесса, имеющая вид производящей функции случайной величины, имеющей отрицательное биномиальное распределение вероятностей.

Ключевые слова: система массового обслуживания, переменная интенсивность, метод производящих функций.

1. Введение

Современные приложения теории массового обслуживания включают исследования математических моделей передачи данных в телекоммуникационных системах и компьютерных сетях связи. В связи с этим возникает необходимость рассмотрения систем массового обслуживания с изменяемыми параметрами функционирования, такими как интенсивность входящего потока, параметры обслуживания и другие [1–5].

В настоящей работе рассматривается математическая модель бесконечнолинейной системы массового обслуживания с входящим пуассоновским потоком с интенсивностью, зависящей от числа занятых приборов.

2. Математическая модель

Рассмотрим систему массового обслуживания, входящий поток которой является пуассоновским с интенсивностью вида $\lambda(i(t)) = a + b \cdot i(t)$, где b – коэффициент, отражающий реакцию интенсивности входящего потока на количество заявок в системе, $i(t)$ – число занятых приборов в системе в момент времени t. В рамках данной работы рассматривается случай, когда $b > 0$. Дисциплина обслуживания определяется тем, что заявка занимает любой из свободных приборов в системе, на котором выполняется ее обслуживание в течение случайного времени, распределенного по экспоненциальному закону с параметром μ. Ставится задача исследования вероятностных характеристик случайного процесса $i(t)$.

3. Система дифференциальных уравнений Колмогорова

Для распределения вероятностей рассматриваемого случайного процесса $i(t)$ составим Δt – методом прямую систему дифференциальных уравнений Колмогорова. По формуле полной вероятности запишем равенства

$$P_0(t+\Delta t) = P_0(t)(1 - a\Delta t) + P_1(t)\mu\Delta t + o(\Delta t),$$
$$P_i(t+\Delta t) = P_i(t)(1 - (a+bi)\Delta t)(1 - i\mu\Delta t) + P_{i-1}(t)[a + b(i-1)]\Delta t +$$
$$+P_{i+1}(t)(i+1)\mu\Delta t + o(\Delta t),\ i \geq 1.$$

Система дифференциальных уравнений Колмогорова примет вид

$$\frac{\partial P_0(t)}{\partial t} = -aP_0(t) + \mu P_1(t),$$
$$\frac{\partial P_i(t)}{\partial t} = -(a + bi + \mu i)P_i(t) + (a - b + bi)P_{i-1}(t) + (i+1)\mu P_{i+1}(t),\ i \geq 1. \quad (1)$$

4. Метод производящих функций

Производящая функция определена в виде

$$F(z,t) = \sum_{i=0}^{\infty} z^i P_i(t).$$

Из системы дифференциальных уравнений Колмогорова (1) для функций $F(z,t)$ получаем линейное дифференциальное уравнение в частных производных первого порядка

$$\frac{\partial F(z,t)}{\partial t} - [(z-1)(bz-\mu)]\frac{\partial F(z,t)}{\partial z} = a(z-1)F(z,t). \quad (2)$$

Для (2) запишем систему дифференциальных уравнений вида

$$\frac{dt}{1} = \frac{dz}{(z-1)(\mu - bz)} = \frac{dF(z,t)}{a(z-1)F(z,t)}. \tag{3}$$

Найдем два первых интеграла этой системы. Один из них найдем из уравнения

$$dt = \frac{dz}{(z-1)(\mu - bz)},$$

$$dt = \frac{dz}{-b(z-1)\left(z - \frac{\mu}{b}\right)},$$

$$-bdt = \frac{dz}{(z-1)\left(z - \frac{\mu}{b}\right)} = \frac{Adz}{z-1} + \frac{Bdz}{z - \frac{\mu}{b}}.$$

Коэффициенты A и B находим методом неопределенных коэффициентов

$$A = \left.\frac{1}{z - \frac{\mu}{b}}\right|_{z=1} = \frac{b}{b-\mu},$$

$$B = \left.\frac{1}{z-1}\right|_{z=\frac{\mu}{b}} = -\frac{b}{b-\mu}.$$

Следовательно,

$$-bdt = \frac{bdz}{(z-1)(b-\mu)} - \frac{bdz}{(b-\mu)\left(z - \frac{\mu}{b}\right)},$$

или

$$-(b-\mu)dt = \frac{dz}{z-1} - \frac{dz}{z - \frac{\mu}{b}}.$$

Интегрируя, получаем

$$-(b-\mu)t = \ln\frac{z-1}{z - \frac{\mu}{b}} - \ln C_1,$$

откуда получаем выражение для C_1

$$C_1 = e^{(b-\mu)t}\frac{z-1}{z - \frac{\mu}{b}}. \tag{4}$$

Второй интеграл найдем из уравнения

$$\frac{dz}{(z-1)(\mu-bz)} = \frac{dF(z,t)}{a(z-1)F(z,t)},$$

откуда получаем выражение

$$F(z,t) = C_2 \left(z - \frac{\mu}{b}\right)^{-\frac{a}{b}}. \tag{5}$$

Подставляя выражение (4), общее решение уравнения (5) можно записать следующим образом

$$F(z,t) = \Phi\left(e^{(b-\mu)t}\frac{z-1}{z-\frac{\mu}{b}}\right)\left(z-\frac{\mu}{b}\right)^{-\frac{a}{b}}, \tag{6}$$

где $\Phi(x)$ – произвольная дифференцируемая функция. Учитывая начальные условия $F(z,0) = \sum_{i=0}^{\infty} z^i P_i(0) = 1$, имеем

$$1 = \Phi\left(\frac{z-1}{z-\frac{\mu}{b}}\right)\left(z-\frac{\mu}{b}\right)^{-\frac{a}{b}} = \Phi\left(1+\frac{\frac{\mu}{b}-1}{z-\frac{\mu}{b}}\right)\left(z-\frac{\mu}{b}\right)^{-\frac{a}{b}},$$

откуда

$$\Phi(x) = \left(\frac{\frac{\mu}{b}-1}{x-1}\right)^{\frac{a}{b}}.$$

Подставляя полученное выражение в (6), получаем вид производящей функции

$$F(z,t) = \left(\frac{\frac{\mu}{b}-1}{\left(e^{(b-\mu)t}\frac{z-1}{z-\frac{\mu}{b}}\right)-1}\right)^{\frac{a}{b}}\left(z-\frac{\mu}{b}\right)^{-\frac{a}{b}} = \tag{7}$$

$$= \frac{\left(\frac{\mu}{b}-1\right)^{\frac{a}{b}}}{\left((z-1)(e^{(b-\mu)t}-1)+\frac{\mu}{b}-1\right)^{\frac{a}{b}}}.$$

5. Математическое ожидание и дисперсия числа заявок в СМО

Используя свойства производящих функций, найдем математическое ожидание и дисперсию числа заявок в исследуемой СМО.

Имеем для математического ожидания

$$m_1(t) = \frac{\partial F(z,t)}{\partial t}\bigg|_{z=1},$$

$$\frac{\partial F(z,t)}{\partial t} = \frac{a(te^{b-\mu}-1)\left(\frac{\mu}{b}-1\right)^{\frac{a}{b}}}{b\left(\frac{\mu}{b}+(z-1)(te^{b-\mu}-1)-1\right)^{\frac{a}{b}+1}}.$$

При $z=1$ получим

$$m_1(t) = \frac{\partial F(z,t)}{\partial t}\bigg|_{z=1} = \frac{a(te^{b-\mu}-1)}{\mu-b}.$$

$$m_2(t) = \frac{\partial^2 F(z,t)}{\partial^2 t}\bigg|_{z=1} + m_1(t),$$

$$\frac{\partial^2 F(z,t)}{\partial^2 t} = \frac{a(te^{b-\mu}-1)^2\left(\frac{a}{b}+1\right)\left(\frac{\mu}{b}-1\right)^{\frac{a}{b}}}{b\left(\frac{\mu}{b}+(z-1)(te^{b-\mu}-1)-1\right)^{\frac{a}{b}+2}}.$$

При $z=1$ получим

$$\frac{\partial^2 F(z,t)}{\partial^2 t}\bigg|_{z=1} = \frac{a(te^{b-\mu}-1)^2\left(\frac{a}{b}+1\right)}{(\mu-b)^2},$$

$$m_2(t) = \frac{\partial^2 F(z,t)}{\partial^2 t}\bigg|_{z=1} + m_1(t) =$$

$$= \frac{a(te^{b-\mu}-1)^2\left(\frac{a}{b}+1\right) + a(te^{b-\mu}-1)(\mu-b)}{(\mu-b)^2}.$$

Дисперсия будет иметь вид

$$D = m_2(t) - m_1^2(t) = \frac{a(te^{b-\mu}-1)(bte^{b-\mu}+\mu-2b)}{(\mu-b)^2}.$$

6. Характеристики системы в стационарном режиме

Рассмотрим стационарный режим функционирования системы.

Выполняя предельный переход при $t \longrightarrow \infty$, получим

$$F(z) = \left(\frac{\frac{\mu}{b} - 1}{\frac{\mu}{b} - z} \right)^{\frac{a}{b}}. \tag{8}$$

Используя несложные математические преобразования, получим вид производящей функции $F(z)$

$$F(z) = \left(\frac{1 - \frac{b}{\mu}}{1 - z\frac{b}{\mu}} \right)^{\frac{a}{b}}. \tag{9}$$

Обозначим $\dfrac{b}{\mu} = \rho$, тогда (9) перепишем в виде

$$F(z) = \left(\frac{1 - \rho}{1 - z\rho} \right)^{\frac{a}{b}}. \tag{10}$$

Полученная производящая функция имеет вид производящей функции случайной величины, имеющей отрицательное биномиальное распределение с параметрами $\dfrac{a}{b}$, $1 - \dfrac{b}{\mu}$. Величина $1 - \dfrac{b}{\mu}$ принимает значения от 0 до 1, так как она имеет смысл вероятности. Отсюда следует, что $b < \mu$.

С помощью формулы (10) запишем выражения для математического ожидания и дисперсии числа занятых приборов в рассматриваемой системе.

Имеем для математического ожидания числа занятых приборов в системе

$$m_1 = \left. \frac{\partial F(z)}{\partial z} \right|_{z=1} = \frac{\frac{a}{b}\rho}{1 - \rho} = \frac{\frac{a}{\mu}}{1 - \frac{b}{\mu}}.$$

$$m_2(t) = \left. \frac{\partial^2 F(z,t)}{\partial^2 t} \right|_{z=1} + m_1(t) = \frac{\frac{a}{b}\rho\left(\frac{a}{b}\rho + 1\right)}{(1-\rho)^2}.$$

Дисперсия будет иметь вид

$$D = m_2(t) - m_1^2(t) = \frac{\frac{a}{b}\rho}{(1-\rho)^2} = \frac{\frac{a}{\mu}}{(1-\frac{b}{\mu})^2}.$$

7. Заключение

В настоящей статье рассмотрена система массового обслуживания, входящий поток которой является пуассоновским потоком с интенсивностью вида $\lambda(i(t)) = a + b \cdot i(t)$. Получены математическое ожидание и дисперсия числа заявок в рассматриваемой системе в стационарном режиме. Производящая функция рассматриваемого случайного процесса имеет вид производящей функции случайной величины, имеющей отрицательное биномиальное распределение с параметрами $\frac{a}{b}$, $1 - \frac{b}{\mu}$.

В дальнейшем планируется провести подобные исследования для случая $b < 0$, а также для систем с непуассоновскими входящими потоками и неэкспоненциальным обслуживанием, когда параметры входящего потока зависят от числа заявок, присутствующих в системе.

ЛИТЕРАТУРА

1. Коротаев И. А., Спивак Л. Р., Системы массового обслуживания в полумарковской случайной среде // Автомат. и телемех. 1992. Выпуск 7. С. 86–92.
2. Зиновьева Л. И., Терпугов А. Ф., Однолинейная система массового обслуживания с переменной интенсивностью, зависящей от времени ожидания // Автомат. и телемех. 1981. № 1. С. 27–30.
3. Таташев А. Г., Система массового обслуживания с переменной интенсивностью входного потока // Автомат. и телемех. 1995. Выпуск 12. С. 78–84.
4. Бондрова О.В. Анализ характеристик незавершенной работы в стационарной СМО с бесконечным накопителем и скачкообразной интенсивностью входного потока // Вестник ВГУ. Серия: Физика. Математика. 2015. № 2. С. 76–91.
5. Головко Н. И., Коротаев И. А. Система массового обслуживания со случайно изменяющейся интенсивностью входного потока // Автоматика и телемеханика. 1990. № 7. С. 80–85.

UDC: 123.456

Multi-user Detection to Improve Downlink Communication of CSS-based LoRa-like Networks

Angesom Ataklity TESFAY[1], Eric Pierre SIMON[1], Laurent CLAVIER [1,2]

[1]University of Lille, CNRS, UMR 8520 - IEMN, F-59000, Lille, France
[2]IMT Lille Douai, France

firstname.surname@univ-lille.fr

Abstract

Low Power Wide Area Networks like LoRa are one of the main building blocks of the Internet of Things. One of the main issues is to scale up the number of devices and one strong limitation comes from the downlink communication. In fact, the access point is constrained by the duty cycle therefore it can not address a large number of devices. We propose a superposition scheme to transmit multiple packets to multiple devices in the same frequency, time slot, and spreading factor. This scheme is applied to the specific physical layer proposed by LoRa, based on the chirp spread spectrum. Our proposal includes the power allocation scheme and the decoding technique that are very specific to this physical layer and show a significant performance improvement, increasing the number of devices that can be connected at least by ten.

Keywords: Chirp Spread Spectrum, Internet of Things, LoRa, Multiuser detector, Power allocation, Scalability

1. Introduction

Nowadays there is a rapid growth of the internet of things (IoT) network and more than 75 billion devices is expected to be connected to the network by 2025 [1]. Most of the IoT network requirements are related to operating in low power, low data rates, and wide-area connectivity [2, 3]. Low Power Wide Area Networks (LPWAN) technologies, such as LoRa, provide a solution to these requirements. However, the huge challenge is to face the scale change in the number of communicating devices.

In this paper we focus on the LoRa downlink. So far, this link is used to send few acknowledgments, not even necessarily for each packet. This link is very important to transmit not only feedback but also to transmit data to the devices and this will certainly require some update in the software as well as capability of the access point to transmit data to the devices. However, several limitations should be solved

first: the complexity at the receiver has to be limited and it has to respect the duty-cycle, which significantly limits the scalability of the network. Authors in [4] have shown that the duty-cycle limits not only the scalability but also the reliability of the network. In [5, 6] the downlink feedback frames are shown to highly impact the network performance. However, no solution to remedy this problem is proposed.

In this paper, we propose to simultaneously transmit multiple frames to multiple end-devices on a single channel (same frequency, same time slot, same spreading factor). As a consequence, we significantly enhance the scalability of the LoRa network. Our idea benefits from the chirp spread spectrum modulation and implements a joint multi-user detection. Multi-user access is, on a single channel, provided by the power domain NOMA (Non Orthogonal Multiple Access) scheme. Our contributions are

1) to propose a superposition transmission scheme for a chirp spread modulation in the downlink,
2) to design a complete multi-user receiving scheme, and
3) to propose a power allocation (PA) scheme to minimize the error probabilities at the receivers.

This paper is organized as follows: the description of LoRa technology is provided in section 2 and in section 3 the proposed scheme is presented. Section 4 analyze the simulation results and conclusions are presented in section 5.

2. Overview on LoRa

LoRa defines a physical Layer based on Chirp Spreading Spectrum (CSS) modulation. This modulation is defined by its spreading factor (SF), ranging from 7 to 12. It provides a trade-off between rate and communication range for a fixed Bandwidth (B) [7]. The symbol consists in a linear frequency change over the symbol duration T_s, where $T_s = 2^{\mathrm{SF}}/B$. The transmitted symbol of the ith user at time qT_s, $q \in \{0, Q-1\}$, with Q the number of symbols transmitted in a packet, is represented by $m_q^{(i)} \in \{0, 2^{\mathrm{SF}}-1\}$. The corresponding modulated chirp is obtained by left-shifting the raw chirp of $\delta_q^{(i)} = m_q^{(i)}/B$ in the time domain as illustrated in Fig. 1. The expression of the coded chirp of user i associated with the qth symbol is:

$$x_q^{(i)}(t) = \begin{cases} \exp\left(2j\pi\left(\frac{B}{2T_s}t^2 + \frac{m_q^{(i)}}{T_s}t\right)\right), & t \in \left[-\frac{T_s}{2}, \frac{T_s}{2} - \delta_q^{(i)}\right[, \\ \exp\left(2j\pi\left(\frac{B}{2T_s}t^2 + (\frac{m_q^{(i)}}{T_s} - B)t\right)\right), & t \in \left[\frac{T_s}{2} - \delta_q^{(i)}, \frac{T_s}{2}\right[. \end{cases} \quad (1)$$

LoRaWAN is an open standard which defines an adapted MAC protocol [8]. Three types of nodes (Class A, B and C) with different specifications are defined in [9]. Class A devices are the lowest energy consuming nodes and the cheapest.

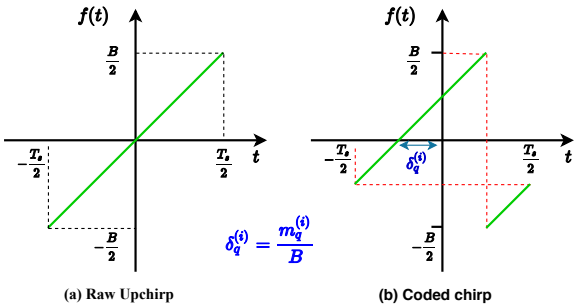


Fig. 1. (a) Raw up-chirp (b) Coded chirp associated with $m_q^{(i)}$.

Transmission is followed by two short downlink windows to receive a response from the gateway. Class B devices allow additional downlink traffic. They are synchronized using periodic beacons sent by the gateway. This is achieved at the expense of additional power consumption in the end nodes. Class C are always listening and can receive packets at any time. Their battery lifetime is significantly less than other devices. The noise level of a receiver at room temperature, when the noise figure NF = 6 dB and $B = 250$ kHz, is $174 + 10\log_{10}(B) + \text{NF} = -114$ dBm, where the first term is the thermal noise in 1 Hz of bandwidth and can only be affected by changing the temperature of the receiver.

LoRa operates in the license-free industrial, scientific and medical (ISM) radio band. In ISM bands, when no listen-before-talk is used, duty-cycle is generally the main restriction of the networks [10], for instance, 1% in EU 868 - 868.6 MHz. This duty-cycle significantly impacts the downlink and the gateway is extremely affected. When multiple uplink frames are received, the gateway cannot send downlink frames to all transmitters, which limits the capacity of downlink transmission and, as a consequence, impacts the scalability of LoRaWAN networks.

3. Proposed Multi-User Scheme

In this paper, to reduce the impact of the duty cycle we propose to transmit N frames simultaneously, with the same spreading factor and on the same frequency band. Class B devices can be used: they can be synchronized and all be in receive mode during the same time frame. The purpose is then to design a communication strategy that allows to superimpose N users in the duration of a single packet. The idea is to generate information streams for N end-devices, modulate using the CSS scheme then add signals in one packet by attributing different powers and add one common preamble at the start of the packet. The information about the number of

users, allocation power scheme and power ordering is added in the preamble. The receivers select and decode the signal which corresponds to their allocated power.

3.1. System Model. A single cell of radius R is considered with a gateway placed at the center. Large number of devices are uniformly distributed within the cell and the gateway has to send information to N of them. The distance from end-device i to the gateway is denoted by $d^{(i)}$. The propagation channel is considered block and flat fading, so a single coefficient, constant on the whole packet. We consider path loss and Rayleigh multi-path fading χ_i. The signal amplitude decays with increasing distance according to ${d^{(i)}}^{-\eta/2}$, where $\eta = 3.5$ is the path loss exponent. The channel attenuation (in amplitude) is expressed as $h^{(i)} = {d^{(i)}}^{-\eta/2} \cdot \chi^{(i)}$. In the following, we are interested in the decoding of symbol q and the user we are trying to decode is denote by j. Therefore, the samples of the received signal corresponding to symbol q of user j sampled at $t = nT$, where $T = 1/B$, $n \in [-\frac{M}{2}, \frac{M}{2} - 1]$, and $M = 2^{\mathrm{SF}}$ is:

$$r_q^{(j)}[n] = h^{(j)} \sum_{i=1}^{\mathrm{N}} \sqrt{p^{(i)}}\, x_q^{(i)}[n] + w_q^{(j)}[n] \tag{2}$$

where $h^{(j)}$ is the channel of user j, $p^{(i)}$ is the power allocated to user i, and $w[n] = w(n - qM) \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_n^2)$ is a complex Gaussian noise.

In order to apply the NOMA scheme the largest power is attributed to the user with the worst channel. The goal the proposed PA is to avoid colliding users give rise to a peak with an amplitude equal to another user or a combination of other colliding users. First, the users are ordered from $l = 1$ to N according to an estimate of the channels from a previous uplink from the strongest to the weakest. For user l we allocate the power:

$$p^{(l)} = \frac{2^{l-1}}{\sum_{l=1}^{\mathrm{N}} 2^{l-1}} p_t \tag{3}$$

where p_t is the total power transmitted by the access point. This guarantees that whatever collision occurs two peaks can not have the same amplitude at the receiver.

3.2. Receiver. The correlation between the received signal and the preamble is calculated first to detect the preamble and by taking the maximum value the user's channel is estimated and this allows us to compute the expected power of the user. To demodulate, the samples of the received signal $r_q^{(j)}[n]$ are multiplied by the conjugate complex form of the up-chirp, $x^*[n]$.The signal obtained after de-chirping which corresponds to the processing of qth symbol is written as:

$$y^{(j)}[n] = r_q^{(j)}[n] x^*[n] = h^{(j)} \sum_{i=1}^{N_u} \sqrt{p^{(i)}} e^{j2\pi \frac{m_q^{(i)}}{M} n} \mathbb{1}_{[-\frac{M}{2}, \frac{M}{2}-1]}(n) + w[n] \tag{4}$$

After the demodulation, Fast Fourier transform (FFT) is applied to the samples of the de-chirped signal $y^{(j)}[n]$ and the conjugate of the channel estimate. The expression is:

$$Y^{(j)}[k] = \mathrm{Re}\Big\{ \sum_{n=-\frac{M}{2}}^{\frac{M}{2}-1} \Big(h^{(j)*} y^{(j)}[n]\Big) e^{-2\imath\pi \frac{nk}{M}} \Big\} = |h^{(j)}|^2 \sum_{i=1}^{\mathrm{N}} \sqrt{p^{(i)}}\delta[k - m_q^{(i)}] + W^{(j)}[k] \quad (5)$$

where $W^{(j)}[k] \sim \mathcal{N}_{\mathbb{C}}(0, |h^{(j)*}|^2\sigma_n^2/2)$ is the FFT of the noise, $\delta[.]$ is Kronecker delta function, $\delta[n] = 1$ for $n = 0$ and $\delta[n] = 0$ for $n \neq 0$.

The classic idea of detection would be to search for the peak having the expected amplitude of the user of interest. However, when collision occurs this approach is not efficient. In this context, collision means two or more users have the same information and their peaks add. Therefore, we are rather interested in a multi-user detection scheme and for this reason first we drive the expression of maximum likelihood in frequency domain. Let us denote the received vector after FFT as $\mathbf{Y}^{(j)} = |h^{(j)}|^2\mathbf{X}^{(j)} + \mathbf{W}^{(j)}$, which counts M components given by (5) with $X^{(j)}[k] = \sum_{i=1}^{N} \sqrt{p^{(i)}}\delta[k - m_q^{(i)}]$. The optimal detector in the maximum likelihood sense maximizes the function

$$\begin{aligned}\Lambda &= \log \mathbb{P}\left(\mathbf{Y}^{(j)} | h^{(j)}, \mathbf{m_q}\right) = \sum_{k=0}^{M-1} \log \mathbb{P}\left(|h^{(j)}|^2 X^{(j)}[k] + W^{(j)}[k] \,\middle|\, h^{(j)}, \mathbf{m_q}\right) \\ &= M \log\left(\frac{1}{\sqrt{\pi |h^{(j)}|^2 \sigma_n^2}}\right) - \frac{\|\mathbf{Y}^{(j)} - |h^{(j)}|^2 \mathbf{X}^{(j)}\|^2}{|h^{(j)}|^2 \sigma_n^2} \end{aligned} \quad (6)$$

When $|h^{(j)}|^2 X^{(j)}[k] + W^{(j)}[k]$ is a Gaussian random variable with mean $|h^{(j)}|^2 X^{(j)}[k]$, variance σ_n^2, and $\mathbf{m_q} = \{m_q^{(1)}, \ldots, m_q^{(\mathrm{N})}\}$.

Maximizing the likelihood function Λ is then equivalent to minimizing the euclidean distance between the transmitted signal $\mathbf{X}^{(j)}$ and the received one $\mathbf{Z}^{(j)}$. The difficulty is that $m_q^{(i)}$ can take any value between 0 and $M - 1$. With N users it makes M^{N} possible combinations which rapidly becomes impossible to implement. Therefore, we propose a new multi-user approach based on peak detection and collision studies.

To reduce the complexity we will proceed in two steps: First, we detect the peaks larger than a given threshold. The goal is to get the peaks including the one from the expected users and the larger ones. For instance if the user is the lth user in terms of allocated power (users are ordered from the strongest to the weakest allocated power), we define a threshold that will allow to detect the l strongest peaks but not the weaker ones. Then, if exactly l peaks are detected, we choose peak l and its

position gives the information of the desired user. However, there is a case where we can miss the information of the desired user, that is when two or more users collided and results in a peak larger than or equal to the desired user's peak. This is a very rare case and can be addressed by the proposed power allocation scheme. If more than l peaks are detected, weaker peaks have probably collided and we choose the closed one from the expected received amplitude. Finally, if we detect less than l peaks, it means that a collision occurred between the l strongest users. In that case we analyze all the possible collisions cases to choose the most likely and make a decision. Recall that we have N users ordered from the strongest allocated power to the weakest. We are considering user l as the user we want to decode. We fix a threshold in order to detect the l strongest peaks but not the $\mathrm{N} - l$ weakest ones. The threshold is tuned by grid-search to $\gamma = h^{(l)}\big(\sqrt{p^{(l+1)}} + \sqrt{p^{(l)}}\big)/2$. If user l is the weakest user we take $p^{(l+1)} = 0$.

Decision Strategy: Let N_{pk} be the number of peaks above the threshold γ. The number of expected peaks is $N_{exp} = l$. The decision rule is the following: If $N_{pk} \geq N_{exp}$, we assume no collision between strong peaks and select the peak that has the closest value to the expected one. But in some rare cases there could be a small possibility to have a peak of collided users above the threshold and perhaps we can miss the desired peak. If $N_{pk} < N_{exp}$, it means a collision has occurred. Then we scan for all possible combinations between the N_{exp}. Let m be such a combination. We create a vector adding the amplitudes of the peaks that collide. We ordered the resulting values (including those that did not collide) and calculate the euclidean distance with the ordered detected peaks. Scanning all possible combinations, we minimize the euclidean distance in (6) to select the most probable one and deduce the estimated information of the desired user.

4. Simulation Results

A Monte Carlo simulations is used to evaluate the performance of the proposed scheme. The channel of the simulated environment is described in 3.1. In the following we used a maximum range $R = 10$ km. However, the channel attenuation has to be drawn in such way that the user can be able to connected to the network with the chosen SF, in other words if their received power is greater than the receiver sensitivity R_s, where, $R_s = -121.5, -124, -127, -129$ dBm for SF $= 7$ up to 10 respectively, $p_t = 14$ dBm, and $B = 250$ kHz. Users that do not respect this condition are discarded and drawn again. The detection of the packet and the channel estimation is performed using the preamble which is common to all user and transmitted with the full power so that it does not generate any errors and the channel estimation is accurate.

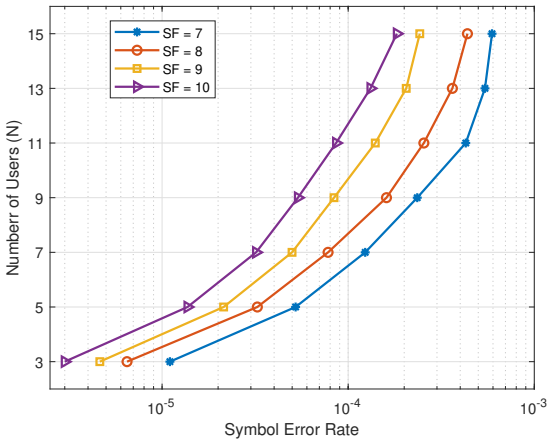


Fig. 2. SER for different N users, SF, and Noise levels, when $B = 250kHz$.

Fig. 2 presents the average symbol error rate (SER) for various number of users N and SF. The result shows that the proposed scheme improves the number of connected devices significantly. Normally the classical receiver can only support one user at time.

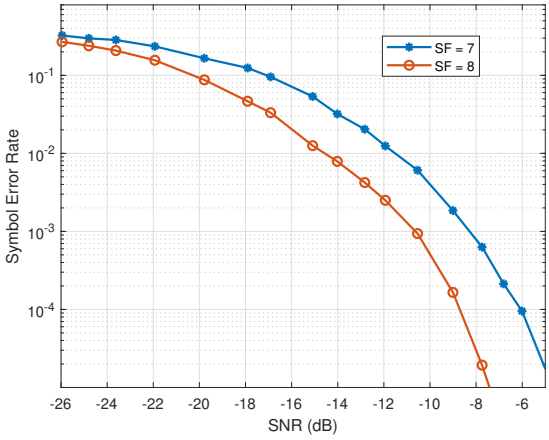


Fig. 3. SER vs SNR of one user in the presence of 9 other users, when SF = 7.

Fig. 3 shows the performance of decoding a single user with a different signal-to-noise ratio (SNR) when there are 9 other interfering users. The figure illustrates that the proposed receiver out-performers the classical receiver.

5. Conclusion

Massive device connectivity in IoT faces the scalability issue. In LoRa-like downlink transmission, the main limiting factor is related to the duty-cycle restriction imposed by the regulatory body. In this paper, we proposed a multi-user detection to improve the performance of the downlink in LoRa-like networks in terms of scalability. The proposed system includes a power allocation scheme and decoding algorithm which are very specific to this physical layer. The results show a significant performance improvement, increasing the number of devices that can be addressed at least by ten.

REFERENCES

1. L. Knud Lasse, State of the iot 2018: Number of iot devices, (Accessed on 05/22/2020) (Aug 2018).
   URL `https://iot-analytics.com/state-of-the-iot-update-q1-q2-2018/`
2. U. Raza, P. Kulkarni, M. Sooriyabandara, Low power wide area networks: An overview, IEEE Communications Surveys Tutorials 19 (2) (2017) 855–873.
3. A. Ikpehai, B. Adebisi, K. M. Rabie, K. Anoh, R. E. Ande, M. Hammoudeh, H. Gacanin, U. M. Mbanaso, Low-power wide area network technologies for internet-of-things: A comparative review, IEEE Internet of Things Journal 6 (2) (2019) 2225–2240.
4. F. Adelantado, X. Vilajosana, P. Tuset-Peiro, B. Martinez, J. Melia-Segui, T. Watteyne, Understanding the limits of lorawan, IEEE Communications Magazine 55 (9) (2017) 34–40. doi:10.1109/MCOM.2017.1600613.
5. M. Centenaro, L. Vangelista, R. Kohno, On the impact of downlink feedback on lora performance, in: 2017 IEEE 28th PIMRC, 2017, pp. 1–6.
6. K. Mikhaylov, J. Petäjäjärvi, A. Pouttu, Effect of downlink traffic on performance of lorawan lpwa networks: Empirical study, in: IEEE 29th PIMRC, 2018, pp. 1–6.
7. Semtech, AN1200.22: LoRa Modulation Basics, Tech. rep., Semtech (2015).
8. Patent, Low power long range transmitter (Jan 2014).
   URL `https://patents.google.com/patent/EP2763321A1`
9. T. M. W. 1.0, What is LoRaWAN, Tech. rep., LoRa Alliance (2015).
   URL `https://lora-alliance.org/resource-hub/what-lorawanr`
10. ETSI, Etsi en 300 220-1 v2.4.1: Electromagnetic compatibility and radio spectrum matters (erm); short range devices (srd); radio equipment to be used in the 25 mhz to 1 000 mhz frequency range with power levels ranging up to 500 mw; part 1: Technical characteristics and test methods, Tech. rep., European Telecommunications Standards Institute (ETSI) (2012).

UDC: 123.456

A Queueing Inventory System with Two Channels of Service

Nisha Mathew [1], V.C.Joshua [2], A. Krishnamoorthy[3]

[1]Department of Mathematics, B.K College Amalagiri, Kottayam, India
[2,3]Department of Mathematics, CMS College, Kottayam, India

nishamathew@cmscollege.ac.in,vcjoshua@cmscollege.ac.in,krishnamoorthy@cmscollege.ac.in

Abstract

We consider a queueing inventory model with positive service time and two service channels. Channel I is a single server facility and channel II is a bulk service facility. There are two types of customers, type-I and type-II. The same type of commodity is served to both types of customers. Channel I provides service to type-I customers and channel II provides service to type-II customers. Service is initiated only if inventory is available. Bulk service is initiated at the end of a random clock or by the accumulation of N type-II customers. The inventory replenishment follows the (s, S) policy with positive leadtime. The service time follows phase type distribution. Steady state analysis of the model is performed. Some performance measures are evaluated.

Keywords: queueing-inventory, lead time, positive service time.

1. Introduction

In most of the real life problems, it takes some time to serve the item to a customer. Such models are called queueing inventory models or inventory with positive service time. Inventory models with positive service time (queueing inventory model) were introduced by Sigman and Simchi-Levi in [1]. A survey of inventory with positive service time is given by Krishnamoorty et al in [2]. Chakravarthy et al in [3] studied a single server queueing model in which the customers are served in batches of varying size. Stochastic inventory system with two types of services that follows (s, S) replenishment policy is given by Anbazhagan et al in [4]. A single server retrial queueing model with two types of customers with service time distribution following phase type is studied by Krishnamoorty et al in [5]. Queueing inventory model with exponential lead time are found in [6], [7], [8].

The modern day retail business is driven by multiple platforms which are either offline or online. Long gone are the days where customers had to be physically

present for shopping. Sometimes customers give orders online through various virtual platforms and on a few ocassions go out for shopping. The sellers move into all the possibilities of customer interaction so that the sales are always boosted. This requires algorithms which could determine the stock pile required by the sellers well in advance, so that the customer demands are met without delay or failure.

This model is motivated by two types of demands that arrive at supermarkets: one is physically arriving customers and the other is online customers. The physically present customers are attended by the system on a FIFO basis. Online customer demands are attended only when the accumulated number of such demands reach a threshhold N or a random clock realizes, whichever occurs first.

2. Model Description

Our model includes one product which is been sold through two different platforms - a physical shop and an online platform. We consider a single server queueing inventory system with two types of customers, namely physical customers (type-I) and online customers (type-II). We assume that both the arrivals of type-I and type-II customers follow independent Poisson process. Let λ_1 be the rate of arrival of type-I customers and λ_2 be that of type-II customers. We also assume that the service time is positive. Type-I customers can form an infinite queue. Type-II customers are served in batches with maximum batch size N. A clock is also set, which starts ticking with the first arrival of type-II customer in every cycle. An (s, S) inventory policy is used. Service of a customer requires an inventory item. Each customer (both type-I and type-II) demands one unit of item, having a random duration of service time. Service time distribution of both type-I and type-II customers are assumed to be of phase type with irreducible representations $\mathrm{PH}(\alpha, T)$ with m_1 phases and $\mathrm{PH}(\beta, U)$ with m_2 phases respectively. The vectors T^0 and U^0 are given by $T^0 = -Te$ and $U^0 = -Ue$

We use the (s, S) replenishment policy here. Lead time is assumed to be exponential with parameter γ. If the server is idle, type-I customers enters into the service. A clock is set for type-II customers. Let θ be the rate of realization (parameter) of the exponental clock. The clock starts at the arrival of first type-II customer. The type-II customers are served only when the order becomes N or the clock time expires, whichever occurs first. They are served in bulk, provided two or more type-II customers have joined before the expiry/realization of the clock. It may happen that no such demand/exactly one demand arrived before clock expiry.

When the service in channel I begins, the inventory level drops by one unit. But when the service in channel II begins, the inventory level drops by n_2 where n_2 is the number of type-II customers present at that time. When the clock expires or when number of type-II customers reaches N, the clock is turned off and the service

of all these type-II customers is provided in a batch, provided the server was idle at that time. On the other hand, if the server is busy at that time, type-II customers are served immediately on completion of service of the current type-I customer. No type II customer is allowed to join the system once the clock expires/ N type II are in the system.

3. Mathematical Formulation

Let

- $N_1(t)$ be the number of type-I customers in the queue at time t
- $N_2(t)$ be the number of type-II customers in the finite buffer at time t
- $B(t)$ be the server status at time t;

$$B(t) = \begin{cases} 0, & \text{if the server is idle} \\ 1, & \text{if the server is busy with a type-I customer} \\ 2, & \text{if the server is busy with a type-II customer} \end{cases}$$

- $C(t)$ be the clock status at time t

$$C(t) = \begin{cases} 0, \text{if the clock is off} \\ 1, \text{if the clock is on} \end{cases}$$

- $J(t)$ be the phase of the service process at time t

Then $\{(N_1(t), N_2(t), B(t), C(t), J(t)); t \geq 0\}$ is a continuous time Markov chain on the state space to be described below. This model can be considered as a Level Independent Quasi-Birth-Death(LIQBD) process and a solution is obtained by Matrix Analytic Method. We define the state space of the QBD under consideration and analyze the structure of its infinitesimal generator.

The state space $\Omega = \Omega_1 \bigcup \Omega_2$ where $\Omega_1 = \{(n_1, n_2, 0, c, i)/n_1 \geq 0; 0 \leq n_2 \leq N; c = 0, 1; 0 \leq i \leq S\}$ and $\Omega_2 = \{(n_1, n_2, b, c, i, j)/n_1 \geq 0; 0 \leq n_2 \leq N; b = 1, 2; c = 0, 1; 0 \leq i \leq S; j = 1, 2, \ldots, m_b\}$.

The transitions are given in the table below.

From	T0	Rate	Description
$(0,0,0,0,i)$	$(0,0,1,0,i-1,j)$	$\lambda_1\alpha_j$	$i \geq 1, j = 1, 2, \cdots m_1$
$(0,n_2,0,1,i)$	$(0,n_2,1,1,i-1,j)$	$\lambda_1\alpha_j$	$i \geq 1, j = 1, 2, \cdots m_1,$ $1 \leq n_2 \leq N-1$

Table 1. Transistion table

From	T0	Rate	Description
$(n_1, n_2, 1, 1, i, j)$	$(n_1+1, n_2, 1, 1, i, j)$	λ_1	$n_1 \geq 0, j = 1, 2, \cdots m_1,$ $1 \leq n_2 \leq N-1$
$(n_1, 0, 1, 0, i, j)$	$(n_1+1, 0, 1, 0, i, j)$	λ_1	$n_1 \geq 0, j = 1, 2, \cdots m_1$
$(n_1, n_2, 1, 0, i, j)$	$(n_1+1, n_2, 1, 0, i, j)$	λ_1	$n_1 \geq 0, j = 1, 2, \cdots m_1$ $1 \leq n_2 \leq N$
$(n_1, 0, 1, 0, i, j)$	$(n_1, 1, 1, 1, i, j)$	λ_2	$n_1 \geq 0, j = 1, 2, \cdots m_1$
$(0, 0, 0, 0, i)$	$(0, 1, 0, 1, i)$	λ_2	$i \geq 0$
$(n_1, n_2, 1, 1, i, j)$	$(n_1, n_2+1, 1, 1, i, j)$	λ_2	$n_1 \geq 0, j = 1, 2, \cdots m_1$ $1 \leq n_2 \leq N-2$
$(0, n_2, 0, 1, i)$	$(0, n_2+1, 0, 1, i)$	λ_2	$1 \leq n_2 \leq N-2$
$(n_1, N-1, 1, 1, i, j)$	$(n_1, N, 1, 0, i, j)$	λ_2	$n_1 \geq 0, j = 1, 2, \cdots m_1$
$(n_1, 0, 2, 0, i, j)$	$(n_1, 1, 2, 1, i, j)$	λ_2	$n_1 \geq 0, j = 1, 2, \cdots m_2$
$(n_1, n_2, 2, 1, i, j)$	$(n_1, n_2+1, 2, 1, i, j)$	λ_2	$n_1 \geq 0,$ $1 \leq n_2 \leq N-2, j = 1, 2, \cdots m_2$
$(n_1, N-1, 2, 1, i, j)$	$(n_1, N, 2, 0, i, j)$	λ_2	$n_1 \geq 0,$ $j = 1, 2, \cdots m_2$
$(0, N-1, 0, 1, i)$	$(0, 0, 2, 0, i-N, j)$	$\lambda_2\beta_j$	$i \geq N, j = 1, 2, \cdots m_2$
$(n_1, n_2, 1, 1, i, j)$	$(n_1, n_2, 1, 0, i, j)$	θ	$n_1 \geq 0, j = 1, 2, \cdots m_1$ $1 \leq n_2 \leq N-1$
$(0, n_2, 0, 1, i)$	$(0, 0, 2, 0, i-n_2, j)$	$\theta\beta_j$	$i \geq n_2,$ $j = 1, 2, \cdots m_2$
(n_1, n_2, b, c, i, j)	(n_1, n_2, b, c, S, j)	γ	$n_1 \geq 0, n_2 \geq 0,$ $b = 0, 1, 2, c = 0, 1, 0 \leq i \leq s$
$(n_1, n_2, 0, 1, 0)$	$(n_1-1, n_2, 1, 1, S-1, j)$	$\gamma\alpha_j$	$n_1 \geq 1,$ $j = 1, 2, \cdots m_1$
$(n_1, n_2, 0, 0, i)$	$(n_1, 0, 2, 0, S-n_2, j)$	$\gamma\beta_j$	$n_1 \geq 0, 1 \leq n_2 \leq N,$ $i < n_2, j = 1, 2, \cdots m_2$
$(0, n_2, 1, 1, i, j)$	$(0, n_2, 0, 1, i)$	T_j^0	$n_2 \geq 1, j = 1, 2, \cdots m_1$
$(n_1, n_2, 1, 1, 0, j)$	$(n_1, n_2, 0, 1, 0)$	T_j^0	$n_2 \geq 1, j = 1, 2, \cdots m_1$
$(n_1, n_2, 1, 1, i, j)$	$(n_1-1, n_2, 1, 1, i-1, k)$	$T_j^0\alpha_k$	$n_1 \geq 1, n_2 \geq 1,$ $i \geq 1, j, k = 1, 2, \cdots m_1$
$(n_1, n_2, 1, 0, i, j)$	$(n_1, 0, 2, 0, i-n_2, k)$	$T_j^0\beta_k$	$i \geq n_2,$ $j = 1, 2, \cdots m_1, k = 1, 2, \cdots m_2$

Table 2. Transistion table

From	T0	Rate	Description
$(0,0,1,0,i,j)$	$(0,0,1,0,i,k)$	T_{jk}	$i \geq 0, j,k = 1,2,\cdots m_1$
$(n_1,n_2,1,1,i,j)$	$(n_1,n_2,1,1,i,k)$	T_{jk}	$n_1 \geq 0$, $j,k = 1,2,\cdots m_1$
$(0,0,2,0,i,j)$	$(0,0,0,0,i)$	U_j^0	$j = 1,2,\cdots m_2$
$(n_1,0,2,0,0,j)$	$(n_1,0,0,0,0)$	U_j^0	$j = 1,2,\cdots m_2$
$(n_1,n_2,2,1,0,j)$	$(n_1,n_2,0,1,0)$	U_j^0	$n_1 \geq 1, 1 \leq n_2 \leq N-1$, $j = 1,2,\cdots m_2$
$(n_1,N,2,0,i,j)$	$(n_1,N,0,0,i)$	U_j^0	$n_1 \geq 0, i \leq N-1$, $j = 1,2,\cdots m_1$
$(n_1,0,2,0,i,j)$	$(n_1-1,0,1,0,i-1,k)$	$U_j^0\alpha_k$	$n_1 \geq 1, i \geq 1$, $j = 1,2,\cdots m_2, k = 1,2,\cdots m_1$
$(n_1,n_2,2,1,i,j)$	$(n_1-1,n_2,1,1,i-1,k)$	$U_j^0\alpha_k$	$n_1 \geq 1, 1 \leq n_2 \leq N-1, i \geq 1$, $j = 1,2,\cdots m_2, k = 1,2,\cdots m_1$
$(n_1,N,2,0,i,j)$	$(n_1,0,2,0,i-N,k)$	$U_j^0\beta_k$	$n_1 \geq 0, i \geq N$, $j,k = 1,2,\cdots m_1$
$(n_1,0,2,0,i,j)$	$(n_1,0,2,0,i,k)$	U_{jk}	$n_1 \geq 0$, $j,k = 1,2,\cdots m_2$
$(n_1,n_2,2,1,i,j)$	$(n_1,n_2,2,1,i,k)$	U_{jk}	$n_1 \geq 0$, $j,k = 1,2,\cdots m_2$

Table 3. Transistion table

The infinitesimal generator Q of the LIQBD describing the above single server queueing inventory system is of the form

$$\begin{pmatrix} B_{00} & B_{01} & O & & \cdots & \cdots & \cdots & \\ B_{10} & A_1 & A_0 & O & \cdots & \cdots & \cdots & \\ O & A_2 & A_1 & A_0 & O & \cdots & \cdots & \\ O & O & A_2 & A_1 & A_0 & O & \cdots & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix} \tag{1}$$

where B_{00}, A_0, A_1, A_2 are all square matrices of appropriate order whose entries are block matrices. A_0 represents the arrival of a customer to the system; that is transition from level $n_1 \to n_1 + 1$. A_2 represents transition from level: $n_1 \to n_1 - 1$, A_1 describes all transitions in which the level does not change (transitions within levels). The structure of the matrices A_0, A_2 are as follows:

$A_0 = \lambda_1 I_K$

where $K = 2m_1(s+1)N + m_2(s+1)(N+1) + (N/2)(N+3)$

$$A_2 = \begin{pmatrix} H_1^0 & & & & \\ & H_1^1 & & & \\ & & \ddots & & \\ & & & H_1^{N-1} & \\ & & & & H_1^N \end{pmatrix}, \text{ where}$$

$$H_1^0 = \begin{pmatrix} O & Y & O \\ O & Z & O \end{pmatrix}, \ H_1^0 \text{ is a square matrices of order } 1 + (s+1)(m_1+m_2)$$

$$H_1^j = \begin{pmatrix} O & O & O \\ O & Y & O \\ O & O & O \\ O & Z & O \end{pmatrix} \text{ for } j = 1 \text{ to } N-1, \text{ and}$$

H_1^j are square matrices of order $(j+1) + (s+1)(2m_1 + m_2)$,
H_1^N is a zero square matrices of order $N + (s+1)(m_1+m_2)$,
$Y = \begin{pmatrix} O & \gamma\alpha & O \end{pmatrix}$, is a matrix of order 1 x $(s+1)(m_1)$,

$$Z = \begin{pmatrix} O & O \\ I_S \otimes T^0 \otimes \alpha & O \\ O & O \\ I_S \otimes U^0 \otimes \alpha & O \end{pmatrix} \text{is a matrix of order } (s+1)(m_1+m_2) \text{ x } (s+1)(m_1)$$

4. Steady-State Analysis

4.1. Stability Condition. The Markov chain with generator Q is positive recurrent if and only if

$$\lambda_1 < \sum_{j=0}^{N} \pi_j H_1^j \mathbf{e} \tag{2}$$

where the stationary vector $\boldsymbol{\pi}$ of A is obtained by solving

$$\boldsymbol{\pi} A = 0; \boldsymbol{\pi}\mathbf{e} = 1. \tag{3}$$

where the matrix A be defined as $A = A_0 + A_1 + A_2$.

4.2. Stationary Distribution. The stationary distribution of the Markov process under consideration is obtained by solving the set of equations

$$\mathbf{x}Q = \mathbf{0}; \mathbf{x}\mathbf{e} = \mathbf{1}. \tag{4}$$

Let $\mathbf{x}$ be decomposed in conformity with Q. Then
$\mathbf{x} = (\mathbf{x_0}, \mathbf{x_1}, \mathbf{x_2}, \dots)$ where $\mathbf{x_i} = (\mathbf{x_{i0}}, \mathbf{x_{i1}}, \dots\dots \mathbf{x_{iN}})$

$$\mathbf{x_{ij}} = (\mathbf{x_{ij0}}, \mathbf{x_{ij1}}, \mathbf{x_{ij2}})$$

for $j = 1, 2, \ldots, N$ whereas for $k = 0, 1, 2$,the vectors

$$\mathbf{x_{ijk}} = (\mathbf{x_{ijk0}}, \mathbf{x_{ijk1}})$$

$\mathbf{x_{ijkl}} = (\mathbf{x_{ijkl1}}, \mathbf{x_{ijkl1}}, \ldots\ldots \mathbf{x_{ijklS}})$ for $l = 0, 1$

$$\mathbf{x_{ijklr}} = (x_{ijklr1}, x_{ijklr2}, \ldots\ldots x_{ijklrt})$$

where $t = m_k$. x_{ijklru} is the probability of being in state (i, j, k, l, r, u) for $i \geq 0 : j = 1, 2, \ldots, N; k = 1, 2; l = 0, 1; 0 \leq r \leq S; u = 1, 2, \ldots, m_k$ and x_{ij0lr} is the probability of being in state $(i, j, 0, k, l, r)$. From $\mathbf{x}Q = 0$, we get the following equations:

$$\mathbf{x_0}B_{00} + \mathbf{x_1}B_{10} = 0 \tag{5}$$

$$\mathbf{x_0}B_{01} + \mathbf{x_1}A_1 + \mathbf{x_2}A_2 = 0 \tag{6}$$

$$\mathbf{x_1}A_0 + \mathbf{x_2}A_1 + \mathbf{x_3}A_2 = 0 \tag{7}$$

$$\mathbf{x_{i-1}}A_0 + \mathbf{x_i}A_1 + \mathbf{x_{i+1}}A_2 = 0, i = 2, 3, .. \tag{8}$$

It may be shown that there exists a constant matrix R such that

$$\mathbf{x_i} = \mathbf{x_{i-1}}R, i = 2, 3, ... \tag{9}$$

The sub vectors $\mathbf{x_i}$ are geometrically related by the equation

$$\mathbf{x_i} = \mathbf{x_1}R^{i-1}, i = 2, 3, ... \tag{10}$$

R can be obtained from the matrix quadratic equation

$$R^2A_2 + RA_1 + A_0 = O \tag{11}$$

5. Performance Measures

In this section we evaluate some performance measures of the system.

1) Expected number of type-I customers in the system

$$E[N_1] = \sum_{i=0}^{\infty} i\mathbf{x_i}\mathbf{e} \tag{12}$$

2) Expected number of type-II customers in the system

$$E[N_2] = \sum_{i=0}^{\infty}\sum_{j=0}^{N} j\mathbf{x_{ij}}\mathbf{e} \tag{13}$$

3) Expected number of items in the inventory

$$E[I] = \sum_{i=0}^{\infty}\sum_{j=0}^{N}\sum_{k=0}^{2}\sum_{l=0}^{1}\sum_{r=0}^{S} r\mathbf{x_{ijklr}e} \tag{14}$$

4) Expected number of customers waiting in the system due to lack of inventory

$$E[W] = \sum_{i=0}^{\infty}\sum_{j=0}^{N}\sum_{l=1}^{2} i\mathbf{x_{ij0l0}e} + \sum_{i=0}^{\infty}\sum_{j=0}^{N}\sum_{l=1}^{2} j\mathbf{x_{ij0l0}e} \tag{15}$$

5) Probability that the server is idle

$$b_0 = \sum_{i=0}^{\infty}\sum_{j=0}^{N}\mathbf{x_{ij0}e} \tag{16}$$

6) Probability that the server is busy with type-I customer

$$b_1 = \sum_{i=0}^{\infty}\sum_{j=0}^{N}\mathbf{x_{ij1}e} \tag{17}$$

7) Probability that the server is busy with type-II customer

$$b_2 = \sum_{i=0}^{\infty}\sum_{j=0}^{N}\mathbf{x_{ij2}e} \tag{18}$$

8) Probability that the clock is on

$$c_1 = \sum_{i=0}^{\infty}\sum_{j=0}^{N}\sum_{k=0}^{2}\mathbf{x_{ijk1}e} \tag{19}$$

9) Expected rate at which replenishment of inventory occurs

$$E_R = \sum_{i=0}^{\infty}\sum_{j=0}^{N}\sum_{k=0}^{2}\sum_{l=0}^{1}\sum_{r=0}^{s} r\mathbf{x_{ijklr}e} \tag{20}$$

10) Probability that the type-II customer is blocked from entering the service

$$p_b = \sum_{i=0}^{\infty}\sum_{j=1}^{N-1}\sum_{k=1}^{2}\sum_{t=0}^{S}\mathbf{x_{ijkot}e} + \sum_{\mathbf{i=0}}^{\infty}\mathbf{x_{iN}e} \tag{21}$$

6. Conclusion

In this paper , we considered a single server queueing inventory model with two channels of service. Service to both channels is provided by a single server. Various perfomance measures are evaluated at steady state conditions. We plan to analyse the problem for cost effectiveness.

REFERENCES

1. Sigman,K., Simchi-Levi, D. : Light Traffic Heutrestic for an M/G/1 Queue with Limited Inventory. Annals of Operation Research , 40 , 371-380 (1992)
2. Krishnamoorthy, A., Dhanya Shajin, Narayanan, V.C : Inventory with Positive Service Time: a Survey, Advanced Trends in Queueing Theory: Series of Books "Mathematics and Statistics", Sciences. ISTE & Wiley, London (2019). Opsearch 48(2), 153-169(2011)
3. Chakravarthy, S.R., Arunava Maity, Gupta,U.C. : An (s,S) Inventory in a Queueing system with batch service facility. Annals of Operation Research, 258, 263–283 (2017).
4. Anbazhagan,N., Vigneshwaran, B., Jeganathan, K. : Stochastic Inventory system with two types of services. International Journals of Advances in Applied Mathematics and Mechanics 2(1), 120-127 (2014)
5. Krishnamoorthy, A., Joshua,V.C., Ambily ,P.M. : A Retrial Queueing system with Abandonment and Search for Priority Customers, DCCN-2017, Springer,CCIS, 700, 98-107 (2017)
6. Krishnamoorthy, A., Binitha Benny, Dhanya Shajin.: A revisit to queueing-inventory system with reservation, cancellation and common life time. Opsearch (2016)
7. Dhanya Shajin, Lakshmy, B., Manikandan, R.: On A Two Stage Queueing-Inventory System With Rejection Of Customers. Neural, Parallel, and Scientific Computations 23 , 111-128 (2015)
8. Krishnamoorthy, A., Manikandan, R., Lakshmy, B.: A revisit to Queueing Inventory system with positive service time. Annals of Operation Research , 233(1) , 221-236 (2013)

UDC: 519.245

On wireless channel modeling with K distribution

S.G. Shorokhov

Peoples' Friendship University of Russia (RUDN University),
6, Miklukho-Maklaya St., Moscow, 117198, Russia

shorokhov-sg@rudn.ru

Abstract

We study modeling of wireless channel with fading and shadowing effects using K distribution with modified Bessel function of the second kind with half integer order. This allows us to obtain probability density function and cumulative distribution function in closed form in terms of elementary functions and simplifies the calculation of miscellaneous channel performance measures. The problem of Monte Carlo simulation using random variables with K distribution is also discussed.

Keywords: wireless channel, K distribution, Bessel function, Monte Carlo method

1. Introduction

K distribution was introduced in [1] for describing the statistics of radiation scattered by media with a wide range of length scales. K distribution can be derived from the product of two random variables, where one variable has a chi distribution and another variable has a complex Gaussian distribution [2]. K distribution is widely used for modeling diverse scattering phenomena such as tropospheric propagation of radio waves, various types of radar clutter, optical scintillation from the atmosphere [3], in synthetic-aperture radar (SAR) imagery [4], radiative heat transfer [5] and also in wireless communication to model composite fading and shadowing effects [6, 7].

One of the most important approaches to stochastic modeling is Monte Carlo simulation. Problems in communication theory form one of the most important domains of application for Monte-Carlo method [8].

We study the problem of wireless channel modeling with K distribution and Monte Carlo method, when modified Bessel function in probability density of K distribution is of half integer order.

The publication has been prepared with the support of the "RUDN University Program 5–100". The research was funded by RFBR, grant No. 19-08-00261.

2. Model of shadowed fading channel

Basically, the model for shadowed fading channel can be described by the following equation [7]

$$r = A\,B\,s + n, \tag{1}$$

where r is the received signal, s is the transmitted signal, n is the Gaussian distributed noise with zero mean, A and B represent the fluctuations in the channel due to fading and shadowing.

Short-term fading in wireless channel (1) can be described using various stochastic models, such as Rayleigh fading, Rician fading, Nakagami fading, etc. [7].

When the envelope of the signal is Rayleigh distributed, its power has an exponential probability density function (PDF), given by

$$f_F\left(p\right) = \frac{1}{p_0} \exp\left(-\frac{p}{p_0}\right) U\left(p\right), \tag{2}$$

where p_0 is the average power of the received signal, $U\left(p\right) = \mathbf{1}_{\{p>0\}} = \begin{cases} 1, & p > 0, \\ 0, & p \leqslant 0. \end{cases}$

In wireless communication average power often varies randomly due to the existence of shadowing by surrounding terrain, mountains, buildings, etc. The density function of the average power can be modeled in terms of lognormal or gamma distribution [9, 10, 11]. In model with gamma distribution, the PDF of the shadowing power is equal to

$$f_S\left(z\right) = \frac{z^{c-1}}{y_0^c\,\Gamma\left(c\right)} \exp\left(-\frac{z}{y_0}\right) U\left(z\right),\ c > 0. \tag{3}$$

Taking into account the simultaneous effect of fading and shadowing on the received signal, the PDF of the received signal power in (1) can be expressed as

$$f\left(p\right) = \int_0^\infty f_F\left(p \mid z\right) f_S\left(z\right) dz, \tag{4}$$

where f_F is the PDF of the power in a short-term fading channel, f_S is the PDF of the mean power.

Rayleigh-lognormal distribution [12], which is a mixture of Rayleigh and lognormal distributions, is probably the most appropriate description of signal envelope in fading-shadowing wireless channels [13]. But the complicated form of its PDF motivates to approximate lognormal distribution by gamma distribution and apply K distribution, which is a mixture of Rayleigh (2) and gamma (3) distributions.

3. K distribution and modified Bessel functions

Random variable has K distribution with shape $\alpha > 0$ and scale $\lambda > 0$, if its PDF and cumulative distribution function (CDF) are equal to

$$f_K(x) = \frac{2}{\lambda\Gamma(\alpha)} \left(\frac{x}{\lambda}\right)^{(\alpha-1)/2} K_{\alpha-1}\left(2\sqrt{\frac{x}{\lambda}}\right) U(x), \tag{5}$$

$$F_K(x) = 1 - \frac{2}{\Gamma(\alpha)} \left(\frac{x}{\lambda}\right)^{\alpha/2} K_{\alpha}\left(2\sqrt{\frac{x}{\lambda}}\right) U(x), \tag{6}$$

respectively, where Γ is the gamma function, $K_\nu(x)$ is the modified Bessel function of the second kind of order ν.

The modified Bessel function of the second kind $K_\nu(x)$ in (5) and (6) is a solution of the modified Bessel's ordinary differential equation and, basically, is impossible to express in terms of elementary functions [14].

But in the case of half integer order ν $\left(\nu = \pm\frac{1}{2}, \pm\frac{3}{2}, ...\right)$ functions $K_\nu(x)$ can be expressed through elementary functions, for example:

$$K_{-\frac{1}{2}}(x) = K_{\frac{1}{2}}(x) = \sqrt{\frac{\pi}{2\,x}} e^{-x},\ K_{\frac{3}{2}}(x) = \sqrt{\frac{\pi}{2\,x}} \left(\frac{1}{x} + 1\right) e^{-x},\ ... \tag{7}$$

Modified Bessel functions of the second kind of higher half integer orders keep the same structure and are represented in the following form [15]

$$K_{n+\frac{1}{2}}(x) = \sqrt{\frac{\pi}{2x}} e^{-x} \sum_{k=0}^{n} \frac{(n+k)!}{k!\,(n-k)!\,(2x)^k},\ n \geqslant 0. \tag{8}$$

For $x > 0$ and $\nu > 0$ function $K_\nu(x)$ is positive and monotonically decreasing.

4. K-fading channel with Bessel function of half integer order

If we set in (5) $b = \frac{2}{\sqrt{\lambda}}$ and $c = \alpha$, then the power PDF in shadowed fading channel (1) can be written in the form [7]

$$f_K(p) = \frac{2}{\Gamma(c)} \left(\frac{b}{2}\right)^{c+1} p^{\frac{c+1}{2}-1} K_{c-1}\left(b\sqrt{p}\right) U(p), \tag{9}$$

where b is a parameter related to the average power, c is a positive parameter related to the effective number of scatterers.

Rayleigh-lognormal and K distributions are similar, but K distribution has a simpler form, which makes it possible to obtain closed-form solutions in the calculation

of bit error rates, diversity effects, etc. The calculations can be further simplified in half integer case $c = n + \frac{1}{2}$, where $n \in \mathbb{N}$ or $n = 0$.

In half integer case PDF (9) takes the following form

$$f_K^{\left(n+\frac{1}{2}\right)}(p) = \frac{2}{\Gamma\left(n+\frac{1}{2}\right)} \left(\frac{b}{2}\right)^{n+\frac{3}{2}} p^{\frac{n}{2}-\frac{1}{4}} K_{n-\frac{1}{2}}\left(b\sqrt{p}\right) U(p). \tag{10}$$

When $n = 0$, we take into account (7) and equality $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ to receive that

$$f_K^{\left(\frac{1}{2}\right)}(p) = \frac{b}{2\sqrt{p}} e^{-b\sqrt{p}} U(p). \tag{11}$$

When $n \in \mathbb{N}$, gamma function for half integer argument is equal to

$$\Gamma\left(n+\frac{1}{2}\right) = \frac{(2n-1)!!}{2^n}\sqrt{\pi},$$

so from (8) we obtain that for $n \in \mathbb{N}$

$$K_{n-\frac{1}{2}}\left(b\sqrt{p}\right) = \sqrt{\frac{\pi}{2b\sqrt{p}}} e^{-b\sqrt{p}} \sum_{k=0}^{n-1} \frac{(n+k-1)!}{k!\,(n-k-1)!\left(2b\sqrt{p}\right)^k},$$

and PDF (10) can be expressed in terms of elementary functions as follows

$$f_K^{\left(n+\frac{1}{2}\right)}(p) = \frac{1}{(2n-1)!!} e^{-b\sqrt{p}} \sum_{k=0}^{n-1} \frac{(2n-k-2)!\,b^{k+2}}{(n-k-1)!\,k!\,2^{n-k}} p^{\frac{k}{2}} U(p). \tag{12}$$

In half integer case CDF of K distribution can also be essentially simplified. Using (8) we receive that for $n \geqslant 0$

$$F_K^{\left(n+\frac{1}{2}\right)}(p) = 1 - \frac{1}{(2n-1)!!} e^{-b\sqrt{p}} \sum_{k=0}^{n} \frac{(2n-k)!b^k}{k!\,(n-k)!2^{n-k}} p^{\frac{k}{2}} U(p). \tag{13}$$

Closed form expression for PDF and CDF for K-fading wireless channel allows to simplify calculation of miscellaneous performance criteria [16].

5. Monte Carlo simulation with K distribution

Monte Carlo method implies generation of random numbers with given probability distribution [17]. The most straightforward way to generate a non-uniform random variable is by inversion of its CDF: if given distribution is characterised by CDF

F with known closed form inverse function (quantile function) F^{-1} and X is a random variable with continuous uniform distribution in the interval (0, 1), then $Y = F^{-1}(X)$ is a random variable with given probability distribution.

For distributions with no closed-form inverse CDF one has to solve equation $F(x) = y$ for $y \in (0, 1)$ using approximations or numerical methods [18].

To find the quantile function of K distribution in general case, one has to solve transcendental equation

$$F_K(x) = 1 - \frac{2}{\Gamma(c)} \left(\frac{b}{2}\right)^c x^{\frac{c}{2}} K_c\left(b\sqrt{x}\right) = y,\ y \in (0, 1). \tag{14}$$

Equation (14) can be solved using various approaches [19].

The basic version of Newton's method for solving (14) gives the following iterative process:

$$x_{l+1} = x_l - \frac{F_K(x_l) - y}{\frac{d}{dx} F_K(x_l)} = x_l - \frac{F_K(x_l) - y}{f_K(x_l)} =$$

$$= x_l - \frac{(1-y)\,\Gamma(c) - 2\left(\frac{b}{2}\right)^c x_l^{\frac{c}{2}} K_c\left(b\sqrt{x_l}\right)}{2\left(\frac{b}{2}\right)^{c+1} x_l^{\frac{c-1}{2}} K_{c-1}\left(b\sqrt{x_l}\right)},\ l \geqslant 0. \tag{15}$$

The iterative process (15) stops when the required accuracy $\varepsilon > 0$ is achieved, i.e. $|F_K(x_l) - y| < \varepsilon$, or when absolute or relative error approximation is below the given tolerance $\epsilon > 0$, i.e. $|x_{l+1} - x_l| < \epsilon$ or $\left|\frac{x_{l+1}}{x_l} - 1\right| < \epsilon$.

The Newton's method of quantile function construction requires an initial starting point $x_0 = x_0(y)$ with good accuracy and a sufficiently efficient algorithm for modified Bessel function K_c and K_{c-1} calculation in (15).

Existence of approximations to the quantile function of K distribution has not been investigated, so the determination of an initial guess requires either application of general considerations [20], or approximation of K distribution by similar distribution with known quantile function. For instance, general K distribution can be approximated by Rayleigh or Weibull distribution, or K distribution with parameter $c = \frac{1}{2}$. In the latter case CDF is equal to $F_K(x) = 1 - be^{-b\sqrt{x}}$ and the quantile function is equal to $x = F_K^{-1}(y) = \left(\frac{1}{b}\ln\left(\frac{1}{b}(1-y)\right)\right)^2$.

Efficient calculation of modified Bessel functions of the second kind $K_\nu(x)$ can be performed in half integer case using formula (8).

Thus, in half integer case the problem of Monte Carlo simulation with K distribution can be easily implemented using CDF inversion approach and Newton's iterative method. Algorithm for Monte Carlo simulation with K distribution and subsequent construction of empirical density function from the simulated values is demonstrated below (algorithm 1).

Algorithm 1: Construction of empirical density with K distribution.

Data: number of trials N, number of bins M, order $c = n + \frac{1}{2}$, $n \in \mathbb{N}$, accuracy $\varepsilon > 0$ (or error tolerance $\epsilon > 0$);

Result: empirical PDF $\rho_K^{\left(n+\frac{1}{2}\right)}(x)$ of random variable with K distribution

for $i \leftarrow 1$ **to** N **do**

  create a pseudo-random number y_i, uniformely distributed on $(0, 1)$;

  `/* solve equation` $F_K^{\left(n+\frac{1}{2}\right)}(x_i) = y_i$ `for` $x_i \in (0, +\infty)$ `*/`

  $\tilde{x}_0 \leftarrow x_0(y_i)$, $l \leftarrow 0$; `/* an initial approximation */`

  repeat `/* Newton's method */`

    $l \leftarrow l + 1$;

    $\tilde{x}_l \leftarrow \tilde{x}_{l-1} - \dfrac{(2n-1)!!(1-y)e^{b\sqrt{\tilde{x}_{l-1}}} - \sum_{k=0}^{n} \frac{(2n-k)!\, b^k}{k!\,(n-k)!\, 2^{n-k}} \tilde{x}_{l-1}^{\frac{k}{2}}}{\sum_{k=0}^{n-1} \frac{(2n-k-2)!\, b^{k+2}}{(n-k-1)!\, k!\, 2^{n-k}} \tilde{x}_{l-1}^{\frac{k}{2}}}$;

  until $|F_K(\tilde{x}_l) - y| < \varepsilon$ $(|\tilde{x}_l - \tilde{x}_{l-1}| < \epsilon)$; `/* accuracy (error tolerance) */`

  $x_i \leftarrow \tilde{x}_l$;

end

$m \leftarrow \max\{x_i\}_{i=1}^{N}$;

divide the half-interval $(0,\, m]$ into M equal parts of length $h = \frac{m}{M}$ and calculate ρ_j as the number of values from the set $\{x_i\}_{i=1}^{N}$, that belong to half-intervals of the form $((j-1)\,h,\, j\,h]$ for $j = \overline{1,\, M}$;

$$\rho_K^{\left(n+\frac{1}{2}\right)}(x) \leftarrow \begin{cases} 0, & x \leqslant 0, \\ \frac{1}{N}\rho_j, & (j-1)\,h < x \leqslant j\,h,\ j = \overline{1,\, M}, \\ 0, & x > m \end{cases}$$

Algorithm 1 for Monte Carlo simulation with K distribution and construction of empirical density function is implemented in Python [21] and produces empirical density function in Fig. 1 similar to theoretical density function of K distribution.

Since samples in Monte Carlo method are generated independently of each other, Monte Carlo simulation is naturally suited to parallel computing. So Monte Carlo simulation with K distribution can be performed using modern multi-core processors or graphics processing units (GPU).

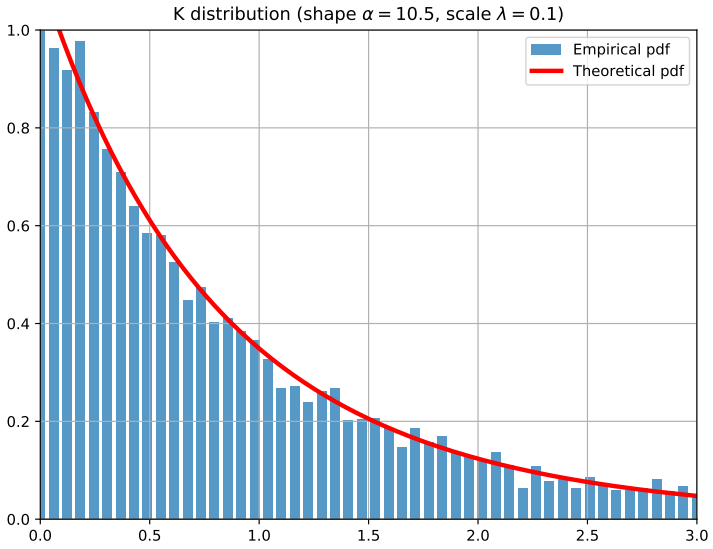


Fig. 1. Empirical and theoretical densities for K distribution

6. Conclusion

Investigation of wireless channel, modelled by K distribution with modified Bessel function of half integer order, gives an opportunity to receive the framework with closed form PDF and CDF expressions and simplifies further investigation and simulation of wireless channel performance.

REFERENCES

1. E. Jakeman, P. N. Pusey, Significance of K Distributions in scattering experiments, Physical Review Letters 40 (9) (1978) 546–550. doi:10.1103/physrevlett.40.546.
2. K. Ward, Compound representation of high resolution sea clutter, Electronics Letters 17 (16) (1981) 561. doi:10.1049/el:19810394.
3. D. A. Abraham, Underwater Acoustic Signal Processing, Springer International Publishing, 2019. doi:10.1007/978-3-319-92983-5.
4. K. Ward, R. Tough, S. Watts, Sea Clutter: Scattering, the K Distribution and Radar Performance, 2nd Edition, Institution of Engineering and Technology, 2013. doi:10.1049/pbra025e.

5. M. Modest, , 3rd Edition, Elsevier, 2013. doi:10.1016/c2010-0-65874-3. URL https://doi.org/10.1016%2Fc2010-0-65874-3
6. M. K. Simon, M.-S. Alouini, Digital Communication over Fading Channels, 2nd Edition, John Wiley & Sons, Inc., 2004. doi:10.1002/0471715220.
7. P. M. Shankar, Fading and Shadowing in Wireless Systems, 2nd Edition, Springer International Publishing, 2017. doi:10.1007/978-3-319-53198-4.
8. Y. Shreider (Ed.), The Monte Carlo Method: The Method of Statistical Trials, Elsevier, 1966. doi:10.1016/c2013-0-01870-1.
9. A. Abdi, M. Kaveh, K distribution: an appropriate substitute for rayleigh-lognormal distribution in fading-shadowing wireless channels, Electronics Letters 34 (9) (1998) 851–852. doi:10.1049/el:19980625.
10. A. Abdi, M. Kaveh, Comparison of DPSK and MSK bit error rates for k and rayleigh-lognormal fading distributions, IEEE Communications Letters 4 (4) (2000) 122–124. doi:10.1109/4234.841317.
11. P. M. Shankar, Error rates in generalized shadowed fading channels, Wireless Personal Communications 28 (3) (2004) 233–238. doi:10.1023/B:wire.0000032253.68423.86.
12. F. Hansen, F. Meno, Mobile fading—rayleigh and lognormal superimposed, IEEE Transactions on Vehicular Technology 26 (4) (1977) 332–335. doi:10.1109/t-vt.1977.23703.
13. G. L. Stüber, Principles of Mobile Communication, 4th Edition, Springer International Publishing, 2017. doi:10.1007/978-3-319-55615-4.
14. B. G. Korenev, Bessel Functions and Their Applications, CRC Press, 2002. doi:10.1201/b12551.
15. I. Gradshteyn, I. Ryzhik, Table of Integrals, Series, and Products, 7th Edition, Elsevier, 2007. doi:10.1016/c2010-0-64839-5.
16. A. Laourine, M. slim Alouini, S. Affes, A. Stephenne, On the capacity of generalized-k fading channels, IEEE Transactions on Wireless Communications 7 (7) (2008) 2441–2445. doi:10.1109/twc.2008.070103.
17. N. Metropolis, S. Ulam, The Monte Carlo method, Journal of the American Statistical Association 44 (247) (1949) 335–341. doi:10.1080/01621459.1949.10483310.
18. J. S. Dagpunar, Simulation and Monte Carlo, Wiley, 2007. doi:10.1002/9780470061336.
19. J. R. Hauser, Numerical Methods for Nonlinear Engineering Models, Springer Netherlands, 2009. doi:10.1007/978-1-4020-9920-5.
20. C. Yu, D. Zelterman, A general approximation to quantiles, Communications in Statistics - Theory and Methods 46 (19) (2016) 9834–9841. doi:10.1080/03610926.2016.1222433.
21. M. Charbit, Digital Signal Processing with Python Programming, John Wiley & Sons, Inc., 2016. doi:10.1002/9781119373063.

UDC: 519.872

Stationary waiting time distribution in the infinite-capacity two-queue single-server resequencing system with HOQ-LIFO-LIFO policy operating in random environment

R.V. Razumchik[1,2]

[1]Institute of Informatics Problems of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow 119333, Russian Federation

[2]Moscow Center for Fundamental and Applied Mathematics, Moscow State University, Moscow 119991, Russia

rrazumchik@ipiran.ru

Abstract

Consideration is given to the waiting time characteristics of the infinite-capacity single-server system with two queues (high and low priority) and negative customers, operating in entirely Markov-modulated environment. Ordinary customers arriving to the system occupy one place in the high priority queue and wait there for service. A negative customer arriving to the system moves one customer from the high to the low priority queue (if there is any) and itself leaves the system without having any further effect on it. It is known that such resequencing of customers during the service process results in hard to analyze delay distributions. When the processes are memoryless in case of HOQ-LIFO-LIFO policy (head of queue customer of the higher priority LIFO queue is moved to low-priority LIFO queue), the variance of the waiting time is invariant with respect to the arrival rate of the negative customers. Pursuing deeper understanding of this effect, in this short note the first step is made: it is shown that the previously developed methodology (based on Kronecker expansions) can be used to obtain the delay distributions in closed-form in terms of Laplace–Stieltjes transform even in the Markov-modulated environment.

Keywords: resequencing buffer, delay analysis, Markov-modulated environment

1. Introduction

In this short note we revisit the problem of the computation of the stationary delay distribution in the entirely Markov-modulated two-queue single-server system

The research was conducted in accordance with the program of Moscow Center for Fundamental and Applied Mathematics.

with negative customers ([3]) and one specific service policy. Negative customers, as can be seen from the system's description (Section 2), can be understood also as resequencing signals and thus the considered system can be called the resequencing queue*. Roughly speaking the system under consideration consists of one server serving customers one-by-one from two queues: high priority queue and low priority queue. The priority of those customers in the low priority queue is relative (i.e. they do not interrupt service of the high-priority queue). Each arriving customer firstly enters the high priority queue and waits there for service. If there is a negative arrival, one customer from the high priority queue is moved (resequenced) to the low priority one, wherefrom they eventually receive service. It is clear that in such a system various rules for scheduling high and low priority queues as well as for moving customers from one queue to the other can be chosen. Two variants of this system, specifically with HOQ-FIFO-FIFO and HOQ-FIFO-LIFO policies†, have already been analyzed in [9, 10]. Here an attempt is made to complement the obtained results with the analysis of the system behaviour under another policy: HOQ-LIFO-LIFO. The motivation behind this study is two-fold. The first is practical. The system described above was thoroughly studied in the memoryless case (see [7, 5, 6, 8] and [4, Chapter 3.4]). In [7] it was shown numerically that the variance of the waiting time of an arbitrary customer usually depends on the combination of the queue scheduling and resequencing order. The only exception (among those policies considered in [7]) is the HOQ-LIFO-LIFO policy for which the variance is invariant with respect to the resequencing rate. By differentiating directly the Laplace–Stieltjes transform of the stationary waiting time W distribution, given in [7, Section 9], it can be shown that the variance $Var(W)$ is equal to

$$Var(W) = \frac{\rho(2-\rho(1-\rho))}{\mu^2(1-\rho)^3},$$

where ρ denotes the system's load and μ — the service‡ rate. This fact did not get any explanation in [7] and still remains unexplained. Pursuing deeper understanding of such behaviour, the next step may be to check whether such invariance holds when the system operates in the Markov-modulated environment (for the definition

*Usually resequencing is due to some disruptive events but it also may be one of the features, which are inherent to the system (for models in the context of queueing theory see, for example, the reviews [1, 11]).

†This abbreviation means that the negative (resequencing) customer moves the customer in the head of the high priority queue to the low priority queue (Head Of Queue, HOQ), the service policy of the high priority queue is FIFO and of the low priority queue is either FIFO or LIFO

‡In [7] it is assumed that service times of customers from both queues are exponentially distributed with the same parameter μ.

of such an environment one can refer to the recent book [2]). Nothing regarding this question is reported below; instead the method to obtain delay distributions in terms of transforms is discussed. The expressions which follow from it can be used for further numerical studies.

The second purpose of this study is purely methodological: it is an open question whether the methodology developed in [9, 10] is general enough for applying in other queueing contexts. Although, due to the lack of space, complete derivations are not presented here, the answer to the question is positive: even in the entirely Markov-modulated environment the stationary waiting time distribution of an arbitrary customer under HOQ-LIFO-LIFO can be obtained in the closed-form, not involving any infinite summations.

2. Model description

Consider a single server queueing system with two infinite buffers: the regular buffer and the resequencing buffer. Two flow of customers arrive at the system: regular and negative (further — resequencing). Regular customers arrive at the system and occupy one place in the regular buffer. Upon arrival each resequencing customer moves one (if there is any) customer from the regular buffer to another buffer (further — resequencing buffer) of infinite capacity and itself leaves the system without having any effect on it. Customers are served one by one from each buffer by a single server. Upon service completion one customer from the regular buffer goes to the server and only if there are no regular customers in the buffer, one customer from the resequencing buffer enters the server. No service interruption is allowed. The HOQ-LIFO-LIFO policy is implemented in the system. It means that the resequencing customer moves the customer in the head of the queue to the resequencing buffer (Head Of Queue, HOQ), the service policy of the regular buffer is LIFO and of the resequencing buffer is LIFO as well.

We assume that regular customers arrive according to a MAP process with generator matrices $(\mathbf{A_0}, \mathbf{A_1})$ and resequencing signals arrive according to a MAP with $(\mathbf{H_0}, \mathbf{H_1})$. The service process is a MAP with $(\mathbf{S_0}, \mathbf{S_1})$. Denoting by $\otimes$ ($\oplus$) the Kronecker product (sum) and by I — the identity matrix of appropriate size, it can be seen that the matrix $\mathcal{A} = \mathbf{A_1} \otimes I \otimes I$ describes the arrival of a customer, the matrix $\mathcal{S} = I \otimes \mathbf{S_1} \otimes I$ — the service of a customer and the matrix $\mathcal{H} = I \otimes I \otimes \mathbf{H_1}$ — the resequencing of a customer and the matrix $\mathcal{L} = \mathbf{A_0} \oplus \mathbf{S_0} \oplus \mathbf{H_0}$ — the phase change when the resequencing is possible.

3. The joint stationary distribution

Usually before deriving the expressions for the waiting time characteristics one has to obtain the expressions for joint stationary distribution of the system's states.

But since the considered policy is work-conserving and the (size-oblivious) scheduling does not affect the number of customers in the system, there is no need in any calculations. The joint stationary distribution for the considered HOQ-LIFO-LIFO system is identical to the one of the HOQ-FIFO-FIFO system studied in [9] and HOQ-FIFO-LIFO system studied in [10]. Thus the joint stationary probabilities π_{ij} $(i, j \geq 0)$ of busy server, i customers in the regular buffer and j customers in the resequencing buffer are considered to be known. Their expressions (in closed form in terms of generating functions) and the necessary and sufficient condition for their existence are given in [9, Sections 3.3, 3.4]. From π_{ij} one can compute the stationary distribution seen by arrivals. Since the PASTA property does not hold for MAP arrivals, then the stationary probabilities $\tilde{\pi}_{ij}$ that after a regular customer arrival there are i $(i \geq 1)$ customer in the regular buffer and j $(j \geq 0)$ in the resequencing buffer, follow from the properties of MAP:

$$\tilde{\pi}_{ij} = \frac{1}{\lambda}\pi_{i-1,j}\mathcal{A},\ i \geq 1,\ j \geq 0,$$

where, as usual, λ denotes the average arrival rate.

4. Stationary waiting time distribution

The waiting time (W) is understood as the time lapse, starting from the instant when regular customer arrives to the system up to the instant when it enters server. In the considered setting it is possible to obtain its stationary distribution only in the terms of Laplace–Stieltjes transform further denoted by $\omega(s) = E(e^{-sW})$. Following the same arguments as in [7, 9, 10], we have

$$\begin{aligned}\omega(s) &= E(e^{-sW}) = \omega_{\mathrm{H}}(s) + \omega_{\mathrm{L}}(s)\\ &= E(e^{-sW} I_{\{\text{served from regular buffer}\}}) + E(e^{-sW} I_{\{\text{served from resequencing buffer}\}}),\end{aligned}$$

where $I_{\{a\}}$ is the indicator of the event a. But in spite of this decomposition, the components $\omega_{\mathrm{H}}(s)$ and $\omega_{\mathrm{L}}(s)$ differ from those obtained in [9, 10]. In the next two subsections we outline the procedures (limiting ourselves only to basic recurrent relations), which eventually allow closed-form expressions for both $\omega_{\mathrm{H}}(s)$ and $\omega_{\mathrm{L}}(s)$. Yet due to the lack of space, complete derivations a left aside and will appear elsewhere.

4.1. Stationary waiting time distribution of the customer that receives service from the regular buffer. For $i \geq 0$ and $k \geq 0$ let $\mathbb{W}_{\mathrm{H}}(t, i, k)$ be the matrix (according to the initial and final phases of the MAPs $(\mathbf{A_0}, \mathbf{A_1})$, $(\mathbf{S_0}, \mathbf{S_1})$ and $(\mathbf{H_0}, \mathbf{H_1})$) of the probabilities that the tagged customer on the $(i+1)^{st}$ position in the regular buffer will enter server in time t, if there are i customers in front and k

customers behind it. Using the first-step analysis for the Laplace–Stieltjes transform $\tilde{\mathbb{W}}_{\mathrm{H}}(s,i,k) = \int_{t=0}^{\infty} e^{-st}\mathbb{W}_{\mathrm{H}}(t,i,k)dt$ we have

$$\tilde{\mathbb{W}}_{\mathrm{H}}(s,i,k) = I_{\{k>0\}}\mathcal{L}(s)\mathcal{S}\tilde{\mathbb{W}}_{\mathrm{H}}(s,i,k-1) + I_{\{i>0\}}\mathcal{L}(s)\mathcal{H}\tilde{\mathbb{W}}_{\mathrm{H}}(s,i-1,k)+ \\ + \mathcal{L}(s)\mathcal{A}\tilde{\mathbb{W}}_{\mathrm{H}}(s,i,k+1) + I_{\{k=0\}}\mathcal{L}(s)\mathcal{S}, \quad (1)$$

where $\mathcal{L}(s) = (sI - \mathcal{L})^{-1}$. From (1) the waiting time of the customer, which enters server from the regular buffer, can be computed as

$$\omega_{\mathrm{H}}(s) = \sum_{i=1}^{\infty}\sum_{j=0}^{\infty}\tilde{\pi}_{ij}\tilde{\mathbb{W}}_{\mathrm{H}}(s,i-1,0)\vec{1}\,. \quad (2)$$

4.2. Stationary waiting time distribution of the customer that receives service from the resequencing buffer. For $i \geq 0$ and $k \geq 0$ let $\mathbb{F}(t,i,k)$ be the matrix (again according to the initial and final phases of the MAPs $(\mathbf{A_0}, \mathbf{A_1})$, $(\mathbf{S_0}, \mathbf{S_1})$ and $(\mathbf{H_0}, \mathbf{H_1})$) of the probabilities that the tagged customer on the $(i+1)^{st}$ position in the regular buffer will enter the resequencing buffer in time t, and at that instant there will be k customers in the regular buffer. Again from the first-step analysis for the Laplace–Stieltjes transform $\tilde{\mathbb{F}}(s,i,k) = \int_{t=0}^{\infty} e^{-st}\mathbb{F}(t,i,k)dt$ we get

$$\tilde{\mathbb{F}}(s,i,k) = I_{\{k>0\}}\mathcal{L}(s)\mathcal{S}\tilde{\mathbb{F}}(s,i,k-1) + I_{\{i>0\}}\mathcal{L}(s)\mathcal{H}\tilde{\mathbb{F}}(s,i-1,k)+ \\ + \mathcal{L}(s)\mathcal{A}\tilde{\mathbb{F}}(s,i,k+1) + I_{\{i=0\}}\mathcal{L}(s)\mathcal{H}. \quad (3)$$

Once the tagged customer enters the resequencing buffer, its remaining waiting time equals to the sum of the remaining service time of the customer in server, sojourn times of k customers remaining in the regular buffer and plus the sojourn times of all those customers, which can arrive during this time period (irrespectively whether they were resequenced or not). Let $\tilde{\mathbb{W}}_{\mathrm{L}}(s,k,j)$ be the matrix (according to the initial and final phases of the MAPs $(\mathbf{A_0}, \mathbf{A_1})$, $(\mathbf{S_0}, \mathbf{S_1})$ and $(\mathbf{H_0}, \mathbf{H_1})$) Laplace–Stieltjes transform of the waiting time of a customer which starts its life in the resequencing buffer in LIFO position j, when the number of customers in the regular buffer is k. For $k \geq 0, j \geq 1$, we have

$$\tilde{\mathbb{W}}_{\mathrm{L}}(s,k,j) = I_{\{k>0\}}\mathcal{L}(s)\mathcal{S}\tilde{\mathbb{W}}_{\mathrm{L}}(s,k-1,j) + I_{\{k=0\}}\mathcal{L}(s)\mathcal{S}\tilde{\mathbb{W}}_{\mathrm{L}}(s,0,j-1)+ \\ + I_{\{k>0\}}\mathcal{L}(s)\mathcal{H}\tilde{\mathbb{W}}_{\mathrm{L}}(s,k-1,j+1) + I_{\{k=0\}}\mathcal{L}(s)\mathcal{H}\tilde{\mathbb{W}}_{\mathrm{L}}(s,0,j)+ \\ + \mathcal{L}(s)\mathcal{A}\tilde{\mathbb{W}}_{\mathrm{L}}(s,k+1,j), \quad (4)$$

where $\tilde{\mathbb{W}}_{\mathrm{L}}(s,0,0) = I$. The solution of $\tilde{\mathbb{W}}_{\mathrm{L}}(s,i,j)$ is not trivial and, as shown in [10], can be found in product form $\tilde{\mathbb{W}}(s,k,j) = \widetilde{\mathbf{G}}(s)^k\widehat{\mathbf{G}}(s)^j$, where $\widetilde{\mathbf{G}}(s)$ and $\widehat{\mathbf{G}}(s)$ form a

pair of coupled matrix quadratic equations, whose minimal non-negative solution can be computed by a simple iterative procedure. Now, based on (4), the waiting time of the customer, which enters server from the resequencing buffer, can be computed as

$$\omega_{\mathrm{L}}(s) = \sum_{i=1}^{\infty}\sum_{j=0}^{\infty}\tilde{\pi}_{ij}\sum_{k=0}^{\infty}\tilde{\mathbb{F}}(s,i-1,k)\widetilde{\mathbf{G}}(s)^{k}\widehat{\mathbf{G}}(s)\vec{1}\,. \tag{5}$$

Both (2) and (5), already in such a form involving multiple infinite summations, can be used for direct numerical implementation (using truncation of the state space) and computation of the moments of the waiting time. But using Kronecker expansions and the methodology from [9], both (2) and (5) can be collapsed into closed-form expressions.

REFERENCES

1. Dimitrov B., Green D., Rykov V., Stanchev P. On performance evaluation and optimization problems in queues with resequencing // Advances in Stochastic Modelling. 2002. P. 55–72.
2. Dudin A.N., Klimenok V.I., Vishnevsky V.M. The theory of queuing systems with correlated flows. Heidelberg, Germany: Springer, 2019. 447 p.
3. Gelenbe E. R'eseaux stochastiques ouverts avec clients negatifs and positifs, et reseaux neuronaux // Comptes-Rendus de l'Academie des Sciences. 1989. V. 309. Serie II. P. 972–982.
4. Pechinkin, A. V., Razumchik R. V. Discrete Time Queuing Systems. Moscow: Fizmatlit. 432 p. ISBN 978-5-9221-1791-3 (in Russian)
5. Pechinkin A.V., Razumchik R.V. The stationary distribution of the waiting time in a queueing system with negative customers and a bunker for superseded customers in the case of the LAST-LIFO-LIFO discipline // Journal of Communications Technology and Electronics. 2012. V. 57. No. 12. P. 1331–1339.
6. Pechinkin A.V., Razumchik R.V. A method for calculating stationary queue distribution in a queuing system with flows of ordinary and negative claims and a bunker for superseded claims // Journal of Communications Technology and Electronics. 2012. V. 57. No. 8. P. 882–891.
7. Pechinkin A.V., Razumchik R.V. On temporal characteristics in an exponential queueing system with negative claims and a bunker for ousted claims // Automation and Remote Control. 2011. V. 72. No. 12. P. 2492–2504.
8. Pechinkin A., Razumchik R. Waiting characteristics of queueing system *Geo*/*Geo*/1 with negative claims and a bunker for superseded claims in discrete time // 2010 International Congress on Ultra Modern Telecommunications and Control Systems and Workshops. 2010. P. 1051–1055.

9. Razumchik R., Telek M. Delay analysis of a queue with re-sequencing buffer and Markov environment // Queueing Syst. 2016. V. 82. No. 1-2. P. 7–28.
10. Razumchik R., Telek M. Delay analysis of resequencing buffer in Markov environment with HOQ-FIFO-LIFO policy // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2017. No. 10497. P. 53–68.
11. Tien Van Do. Bibliography on G-networks, negative customers and applications // Mathematical and Computer Modelling. 2011. V. 53. No. 1. P. 205–212.

UDC: 004.738, 629.735

Concept of UFP based WBAN Data Acquisition Network

S. Vladimirov[1], V. Vishnevsky[2], A. Larionov[2], R. Kirichek[1]

[1]St.Petersburg State University of Telecommunications, 22 Bolshevikov Pr., St.Petersburg, 193232, Russian Federation

[2]V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, 65 Profsoyuznaya st., Moscow, 117997, Russian Federation

vladimirov.opds@gmail.com, larioandr@gmail.com, kirichek@sut.ru

Abstract

The paper defines a general concept of a system for collecting information from wireless body area networks based on unmanned flying platforms. Possible approaches to its implementation are considered. A list of problems to be solved and their features in the framework of building a network is given. Variants of optimal network radio technologies and topologies, antenna devices and types of unmanned flying platforms are proposed. Recommended options for implementing the interaction protocol and methods for organizing safe data transfer taking into account the characteristics of the problem being solved.

Keywords: WBAN, unmanned flying platform, data acquisition network

1. Introduction

Wireless body area networks (WBANs) are an important variety of sensor networks. They consist of wearable body sensor units (BSUs) that get sensor readings from a person and transmit them to the wearable body control unit (BCU), which processes the readings [1]. Usual BCU is a general-purpose user device, for example, a smartphone in a Wi-Fi or Bluetooth network, or a specialized microcontroller based device in the WPAN [1, 2].

The main application of WBAN is monitoring of human health indicators for various applications: medical institutions, professional and amateur sports, the army, and rescue services. In all these cases, there is a certain number of people – carriers of WBAN networks – in a limited area. The information from WBANs is required to be quickly transferred to a data collecting center (DCC). For example, in sporting events, timely information on the health of athletes received by referees can prevent accidents and allow doctors to quickly respond to possible injuries and precritical conditions [2, 3].

The reported study was funded by RFBR according to the research project No.20-37-70059.

2. Collecting data from WBAN

If the WBAN is geographically located in the coverage area of a larger radio access network, it makes sense to use this network to transmit live data. However, if the WBAN location is not covered by the existing radio access network, or the data cannot be transmitted over open networks, it is necessary to deploy a special wireless network – data acquisition network (DAN) – to collect data from the WBAN.

To ensure the best conditions for receiving signals, it is logical to place the acquisition node high above the ground. The height of the collection node depends on the maximum working range of the applied wireless technology, which determines the maximum distance from the acquisition node to the most remote WBAN from which information can be read.

In order not to depend on the presence of buildings or poles in the territory of DAN, it is convenient to create the acquisition node as an unmanned flying platform (UFP), as shown in Fig. 1. This allows to place the acquisition node at the optimal location and height and, if necessary, change its location during operation [4, 5].

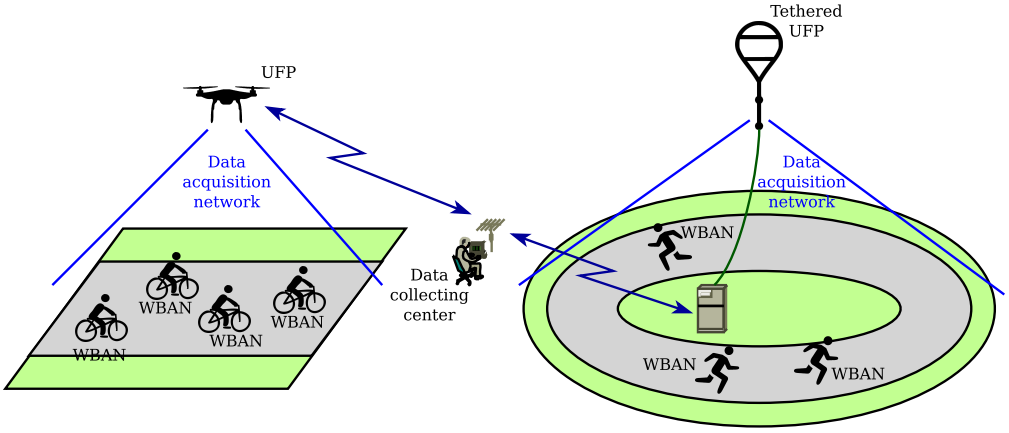


Fig. 1. Data acquisition network based on unmanned flying platform

Constructing parameters of UFP-based DAN and related issues are presented in Table 1. In general, their relationship can be represented as a function

$$[z_1, z_2, \ldots, z_n] = F(p_1, p_2, \ldots, p_m), \tag{1}$$

where z_i — issues to solve, p_j — network parameters.

3. Issues to solve

3.1. Network technology and topology. The choice of radio technology and the topology of the deployed DAN primarily depends on the size and topography

Parameter p_j	Issue z_i
Territory	Network technology and topology
Number of WBANs	Type of UFP
Distribution of WBANs	Antennas for WBAN and UFP
Moving o WBANs	Interconnecting protocol
Acquisition frequency	Security of data transmission

Table 1. Network parameters and issues to solve

of the provided territory. For example, a marathon, bicycle racing or paramilitary sports are usually held over a wide area, so the use of distributed DANs based on the widely used Wi-Fi technology will require additional deployment costs due to the limited communication range and significant energy consumption of the acquisition nodes. Using cellular communication technology based on a separate base station may be convenient, but it will require a deployment license. Thus, it is advisable to use medium and long range low power technologies operated in unlicensed frequency ranges [6, 7]. Since the idea of collecting information from the WBAN does not imply the exchange of a large amount of information, the requirements for the data transfer rate in DANs are not significant. Depending on the DAN coverage area, either a star-shaped network topology with one acquisition node or a distributed multi-node network should be used. The deployment height of the UFP with the acquisition node will depend on the maximum working distance of the radio link, which determines the maximum distance from the acquisition node to the most remote WBAN from which information must be read.

3.2. Type of UFP. By design, there are UFPs heavier than air – aircraft and multirotor platforms, and platforms that are lighter than air – balloon type [8]. Multirotor and balloon UFPs can hang for a while at a given point. Balloon UFPs are convenient for long-term placement of the acquisition node without the need for a quick change of position, while multirotor UFPs are more universal and allow moving the acquisition node at a fairly high speed, serving WBANs along a certain route. At the same time they have increased requirements on power sources.

In terms of placement height, two types of UFPs are usually distinguished: low-altitude platforms LAP, operating at heights of up to several kilometers, and high-altitude platforms HAP, operating in the altitude range from 17 to 21 km [8].

According to the method of communication with the ground, the platforms are divided into tethered UFP, in which communication is provided via a connecting cable [9, 10], and autonomous UFP, using the radio channel. Tethered systems traditionally have great standalone flight time, as they can be powered from a stationary power

source located on the ground, but they are not mobile and have a limited working height. Autonomous UFP with a radio channel allows the use of highly mobile acquisition nodes with a limited flight time.

To solve the problem under consideration, the use of multirotor and balloon LAPs seems to be optimal. The method of communication with the ground should be selected based on the characteristics of the specific situation.

3.3. Antennas for WBAN and UFP. WBAN antennas are limited by the linear dimensions of the wearable BCU, and their location relative to the acquisition node is not known in advance. Therefore it is advisable to use compact omnidirectional antennas, such as linear and helical monopole antennas and planar antennas [11, 12].

UFP antennas located above the WBAN (Fig. 1) have smaller size restrictions, so the best option would be to use efficient large aperture antennas with a directional pattern towards the earth, such as dipoles [13] and dual-rhombic loop antennas [14] of a flat design suitable for placement on the lower part of the UFP.

3.4. Interconnecting protocol. Collecting information from WBAN involves the data transmission in a previously known format determined by the type of sensors used. Thus, for organizing data acquisition, it seems optimal to use a specially developed data transfer protocol working on top of the data link layer protocol, which will provide a minimum amount of service information, in contrast to working over TCP/IP protocols.

Imagine the DAN as a redundant dynamic system consisting of WBAN, mobile UFPs and stationary nodes forming a WBAN–DCC network. Information is transmitted through a three-stage chain: WBAN – UFP – stationary node – DCC. At each stage, various data transmission technologies can be used. From the point of view of organizing the collection of information, the main stage is the WBAN–UFP interaction in the DAN area.

Since the number of WBANs served by one UFP at a gathering area and the number of UFPs simultaneously operating in one gathering area can vary, it is necessary to provide a mechanism for registering WBANs at UFPs. The entire UFP operating time is divided into two-window cycles, including the registration window (channel) W_{reg}, during which WBANs are registered at UFP, and the data transmission window W_{data}, in which the UFP polls registered WBANs. The W_{reg} window uses predefined transmission parameters that are the same for all WBAN and UFP in the network. In the W_{data} window, nodes transmit with parameters that are separate for each UFP. This allows several UFPs to operate simultaneously on the same territory without interfering with each other.

During W_{reg} WBANs register on the UFP using time contention with the CSMA/CA access method. The start of the registration window is specified by the UFP with the Registration Request (RReq) packet containing the UFP identifier

and data channel settings. To avoid registration collisions, WBANs that receive an RReq and wish to transmit information to the UFP send a Registration Reply (RRep) packet with a WBAN identifier, the size of the transmitted information in bytes and a data urgency indicator after a random period of time. After receiving the RRep, the UFP sends the next RReq with the identifier of the last registered WBAN. If a collision or error occurred while transmitting the RRep, the UFP repeats the previous RReq, confirming the registration error. In total, in the W_{reg} window, the UFP sends N_{req} RReq requests, and $N_{reg} \leq N_{req}$ WBAN can register. Upon successful registration, the WBAN switches the transmitters to the data channel settings. An example of the acquisition system operation in the registration window with $N_{req} = 8$ is shown in Fig. 2. Given three errors, $N_{reg} = 5$ WBANs were registered.

The duration of one W_{reg} window is equal to

$$T_{W_{reg}} = N_{req} \cdot (\frac{(L_{RReq} + L_{RRep})}{B} + t_{proc}). \quad (2)$$

where L_{RReq} – size of RReq in bits; L_{RRep} – size of RRep in bits; B – channel data rate in bps; t_{proc} – processing time of RReq by WBAN and RRep by UFP in seconds.

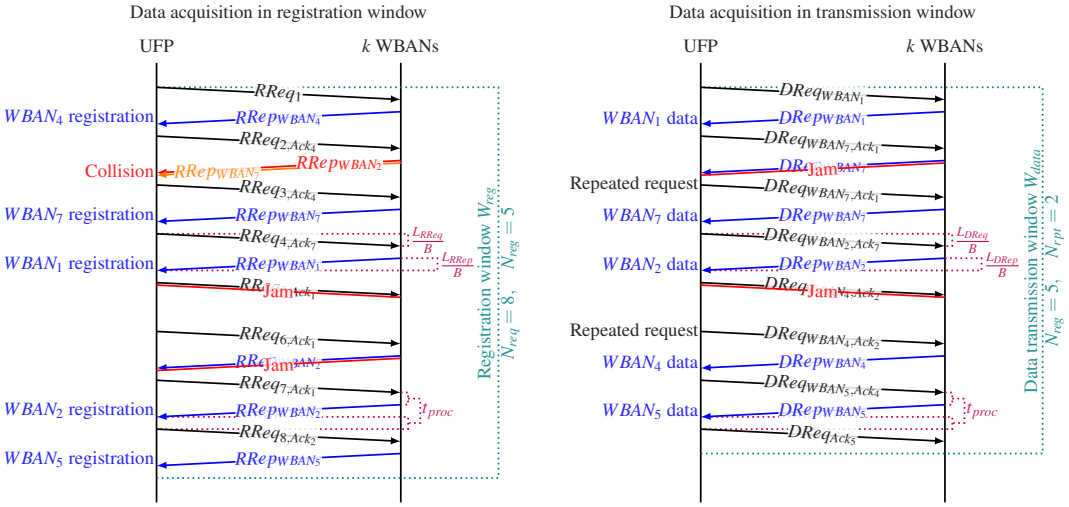


Fig. 2. An example of the acquisition system operation

After receiving all registration requests, UFP generates a N_{reg} length polling sequence according to the urgency indicators from the requests. During the next W_{data} window, UFP polls the WBAN by sending Data Request (DReq) packets indicating the identifier of the polled WBAN, for example, WBAN_1. WBAN_1, which received DReq with its identifier, responds with a Data Reply (DRep) packet. UFP

checks the DRep by comparing the identifiers and, if necessary, the checksum, and then responds with the next DReq packet, in which it confirms the reception of data and requests the next WBAN in the sequence, for example, WBAN_2. If the DRep packet failed or was lost, the UFP reports this by repeating the previous DReq. If after 2 retries it was not possible to receive data from WBAN_1, it is marked as unavailable and is no longer interrogated until the next registration request from it. Also, the DCC is notified of an unsuccessful attempt to receive data from WBAN_1. An example of data acquisition in transmission window with $N_{reg} = 5$ is shown in Fig. 2. We believe that the UFP has formed a WBAN polling sequence: 1, 7, 2, 4, 5. Due to two transmission errors, $N_{rpt} = 2$ retries were performed.

The duration of W_{data} window depends on retry number N_{rpt} and equal to

$$T_{W_{data}} = (N_{reg} + N_{rpt}) \cdot \left(\frac{(L_{DReq} + L_{DRep})}{B} + t_{proc} \right) + \left(\frac{L_{DReq}}{B} + 0.5 t_{proc} \right), \quad (3)$$

where L_{DReq} – size of DReq in bits; L_{DRep} – size of DRep in bits; t_{proc} – processing time of DReq by WBAN and DRep by UFP in seconds; the second summand takes account of the final data packet confirmation.

At the end of the data transmission window, UFP and WBAN switch to registration mode and the cycle repeats.

The proposed protocol packet format is shown in Fig. 3.

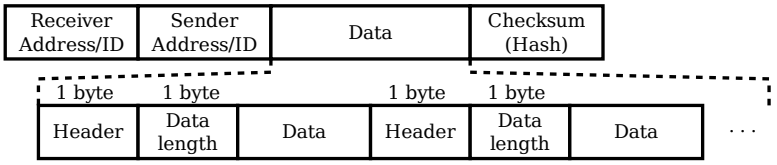


Fig. 3. Data acquisition network protocol packet format

At the beginning of the packet there are addresses/identifiers of receiver and sender. The content of the data field depends on the direction of transmission. The UFP sends data requests to the WBANs, while the WBAN responds with sensor readings. For greater versatility, the data field is divided into separate blocks consisting of three fields: header, data size in bytes, transmitted data. The header indicates the UFP command code or WBAN sensor number. The sizes of the header fields and lengths should be optimally set equal to 1 byte, as shown in Fig. 3, based on a limited number of possible values.

3.5. Security of data transmission. For the safe interaction of the DAN nodes, WBAN and the DCC, it is necessary to solve the problems of mutual identification of participants, encryption and verification of data integrity.

It is proposed to use various identification methods as device addresses/identifiers. To ensure maximum versatility and security, globally unique identifiers should be used: controller MAC addresses, digital object identifiers DOI [15], hardware identifiers based on degraded flash memory [16].

For encryption and checking the integrity of the transmitted data, it seems optimal to use a symmetric cipher for data exchange, asymmetric cipher for exchanging encryption keys and hash functions for checking integrity. The choice of specific algorithms depends on the DAN requirements and the equipment used.

4. Conclusion

The paper proposes a general concept of possible options for implementing a system for collecting information from WBANs based on unmanned flying platforms. The basic conceptual approaches to solving network design problems and their features are considered. It is planned to continue work within the framework of the considered tasks.

REFERENCES

1. Movassaghi S., Abolhasan M., Lipman J., Smith D., Jamalipour A. Wireless Body Area Networks: A Survey // IEEE Communications Surveys & Tutorials. 2014. V. 16. No. 3. P. 1658–1686.
2. Ghamari M., Janko B., Sherratt R. S., Harwin W., Piechockic R., Soltanpur C. A Survey on Wireless Body Area Networks for eHealthcare Systems in Residential Environments // Sensors 2016. V. 16. Art. 831.
3. Kirichek R., Pirmagomedov R., Glushakov R., Koucheryavy A. Live Substance in Cyberspace - Biodriver System // 18th International Conference on Advanced Communication Technology (ICACT). 2016. P. 274–278.
4. Ha I., Cho Y. Unmanned Aerial Vehicles-based Health Monitoring System for Prevention of Disaster in Activities of the Mountain // International Journal of Control and Automation 2016. V. 9. No. 9. P. 353–362.
5. Kirichek R. The model of data delivery from the wireless body area network to the cloud server with the use of unmanned aerial vehicles // Proceedings 30th European Conference on Modelling and Simulation (ECMS). 2016. P. 603–606.
6. Wu F., Wu T., Yuce M. R. An Internet-of-Things (IoT) Network System for Connected Safety and Health Monitoring Applications // Sensors. 2019. V. 19. Iss. 1. Art. 21.
7. Olatinwo D. D., Abu-Mahfouz A., Hancke G. A Survey on LPWAN Technologies in WBAN for Remote Health-Care Monitoring // Sensors. 2019. V. 19. Iss. 23. Art. 5268.

8. Alsamhi S. H., Rajput N. S. An Intelligent HAP for Broadband Wireless Communications: Developments, QoS and Applications // International Journal of Electronics and Electrical Engineering. 2015. V. 3. No. 2. P. 134–143.
9. 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. 2018. V. 919. P. 561–569.
10. Vishnevskiy V. M., Shirvanyan A. M., Tumchenok D. A. Mathematical Model of the Dynamics of Operation of the Tethered High-altitude Telecommunication Platform in the Turbulent Atmosphere // 2019 Systems of Signals Generating and Processing in the Field of on Board Communications. 2019. P. 8706784.
11. Wang J. C., Lim E. G., Leach M., Wang Z., Man K. L., Huang Y. Conformal Wearable Antennas for WBAN Applications // Proceedings of the International MultiConference of Engineers and Computer Scientists, IMECS 2016. 2016. V. II. P. 651–654.
12. Xue S., Yi Z., Xie L., Wan G., Ding T. A Displacement Sensor Based on a Normal Mode Helical Antenna // Sensors. 2019. V. 19. Iss. 17. Art. 3767.
13. Dw E. F., Pratama H., Ihsan N., Rahmatia S., Wulandari P. Design and Performance Investigation Of Dipole Antenna Using Aluminum and Iron At 644 MHz - 736 MHz // International Conference on Engineering, Technologies and Applied Sciences. Kuala Lumpur, Malaysia, January 2017.
14. Li R., Traille A., Laskar J., Tentzeris M. M. Bandwidth and Gain Improvement of a Circularly Polarized Dual-Rhombic Loop Antenna // IEEE Antennas and Wireless Propagation Letters. 2007. V. 5. P. 84–87.
15. Albahri M., Kirichek R., Muthanna A., Ateya A. A., Borodin A. Combating Counterfeit for IoT System Based on DOA // 2018 10th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). 2018. P. 8631257.
16. Vladimirov S., Pirmagomedov R., Kirichek R., Koucheryavy A. Unique degradation of flash memory as an identifier of ICT device // IEEE Access. 2019. V. 7. P. 107626–107634.

УДК: 004

Влияние технологий 5G на развитие цифровых экосистем умных городов: наукометрический и патентный анализ

Д.М. Кочетков[1], И.А. Кочеткова[2], Е.Д. Макеева[2]

[1]НИУ «Высшая школа экономики», Москва, Россия
[2]Российский университет дружбы народов (РУДН), Москва, Россия

Аннотация

Термин «умный город» в последнее время получил широкое распространение в академическом и политическом дискурсе. Тем не менее, по нашему мнению, это скорее маркетинговый термин, объединяющий ряд технологических (и не только) областей: Интернет вещей (IoT), дополненная и виртуальная реальность (AR / VR), сети связи. Сети последнего поколения необходимы для развития цифровых экосистем умных городов. Мы предположили, что умный город и сети 5G формируют развивающуюся технологическую область. Целью нашей работы является изучение структуры разработки и внедрения новых технологий для городской среды на примере технологий 5G. Для анализа новых технологий в выбранной предметной области нами было проведено исследование патентных ландшафтов и наукометрический анализ тематической области. Объектом наукометрического анализа является изучение закономерностей цитирования. Использование патентного ландшафта основано на информационных системах и базах данных патентной информации, разработанных патентными ведомствами и коммерческими компаниями, и состоит в визуализации логических связей между различными показателями патентной активности, с одной стороны, и технологическими и рыночными тенденциями, с другой. В совокупности наукометрический и патентный ландшафт показывают наиболее перспективные направления технологических исследований. Результаты исследования могут быть использованы в дальнейших теоретических и прикладных исследованиях, при формировании государственной политики в области исследований и разработок, а также при принятии решений в области управления городским хозяйством.

Ключевые слова: умный город, 5G, наукометрический анализ, патентный ландшафт

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта №18-00-01040 КОМФИ «Влияние новых технологий на городскую среду и качество жизни городских сообществ».

1. Введение

74 % населения Европы проживает в городах, в Северной и Южной Америке доля городского населения превышает 80 % [1]. Страны Азии и Африки пока отстают по доле городского населения, но в перспективе 80% всего городского населения в мире будет проживать в развивающихся странах [2]. Столь стремительные темпы урбанизации, с одной стороны, способствуют прогрессу, с другой стороны, представляют серьезную угрозу устойчивому развитию. Город как объект исследования привлекает внимание широкого спектра социальных наук, включая экономику, социологию и географию. Феномен постиндустриального умного города находится на пересечении социальных наук и инжиниринга. Несмотря на то, что термин уже достаточно давно находится в центре научного и политического дискурса, единого подхода к его определению нет. Анализ различных определений умного города выходит за рамки данного исследования, заинтересованного читателя мы отсылаем к анализу International Telecommunication Unit [3]. Мы опирались на одно из самых широко распространенных определений [4]: Умный устойчивый город – это инновационный город, который использует информационно-коммуникационные технологии (ИКТ) для повышения уровня жизни, эффективности деятельности и услуг в городах, а также конкурентоспособности при параллельном обеспечении удовлетворения потребностей настоящего и будущего поколений в отношении экономических, социальных, экологических и культурных аспектов. Изначально умный город характеризовался как повсеместно доступный, то есть пользователь может получить доступ к информации через Интернет, но что более важно, она доступна в любое время и в любом месте (на мобильном устройстве) [5]. Таким образом, беспроводные сети играют ключевую роль в развитии умного города. Сети пятого поколения (5G) предоставляют абсолютно новые возможности в этом отношении. Сетевая архитектура умных городов тесно связана с развитием гибких сетевых узлов на основе архитектуры программно-определяемой сети (SDN) и виртуализации сетевых функций (NFV) для оптимальной обработки функций узла и повышения эффективности работы сети [6]. Задача исследования состоит в определении исследовательской ниши, которая возникает на пересечении исследований умных городов и беспроводных сетей последнего поколения. Мы выявили ключевые исследовательские тренды в этой области на основе наукометрического анализа. Тем не менее, мы понимали, что для развивающихся областей характерна скорее повышенная патентная активность, чем большое количество научных публикаций. Поэтому анализ патентного ландшафта дополняет результаты наукометрического анализа.

2. Данные и методы

Мы использовали программное обеспечение *VOSviewer* для наукометрического анализа. Это компьютерная программа, которая позволяет создавать, визуализировать и анализировать сетевые карты, построенные на библиографических данных. *VOSviewer* также имеет функционал глубокого анализа текста, который можно использовать для построения и визуализации сетей ключевых слов, извлеченных из научной литературы [7]. *VOSviewer* уделяет особое внимание визуализации библиометрических карт. Функциональность VOSviewer особенно полезна для отображения больших библиометрических карт простым эффективным способом. Это особенно полезно для карт, содержащих достаточно большое количество элементов (как минимум 100 элементов) [8, 9]. *VOSviewer* по умолчанию применяет нормализацию силы связи [10]. Затем, программа использует технику картирования, подробно описанную в работе [11]. Наконец, *VOSviewer* распределяет узлы сети по кластерам; метод кластеризации представлен в статьк [12]. Для целей настоящего исследования мы в основном использовали функциональность интеллектуального анализа текста для построения сетей ключевых слов, извлеченных из метаданных. Для анализа мы использовали данные, полученные из БД *Scopus*[*]. Запрос, по ключевым словам, "*smart cities*", "*smart sustainable city*", "*SSC*", "*5g and fifth generation*" дал в результате 239 документов начиная с 2016 года. В свою очередь, тот же запрос в БД *Web of Science* выдал только 141 результатов, что кажется несколько недостаточным для полноценного сетевого анализа. Далее, мы ограничили запрос первичными документами "*article*", "*review*", "*conference paper*" (учитывая то, что тематика охватывает публикации по компьютерным наукам, материалы конференций имеют большое значение). В итоге у нас получился набор из 202 документов. Как бы то ни было, без анализа патентного ландшафта картина будет неполной. Мы нашли 331 международный патент по тематике за указанный период. Анализ патентного ландшафта проводился на основе данных патентной базы *Questel Orbit*[†] , которая объединяет более 100 различных баз. Это крупнейший в мире патентный фонд, который содержит свыше 60 миллионов документов из 95 стран и Международных Патентных ведомств. Также в базе доступна максимально полная информация о родственных патентах, включая их юридический статус.

3. Результаты

Массив для анализа содержит 1842 ключевые фразы. На первом этапе мы с помощью тезауруса объединили синонимы, а также установили порог в 10

[*] URL: https://www.scopus.com/ (дата обращения 12.07.2020г.)
[†] URL: https://www.orbit.com/ (дата обращения 12.07.2020г.)

появлений. В итоге мы получили семантическое ядро из 19 ключевых фраз. Результаты сетевого анализа представлены на рисунке 1. Термины распреде-

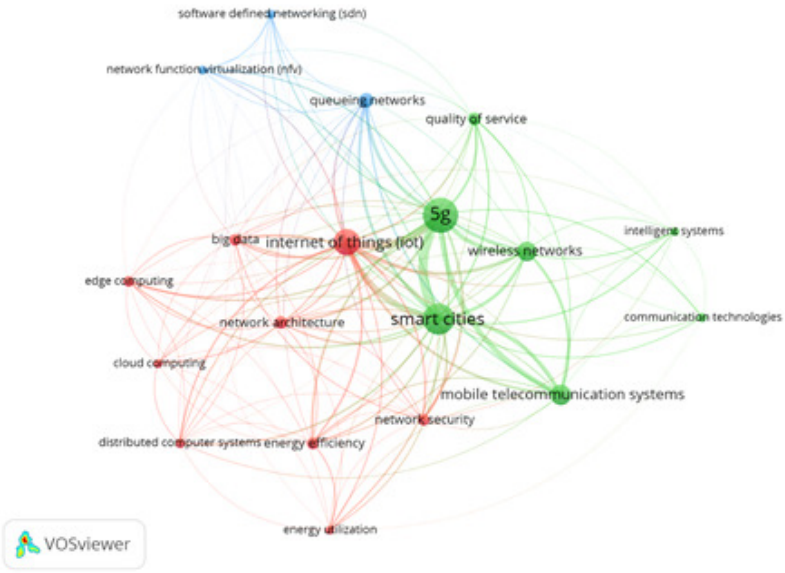


Рис. 1. Сеть ключевых слов. Разработано авторами с помощью программного обеспечения *VOSviewer*.

лились между 3 кластерами. Самый большой кластер сформировался вокруг интернета вещей и его практических применений в контексте умного города [13, 14, 15]. Это косвенно подтверждает нашу гипотезу, что интернет вещей в какой-то степени является «ядерным» термином или технологическим ядром для умного города. Сам термин «умный город» носит окраску политического и медийного дискурса. С точки зрения технологической перспективы, интернет вещей – это глобальная инфраструктура для информационного общества, которая обеспечивает возможность предоставления более сложных услуг путем соединения друг с другом (физических и виртуальных) вещей на основе существующих и развивающихся функционально совместимых ИКТ [16]. Сюда же относятся большие данные [17], которые также являются ключевой технологией для умного города. Именно необходимость постоянного доступа к данным и передачи больших объемов информации делает развитие беспроводных сетей пятого поколения насущным. Соответственно, будущие сценарии развития International Mobile Telecommunications (IMT) включают в себя [6]:

1) Сверхширокополосная мобильная связь (eMBB). Мобильная широкополосная связь охватывает сценарии использования, ориентированные на человека и обеспечивающие доступ к мультимедийному контенту, услугам и данным.
2) Сверхнадежная межмашинная связь с низкими задержками (URLLC). В данном сценарии использования предъявляются жесткие требования к таким показателям, как пропускная способность, задержка и готовность.
3) Массовая межмашинная связь (mMTC). Данный сценарий использования характеризуется большим количеством подключенных устройств, как правило, передающих относительно небольшой объем данных, не столь чувствительных к задержке. Необходимо обеспечить небольшую стоимость и продолжительное время заряда батареи.

Второй кластер (7 терминов) строится непосредственно вокруг умного города и 5G [18, 19]. Третий кластер совсем небольшой (3 термина) и стоит несколько особняком. Он включает в себя технические термины, которые непосредственно относятся к сетевой инфраструктуре умного города [20, 21, 22]. Собственно, он состоит из:

1) Программно-определяемая сеть (SDN), представляющая собой набор методов, которые позволяют пользователям напрямую программировать, контролировать и управлять сетевыми ресурсами, что облегчает динамическое и масштабируемое проектирование, доставку и эксплуатацию сетевых служб [23].
2) Виртуализация сетевых функций (NFV) - принцип отделения сетевых функций от оборудования, на котором они работают, с помощью абстракции виртуального оборудования. [24].
3) Сети массового обслуживания, то есть сети, основанные на математическом аппарате теории массового обслуживания [25].

Связующим звеном между вторым и третьим кластером является качество услуг (quality of service, QoS) [26]. Как уже говорилось выше, патентная активность в анализируемой области опережает количество публикаций. Мы сравнили количество патентов, публикаций и цитирований в динамике (рис. 2). Тем не менее, график показывает, что пик патентной активности был пройден в 2017г. Это один из признаков того, что технология переходит от фазы роста к фазе зрелости. Лидерами в области патентов в этой области являются Китай и США, также патентная активность наблюдается в Европе, Индии, Корее, Японии, Австралии, Канаде и Бразилии (рис. 3). Мы анализировали только международные патенты. Некоторые страны (например, Россия) регистрируют патенты только в национальных юрисдикциях, поэтому на карте они не представлены. Такой

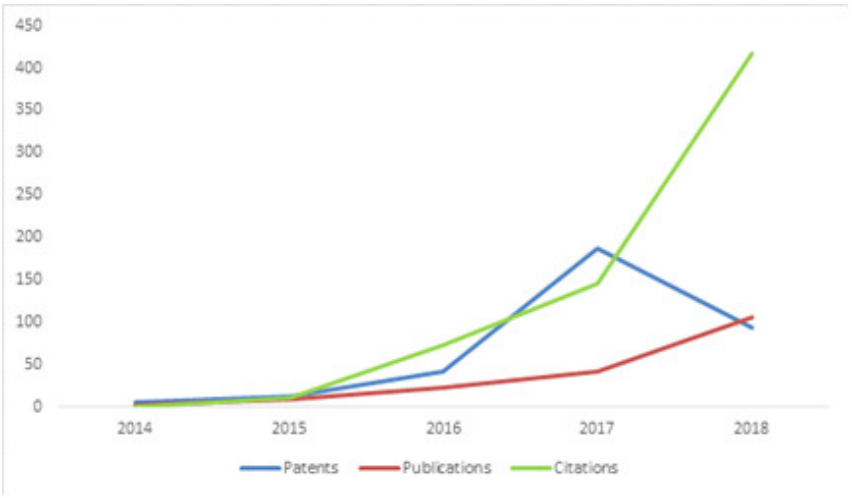


Рис. 2. Анализ ключевых наукометрических показателей. Разработано авторами на основе данных *Scopus* и *Orbit Intelligence*.

подход ведет к изначальному суживанию потенциального рынка для изобретений. Основными игроками на этом рынке являются компании *Samsung Electronics*

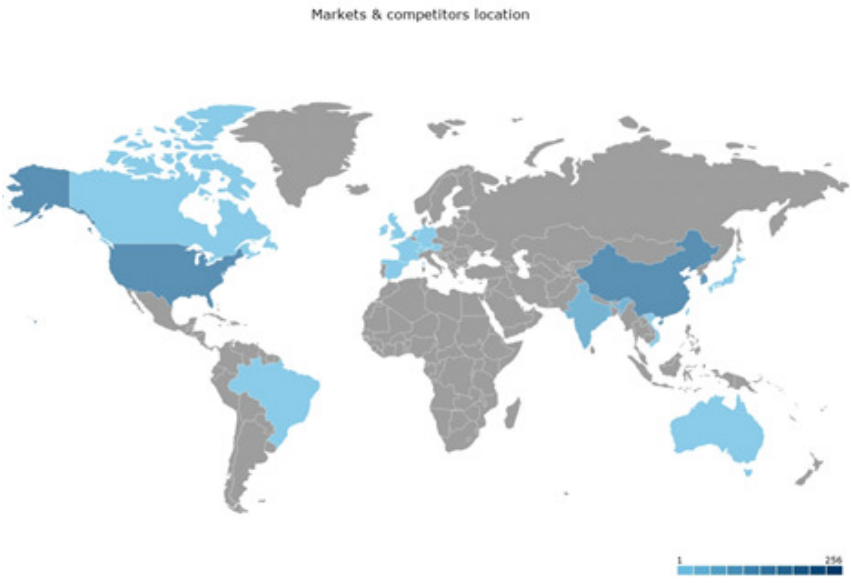


Рис. 3. Рынки и местоположение конкурентов. Получено из *Questel Intelligence*.

и *Huawei* (рис. 4a). Обе компании инвестировали в исследования и разработки существенно больше конкурентов (рис. 4b). Тем не менее, негативный фон, создаваемый вокруг *Huawei* в западных странах, может существенно помешать дальнейшему продвижению компании на европейском и североамериканском рынках. Что касается распределения инвестиций по годам, то график 4b подтверждает выводы из графика 3: пик инвестиций в исследования и разработки уже пройден.

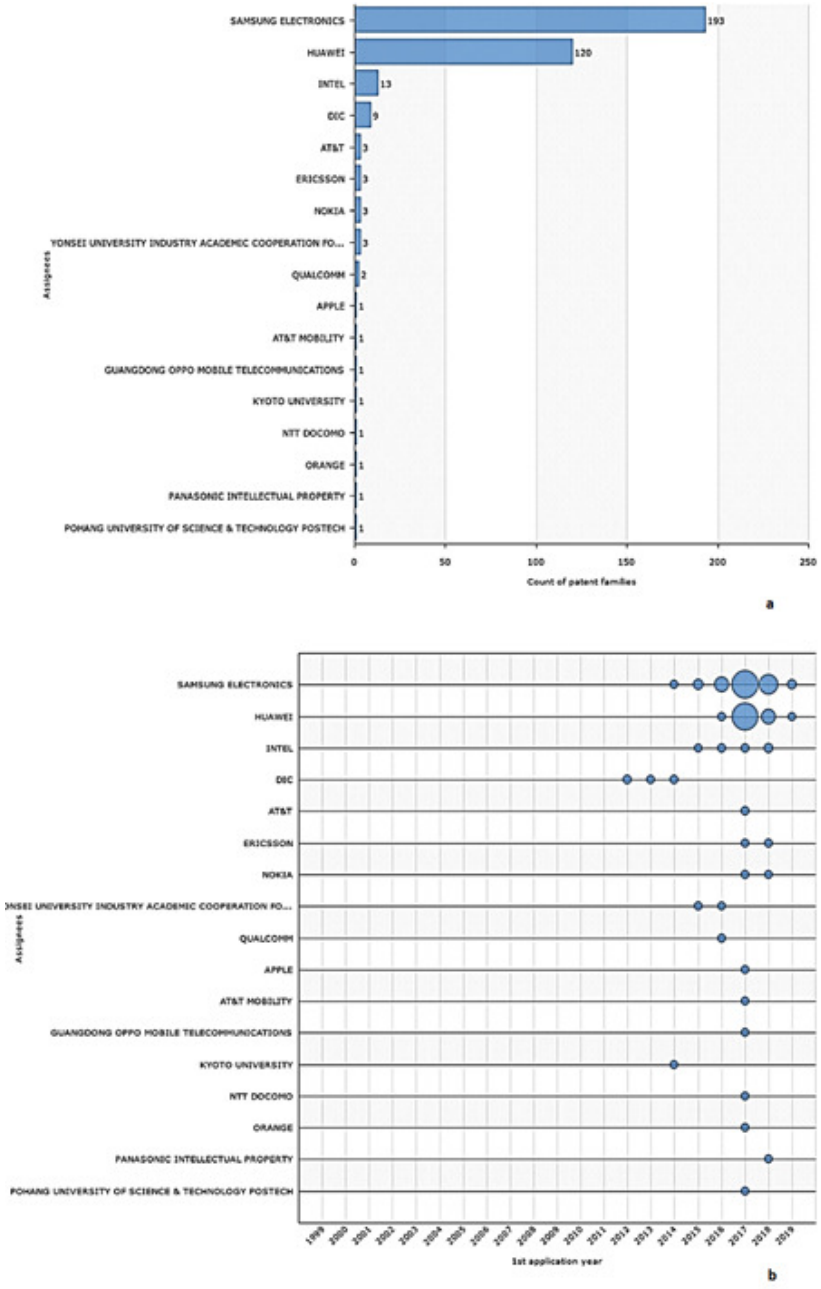


Рис. 4. а) Ключевые игроки; б) Инвестиционные тренды для ключевых игроков. Получено из *Questel Intelligence*.

4. Заключение

Умный город как объект исследования находится на пересечении инжиниринга, компьютерных и социальных наук. Определение умного города изначально связано с ИКТ инфраструктурой. Беспроводные телекоммуникации играют важную роль в развитии вертикально интегрированных городских индустрий. Мы использовали наукометрический анализ и анализ патентных ландшафтов для определения структуры этой технологической области и потенциала ее развития. Наукометрический анализ выявил три кластера ключевых фраз, строящихся вокруг терминов «интернет вещей», «умный город» и «5g», и сетевых технологий («SDN», «NFV», «сети массового обслуживания»). Пик патентной активности, как и инвестирования в RD, уже пройден, хотя патентов по-прежнему больше, чем научных публикаций. Это показывает на переход технологии от фазы роста к фазе зрелости (замедления роста), когда появляются первые признаки насыщения спроса. На рынке интеллектуальной собственности лидируют США и Китай, среди организаций это Samsung Electronics и Huawei. Тренд инвестирования подтверждает точность определения фазы жизненного цикла технологии. Безусловно, беспроводные сети всегда будут играть важную роль в контексте умных городов. Тем не менее, крупномасштабные инвестиции в исследования и разработки по 5G для умных городов представляются рискованными ввиду насыщенного конкурентного рынка.

ЛИТЕРАТУРА

1. World Urbanization Prospects: The 2018 Revision (ST/ESA/SER.A/420). New York, 2019. 103 p.
2. World Economic and Social Survey 2013 Sustainable Development Challenges. 2013. Vol. E/2013/50/. 181 p.
3. ITU Telecommunication Standardization Sector (ITU-T). ITU-T Y.4050-Y.4099 – Smart sustainable cities – an analysis of definitions // Series Y: Global Information Infrastructure, Internet Protocol Aspects and Next-Generation Networks, Internet of Things and Smart Cities. 2015. 69 p.
4. ITU Telecommunication Standardization Sector (ITU-T). Y.4900 Series – Key performance indicators definitions for smart sustainable cities // Series Y: Global Information Infrastructure, Internet Protocol Aspects and Next-Generation Networks, Internet of Things and Smart Cities. 2015. 98 p.
5. Ajit Jaokar. Big Data for Smart cities . // Smart Cities Industry Summit. London, 2012.
6. ITU Telecommunication Standardization Sector (ITU-T). Recommendation ITU-R M.2083-0 - IMT Vision – Framework and overall objectives of the future

development of IMT for 2020 and beyond // M Ser. Mobile, radiodetermination, Amat. Relat. Satell. Serv. 2015.
7. VOSviewer - Visualizing scientific landscapes [Electronic resource]. URL: https://www.vosviewer.com/ (accessed: 28.08.2019).
8. van Eck N.J., Waltman L. Visualizing Bibliometric Networks // Measuring Scholarly Impact. 2014. 285–320 p.
9. van Eck N.J., Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping // Scientometrics. 2010. Vol. 84, № 2. P. 523–538.
10. van Eck N.J., Waltman L. How to normalize cooccurrence data? An analysis of some well-known similarity measures // J. Am. Soc. Inf. Sci. Technol. 2009. Vol. 60, № 8. P. 1635–1651.
11. van Eck N.J. et al. A comparison of two techniques for bibliometric mapping: Multidimensional scaling and VOS // J. Am. Soc. Inf. Sci. Technol. 2010. Vol. 61, № 12. P. 2405–2416.
12. Waltman L., van Eck N.J., Noyons E.C.M. A unified approach to mapping and clustering of bibliometric networks // J. Informetr. 2010. Vol. 4, № 4. P. 629–635.
13. Albreem M.A.M. et al. Green internet of things (IoT): An overview // 4th IEEE International Conference on Smart Instrumentation, Measurement and Applications, ICSIMA 2017. Department of Electronics and Communication Engineering, ASharqiyah University, Ibra, Oman: Institute of Electrical and Electronics Engineers Inc., 2018. Vol. 2017-Novem. P. 1–6.
14. Satyakrishna J., Sagar R.K. Analysis of smart city transportation using IoT // 2nd International Conference on Inventive Systems and Control, ICISC 2018. Department of Computer Science and Engineering, AMITY University, Uttar Pradesh Noida, India: Institute of Electrical and Electronics Engineers Inc., 2018. P. 268–273.
15. Poncha L.J. et al. 5G in a convergent internet of things Era: An Overview // 2018 IEEE International Conference on Communications Workshops, ICC Workshops 2018. Dept. of Computer Engineering, Middle East Technical University, Northern Cyprus Campus, Mersin 10, Turkey: Institute of Electrical and Electronics Engineers Inc., 2018. P. 1–6.
16. ITU Telecommunication Standardization Sector (ITU-T). Recommendation ITU-T Y.2060: Overview of the Internet of things // Series Y: Global information infrastructure, internet protocol aspects and next-generation networks - Frameworks and functional architecture models. 2012. 22 p.
17. Shi W. et al. Edge Computing-An Emerging Computing Model for the Internet of Everything Era // Jisuanji Yanjiu yu Fazhan/Computer Res. Dev. Department of Computer Science, Wayne State University, Detroit, 48202, United States: Science Press, 2017. Vol. 54, № 5. P. 907–924.

18. Rao S.K., Prasad R. Impact of 5G Technologies on Smart City Implementation // Wirel. Pers. Commun. Tata Consultancy Services Ltd, SJM Towers, Sheshadri Road, Gandhinagar, Bangalore, 560009, India: Springer New York LLC, 2018. Vol. 100, № 1. P. 161–176.
19. Usman M. et al. Integrating smart city applications in 5G networks // 2nd International Conference on Future Networks and Distributed Systems, ICFNDS 2018. College of Science and Engineering, Hamad Bin Khalifa University (HBKU), Doha, Qatar: Association for Computing Machinery, 2018.
20. Khan M.A. et al. Mobility management approaches for SDN-enabled mobile networks // Ann. des Telecommun. Telecommun. DAI Labor, TU Berlin, Berlin, Germany: Springer-Verlag France, 2018. Vol. 73, № 11–12. P. 719–731.
21. Velasco L., Ruiz M. Flexible Fog Computing and Telecom Architecture for 5G Networks // 20th International Conference on Transparent Optical Networks, ICTON 2018. Universitat Politècnica de Catalunya (UPC), Barcelona, Spain: IEEE Computer Society, 2018. Vol. 2018-July.
22. Oproiu E.-M. et al. 5G Network Architecture, Functional Model and Business Role for 5G Smart City Use Case: Mobile Operator Perspective // 12th International Conference on Communications, COMM 2018. Technology Department, Orange Romania, Bucharest, Romania: Institute of Electrical and Electronics Engineers Inc., 2018. P. 361–366.
23. ITU Telecommunication Standardization Sector (ITU-T). Framework of software-defined networking // Series Y: Global Information Infrastructure, Internet Protocol Aspects and Next-Generation Networks Future. 2014. 40 p.
24. ETSI Industry Specification Group (ISG). Gs Nfv 003 - V1.4.1 - Network Functions Virtualisation (NFV); Terminology for Main Concepts in NFV. 2018.
25. Kleinrock L. Queueing Systems, Volume I. New York (NY): Wiley-Interscience, 1975. 448 p.
26. Karadimce A., Marina N. Smart Mobile City Services in the 5G Era // 10th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, ICUMT 2018. University of Information Science and Technology St. Paul the Apostle, Ohrid, Macedonia: IEEE Computer Society, 2019. Vol. 2018-Novem.

UDC: 004.7

Development and Investigation of model network IMT2020 with the use of MEC and Voice Assistant technologies

M. Makolkina[1,2], N. Shipota[1], A. Koucheryavy[1]

[1]The Bonch-Bruevich Saint-Petersburg State University of Telecommunications, St.-Petersburg, Russian Federation

[2]Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation

makolkina@list.ru, shypotanikoly@gmail.com, akouch@mail.ru

Abstract

Humanity is at a new stage in the development of information and computer technologies. The digitalization of most areas of activity creates a contradiction. On the one hand, to start the operation of the device, you must promptly enter information. Only the information entered by the user, allows the software agent to perform tasks. On the other hand, the technological diversity of modern devices requires new complex skills from the user. Consequently, the search for a solution to the problem of the quick introduction of information by a person without special and long-term preparation becomes relevant. The technology of voice input of information makes it possible to solve this problem. It underlies the work of virtual voice assistants, the use of which is constantly growing. The paper developed a simulation model using voice assistants (VA) based on software-defined (SDN) networks and mobile edge computing (MEC) technology. The analysis of the scope of voice assistants. An experimental study is given of the influence of the dependence of the execution time of processes and the total delay on the method of processing speech information of devices. The dependence of the total delay on the type of traffic and computing device for the final processing of packets.

Keywords: voice assistant, mobile edge computing, software-defined network, model network, IMT-2020

1. Introduction

Voice assistants have become a necessary part of the functioning of various online technologies: in distance learning, in online consulting in the field of medicine, in

geolocation and GPS navigation, in robotics, in the field of business management as virtual assistants [1–5]. Voice Assistant is now one of the essential components of a software-defined network management system [6].

Voice assistants are also actively used when working with clients to advise and advertise products. They have an impact on the economic results of companies and reduce their costs. This technology has become necessary when interacting with devices operating on the "smart home" platform [7].

A large number of diverse VA applications require a certain organization of the communication network structure. Particular requirements are placed on providing performance indicators for the quality of experience of VA services in various areas of human life. Today, SDN and MEC technologies are most suitable for VA services.

SDN technology is one of the 5G/IMT-2020 core technologies [8]. Its use provides: programmability and flexibility, adaptability of network management, increased reliability, etc. SDN has such advantages as: the ability to dynamically configure traffic flows throughout the network in accordance with changing needs; the administrator can configure, manage, provide protection, quickly optimize network resources using SDN programs, independently develop these algorithms; SDN is based on open standards; network management is provided not by devices and protocols of certain manufacturers, but by software SDN controllers. SDN greatly simplifies network design and operation.

MEC technology also has several advantages: it reduces the time interval of circular delay inside the system; increase network bandwidth; realizing the potential of introducing new applications and services based on network structure data; reduced load on the core network. Among the important characteristics are also called the prompt provision of information about the state of the network. This can be used to provide services, focusing attention on information about the network structure in real time.

SDN and MEC are complementary technologies with a common goal: the application of specific control principles to the data plane.

This article developed a model network based on SDN and MEC technologies using VA. Developed various options for organizing the structure 5G/IMT-2020 networks using VA. A comparative analysis of three options for the implementation of the network structure according to the results of an experimental study. An investigation of the work of VA on this model network was carried out, with a different number of simultaneously processed packets and the effect of the number of packets on the delivery delay.

2. Goal of investigation

The goal of this investigation is to study the technology of voice assistant based on the SDN network using MEC. Including the study of the dependence of the execution time of processes and the total delay in the delivery of data from the method of processing voice information.

3. Related works

Currently, the future of networks is the transition to the fifth generation 5G / IMT-2020 communication network. Today, a number of works is known dedicated to the study of architecture, requirements for construction, technology and organization of such networks [9,10]. These networks will be able to provide a seamless connection between various devices and numerous applications as shown in work [11]. In article [12], the authors develop and explore approaches to centralized management of IoT devices when implementing the concept of a smart home in fifth-generation communication networks. A number of researchers are working on the study of the network structure and methods of distributing information over the network in order to reduce the load on the network core and thereby increase the quality of service indicators for applications that are especially demanding for delays [13, 14]. The advantages of SDN technology and the possibility of its application for implementing applications of the Internet of Things, augmented reality, Tactile Internet, etc. made it main in the implementation of modern services. Many researchers create model networks on the basis of which they study the features of using this technology for various applications and together with other well-proven technologies, for example, MEC [15–18].

4. Development of a simulation model using Voice assistant

The types of voice assistants are quite difficult to systematize, due to the fact that even those created to provide assistance in a certain type of activity, designed to perform unique tasks, they can support conversation, report news, read weather information and perform other similar functions, characterized as related not the main ones. At the same time, in our opinion, two large groups of assistants can be distinguished. These are voice assistants with wide functionality and those developed for work in narrow-profile areas.

Voice assistants with a wide range of functions: Amazon Alexa, Microsoft Cortana, Yandex Alice, Google Assistant. Amazon Alexa has extensive functionality: sets an alarm and opens the blinds, turns on the light and sets the air conditioning, reads books and conducts workouts, draws up a schedule and reminds you of visits, calls, necessary purchases, selects and voices news information.

Assistants with limited functionality are trained to assist in activities in a specific area: in business, in banking, in household management, in gps navigation and geolocation, in services for recognizing music and songs.

By systematizing voice assistants according to their functional tasks, we can distinguish services designed for music recognition. As an example, let's take the following: Shazam, SoundHound, MusiXmatch, Midomi, TrackID, BeatFind, MusicID, Spotify, AudioTag, Audiggle.

As you can see, voice assistants have firmly entered various spheres of people's lives and require special attention when organizing a communication network for the timely delivery of information to the user.

In this paper, we will examine the interaction of Voice Assistant and the MEC system. The algorithm of the voice assistant will be as follows. Initially, through a microphone, speech is delivered to a portable device. On this device, the processes of obtaining a voice signal, its filtering and digitization are performed. The presented structure includes a mobile device with a built-in microphone, as well as a voice assistant responsible for determining the actions to be performed on the commands received from the user. The processes for receiving a voice signal, its filtering and digitization in all three versions were carried out on a mobile device.

Three options were developed for organizing the network structure.

The first version of the algorithm for organizing the network structure is VA located on a remote server. The user's mobile device does not extract keywords from the received signal. The steps of highlighting keywords and comparing them with the commands that are performed are performed on the remote server. The second option, the selection of keywords is carried out on a mobile device, and their comparison with the executed commands is performed on a remote server with a voice assistant. In the third option, VA is located on the mobile device and, therefore, all processes are performed on it. Assume that only the remote server can execute the command requested by the user. This means that in the third version of the algorithm, despite the fact that all processes are performed on a mobile device, the need to transfer service data containing information about the command required to be executed is saved to the remote server.

The experimental part of this study was carried out using the AnyLogic 7 simulation software package.

Fig. 1 presents a simulation model of the first and second method of networking.

Fig. 2 shows the structure of the studied network. For the experiment, data packets of different types and sizes were prepared. The first data packet consists of 218 bytes and simulates traffic from the voice assistant. The second type of packet transmits data after the database forms key phrases and selects the payload,

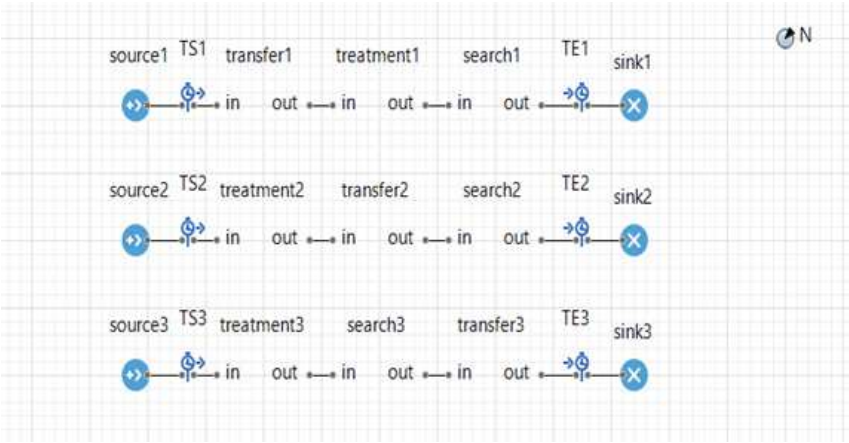


Fig. 1. SDN Network simulation model using VA

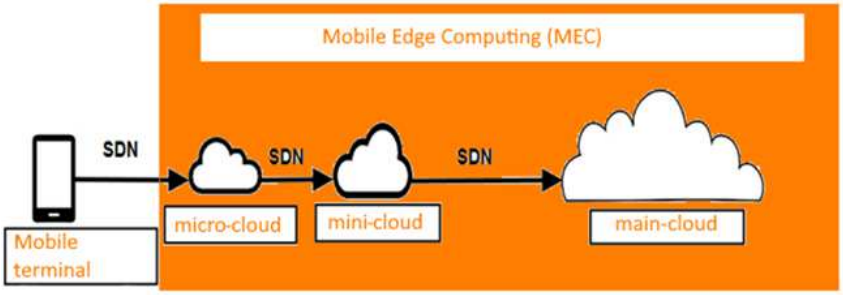


Fig. 2. The structure of the network

therefore, it has a shorter length of 138 bytes. The third data packet relates to the transmission of commands for performing certain actions and is 98 bytes.

The first data packet size is incoming traffic from the voice assistant - 218 bytes, of which 160 Bytes related to the payload, and 58 bytes to the headers (HTTP2.0 protocol).

The second size of the data packet occurs after the formation of the database of key phrases and the choice of the payload, and therefore should have a shorter data packet length. In this regard, a value equal to a sample of the G.711 codec was determined, that is, equal to 80–, respectively, the packet size is 138 bytes, of which 80 bytes were related to the payload, and 58 bytes were related to headers.

The third data packet relates to the transfer of commands to perform certain actions. The size of this packet is 98 bytes, of which 40 bytes related to the payload, and 58 bytes to the headers.

The data processing speeds on the mobile device, micro-cloud, mini-cloud and main cloud were 2 Mbps, 5 Mbps, 8 Mbps 25 Mbps, respectively [13].

5. The results of the study of the influence of the dependence of the execution time of processes and the total delay on the method of processing voice information

Tab. 1 shows the results with generalized data on the dependence of the execution time of processes and the total delay on the method of processing voice information for the first experiment.

Network Algorithm	Transfer Delay	Treatment Delay	Search Delay	Common Delay
1 variant	20 ms	10 ms	7 ms	37 ms
2 variant	13 ms	26 ms	7 ms	46 ms
3 variant	9 ms	26 ms	17 ms	52 ms

Table 1. Results of the first experiment

In the first version of the algorithm, we considered a scenario when the delay time of data transfer on a mobile device and on a server was comparable. As a result, the delay in the transmission of data, but the delay in the processing of data packets, which are related to the allocation and matching processes, was more important in the overall delay.

In the first version, both processes were performed on a remote cloud with great computational characteristics, and in the third version of the algorithm, on a mobile device. Moreover, the difference in data transmission delays was less significant compared with the delay in calculating the total network delay.

Tab. 2 shows the results of the second experiment.

Network Algorithm	Transfer Delay	Treatment Delay	Search Delay	Common Delay
1 variant	20 ms	3 ms	2 ms	25 ms
2 variant	13 ms	7 ms	2 ms	22 ms
3 variant	9 ms	7 ms	4 ms	20 ms

Table 2. The results of the second experiment

In the second experiment, the entire processing speed of computing devices was divided not by 30 simultaneously processed packets, but by 8. As a result, when forming the overall delay, the execution time of the processes for extracting key phrases and determining correspondence was less important than the transmission time of packets.

Consequently, the delay in data transmission will have a greater effect on the overall delay than the parameters of the time interval for performing the processes of extracting key phrases in the user's request, their comparison with the database of commands.

Next, a network model was developed, in addition to voice assistant traffic, IoTDM and VLC traffic was present. The network architecture is shown in Fig. 3.

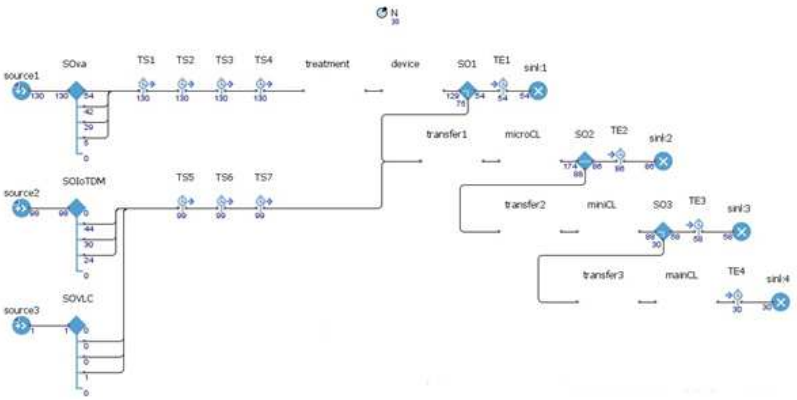


Fig. 3. Layered architectures for processing voice assistant traffic, IoTDM and VLC

The intensity of traffic generation is taken as a constant. Its value was calculated by dividing the total number of selected flows over a certain period of time. The packet sampling time was 5 minutes or 300 seconds. During this period, 8001 IoTDM traffic packets and 168 VLC traffic were selected.

Tab. 3 shows the dependence of the total delay on the type of traffic and the computing device for the final processing of packets.

Common delay, ms				
Types of traffic generators	Mobile terminal	Micro-cloud	Mini-cloud	Main-cloud
Voice assistant	43	49	54	57
Video translation, VLC	-	62	67	71
Internet of things, IoTDM	-	13	18	22

Table 3. The dependence of the total delay on the type of traffic and the computing device for the final processing of packets

Fig. 4 shows a simulation model for processing voice assistant traffic.

Similar models have been developed to handle voice broadcast video assistant (VLC) traffic and IoTDM traffic.

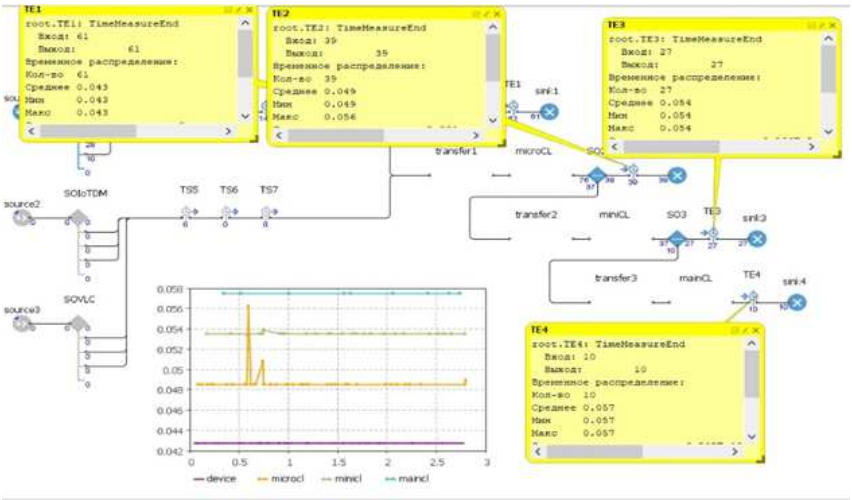


Fig. 4. Simulation model for processing voice assistant traffic

Fig. 5 shows the dependence of the total delay on the type of terminal processing device for four different processing levels.

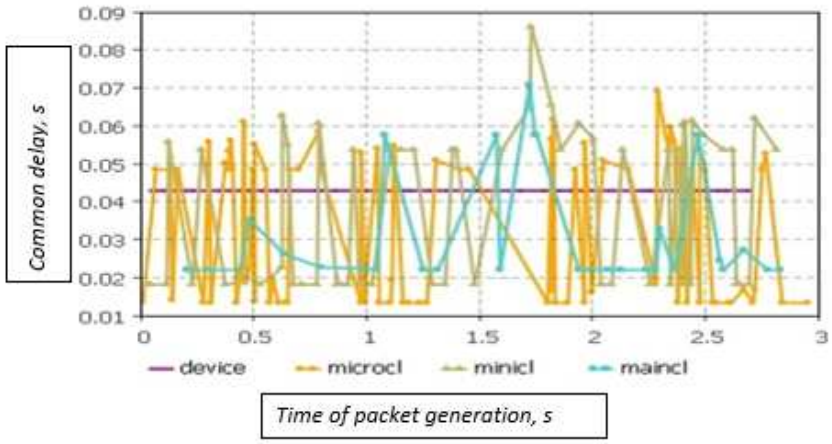


Fig. 5. Dependence of the total delay on the terminal processing device

The abscissa axis takes the time of generation of the packet - at what point in time the model was generated. The y-axis is the total delay. The graph shows a comparison of the delay time for various network units with the processing time. As

you can see from the graph, the use of a multi-level architecture for voice assistant traffic processing and SDN technology allows us to fulfill the requirements for ensuring delay in the implementation of various services using.

6. Conclusion

The use of SDN and MEC technologies for the implementation of services in modern communication networks improves the quality of service, unloads the core of the network and rational use of resources.

As part of this investigation, options were developed for organizing the network structure for the implementation of various services using a voice assistant. A comparative analysis of the three options for networking based on the results of an experimental study. While processing thirty data packets in the first experiment and eight packets data, respectively, in the second. The dependencies of the execution time of processes and the total delay on the method of processing speech information were investigated. The experimental results showed that, depending on the type of application and its requirements for network characteristics, it is advisable to use various network structures and multilevel traffic offloading architectures.

Acknowledgement

The publication has been prepared with the support of the "RUDN University Program 5-100".

REFERENCES

1. Hoy, M. B. Alexa, Siri, Cortana, and more: an introduction to voice assistants. Med. Ref. Serv. Q. 37, 81–88 (2018).
2. Chung, A. E., Griffin, A. C., Selezneva, D. & Gotz, D. Health and fitness apps for hands-free voice-activated assistants: content analysis. JMIR Mhealth Uhealth 6, e174 (2018).
3. Bickmore, T. W. et al. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of Siri, Alexa, and Google Assistant. JMIR 20, e11510 (2018).
4. Boyd, M. & Wilson, N. Just ask Siri? A pilot study comparing smartphone digital assistants and laptop Google searches for smoking cessation advice. PLoS ONE 13, e0194811 (2018).
5. Stone Temple. Rating the Smarts of the Digital Personal Assistants in 2018 https://www.stonetemple.com/digital-personal-assistants-study/ (accessed, 12 February 2019).

6. W. Bekri, R. Jmal, L. Ch. Fourati. Internet of Things Management Based on Software Defined Networking: A Survey. International Journal of Wireless Information Networks volume 27, pages385–410(2020).
7. C.Badii, P.Bellini, D.Cenni, A.Difino, P.Nesi, M.Paolucci. Analysis and assessment of a knowledge based smart city architecture providing service APIs. Future Generation Computer Systems. Vol. 75, October 2017, Pages 14-29.
8. Recommendation Y.3100: Terms and definitions for IMT-2020 network. Geneva, September 2017.
9. Yastrebova A., Kirichek R., Koucheryavy Ye., Borodin A., Koucheryavy A. Future networks 2030: architecture & requirements. In: 2018 10th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT) 2018
10. Ateya A.A., Muthanna A., Koucheryavy A. 5G framework based on multi-level edge computing with D2D enabled communication. In: 20th International Conference on Advanced Communication Technology (ICACT) conference proceedings. 2018. pp. 507-512.
11. Arash Asadi, Qing Wang, Dr. Vincenzo Mancuso A Survey on Device-to-Device Communication in Cellular Networks // IEEE Wireless Communications. 2014. P. 1–19.
12. A. Muthanna, R. Gimadinov, R. Kirichek, A. Koucheryavy. Software Development for The Centralized Management of IoT-Devices in The "Smart Home" Systems. 2017 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering.
13. Development of intelligent core network for tactile internet and future smart systems. // Ateya A., Muthanna A., Gudkova I., Abuarqoub A., Vybornova A., Koucheryavy A. Journal of Sensor and Actuator Networks. 2018. V. 7(1). P. 1.
14. A. Khakimov, A. Muthanna, R. Kirichek, A. Koucheryavy, Muthanna Mohammed Saleh Ali. Investigation of Methods for Remote Control IoT- Devices Based on Cloud Platforms and Different Interaction Protocols. 2017 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering.
15. Kirichek R., Vladyko A., Zakharov M., Koucheryavy A. ≪Model networks for Internet of Things and SDN≫, in Proc. 18th ICACT, pp. 76-79, 2016.
16. Vladyko, A., Muthanna, A., Kirichek, R.: Comprehensive SDN Testing Based on Model Network. In: Galinina, O., Balandin, S., Koucheryavy, Y. (eds.) Internet of Things, Smart Spaces, and Next Generation. LNCS, vol. 9870, pp. 539-549. Springer International Publishing (2016).
17. Khan P.W., Abbas Kh., Shaiba H.A., Mutkhanna A.S.A., Abuarqoub A., Khayyat M. Energy efficient computation offloading mechanism in multi-server mobile edge

УДК: 004

5G: патентный ландшафт

Д.М. Кочетков[1] and М.О. Альмаганбетов[2]

[1]НИУ «Высшая школа экономики», Москва, Россия
[2]LexisNexis Eastern Europe, Москва, Россия

Аннотация

Отчеты о патентных ландшафтах (PLR) уже давно используются в бизнесе, науке и RD. Они позволяют принимать решения, основанные на данных и минимизировать риски стратегического технологического выбора. Авторы проанализировали патентный ландшафт в области беспроводных сетей связи 5G с помощью программного продукта PatentSight®. Среди компаний однозначным лидером по Индексу патентного портфеля является Samsung Electronics, среди стран США и Китай. Тем не менее, патенты с высокой технологической релевантностью и/или коммерческой ценностью могут появляться в небольших технологичных компаниях или как «побочный продукт» других направлений деятельности корпораций. Результаты исследования представляют интерес для исследователей и практиков, занимающихся развитием и внедрением технологий 5G.

Ключевые слова: патентный ландшафт, 5G, Индекс патентных активов, технологическая релевантность, рыночное покрытие, конкурентное влияние

1. Введение

Патентный ландшафт — это систематическое исследование патентной документации, которое позволяет выявить и визуализировать тренды в бизнесе, науке и технологии. Отчеты о патентных ландшафтах обычно сосредоточены на одной отрасли, технологии или географическом регионе. Отчеты о патентных ландшафтах (PLR) поддерживают принятие обоснованных решений и предназначены для эффективного решения проблем, связанных с высокими рисками в различных областях технологии, что повышает степень доверия. Благодаря патентной аналитике и PLR эти критические решения могут приниматься с использованием фактических данных, которые обеспечивают осознанный выбор и снижают риски, связанные с принятием решений [1].

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта №18-00-01040 КОМФИ «Влияние новых технологий на городскую среду и качество жизни городских сообществ».

Данное патентное исследование посвящено технологии 5G. 5G означает сотовую беспроводную связь пятого поколения, первоначальные стандарты для нее были установлены в конце 2017 года. Начиная с 2019г. мобильные операторы начали развертывать сети 5G по всему миру. Основным преимуществом новых сетей является то, что они будут иметь большую пропускную способность, обеспечивая более высокую скорость передачи информации. Благодаря увеличенной пропускной способности сети, мобильные операторы смогут конкурировать с интернет-провайдерами, предоставляющими услуги кабельного интернета. Также беспроводные сети последнего поколения открывают целый спектр новых применений в интернете вещей (IoT), расширенной/дополненной реальности, M2M системах [2].

В исследовании использовалась патентная аналитика, предоставленная компанией LexisNexis (основана в 1977г., является частью крупнейшего транснационального информационного холдинга RELX Group). LexisNexis IP является мировым лидером в области информационных решений и услуг для удовлетворения потребностей рынка интеллектуальной собственности, государственных учреждений и академических организаций. Компания агрегирует данные из 111 патентных ведомств по всему миру [3].

2. Данные и методы

В исследовании были использованы данные о 4069 патентных семействах, связанных с технологией 5G. Поиск проводился по ключевым словам “5G” и “fifth generation” с фильтром по домену (IPC) H04W (Wireless Communication Networks [2009.01]). Патентное семейство – набор патентов, полученных в разных странах для защиты одного изобретения (когда первая заявка в стране – приоритет – затем распространяется на другие патентные ведомства) [4] стр. 60. Источником данных является программный продукт PatentSight®[5]. от компании LexisNexis. Данный инструмент предоставляет объективные метрики глобальной технологической силы и влияния, разработанные немецкими учеными во главе с Н.Омландом и Х. Эрнстом [6].

Индекс патентных активов (Patent Asset Index ™, PAI) учитывает как количество активно охраняемых изобретений, так и их качество. Метод был разработан и апробирован в научных исследованиях и уже несколько лет используется ведущими компаниями во многих отраслях. Индекс патентных активов — это общая сила патентного портфеля. Он основан на индикаторах технологической релевантности (Technology Relevance ™) и рыночного покрытия (Market Coverage ™) патентов компании, которые позволяют проанализировать конкурентное влияние портфеля.

Технологическая релевантность – цитирование по всему миру, полученное из более поздних патентов, с учетом возраста, практики патентного ведомства и области технологий. Рыночное покрытие представляет собой размер рынка, защищенного активными патентами и находящимися на рассмотрении патентными заявками на определенное изобретение. Наконец, конкурентное влияние (Competitive Impact ^{TM}) оценивает силу индивидуального патента, т.е. относительную коммерческую ценность патента.

3. Результаты

На 1. представлены результаты размера портфеля, Индекса патентных активов и конкурентного влияния или относительной коммерческой ценности. Безусловным лидером рынка является компания Samsung Electronics, которая обладает крупнейшим патентным портфелем в области 5G с наибольшим Индексом патентных активов. Обратной стороной медали является низкая селективность портфеля. В тоже время мы видим две компании с небольшими патентными портфелями, но крайне высокой ценностью (ATT и INTERDIGITAL). Samsung

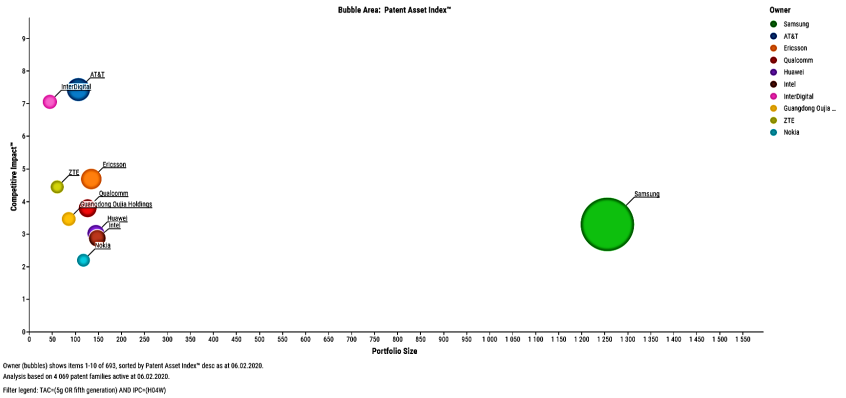


Рис. 1. Размер портфеля (по горизонтальной оси), конкурентное влияние (по вертикальной оси) и Индекс патентных активов (размер пузырька). Создано с помощью PatentSight®.

лидирует не только в целом в области, но и во всех подобластях (уровень 4 классификации IPC) (рис. 2). Однако если мы посмотрим на индикатор технологической релевантности, то мы не увидим в списке компаний с крупнейшими патентными портфелями (рис. 3). Часто наиболее цитируемые в дальнейшем патенты появляются в небольших технологичных компаниях или как «побочный продукт» других направлений деятельности (например, Toyota). Наконец, по про-

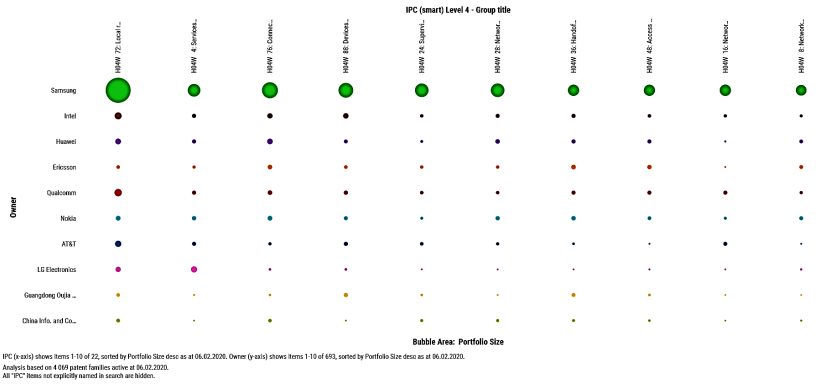


Рис. 2. Размер патентного портфеля в разрезе 4 уровня классификации IPC. Создано с помощью PatentSight®.

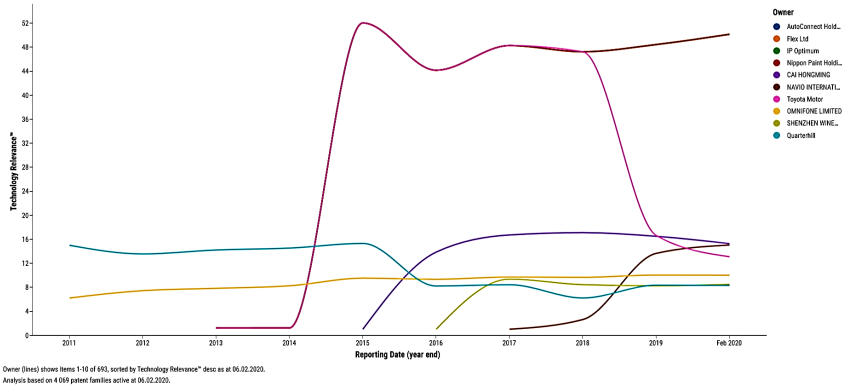


Рис. 3. Технологическая релевантность патентного портфеля. Создано с помощью PatentSight®.

исхождению патентных семейств с заметным отрывом лидируют две юрисдикции – США и КНР (рис. 4).

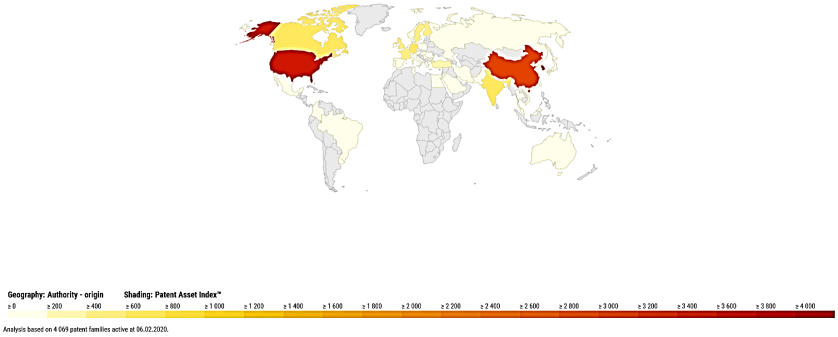


Рис. 4. Патентные семейства в географическом разрезе по юрисдикции происхождения. Создано с помощью PatentSight®.

4. Заключение

Отчеты о патентных ландшафтах уже давно используются в бизнесе, науке и RD. Они позволяют принимать решения, основанные на данных и минимизировать риски. В данном исследовании мы проанализировали патентный ландшафт в области беспроводных сетей связи 5G. Среди компаний однозначным лидером по Индексу патентного портфеля является Samsung Electronics, среди стран США и Китай. Тем не менее, патенты с высокой технологической релевантностью и/или коммерческой ценностью могут появляться в небольших технологичных компаниях или как «побочный продукт» других направлений деятельности корпораций.

Проведенный анализ показывает, что семейство технологий 5G вступило в фазу зрелости, т.е. дальнейшие изменения будут носить инкрементный характер. На этом этапе рынок уже сформирован, акцент смещается на коммерциализацию и внедрение. На пороге новое поколение систем мобильной связи [7]. Тем не менее, новое поколение сетей связи (6G) будет опираться на уже имеющийся научный задел, связанный с пониманием как целевых приложений, так и наиболее перспективных технологий-кандидатов. Исторический анализ развития технологии на основе наукометрического анализа и патентных ландшафтов позволяет принимать обоснованные стратегические решения для технологических прорывов в будущем.

ЛИТЕРАТУРА

1. Trippe A. Guidelines for Preparing Patent Landscape Reports. 2015.

2. Looper C. de. What Is 5G? The Next-generation Network Fully Explained | [Electronic resource] // Digital Trends. 2020. URL: https://www.digitaltrends.com/mobile/what-is-5g/ (accessed: 14.07.2020).
3. About Us [Electronic resource] // LexisNexis. 2020. URL: https://www.lexisnexisip.com/about-us/ (accessed: 14.07.2020).
4. OECD Science, Technology and Industry Scoreboard 2001. OECD, 2001.
5. Patent Research and Analytics Products | LexisNexis PatentSight® [Electronic resource]. URL: https://www.lexisnexisip.com/products/patent-sight/ (accessed: 14.07.2020).
6. Ernst H., Omland N. The Patent Asset Index – A new approach to benchmark patent portfolios // World Pat. Inf. 2011. Vol. 33, № 1. P. 34–41.
7. Giordani M. et al. Towards 6G Networks: Use Cases and Technologies [Electronic resource]. 2019. P. 1–7.

УДК: 519.248

Профилактическое обслуживание привязного модуля высотной телекоммуникационной платформы

В.Вишневский[1], В. Рыков[2,3], М. Финкельштейн[4]

[1]Институт проблем управления (ИПУ) имени В.А. Трапезникова РАН,, Профсоюзная. 65,Москва, 117997, Россия
[2]Российский государственный университет нефти и газа (НИУ) имени И.М. Губкина, Ленинский пр-т, 65, Москва, 119991, Россия
[3]Институт проблем передачи информации (ИППИ) РАН, Большой Каретный переулок, д.19 стр. 1, Москва, 127051, Россия
[4]Университет Свободного Штата, Проспект Нелсона Мандела, 205, Блюмфонтейн, 9300, Южная Африка

vishn@inbox.ru, vladimir_rykov@mail.ru, FinkelM@ufs.zc.za

Аннотация

Решается задача оценки эффективности профилактического обслуживания привязных высотных телекоммуникационных платформ, отказы которых зависят как от числа, так и расположения отказавших компонент.

Ключевые слова: Высотные привязные телекоммуникационные платформы, профилактическое обслуживание

1. Введение

Привязные телекоммуникационные платформы занимают ведущие позиции в современной структуре связи [1, 2, 3]. Они предназначены для долговременного использования и широко применяются как в повседневной жизни, так и в военных целях. Структура таких платформ включающая несколько, скажем n двигателей позволяет им работать даже если несколько из них, скажем k откажут (см. рис. 1).

Однако отказ одного или нескольких двигателей ведёт к увеличению нагрузки на остальные, что приводит к возможности более быстрого их отказа. Кроме того, отказ всей системы зависит от расположения отказавших двигателей, например, отказ рядом расположенных двигателей с большей вероятностью приводит к

Работа выполнена при финансовой поддержке РФФИ, проекты №20-01-00575А и №19-29-06043.

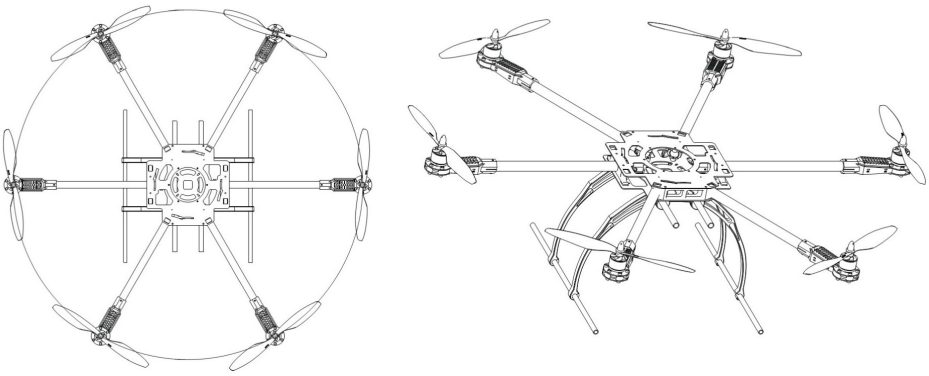


Рис. 1. Структура шестироторной платформы

отказу системы, чем отказы далеко отстоящих двигателей. Исследование надёжности таких систем и разработка методов её повышения и поддержания на заданном достаточно высоком уровне является одной из актуальных задач [4, 5].

Архитектура и условия эксплуатации привязных платформ позволяют моделировать их надёжность с помощью систем k-из-n с зависимыми отказами. Благодаря широкому применению моделей таких систем их изучению посвящены многочисленные исследования. Ранние исследования в этом направлении имели дело с однородными бинарными моделями, компоненты которых принимали два состояния: исправное и отказовое. Имеется обширная литература по исследованию таких систем, обзор которой см., например, в [6], [7], а также [4]).

В дальнейшем исследования таких систем получили значительное развитие, появились исследования восстанавливаемых систем k-из-n с не показательно распределёнными длительностями ремонта (библиографию см., например, в [8]). Эти исследования позволили поставить и развить одну из принципиальных проблем надёжности систем — чувствительность их характеристик к виду исходных распределений (см., например, [9] и приведённую там библиографию).

Проблематика повышения надёжности систем путём проведения профилактического обслуживания имеет долгую историю, о чём можно судить, например по работе, [10]. Достаточно подробный обзор методов профилактического обслуживания можно найти в монографии Герцбаха [11]m Some recent developments on optimal maintenance policies can be found in [12], [13].

В настоящей работе исследуются стратегии назначения профилактического обслуживания системы k-из-n по наблюдениям за её состоянием. В следующем разделе приводится постановка задачи и основные обозначения, затем в разделе 3 находятся условия эффективности профилактического обслуживания

для однородной системы. Более сложным проблемам назначения профилактического обслуживания для системы, отказы которой зависят как от числа, так и расположения отказавших компонент посвящены разделы 4 и 5. В заключении приводятся направления дальнейших исследований.

2. Постановка задачи. Обозначения

Рассмотрим модель мультикоптера, которая представляет собой часть архитектуры привязной высотной телекоммуникационной платформы [4]. В данном случае мультикоптер рассматривается как система горячего резервирования, состоящая из n компонент (роторов), отказы которых могут зависеть как от числа отказавших роторов, так и от их расположения в системе. (см. рис. 1).

Обозначим через $A_i : i = 1, 2, \dots$ случайное время безотказной работы (в.б.р.) компонент системы, а через $A(t) = \mathbf{P}\{A_i \le t\}$ их общую функцию распределения (ф.р.). При отказе системы она восстанавливается, длительности восстановления после каждого отказа системы – н.о.р. с.в. $B_i^{(0)} : i = 1, 2, \dots$ с распределением $B_0(t) = \mathbf{P}\{B_i^{(0)} \le t\}$ и средним значением $b_0 = \mathbf{E}[B_0] = \int_0^\infty (1 - B_0(t))dt$. Обозначим далее через $\mathbf{j} = (j_1, j_2, \dots j_n)$ состояние системы, где $j_i = 0$, если i-ая компонента находится в работоспособном состоянии и $j_i = 1$, если i-ая компонента отказала, а через $E = \{\mathbf{j} = (j_1, j_2, \dots j_n) : \ (j_i \in (0, 1))\}$ множество состояний системы, а через E_0 и $\bar{E}_0$ подмножества её работоспособных и отказовых состояний.

Для увеличения надёжности системы (её коэффициента готовности) предполагается проведение профилактического обслуживания по наблюдению за состоянием системы. Возможны различные стратегии профилактического обслуживания, которые определяются выбором подмножества состояний системы для начала проведения профилактических работ. Для l-ой стратегии профилактического обслуживания обозначим через E_l подмножество опасных (предотказовых) состояний, при попадании в которые предполагается начало проведения профилактики. Длительности проведения профилактического обслуживания — н.о.р. с.в. $B_i^{(l)}$ с ф.р. $B_l(t) = \mathbf{P}\{B_i^{(l)} \le t\}$ и средним значением $b_l = \mathbf{E}[B_l] = \int_0^\infty (1 - B_l(t))dt$. Предполагается, что время проведения профилактик в среднем меньше времени ремонта системы в случае её отказа, $b_l < b_0$, но зависит от состояния системы для начала проведения профилактики. Все последовательности с.в.: в.б.р. компонент системы, длительности ремонтов и профилактик предполагаются независимыми.

В работе решается задача оценки качества различных стратегий профилактического обслуживания по критерию *коэффициент готовности* системы. Для решения поставленной задачи определим регенерирующий случайный процесс

$J = \{\mathbf{J}(t) : \ t \geq 0\}$ с множеством состояний E соотношением

$$\mathbf{J}(t) = \mathbf{j}, \ \text{если в момент времени } t \text{ система находится в состоянии } \mathbf{j} \in E$$

и обозначим через G_0 время до первого отказа (в.б.р.) систем, а через G_l время до начала проведения профилактического ремонта при выборе l-ой стратегии профилактического обслуживания,

$$G_0 = \inf\{t : \ \mathbf{J}(t) \in \bar{E}_0\}, \quad G_l = \inf\{t : \ \mathbf{J}(t) \in E_l\},$$

а через $\Pi_0 = G_0 + B_0$ и $\Pi_l = G_l + B_l$, периоды регенерации процесса без проведения и при наличии профилактик (типа l). Тогда коэффициент готовности системы вычисляется как отношение длительности рабочего периода к длительности периода регенерации

$$K_{\text{гот.},0} = \frac{\mathbf{E}[G_0]}{\mathbf{E}[\Pi_0]}, \quad K_{\text{гот.},l} = \frac{\mathbf{E}[G_l]}{\mathbf{E}[\Pi_l]},$$

Таким образом, так как при любом из режимов $l = 0, 1, 2, \ldots$ длительности ремонта и профилактического обслуживания предполагаются известными, для решения задачи необходимо вычислить среднее значение рабочего времени системы G_l при отсутствии $l = 0$ и использовании l-ой стратегии профилактического обслуживания.

Остановимся вначале на задаче оценки эффективности профилактического обслуживания для системы k-из-n при наличии единственной стратеги профилактического обслуживания.

3. Профилактическое обслуживание системы k-из-n

Рассмотрим сначала однородную систему типа k-из-n, для которой множества работоспособных и отказовых состояний имеют вид

$$E_0 = \{1, 2, \ldots, k-1\} \quad \bar{E}_0 = \{k, k+1, \ldots, n\},$$

а в качестве подмножества состояний E_l для начала проведения профилактик назначим подмножество из единственного состояния, $E_1 = \{k-1\}$.

Для исследования поведения процесса обозначим через T_i время между отказами i-ой и $i-1$-ой компонентами системы на цикле регенерации. Тогда

$$\Pi_0 = T_1 + T_2 + \ldots T_{k-1} + T_k + B_0, \quad \Pi_1 = T_1 + T_2 + \cdots + T_{k-1} + B_1.$$

Для вычисления распределений

$$F_i(x) = \mathbf{P}\{T_i \leq x\}$$

их средних значений $m_i = \int_0^\infty (1 - F(t))dt$ и средних длительностей достижения множеств E_0 и E_1 в работе предлагается алгоритм.

Алгоритм. Исходные данные:

$$n,\ k,\ I_0 = \{1, 2, \ldots, n\},\ A_i^{(0)} = A_i,\ \ b_0 = \mathbf{E}[B^{(0)}],\ \ b_1 = \mathbf{E}[B^{(1)}].$$

Шаг 1. Положим

$$T_0 = \min\{A_i^{(0)} :\ i \in I_0\}, \quad i_0 = \arg\min\{A_i^{(0)} :\ i \in I_0\},$$

Найти

$$F_0(x) = \mathbf{P}\{T_0 \le x\} = 1 - (1 - A(x))^n. \tag{1}$$

Шаг 2. Для $j = \overline{1,k}$ положим

$$I_j = I_{j-1} \setminus \{i_{j-1}\}, \qquad A_i^{(j)} = A_i^{(j-1)} - T_{j-1},$$
$$T_j = \min\{A_j^{(i)} :\ i \in I_j\}, \qquad i_j = \arg\min\{A_i^{(j)} :\ i \in I_j\}.$$

Считать

- распределение $A^{(j)}(x)$ остаточного времени обслуживания $A^{(j)}$

$$A^{(j)}(x) \ = \ \mathbf{P}\{A_i^{(j)} \le x\} = \int_0^\infty \mathbf{P}\{A_i^{(j-1)} - T_{j-1} \mid T_{j-1} = x\} dF_{j-1}(x) =$$

$$= \ \int_0^\infty A^{(j-1)}(x + u) dF_{j-1}(u) \ \ (j = \overline{1,k}), \tag{2}$$

- распределение их минимума

$$F_j(x) \ = \ \mathbf{P}\{T_j \le x\} = 1 - (1 - A^{(j)}(x))^{n-j} \ \ (j = \overline{1,k}), \tag{3}$$

Шаг 3. Считать средние длительности до отказов

$$m_j = \mathbf{E}[T_j] = \int_0^\infty (1 - F_j(x) dx \ \ (j = \overline{1,k}),$$

Шаг 4. Выдача результатов: $m_j\ (j = \overline{1,k})$
Конец.

Результатом работы алгоритма являются распределения $F_i(x)$ с.в. T_i и их средние значения m_i для любых исходных распределений $A(x)$ и заданных

множеств E_0 и E_1, что позволяет сравнивать различные стратегии проведения профилактик и выбрать наилучшую.

Для оценки качества профилактического обслуживания заметим, что коэффициенты готовности системы при проведении профилактических работ и их отсутствии выражаются в терминах средних интервалов между достижениями множеств E_1 и $E_)$ в виде

$$K_{\text{гот.},0} = \frac{m_1 + \ldots m_k + m_k}{m_1 + \ldots m_k + m_k + b_0}, \quad K_{\text{гот.},1} = \frac{m_1 + \cdots + m_{k-1}}{m_1 + \cdots + m_{k-1} + b_1},$$

Так как профилактическое обслуживание эффективно, если $K_{\text{гот.},0} < K_{\text{гот.},1}$, то из ннеравенства

$$\frac{m_1 + \cdots + m_k + m_k}{m_1 + \cdots + m_k 2 + m_k + b_0} > \frac{m_1 + \cdots + m_{k-1}}{m_1 + \cdots + m_{k-1} + b_1},$$

следует, что необходимым и достаточным условием проведения профилактического обслуживания по достижению состояния $k-1$ является условие

$$\frac{b_0 - b_1}{b_1} > \frac{m_k}{m_1 + \cdots + m_{k-1}} \quad \text{или} \quad \frac{b_0}{b_1} > 1 + \frac{m_k}{m_1 + \cdots + m_{k-1}}. \tag{4}$$

Аналитические выражения для m_i доступны не всегда, Однако их численный анализ не представляет труда. Для демонстрации конкретных результатов рассмотрим пример системы 3-из-6 в предположении о показательном распределении в.б.р. компонент системы.

Пример: система 3-из-6

Рассмотрим простейшую модель системы k-из-n при $k = 3,\ n = 6$ и показательных распределениях в.б.р. с ф.р. $A(x) = \mathbf{P}\{A_i \le x\} = 1 - e^{-\alpha x}$. Отказавшая система отправляется на ремонт и ремонт длится в среднем время b_0. Начало проведения проведения профилактического обслуживания предлагается по достижению состояния 2, когда отказывают только две компоненты системы (множество E_1). В этом случае профилактическое обслуживание длится в среднем время b_1, а после окончания ремонта и профилактики система также становится "как новая".

В предположении, что в.б.р. компонент системы имеют показательные распределения остаточные в.б.р. не зависят от моментов отказа компонент и, следовательно, с.в.T_i равны $T_i = \min\{A_1, A_2, \ldots A_{6-i}\}$, так что

$$F_i(x) = \mathbf{P}\{T_i \le x\} = 1 - \mathbf{P}\{T_i > x\} = 1 - (1 - A(x))^{6-i} = 1 - e^{(6-i)\alpha x}.$$

Таким образом имеем $m_i = \mathbf{E}[T_i] = [(6-i)\alpha]^{-1}$. Откуда следует, что необходимым и достаточным условие (4) проведения профилактического обслуживания по достижению состояния 2 является условие

$$\frac{b_0}{b_1} > \frac{m_1 + m_2}{m_1} = \frac{9}{4} = 2.25.$$

4. Профилактическое обслуживание системы, отказ которой зависит от расположения отказывающих компонент

Если множества работоспособных и отказовых состояний зависят от расположения отказывающих компонент, то сравнение различных стратегий профилактического обслуживания со стратегий отказа от профилактического обслуживания зависит от конкретных условий эксплуатации системы. Поэтому в общем случае можно предложить только методологию такого сравнения, которую продемонстрируем на рассматриваемом ниже примере решения этого вопроса для конкретной модели 3-из-6.

Пример: система 3-из-6

Продолжим исследование модели типа 3-из-6 с показательно распределёнными в.б.р. компонент в предположении, что система отказывает, когда отказывают два рядом расположенных двигателя или когда отказывают три любых двигателя. Для удобства перенумеруем состояния системы в двоичном коде, то есть поставим в соответствие состоянию $\mathbf{j} = (j_1, j_2, \dots j_6)$ его номер по формуле

$$j = |\mathbf{j}| = \sum_{0 \le i \le 6} j_i 2^{6-i}.$$

Тогда множество работоспособных состояний $\bar{E}_0$ состоит из состояний с номерами

$$\bar{E}_0 = \{0, 1, 2, 4, 5, 8, 9, 10, 16, 17, 18, 20, 32, 34, 40\}$$

Рассмотрим две возможные стратегии профилактического обслуживания, когда профилактика назначается:

- при отказе любого из двигателей системы, то есть по достижении множества $E_1 = \{1, 2, 4, 8, 16, 32\}$ (Стратегия 1), или

- при отказе любых двух не рядом стоящих двигателей, то есть по достижении множества $E_2 = \{, 9, 10, 17, 18, 20, 34, 40\}$ (Стратегия 2)

и сравним их

- с режимом работы системы до отказа, то есть до достижения множества $E_0 = E \setminus (E_1 + E_2)$ (Стратегия 0).

В этом случае действуя в соответствии с алгоритмом найдём, что время T_1 достижения множества E_1 равно $T_1 = \min\{A_i : i = \overline{1,6}\}$ и имеет распределение

$$F_1(t) = \mathbf{P}\{T_1 \le t\} = 1 - (1 - A(t))^6 \quad \text{с} \quad m_1 = \mathbf{E}[T_1] = \int_0^\infty (1 - A(t))^6 dt.$$

Время достижения множества E_2 включает в себя время T_1 до отказа одного из двигателей плюс время T_2 до отказа второго, расположенного не рядом с первым, которое равно минимуму из остаточных длительностей работающих компонент. При показательно распределённых в.б.р. компонент распределение последних не зависит от момента отказа предыдущей и имеет вид

$$F_2(t) = \mathbf{P}\{T_2 \le t\} = 1 - (1 - A^{(1)}(t))^3 dt \quad \text{с} \quad m_2 = \mathbf{E}[T_2] = \int_0^\infty (1 - A^{(1)}(t))^3 dt.$$

Таким образом, условие предпочтения первой стратегии перед второй $K_{\text{гот.},1} > K_{\text{гот.},2}$ имеет вид

$$\frac{m_1}{m_1 + b_1} > \frac{m_1 + m_2}{m_1 + m_2 + b_2} \quad \text{или} \quad \frac{b_2}{b_1} > \frac{m_1 + m_2}{m_1},$$

то есть время проведения профилактики при второй стратегии должно значительно превышать время профилактики при первой стратегии. Например при показательно распределённых в.б.р. компонентов системы условие предпочтения первой стратегии перед второй принимает вид

$$\frac{b_2}{b_1} > 1 + \frac{m_2}{m_1} = 3,$$

то есть время профилактики при второй стратегии должно превышать время профилактики при первой стратегии по крайней мере в три раза.

Сравним теперь каждую из стратегий профилактического обслуживания с режимом работы системы до отказа с последующим ремонтом. Время до отказа системы после достижения множества состояний $E_1 \cup E_2$ включает в себя время T_3 до отказа любого из оставшихся 4-х двигателей, то есть равно $T_3 = \min\{A_i^{(2)} : i = \overline{1,4}\}$, где $A_i^{(2)} : i = \overline{1,4}$— остаточные длительности работы оставшихся после отказов двух не рядом расположенных двигателей. Их распределение вычисляется, как и ранее, согласно алгоритму по формуле (2). Таким образом, распределение с.в. T_3 имеет вид

$$F_3(t) = \mathbf{P}\{T_3 \le t\} = 1 - (1 - A^{(2)}(t))^4 \quad \text{с} \quad m_3 = \mathbf{E}[T_3] = \int_0^\infty (1 - A^{(2)}(t))^4 dt.$$

При этом условие предпочтительности первой стратеги профилактического обслуживания перед использованием системы до отказа с последующим ремонтом имеет вид

$$\frac{m_1}{m_1+b_1} > \frac{m_1+m_2+m_3}{m_1+m_2+m_3+b_0} \quad \text{или} \quad \frac{b_0}{b_1} > \frac{m_1+m_2+m_3}{m_1}.$$

Аналогично условие предпочтения второй стратегии перед использованием системы до отказа с последующим ремонтом имеет вид

$$\frac{m_1+m_2}{m_1+m_2+b_2} > \frac{m_1+m_2+m_3}{m_1+m_2+m_3+b_0} \quad \text{или} \quad \frac{b_0}{b_2} > \frac{m_1+m_2+m_3}{m_1+m_2}.$$

Таким образом, например при показательно распределённых в.б.р. компонент системы условие предпочтения первой стратегии перед работой системы до отказа с последующим ремонтом имеет вид $\frac{b_0}{b_1} > \frac{9}{2}$, а условием предпочтения второй стратегии перед работой системы до отказа с последующим ремонтом является неравенство $\frac{b_0}{b_1} > \frac{3}{2}$.

5. Профилактическое обслуживание системы с зависимыми отказами её компонент

Предположим теперь, что отказ одной из компонент системы ведёт к увеличению нагрузки на остальные её компоненты. Формально это предположение можно сформулировать следующим образом. В результате отказа одной из компонент системы остаточное в.б.р. $A_{\text{res.new}}$ каждой из оставшихся компонент системы имеет распределение

$$\mathbf{P}\{A_{\text{res.new}} \le t\} = A_{\text{res.old}}(c_i t),$$

где $1 < c_1 < \ldots c_i < \ldots c_k$ и $A_{\text{res.old}}(t)$ вычисляется в соответствии с процедурой, задаваемой формулами (1–2).

Напоминая, что T_i обозначает интервал времени между $i-1$-ым и i-ым отказами. Для её вычисления имеем, как и ранее,

$$F_1(t) = \mathbf{P}\{T_1 \le t\} = \mathbf{P}\{\min_{1\le i\le n} A_i \le t\} = 1-(1-A(t))^n.$$

После отказа первой из компонент оставшиеся $n-1$ компонента работают с усиленной нагрузкой и ф.р. в.б.р. $A^{(1)}$ каждой из них имеет вид

$$A^{(1)}(t) = \mathbf{P}\{A^{(1)} \le t\} = A_{\text{res}}(c_1 t),$$

где распределение $A_{\text{res}}(t)$ вычисляется в соответствии с формулой (2) алгоритма. Следовательно,

$$F_2(t) = \mathbf{P}\{\min_{1\le i\le n-1} A_i^{(1)} \le t\} = 1-(1-A^{(1)}(c_1 t))^{n-1}.$$

Аналогично, ф.р. времени между $(i-1)$-ым и i-ым отказами имеет вид

$$F_i(t) = \mathbf{P}\{\min_{1\le l\le n-(i-1)} A_l^{(i-1)} \le t\} = 1 - (1 - A^{(i-1)}(c_{i-1}t))^{n-(i-1)}.$$

Таким образом, предложенная процедура позволяет вычислить все распределения $F_i(t)$ $(i = \overline{1,k})$ и их средние значения

$$m_i = \int_0^\infty (1 - F_i(t))dt = \int_0^\infty (1 - A^{(i-1)}(c_{i-1}t))^{n-(i-1)}dt,$$

которые, как и ранее, используются для проверки эффективности проведения профилактики. Для получения окончательных результатов обратимся снова к примеру системы 3-из-6 с показательным распределение в.б.р. её компонент.

Пример: система 3-из-6

Продолжим исследование примера 3-из-6 при показательных распределениях в.б.р. компонентов системы $A(t) = 1 - e^{-\alpha t}$. при этом остаточные длительности не зависят от предшествующих моментов отказов и также имеют показательное распределение). Имеем

$$F_1(t) = 1 - (1 - A(t))^6 = 1 - e^{-6\alpha t} \quad \text{и} \quad A^{(1)}(t) = 1 - e^{-6c_1\alpha t}$$

Таким образом,

$$F_2(t) = 1 - (1 - A^{(1)}(t))^5 = 1 - e^{-30c_1\alpha t} \quad \text{и} \quad A^{(2)}(t) = 1 - e^{-30c_1c_2\alpha t}$$

И далее

$$F_3(t) = 1 - (1 - A^{(2)}(t))^4 = 1 - e^{-120c_1c_2\alpha t}$$

Таким образом,

$$m_1 = \frac{1}{6\alpha},\ m_2 = \frac{1}{30c_1\alpha},\ m_3 = \frac{1}{120c_1c_1\alpha}$$

и условие (4) эффективности проведения профилактики в этом случае приобретает вид

$$\frac{b_0}{b_1} > 1 + \frac{m_3}{m_1 + m_2} = 1 + \frac{1}{4c_2(1 + 5c_1)}.$$

6. Заключение

В работе приводится алгоритм оценки эффективности проведения профилактического обслуживания привязного модуля высотной телекоммуникационной платформы с произвольными распределениями и.б.р. его компонент. Приведены численные расчёты в частном случае пуассоновского потока отказов его компонент.

Работы в этом направлении предполагается продолжать привлекая модели с зависимыми отказами, в том числе, например, в рамках моделей многорерных распределений типа Маршалла-Олькина.

ЛИТЕРАТУРА

1. M.A. Khan, R. Hamila, M.S. Kiranyaz, A.M. Gabbou. A Novel UAV – Aided NetWork Architecture Using WiFi Direct?, IEEE Access, V. 7, pp. 67305-67318.
2. V. M. Vishnevskiy, A. M. Shirvanyan and D. A. Tumchenok. Mathematical Model of the Dynamics of Operation of the Tethered High-Altitude Telecommunication Platform in the Turbulent Atmosphere. //Systems of Signals Generating and Processing in the Field of on Board Communications, IEEE Xplore, 2019, pp. 1-7. DOI: 10.1109/SOSG.2019.8706784
3. V.M. Vishnevsky. D.V. Efrosinin, A. Krishnamoorthy. Principles of Construction of Mobile and Stationary Tethered High-Altitude Unmanned Telecommunication Platforms of Long-Term Operation. // Communications in Computer and Information Science, 2018.Volume 919. Springer, Cham, Pp. 561-569. DOI:10.1007/978-3-319-99447-5.
4. В.М. Вишневский, Д.В. Козырев, В.В. Рыков, З.Ф. Нгуен. Моделирование надёжности беспилотного высотного модуля привязной телекоммуникационной платформы / Информационные технологии и вычислительные системы, Выпуск 4, 2020 (в печати).
5. Kozyrev, D.V., Phuong, N.D., Houankpo, H.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, 1141 CCIS, pp. 646-656. 2019. DOI: 10.1007/978-3-030-36625-4_52
6. S. R. Chakravarthy, A. Krishnamoorthy and P. V. Ushakumari (2001). 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(4): 361-380.
7. K.S. Trivedi, Probability and Statistics with Reliability, Queuing and Computer Science Applications // Wiley, New York, 2002.

8. V. Rykov. On Reliability of Renewable Systems. // In Reliability Engineering. Theory and Applications (Edds.by Ilia Vonta and Mangey Ram) CRC Press. 2018, pp. 173-196.
9. Rykov, V., Kozyrev, D., Filimonov, A., Ivanova, N. On Reliability Function of a k-out-of-n System With General Repair Time Distribution // Cambridge University Press: Probability in the Engineering and Informational Sciences, 1-18. 2020. DOI: 10.1017/S0269964820000285
10. В.В. Рыков. Определение оптимального времени между профилактиками. Труды ЦНИИКА вып.12, 1965, с.258-266
11. I. Gertsbakh. Reliability theory with applications to preventive maintanance. Springer, 2000 219 pp. Русский превод: И. Герцбах. Теория надёжности с приложениями к профилактическому обслуживанию. Изд. «Нефть и газ». М. 2003. 263с.
12. M. Finkelstein, G. Levitin (2019). Preventive maintenance for homogeneous and heterogeneous systems. Applied Stochastic Modelling in Business and Industry, 35, 908-920.
13. M. Finkelstein, J.H. Cha and G. Levitin (2020). On a new age-replacement policy for items with observed stochastic degradation. Quality and Reliability Engineering International, 36, 1132-1146.

UDC: 004.056

The Markov Model for a Multiphase Security System with the Partial Concurrent Service

Yermakov A.S.[1], Shukmanova A.A.[1], Seilova N.A.[2]

[1]Caspian Public University
[2]Kazakh National Technical University
ermakov_as@mail.ru

Abstract

A cryptographic protocol is not only an algorithm but also a procedure for exchanging data between different network subscribers, associated with the message exchange at different levels. In this case, the problem of minimizing the number of protocol implementation levels depending on the task size is relevant. The algorithm graph for finding the state probabilities of the multiphase mixed-type queuing system is derived. The characteristic parameters are determined: performance, downtime rate, and balancing rate. The example of calculating system performance is considered.

Keywords: cryptographic protocol; protocol implementation; multiphase mixed-type queuing system

1. Introduction

A cryptographic protocol is not only an algorithm but also a procedure for exchanging data between different network subscribers, associated with the message exchange at different levels. In this case, the problem of minimizing the number of protocol implementation levels depending on the task size is relevant. The algorithm graph for finding the state probabilities of the multiphase mixed-type queuing system is derived. The characteristic parameters are determined: performance, downtime rate, and balancing rate. The example of calculating system performance is considered.

2. Flow-charts of information security algorithms and their Markov models

Queuing models at the level of processing operations by the processor and input/output through external devices are considered. Operations are interpreted

at the assembly language level. Communication between processing devices and input/output devices provides for three modes:

1. Synchronous communication – Synchronous input and synchronous output with processing via the processor. This option includes the symmetrical encryption and decryption system based on a common key.

2. Synchronous and asynchronous communication - Some operations are performed synchronously while the rest of them are performed in asynchronous mode. Provision is made for subdividing the encryption and decryption process into two stages - public and secure - using keys of different security levels.

3. Asynchronous communication – All pairwise actions are performed in parallel. It is typical for dialogue systems.

From the perspective of security, the greatest depth of secrecy is provided in the first mode, when communication between users and the environment is only at the auxiliary level. Finally, extending the access environment to the object protected by the security system reduces its security level.

3. Multiphase object security model

Models of communication between Alice, Bob, and Trent [1] are considered, and their analogs are built based on queuing systems to quantitatively determine their characteristics.

Alice is represented as the first phase to generate requests. Requests come from Alice to Trent for forwarding to Bob with rate μ_1 and consist of transmitting messages to the receiver Bob through Trent. Trent encrypts the prepared messages using the secret key kd with rate μ_2 and transmits them to Bob who decrypts them with rate μ_3 using the secret key ke. The model structure is shown in Fig. 1. The given

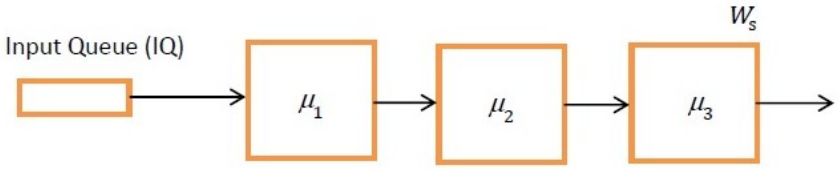


Fig. 1. Multiphase Information Security Model

parameters are used to establish the cryptosystem class [1]. Specifically, encryption and decryption in symmetric cryptosystems are carried out using the same secret key. The condition $kd = ke$ feature cryptosystems with a secret key. Due to the condition $kd \neq ke$ public-key cryptosystems are known as asymmetric cryptosystems.

The processor processing rate value is B_p, input rate – B_I, output rate – B_O. The specified values are measured in units of [words/sec]. The rate of the first and third phases for input/output.The rate of the first and third phases for input/output

$$\mu_i = \frac{B_i}{\theta_i}[words/sec], i = 1, 3$$

The input/output rate here is determined by datasheet specifications in units of [words/sec], a and the volume of work θ_i - by the input/output format and the amount of auxiliary processing per one 16-bit data word. The service rate in the second phase is determined by the speed of the processor and the processing work volume R per word.

$$\mu_2 = \frac{B_p}{R}[words/sec]$$

The models under consideration interpret the microprocessors' functionality at the instruction assembler level in the mode of word-by-word 16-bit information security [2]. The following indicators are used as the main target functions:

– System performance;
– Main equipment loading;
– Data protection level.

The security system performance W_s is determined by load factors η_i. The load factors are obtained through the phase downtime probabilities P_{0j} where $n = \overline{1,k}$ is the downtime probability of the phase j.

4. Markov model of noncurrent input and concurrent output with processing via the processor

State-of-art microprocessor-based tools offer various options to coordinate the equipment operating within a common interaction system. This primarily includes the system for ensuring the safety of data when the computer power is turned off. A number of control coordination modes based on Markov models of the computer systems' operation processes were considered in [4] for arrays of input/output data sets.

Multiphase models of input/output control and processing of data sets based on input data machine words are considered herein. The multiphase model structure Synchronous Decryption and Asynchronous Encryption (SDAE) [1] is shown in Fig.2. The diagram of the equipment operating when buffering management on allocated registers of shared memory. Obviously, there may be several options presented. Methods of the buffering option control whilst word by word information output

(that is, in the absence of saving buffered data during output in the event of a power outage) are considered.

Referred option corresponds to Bob's back transfer of the message about the results of processing Alice's message. The transmission corresponds to communication according to the scheme "synchronous reception over a secure channel from Bob, decryption by Trump and secure transmission to Alice."

In the models under consideration, one of the key functions is the interface data conversion for transfers between information input and output. This function, among others, is taken into account in the presented model when it is parameterized [3]

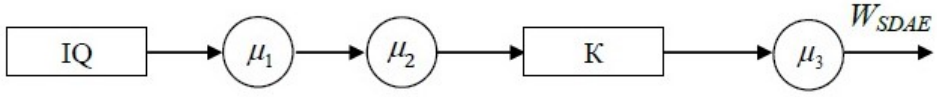


Fig. 2. SDAE Model

At the first stage, Trump receives the encrypted message from Bob with the rate μ_1 and proceeds to its decryption performed in the second phase with the rate μ_2. The decrypted message is asynchronously transmitted to Alice over the secure channel with the rate μ_3 (channel not shown).

The multiphase interpretation of the encryption processes under discussion is applied when used for the analysis of the apparatus of the Markov chain theory.

The first phase can be in one of three states $i = \{1, 0, \beta\}$:

$i = 1$ – Trump receives the encrypted message from Bob and proceeds to its decryption;

$i = 0$ – Trump locks reception of the next message;

$i = \beta$ – Locking while servicing the previous message in the third phase at the end of receiving the next message from Bob.

The second service phase can be in one of the following states: $j = \{1, 0, \beta\}$:

$j = 1$ – Trump decrypts the message received from Bob;

$j = 0$ – Locking decryption of the message received from Bob;

$j = \beta$ – Locking after the end of decryption of the message received from Bob while servicing the previous message in the third phase.

Finally, the third phase can be in one of two states: $h = \{1, 0\}$. The state $h = 1$ corresponds to Alice's receipt of the message from Trump through the channel agreed with him, and the state $h = 0$ indicates the absence of messages.

The graph of states and transitions is shown in Fig. 3. The model is represented by the multiphase queuing system with phases of input, processing, and output. The phases of input and processing are serviced in the mode of direct communica-

tion between the input device and the main memory, that is, they operate in the synchronous mode. Processing and output operate asynchronously through buffer memory.

In the diagram shown in Fig. 3, the buffer memory states are represented by the number of words that have been processed in the processor and are waiting for output.

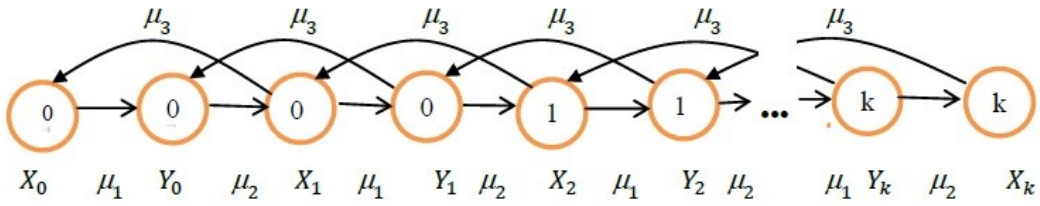


Fig. 3. Diagram of multiphase character-oriented service based on the principle of concurrent processing and output

5. Derivation of a system of balance equations for the Markov information security process using the SDAE algorithm

On the basis of assumptions about random values distribution law, the probabilities of states P_{ijk}(n) are introduced, where i, j and k are states of the 1st, 2nd and 3rd phases, respectively; $n = 0, 1, 2, \ldots, k$ is number of requests in the buffer. The first phase can be in the states $i = \{1, 0, \beta\}$. The state is servicing; $i = 0$ is locking the first phase while servicing in the second; $i = \beta$ the state of locking a phase by buffer occupancy. For the second phase, the state is determined by the set $j = \{1, 0, \beta\}$, where j = 1 is the same to the first phase; the state j = 0 is locking by the first phase, and the state $j = \beta$ is locking by buffer occupancy. The third phase can be in states k = 1,0, same as previous – in states of service and downtime waiting for a request. For the purpose of simplification, it is assumed that $X_n = P_{101}$ (n), $Y_n = P_{011}(n)$ for $n = 1, 2, ...k$; $X_0 = P_{100}$; $Y_0 = P_{010}$; $X_k = P_{101}(k)$ and $Y_k = P_{011}(k)$, $X_\beta = P_{\beta\beta1}$.The indicated designations are used in the structure illustrated in Fig. 2 and further in the text.

In the diagram, synchronous links between phases define unbuffered communication when service is allowed in one of the two phases. Asynchronous servicing in phases is connected by a finite capacity buffer k.

The original system of equations is not trivial, with an unambiguous derivation of the solution. Therefore, an algorithmic approach is presented for consideration.

In accordance with this approach, it is envisaged to derive a solution from the initial part, regular and final components of the original system of balance equations (1).

The initial part of this system is:

$$\left.\begin{array}{c} \mu_1 P_{100} = \mu_3 P_{101} \\ \mu_2 P_{010} = \mu_1 P_{100} + \mu 3 P_{011} \\ (\mu_1 + \mu_3) P_{101} = \mu_2 P_{010} + \mu_3 P_{101}(1) \\ (\mu_2 + \mu_3) P_{101} = \mu_1 P_{101} + \mu_3 P_{011}(1) \end{array}\right\} \quad (1)$$

The sum of the first two equations gives the equality as follows:

$$\mu_2 P_{010} = \mu_3 (P_{101} + P_{011}),$$

and the sum of the two subsequent equations implies the equality as follows:

$$\mu_2 P_{011} = \mu_3 (P_{011}(1) + P_{101}(1)),$$

With regard to the regular part, the representation of the balance equations is as follows:

$$\left.\begin{array}{c} (\mu_1 + \mu_3) P_{101}(n) = \mu_2 P_{011}(n-1) + \mu_3 P_{101}(n+1) \\ (\mu_2 + \mu_3) P_{101}(n) = \mu_1 P_{101}(n) + \mu_3 P_{011}(n+1) \end{array}\right\} \quad (2)$$

For $n = \overline{1, k-1}$ in the current representation we get the following:

$$\left.\begin{array}{c} (1 + \rho_1) X_n = \rho_2 Y_{n-1} + X_{n+1} \\ 1 + rho_2) Y_n = rho_1 X_n + Y_{n+1} \end{array}\right\} \quad (3)$$

where $\rho_1 = \dfrac{\mu_1}{\mu_3}$; $\rho_2 = \dfrac{\mu_2}{\mu_3}$. Using the previously derived relation for the probability P_{011} in pairs with the equations of the regular and closing parts, we obtain SDAE systems for this probability element $\rho_2 Y_{n-1}$ in accordance with the above equation gives the following:

$$\rho_2 Y_n = Y_{n+1} + X_{n+1}$$

The representation of the element $\rho_2 Y_{n-1}$ in accordance with the above equation gives the following:

$$\rho_2 Y_{n-1} = Y_n + X_n$$

that provides the following representation of the regular part of the system of balance equations in the following form

$$\begin{array}{c} (1 + \rho_1) X_n = Y_n + X_n + X_{n+1}, \\ (1 + \rho_2) Y_n = \rho_1 X_n + Y_{n+1} \end{array}$$

Solving the resulting system with respect to $(n+1)$-th terms, after some transformations we obtain the following system of equations for the regular part:

$$\left.\begin{array}{c} X_{n+1} = \rho_1 X_n - Y_n \\ Y_{n+1} = (1+\rho_2) Y_n - \rho_1 X_n \end{array}\right\} \quad (4)$$

Finally, the closing part equations of the balance system have the following form:

$$\begin{array}{c} (\mu_1 + \mu_3) P_{101}(k) = \mu_2 P_{011}(k-1) + \mu_3 P_{\beta\beta 1} \\ (\mu_2 + \mu_3) P_{011}(k) = \mu_1 P_{101}(k) \\ \mu_3 P P_{\beta\beta 1} = \mu_2 P_{011}(k) \end{array}$$

In the above designations, this part is represented as follows:

$$\left.\begin{array}{c} (\rho_1 + 1) X_k = \rho_2 Y_{k-1} + X_\beta \\ (\rho_2 + 1) Y_k = \rho_1 X_k \\ X_\beta = \rho_2 Y_k \end{array}\right\} \quad (5)$$

Thus, the main components of the recurrent system of balance equations for the system with the partial concurrent operations of information security are derived.

The system for calculating the characteristics of the model under consideration is built based on the previous material.

The basic parameters include the initial values of the variables of the equations system X_0, Y_0, the buffer capacity k, and the values of the service parameters μ_i. Performance characteristics include system performance and phase load factors.

Performance determines the stream rate of requests served by the system starting from the arrival of a request to the system input, more specifically from sending Bob the answer. It is measured in the number of words per second at the system output and is determined by the condition of the balance of the system.

$$W_{SDAE} = \eta_i \mu_i, i = 1, 3,$$

where i are phase numbers, η_i are phase load factors. They are determined by the simulation results for specific purposes and depend on the conditions of the system functioning

$$(1 - P_{010} - P_{011}(m) - P_{\beta\beta 1}) = (1 - P_{100} - P_{101}(m) - P_{\beta\beta 1} = (1 - P_{100} - P_{010} \quad (6)$$

The system of balance equations is given below.

$$\left.\begin{array}{c}\mu_1 P_{100} = \mu_3 P_{101} \\ \mu_2 P_{010} = \mu_3 (P_{101} + P_{011}) \\ \mu_2 P_{011} = \mu_3 (P_{011}(1) + P_{101}(1)) \\ \mu_1 P_{101}(n) = \mu_3 P_{011}(n) + \mu_3 P_{101}(n+1)) \\ (\mu_2 + \mu_3) P_{011}(n) = \mu_1 P_{101}(n) + \mu_3 P_{011}(n+1), n = \overline{1, k-1} \\ (\mu_1 + \mu_3) P_{101}(k) = \mu_2 P_{011}(k-1) + \mu_3 P_{\beta\beta 1} \\ (\mu_2 + mu_3) P_{011}(k) = \mu_1 P_{101}(k) \\ \mu_3 P_{\beta\beta_1} = \mu_2 P_{011}(k)\end{array}\right\} \quad (7)$$

When presented in a more convenient expression the indicated system has the form for the case k=4

$$\begin{array}{c}\mu_1 X_0 = \mu_3 X_1; \\ \mu_2 Y_0 - \mu_1 X_0 + \mu_3 Y_1; \\ (\mu_1 + \mu_3) X_1 = \mu_2 Y_0 + \mu_2 X_2; \\ (\mu_2 + \mu_3) Y_1 = \mu_1 X_1 + \mu_3 Y_2; \\ (\mu_1 + \mu_3) X_2 = \mu_2 Y_1 + \mu_3 X_3; \\ (\mu_2 + \mu_3) Y_2 = \mu_1 X_2 + \mu_3 Y_3; \\ (\mu_1 + \mu_3) X_3 = \mu_2 Y_2 + \mu_3 X_4; \\ (\mu_2 + \mu_3) Y_3 = \mu_1 X_3 + \mu_3 Y_4; \\ (\mu_1 + \mu_3) X_4 = \mu_2 Y_3 + \mu_3 X_\beta; \\ (\mu_2 + \mu_3) Y_4 = \mu_1 X_4; \\ \mu_3 X_\beta = \mu_2 Y_4.\end{array}$$

For this system, an example has calculated. The conditions are the parameters of the service rates $\mu_i = 2$ the buffer capacity k.According to the initial conditions, we have $\rho_1 = \rho_2 = 1$. From the first equation $X_1 = X_0$. . From the second equation we deduce$Y_1 = Y_0 - X_0$, then $X_2 = 2X_0 - Y_0$ and $Y_2 = 2Y_0 - 3X_0$., and from now in due order. See Table 1. The table shows the procedure for calculating the particular case of the system with partial concurrent service operations and the procedure for finding the probabilities of states. The required characteristics are determined by probabilities.

REFERENCES

1. Mao, Wenbo. Modern Cryptography: Theory and Practice: Transl. from Eng. – M.: Publishing house "Williams", 2005 – p. 768

Fig. 4. Table 1

	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4	X_β
X_0	$Y_0 - X_0$	$2X_0 - Y_0$	$2Y_0 - 3X_0$	$5X_0 - 3Y_0$	$5Y_0 - 8X_0$	$13X_0 - 8Y_0$	$13Y_0 - 21X_0$	$34X_0 - 21Y_0$
$X_\beta = Y_4$					$Y_4 = X_4$			$55X_0 = 34Y_0$
$34X_0 - 21Y_0 = 3Y_0 - 21X_0$					$26Y_0 - 42X_0 = 13X_0 - 8Y_0$			$55X_0 = 34Y_0$

2. V.Y. Khartov. AVR microcontrollers. Workshop for beginners. – M.: Publishing house of Moscow State Technical University named after N.E. Bauman, 2012 – p. 280
3. V.M. Vishnevsky, A.I. Lyakhov, S.L. Portnoy, I.V. Shakhnovich. Broadband wireless information transmission networks. - M .: "Technosphere", 2005 – p. 592
4. A.S. Yermakov. Performance evaluation models for synchronous/asynchronous execution of processes information input/output and processing. KazNU Bulletin, Almaty, No. 1 (60), 2009, pp.16-23.

УДК: 519.718

Оценка комплектов ЗИП для распределённой коммуникационной сети метеостанций минимальной конфигурации

Головинов Е.Э.[1], Аминев Д.А.[1,2], Татунов С. Ю.[3], Полесский С.Н.[3], Козырев Д.В.[2,4]

[1]ФГБНУ «ВНИИГиМ им. А.Н. Костякова», ул. Большая Академическая, 44 корпус 2, Москва, Россия
[2]Институт проблем управления им. В.А. Трапезникова РАН, ул. Профсоюзная, 65, Москва, Россия
[3]Национальный исследовательский университет «Высшая школа экономики», ул. Мясницкая, д. 20, 101000 Москва, Россия
[4]Российский университет дружбы народов, ул. Миклухо-Маклая, д. 6, Москва, 117198, Россия

evgeny@golovinov.info, aminev.d.a@ya.ru, spolessky@hse.ru, bestdk2@gmail.com, kozyrev-dv@rudn.ru

Аннотация

Рассматривается топология распределённой коммуникационной сети метеостанций (РКСМ) минимальной конфигурации и структура её аппаратуры. Сформулированы критерии работоспособности РКСМ. Рассмотрены теоретические основы и методика оценки комплектов ЗИП. По данным эксплуатационных интенсивностей отказов составных элементов проведена оценка комплектов ЗИП для составных частей аппаратуры РКСМ минимальной конфигурации. Исследованы зависимости коэффициента готовности от времени восстановления и от количества запасных частей.

Ключевые слова: надежность, агрометеопараметры, метеостанция, резервирование, интенсивность отказов, работоспособность, запасные части.

1. Введение

В реализации программ цифровизации сельского хозяйства важной задачей является оснащение сельскохозяйственных полей автоматическими средствами мониторинга агрометеопараметров, таких как температура, влажность приземного и слоя атмосферы, осадков, влажности почвы на различных уровнях и пр.

Работа выполнена при финансовой поддержке РФФИ в рамках научного проекта № 19-29-06043.

[1]. Метеостанции регистрируют такие агрометеопараметры и передают их на мониторинговые серверы для последующего анализа. Так как по этим агрометеопараметрам определяется расход воды, качество, эффективность и скорость выращивания агрокультур, и, следовательно урожайность в целом, их высокая важность диктует к метеостанциям жесткие требования по надежности [2, 3, 4] и времени восстановления работоспособности метеостанциии в случае отказа.

Из практики агрометеоизмерений для восстановления отказавшей в РКСМ одной метеостанции допустимо время до 1 дня. Затем эта метеостанция может быть отремонтирована в течение недели.

Для обеспечения бесперебойности работы РКСМ одним из способов обеспечения надежности являются комплекты ЗИП. Однако основная методологическая база для расчета показателей достаточности данных комплектов построена на оценке одиночных ЗИП или групповых ЗИП без учета возможности объединения всех ЗИП в единый склад, что позволит снизить затраты на поддержание по сравнению со стандартной схемой, когда на каждом объекте есть свой набор ЗИП. Основными источниками данных о методах расчета показателей ЗИП являются стандарт ГОСТ 27.507 – 2015 [5], учебник Ушакова И.А. [6], книга Черкесова Г.Н. [7], а также американский стандарт MIL-HDBK-472 [8]. Однако их использование не оптимально для расчета территориально распределенных ЗИП. Поэтому возникает необходимость создания методики для оценки комплектов ЗИП для территориально-распределенных систем, таких как РКСМ.

Минимальная конфигурация РКСМ может состоять из точки доступа мобильной связи и трех МС, две из которых ($МС_1$ и $МС_2$) соединены с точкой доступа напрямую, а одна, удалённая от точки доступа $МС_3$, через станцию $МС_2$ по WiFi (рис. 1а). Структурная схема аппаратуры МС представлена на рис. 1б.

Микроконтроллер (МК) принимает и обрабатывает данные от датчиков метеопараметров и GPS приемника, управляет передачей данных по GSM модему и модулю WiFi. В минимальной конфигурации достаточно двух датчиков: влажности почвы и влажности и температуры приземного слоя атмосферы. Телеметрия и данные местонахождения передаются посредством GSM модема через сеть Интернет на мониторинговый сервер для дальнейшей обработки и анализа. Наблюдать за местонахождением МС и считывать метеопараметры можно в любой момент времени. Источник питания (ИП) представляет собой типичную аккумуляторную батарею [2].

2. Методика оценки комплектов ЗИП для территориально-распределенных систем

ГОСТ 27.507–2015 описывает четыре типа стратегии пополнения ЗИП: периодическое пополнение, периодическое пополнение с экстренными доставками,

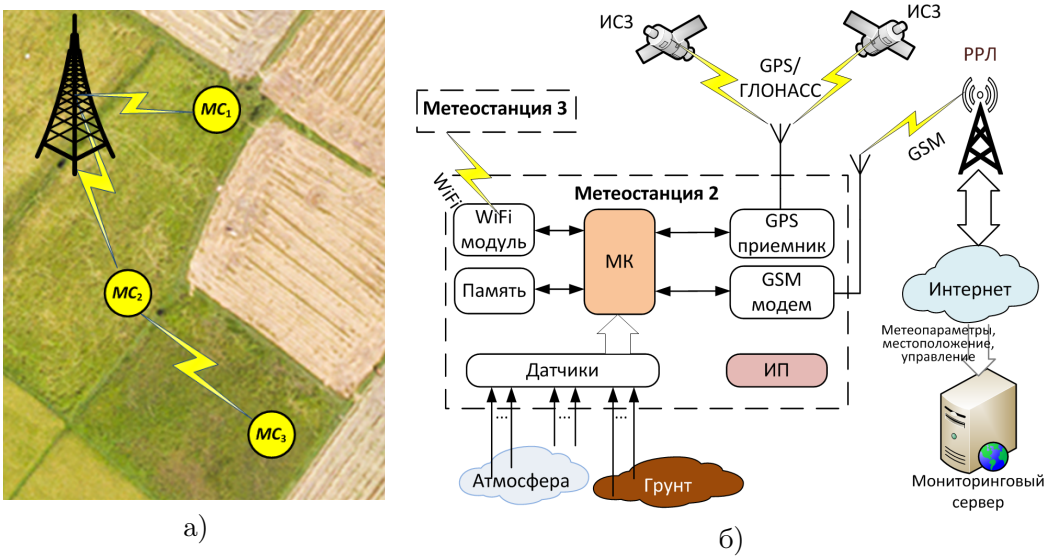


Рис. 1. РКСМ минимальной конфигурации на местности (а), структура аппаратуры МС2 (б)

непрерывное пополнение, пополнение по уровню неснижаемого запаса. По данному стандарту оценка запаса происходит в два шага: корректировка параметра T для всех одиночных ЗИП и итоговая оценка групповых и одиночных ЗИП. Основные показатели в данном стандарте указывают коэффициент готовности и среднее время задержки в удовлетворении заявок:

$$K_r = e^{-\sum_{i=1}^{N} R_i}, \tag{1}$$

$$\Delta t = \frac{\sum_{i=1}^{N} R_i}{\sum_{i=1}^{N} m_i \lambda_i}, \tag{2}$$

где R_i — показатель недостаточности запаса типа i, и способ его расчета зависит от стратегии.

К положительным сторонам метода, описанного в ГОСТ, необходимо отнести простоту вычислений и их скорость. Однако не все стратегии пополнения применимы к групповым запасам, групповые ЗИП полностью зависят от условия одинаковости изделий, отсутствуют методики резервирования запасов.

В [6] описывается схожий способ расчета, но с учетом ограничений: длительность безотказной работы/хранения/ремонта для каждого типа ЗИП распределена по экспоненциальному закону, элементы отказывают независимо друг от друга, во время простоя отказ невозможен. Помимо одиночных и групповых ЗИП

были введены ремонтные комплекты, а также две дополнительных непрерывных стратегии пополнения. В остальных деталях методы [6] аналогичны ГОСТу, но с дополнениями по типу большего количества стратегий пополнения и введения ремонтных комплектов чтобы учесть необходимость резервирования запаса. Однако введенные ограничения уменьшают спектр задач, которые можно решить, используя такую методологию.

В [7] надежность систем полностью обеспечивается с помощью ЗИП. Структуры ЗИП аналогичны ГОСТ, но с добавлением многоуровневых ЗИП — расширенных двухуровневых систем, где каждый следующий уровень обеспечивает запасом предыдущий. Введен новый показатель — вероятность достаточности запаса, описывающий вероятность пополнения запаса в заданном интервале. Рассматриваются дополнительные режимы, такие как режим непрерывного и режим многократного циклического применения, имеющие различные модели отказа запаса, а также расчет ЗИП для систем с применением резервирования. Однако для подхода характерна низкая универсальность — для каждой задачи необходимо формировать отдельное решение, что увеличивает сложность вычислений.

Американский военный стандарт MIL-HDBK-472 опирается на известные показатели надежности существующих деталей и систем, подразумевает использование системы складов для обеспечения достаточности: центральный, региональный и индивидуальный склады. Основная задача, рассматриваемая в стандарте — построение надежной системы с ограниченным или минимальным ЗИП, иначе говоря, основной упор идет на повышение показателей надежности отдельных систем, а не создание сложных систем ЗИП.

Из всех рассмотренных подходов, ни один не может быть использован в полной мере для расчета территориально-распределенной системы. Это связано с четкой зависимостью типа запасных частей, их стратегии пополнения и количества таких частей. В то время как в территориально-распределенной системе запасных частей одного типа может быть несколько, и каждая часть может пополняться по различным стратегиям. Поэтому оптимальным решением будет разработка новой методики для оптимизации ЗИП таких систем.

Подход, при котором для различных территориально-распределенных объектов используется единый запас (склад) позволяет уменьшить количество запасных частей по сравнению с подходом, использующим для каждого эксплуатационного объекта отдельный набор ЗИП. Подобный подход используется в территориально-распределенных системах (рис. 2) — системах, в которых надежность и работоспособность нескольких сильно разнесенных по местности эксплуатационных объектов поддерживается одним комплектом ЗИП. Кроме того, один и тот же тип запаса можно поставлять по разным стратегиям: пери-

одическое пополнение, с экстренными доставками, непрерывное. Несмотря на то, что внутри склада запас должен быть учтен для каждого объекта, нельзя забывать, что для каждого объекта могут быть одинаковые детали запаса и для расчета оптимального комплекта, необходимо учесть данное условие [9].

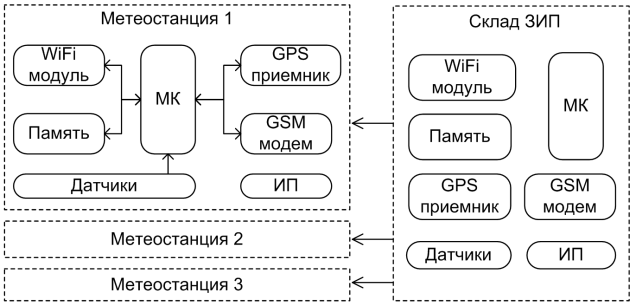


Рис. 2. Схема территориально-распределенной системы ЗИП (на примере РКСМ минимальной конфигурации)

При применении любой методики, необходимо определить какие входные параметры необходимы для расчета: требуемое значение показателя достаточности для проведения оптимизации, вид затрат на запас и их единица измерения, количество типов запаса, описание запаса каждого типа, точность вычисления, количество эксплуатационных объектов. Кроме того, у запасов есть свои параметры: количество частей определенного типа в изделии (m), интенсивность замены части, затраты на запасную часть (c), число заявок, в среднем поступающее на запас каждого типа за период пополнения (α), параметры, описывающие стратегию пополнения (T, β), начальный уровень запаса для каждого типа частей (n). Все показатели для формализации записи и удобства доступа, необходимо записать в таблицу, отдельно для каждого объекта.

После учета всех эксплуатационных объектов, необходимо разделить типы запаса, на те которые уникальны для каждого объекта и на общие. Общие части можно рассчитать по известной методике из ГОСТ [5], но для расчета уникальных для объекта частей, нужно действовать по новой методике:

1) Рассчитать показатель D_0 по формуле:

$$D_0 = \begin{cases} \Delta t^{\text{тр}} \sum_{i=1}^{N} m_i \lambda_i, \text{ если задано значение } \Delta t^{\text{тр}}; \\ -\ln K_{\text{г}}^{\text{тр}}, \text{ если задано значение } K_{\text{г}}^{\text{тр}}. \end{cases} \tag{3}$$

2) В зависимости от исходных значений считается показатель α:

$$\alpha_i = m_i \lambda_i T_i \tag{4}$$

$$\alpha_i = \lambda_i T_i. \tag{5}$$

3) В зависимости от стратегии пополнения запаса рассчитывается нулевой уровень запаса (n) при котором выполняется:

$$R_i \leq D_0. \tag{6}$$

a. Периодическое пополнение:

$$R_i = -\ln\left\{1 - \frac{1}{a_i}\Big[e^{-a_i}\sum_{\gamma=n_i+2}^{\infty}(\gamma - n_i - 1)\frac{a_i^{\gamma}}{\gamma!}\Big]\right\}. \tag{7}$$

b. Периодическое с экстренной доставкой:

$$R_i = -\ln\left\{1 - \frac{T_{\text{эд}\,i}}{T_i}\Big[e^{-a_i}\sum_{l=1}^{\infty}\sum_{\gamma=l(n_i+1)}^{\infty}\frac{a_i^{\gamma}}{\gamma!}\Big]\right\}. \tag{8}$$

c. Непрерывное пополнение:

$$R_i = -\ln\left[1 - \frac{a_i^{n+1}}{(n_i+1)!\sum_{\gamma=0}^{n+1}\frac{a_i^{\gamma}}{\gamma!}}\right]. \tag{9}$$

4) Определяется значение для каждого типа:

$$R_i(n_i^0 + 1). \tag{10}$$

5) Определяется значение для каждого типа:

$$\Delta_i = \frac{R_i(n_i^0) - R_i(n_i^0 + 1)}{c_i}. \tag{11}$$

6) Вычисляется оптимальный по затратам запас:

$$R_{\Sigma}^0 = \sum_{i=1}^{N} R_i. \tag{12}$$

Если условие $R_{\Sigma}^0 \leq D_0$ выполняется рассчитанные значения n являются оптимальными, если нет, то добавляется запасная часть того типа на которой условие не выполняется и расчет повторяется заново.

7) Коэффициент готовности для уникальных частей:

$$K_{\text{г}} = e^{-\sum_{i=1}^{N} R_i(n_i)}. \tag{13}$$

Далее проводим расчет для групп общих для каких-либо эксплуатационных объектов запасных частей. Основой для расчета этих групп является формула полной вероятности и теорема Байеса элементарной теории вероятностей,

которая позволяет определить вероятность какого-либо события при условии, что произошло другое статистически взаимозависимое с ним событие.Так как коэффициент готовности является вероятностью того, что нужная запасная часть в момент запроса имеется в комплекте ЗИП и одна и та же запасная часть может использоваться для нескольких эксплуатационных объектов, то общая вероятность является суммой вероятностей того, что нужная запасная часть имеется хотя бы для одного из поддерживаемых эксплуатационных объектов. Рассмотрим случай, когда имеется тип запасных частей сразу для двух эксплуатационных объектов. Допустим, имеется отдельно запас для одного и другого объекта. Тогда общая вероятность того, что во время запроса нужная запасная часть имеется в запасе складывается из вероятности того, что она имеется в обоих запасах, имеется в 1 и не имеется в 2 и имеется в 2, но не имеется в 1. Формула для двух эксплуатационных объектов имеет вид:

$$K_{\text{г общ}} = K_{\text{г }1} \cdot K_{\text{г }2} + K_{\text{г }1}(1 - K_{\text{г }2}) + K_{\text{г }2}(1 - K_{\text{г }1}), \tag{14}$$

где $K_{\text{г }1} \cdot K_{\text{г }2}$ — вероятность того, что оба запаса готовы; $K_{\text{г }1}(1 - K_{\text{г }2})$ — вероятность того, что готов один и не готов второй и $K_{\text{г }2}(1 - K_{\text{г }1})$ — вероятность того, что готов второй и не готов первый.

Однако, если эксплуатационных объектов больше двух, то и формула общей вероятности увеличивается. При этом появляется возможность рассчитать коэффициенты готовности для первого и второго объекта с учетом различных времен доставки и стратегий пополнения. Соответственно оптимизация проводится таким образом, чтобы итоговый общий коэффициент готовности был равен или близок к заданному. То есть итерационно добавляется одна запасная часть, проводится расчет коэффициента готовности по алгоритму, приведенному выше, проводится расчет общего коэффициента готовности по формуле общей вероятности и происходит сравнение с заданным коэффициентом. Добавление запасной части происходит для того типа, для которого показатель Δ_i является наибольшим. Таким образом обеспечивается оптимальное по затратам добавление запасных частей в комплект. Чтобы получить оптимальный по затратам комплект, необходимо достичь необходимого произведения коэффициентов готовности уже оптимизированного комплекта уникальных запасных частей и общих запасных частей. Если он равен заданному, то получен оптимизированный комплект для территориально-распределенной системы.

Итоговый коэффициент готовности системы — это произведение коэффициентов готовности уникальных частей и коэффициентов готовности общих групп. Полученный коэффициент должен быть с заданной точностью близок к заданному в задаче и тогда полученный комплект запаса является искомым.

3. Оценка комплектов ЗИП для РКСМ минимальной конфигурации

Используя описанную методику, составляется экземпляр таблицы (таблица 1) для эксплуатационного объекта распределенной сети [3] (метеорологической станции).

Название ЗЧ	i	m, шт. для МС	m, шт. для РКСМ	$\lambda \cdot 10^{-6}$, 1/ч	c, ед.затрат	α, стратегия пополнения	T, часы
Микроконтроллер	1	1	3	0.02	2000	3	168
Память RAM	2	1	3	0.03	1000	3	168
Датчики	3	1	3	0.03	1200	3	168
GPS модулей	4	1	3	5	800	3	168
GSM модуль	5	1	3	10	800	3	168
WiFi модуль	6	1	3	10	800	3	168
ИП	7	1	3	1	50	3	168

Таблица 1. Типы и параметры запасов РКСМ

Необходимо достичь коэффициента готовности $K_{\text{г}} = 0,999$. Затраты представлены в виде стоимости в рублях, начальное количество запасных частей равно нулю. Так как в предложенной минимальной конфигурации из трех станций нет уникальных частей, то комплект ЗИП-О всех трех частей объединяется без изменений.

При расчете каждой строки получаем следующий набор запасных частей (расчетный запас): GPS модулей (1 шт.), GSM модуль (1 шт.), WiFi модуль (1 шт.), ИП (1 шт.).

Итоговое значение:

$$K_{\text{г общ}} = K_{\text{г }1} \cdot K_{\text{г }2} \cdot K_{\text{г }2} = 0,9999311144362.$$

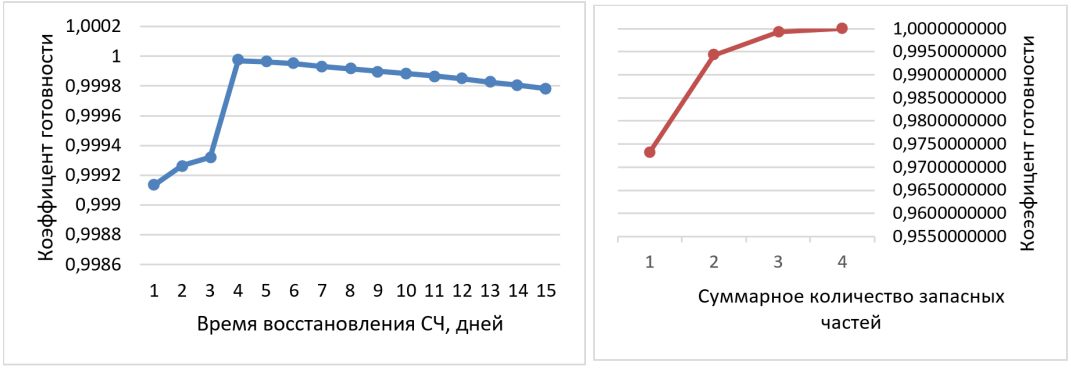
Результаты расчёта показателей достаточности для комплекта ЗИП РКСМ раскрыты в таблице 2.

Полученное значение удовлетворяет предъявленным требованиям, однако при использовании выбранной методики можно обнаружить, что предъявленные требования можно выполнить и с меньшим количеством затрат на запасные части и меньшим ожиданием восстановления (рис. 3).

Как видно из графика, для обеспечения требуемого времени восстановления в семь дней необходимо 4 запасных части, однако в случае необходимости, можно снизить время восстановления до одного дня без понижения показателя готовности ниже заданного предела, а также возможен выбор между непрерывной стратегией пополнения и стратегии пополнения с периодическими доставками.

Показатели достаточности	Расчетные значения для стратегии непрерывного пополнения (1 день)	Расчетные значения для стратегии непрерывного пополнения (7 дней)	Расчётные значения для стратегии с периодическими доставками (7 дней)	Требуемые значения
Среднее время задержки в удовлетворении заявок на ЗЧ, Δt, ч.	4,68067265	0,88046954	0,379668747	12,78758095
Коэффициент готовности, $K_г$, отн.ед.	0,99963385	0,9999311	0,999970295	0,999
Суммарный уровень недостаточности для n, $\sum R(n)$, отн.ед.	0,00036622	0,000068888	0,000029705	0,0010005
Суммарный уровень недостаточности для $n+1$, $\sum R(n+1)$, отн.ед.	0,000000065	0,000000045	0,0000000114	Не задано
Суммарные затраты для комплекта	1650	2450	2450	Не задано
Суммарное количество запасных частей в комплекте, шт.	3	4	4	Не задано

Таблица 2. Показатели достаточности для комплекта ЗИП РКСМ



а) б)

Рис. 3. Зависимость коэффициента готовности от времени восстановления (а) и от количества запасных частей (б)

4. Заключение

Предложенная методика по оценке комплекта ЗИП показывает лучшие результаты по сравнению с рассмотренными ранее методиками за счет учета

особенностей территориально-распределенных систем. Использование методики возможно как в ручном, так и в автоматизированном расчете.

Полученное в результате оценки количество запасных частей необходимо для достижения максимального коэффициента готовности комплекта, добавление других запасных частей будет избыточным. Итоговый $K_{\text{г общ}}$, рассчитанный по формуле (14) составил $0,9999311$, что удовлетворяет требованиям, предъявляемым к аппаратуре такого класса.

Использование полученного в ходе оценки комплекта ЗИП для РКСМ минимальной конфигурации позволит повысить её надёжность и обеспечить восстановление работоспособности в случае неисправности за 24 часа.

ЛИТЕРАТУРА

1. Бородычев В.В., Лытов М.Н, Головинов Е.Э. Мониторинг и управление орошением в режиме реального времени: монография. – М.: МЭСХ, 2017. – 154с. ISBN: 978-5-9909008-9-9
2. Головинов Е.Э., Аминев Д.А., Бакиров Ш.М. Анализ элементной базы для реализации мобильного измерительного агрометеокомплекса // Проектирование и технология электронных средств – Владимир: №3, 2017. С. 33-40.
3. Головинов Е.Э., Аминев Д.А., Козырев Д.В., Ларионов А.А. Модель надёжности распределённой коммуникационной сети метеостанций минимальной конфигурации // Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь (DCCN-2019) материалы XXII Международной научной конференции. Российский университет дружбы народов. 2019. С. 484–491. eLIBRARY ID: 41384277
4. Kozyrev D.V., Phuong N.D., Houankpo H.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, vol 1141, 2019. Springer, Cham. Pp.646-656. DOI: 10.1007/978-3-030-36625-4_52
5. ГОСТ 27.507 –2015. Надежность в технике. Запасные части, инструменты и принадлежности. Оценка и расчет запасов, 2015.
6. Ушаков И.А. Курс теории надежности систем // Учебное пособие. М.: Дрофа, 2008. - 239 с.
7. Черкесов И.А. Оценка надёжности систем с учётом ЗИП // СПБ.: БХВ-Петербург, 2012. - 480 с.
8. MIL-HDBK-472 Maintainability Prediction // SESS (Systems Engineering Standards and Specifications), Validation Notice 2 - May 26, 2020.
9. Liu, R.-Y.,Li, Q.-M.,Wang, S.,Li, H. Analytical algorithm of spare demand for voting system of any life distribution units // Systems Engineering and Electronics, 2016 - P.714–718.

УДК: 004.75

Беспроводная сенсорная сеть для интенсивного сбора данных на основе технологии LoRaWAN и распределенного алгоритма сжатия информации

Юрий Рассадин[1] and Сергей Душин[1]

[1]Институт проблем управления имени В.А. Трапезникова РАН, Москва, Россия

rassadin@ipu.ru, s.dushin@inbox.ru

Аннотация

В работе предлагается способ построения беспроводной сенсорной сети на основе технологии LoRaWAN, позволяющий повысить интенсивность сбора данных по сравнению со стандартными реализациями таких сетей и сохранить энергоэффективность беспроводных автономных датчиков. Основной особенностью сети, позволяющей достигнуть этих результатов, а также оптимизировать использование физического канала связи, является использование алгоритма распределенного сжатия данных в сетях LPWAN.

Эффективность предложенного решения исследована на примере сети температурных датчиков, используемых для интенсивного измерения температуры в задаче идентификации топологии тепловой системы офисного здания. Сеть построена на основе программного обеспечения с открытым исходным кодом, в частности, серверного стека протоколов ChirpStack, стека LoRaWAN end-device от Semtech, и разработанных авторами дополнительных модулей.

Ключевые слова: беспроводные сенсорные сети, высокоинтенсивный сбор данных, микроклимат зданий, LoRaWAN

1. Введение

Порой кажется, что современные технологии проникли во все сферы нашей жизни. Тем сильнее ощущается контраст с реальным уровнем технологической оснащенности большинства эксплуатируемых на сегодня зданий, ведь внедрение в них современных технологий автоматизации и интеллектуального управления представляет из себя весьма трудоемкую задачу. В частности, при автоматизации энергетических, тепловых и климатических систем зданий или модернизации уже существующих автоматизированных систем, возникают различные трудности, связанные с тем, что необходимая функциональность не закладывалась на

этапе проектирования здания. Типовыми проблемами являются обеспечение питания элементов системы, наличие и доступ к кабельным линиям связи, а также непосредственная достижимость ключевых узлов системы. При этом зачастую отсутствует точная схема объекта, а задачи эксплуатационных служб решаются в условиях неопределенностей и изменяющихся параметров, что влияет на эффективность принимаемых решений.

В случае невозможности подключения к надежным сетям питания и отсутствия кабельной инфраструктуры сети связи, безальтернативным является использование беспроводных автономных датчиков. Для подобных устройств важнейшей проблемой является обеспечение энергоэффективности, что во многом исключает применение высокоскоростных технологий беспроводной передачи информации, например стандартов семейства IEEE 802.11 (Wi-Fi). Существующие решения, базирующиеся на современных беспроводных LPWAN технологиях, таких как LoRaWAN, SigFox и NB-IoT, довольно эффективно решают типовые задачи сбора данных и управления оборудованием [1]. Общим минусом применения LPWAN технологий для построения беспроводной сенсорной сети является низкая пропускная способность используемых каналов связи и ограничения на интенсивность обмена данными, направленные на обеспечение приемлемого уровня времени работы датчиков от батареи [2, 3].

В реальных же приложениях Интернета Вещей зачастую требуется собирать значительно больше данных, нежели могут обеспечить современные LPWAN технологии, в настоящее время к таким проблемам проявляется научный интерес [4]. Ещё одним интересным примером такой задачи является обеспечение высокочастотного измерения температуры тепловых труб, которое необходимо для идентификации топологии теплосети при помощи тепловых волн. В частности, для регистрации изменения температуры необходима система сбора, которая обеспечивает регистрацию температуры с частотой не ниже 1Гц.

В данной работе мы рассматриваем способ обеспечения высокоинтенсивного сбора данных с беспроводных автономных датчиков, позволяющий решить обозначенную задачу интенсивного измерения температуры для идентификации топологии тепловой подсистемы здания, а также во множестве других приложений, требующих интенсивного сбора данных климатических датчиков.

Дальнейшее изложение имеет следующую структуру. Во втором разделе описана основная идея работы, алгоритм сжатия данных при коммуникации периферийных устройств и центрального сервера, который позволяет применить LoRa-канал при повышенных требованиях по частоте опроса. Третья часть посвящена описанию архитектуры сети, серверной части и программного обеспечения оконечных устройств. В четвертом разделе рассматриваются посвящен вопро-

сы развертывания сети, описанию оборудования и направлениям дальнейших исследований.

2. Способ повышения интенсивности сбора данных

Как и в традиционных сетях связи, мощным инструментом повышения эффективности использования канала связи и оптимизации энергопотребления является применение алгоритмов сжатия информации [5]. Однако применение в LPWAN сетях традиционных подходов к сжатию данных приводит к неприемлемому увеличению времени задержки данных и повышению энергопотребления на стороне датчика, так как требуется дополнительная обработка данных. В связи с этим, как правило, информация передается в несжатом виде, а при разработке алгоритмов сжатия данных для LPWAN должны учитываться требования и специфические условия, необходимые для работы в этих сетях.

Суть предлагаемого подхода состоит в синхронном предсказании измеряемого сигнала как на стороне сервера, так и на стороне датчика [6]. При этом информация с датчика посылается при превышении ошибки предсказания заданного порога и используется для корректировки предсказания на сервере. В случае, если датчик не посылает корректирующий отсчет, сервер считает, что отсчет предсказан достаточно точно и использует его как информацию, полученную в реальном времени. Предсказание осуществляется при помощи алгоритма линейного прогнозирования на основе рекурсии Левинсона-Дурбина. Важно также, что коэффициенты фильтра предиктора [7, 8] рассчитываются на стороне сервера для экономии заряда батареи и передаются на датчик в случае превышения порогового значения ошибки. Схематично данный алгоритм представлен на рисунке 1.

В этих условиях минимизируется использование физического передатчика (наиболее энергопотребляющая функциональная часть автономного датчика), а сервер использует свою оценку как данные в реальном времени до тех пор, пока не придет корректирующая информация от датчика.

3. Архитектура сети

Так же как и стандартная архитектура LoRaWAN сети [9], модернизированная система имеет серверную часть и клиентскую часть. Основные функциональные блоки представлены на рисунке 2.

Серверная часть состоит из стандартных LoraWAN компонентов (network bridge, network server, application server) и дополнительных модулей, обеспечивающих сжатие информации, повышение энергоэффективности датчиков и эффективность использования канала. Каждый интервал времени модуль формирования отсчетов реализует предсказание отсчетов, в случае если данные

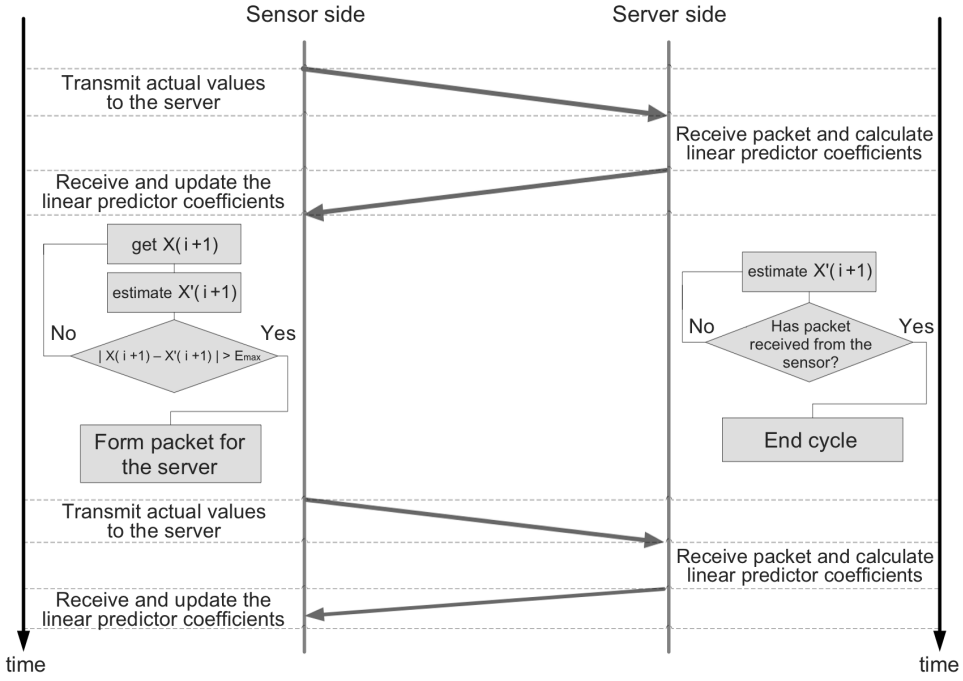


Рис. 1. Алгоритм распределенного сжатия данных для LPWAN

от датчика не пришли, либо пересылку данных, поступивших непосредственно от датчика. В случае поступления данных от датчика и, соответственно, превышения заданного порога ошибки, модуль пересчета коэффициентов фильтра предиктора формирует новые коэффициенты фильтра согласно алгоритму Левенсона-Дурбина. Модуль управления передачей обеспечивает пересылку коэффициентов фильтра предиктора на датчик, а также прием и обработку пакетов от датчика.

На стороне датчика реализуются необходимые для работы в LoRaWAN физический и канальный уровни, а также дополнительные модули. В частности, модуль предсказания рассчитывает прогнозное значение по формуле, идентичной той, которая используется на сервере. Далее предсказанное значение сравнивается с результатом реальных измерений. Как известно, термисторы нелинейно меняют свою характеристику в зависимости от температуры. Поэтому значения температуры хранятся в виде табличной функции, а промежуточные значения вычисляются линейно, $t(r) = t_1 + (t_2 - t_1)(r - r_1)/(r_2 - r_1)$. По таблице организован бинарный поиск, в качестве нулевой итерации использовано сопротивление,

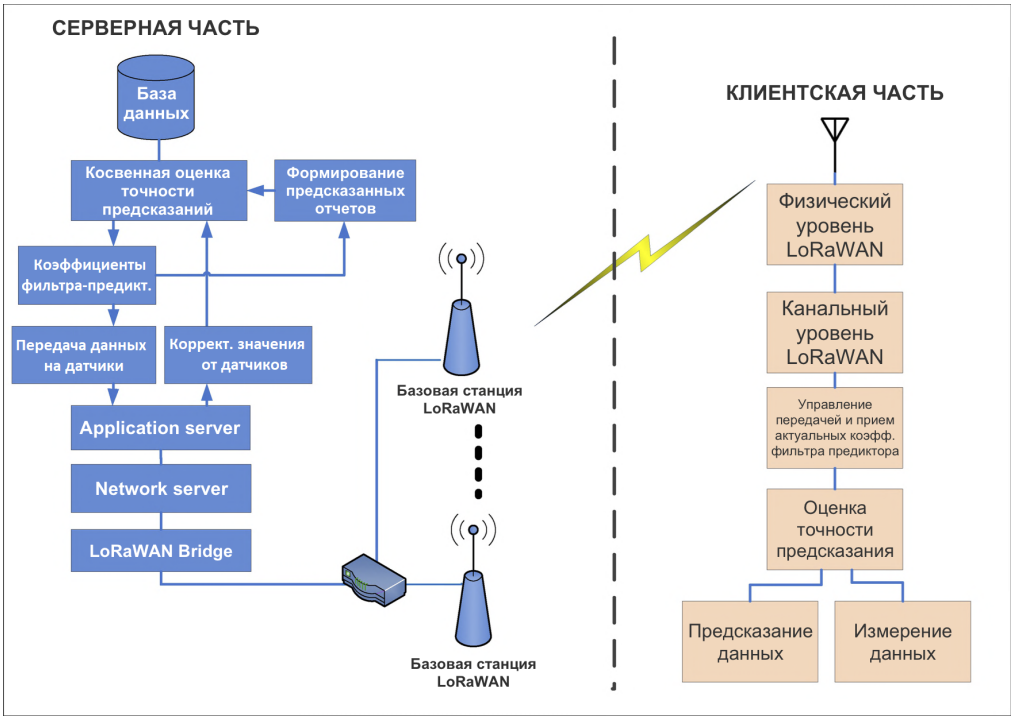


Рис. 2. Модифицированная структура LoRaWAN сети

соответствующее температуре 25°C. Модуль управления обеспечивает отправку данных, если ошибка предсказания [5] больше, чем установленный порог, а также прием коэффициентов фильтра предиктора от сервера и их обновление.

4. Реализация сети для задачи интенсивного измерения температуры тепловых труб в произвольных узлах

Процесс построения описываемой сети можно условно разделить на три этапа. Первый заключался в выборе оборудования, способного решить стоящие перед авторами задачи. На втором этапе создавалось программное обеспечение для оконечных устройств и для серверной части, тестирование и отладка их взаимодействия. К третьему, завершающему этапу следует отнести работы по разворачиванию сети в конкретном здании, а именно: масштабирование результатов второго этапа на здание целиком, обеспечение надежного покрытия, пробные запуски.

Необходимым оборудованием являются сервер, базовые станции и сетевая инфраструктура для них, а также температурные датчики. В качестве централь-

ного узла, мы использовали виртуальную машину с операционной системой OpenSUSE. Базовые станции выбирались соответствующими российским региональным стандартам. Доступным решением является шлюз MikroTik R11e-LR8 на базе процессора Semtech SX1301. В процессе разработки мы также использовали шлюз Вега БС-1, который впоследствии будет интегрирован в рабочую сеть. Температурный датчик, используемый в проекте, управляется микроконтроллером семейства STM32L151 [10], который заявлен Semtech как энергоэкономичный. Безальтернативным был выбор производителя LoRa-передатчика, мы остановились на Semtech SX1272 [11]. Модуль измерения температуры представляет собой делитель напряжения с терморезистором с отрицательным температурным коэффициентом B57861. В качестве питающего элемента может использоваться батарея напряжением от 3.3V до 5V (в нашем случае – 3.6V).

4.1. Серверная часть . Для построения серверной части использован открытый программный комплекс ChirpStack, вычислительный модуль алгоритма сжатия реализуется средствами предоставляемого API. В качестве языка программирования мы используем Python, на котором написан фильтр-предиктор, а также WebSocket клиент и сервер. Компоненты ChirpStack, gateway bridge, network server и application server мы разместили на центральной серверной машине для того, чтобы переложить на нее основную вычислительную нагрузку в соответствии с основной идеей работы.

4.2. Клиентская часть. При создании клиентской прошивки мы опирались на среду разработки и библиотеки от компании Semtech, находящиеся в открытом доступе. Компания предоставляет полную свободу разработчикам, в том числе в области коммерческого использования получаемых результатов. Разработка прошивки датчика велась на языке C в IDE-среде ac6 System Workbench for STM32 на основе открытого проекта LoRaWAN end-device stack. Энергоэффективность достигается не только за счет сокращения количества сеансов передачи данных, но и управлением питания периферийных устройств. Модем SX1272 способен самостоятельно управлять высокочастотным модулем передатчика, но направленность на максимальную экономию энергии продиктовала решение управлять им напрямую через микроконтроллер.

4.3. Схема размещения оборудования. По зданию должны быть распределены два типа устройств, датчики и базовые станции. Выбор в пользу протокола LoRaWAN был обусловлен большой свободой и удобством в размещении оконечных устройств, поэтому размещение датчиков мы считаем свободным, зависящим от конечной задачи. В случае идентификации тепловой сети место размещение датчиков – трубы отопления в различных помещениях, расположенные преимущественно около окон. В протокол LoRaWAN заложен механизм управления несколькими шлюзами, подключенными к одному сетевому серверу,

поэтому самой сложной задачей при развертывании подобной сети является поиск неизбыточного расположения базовых станций. Здание, где разворачивалась опытная сеть, вытянутое, поэтому мы решили располагать базовые станции у окон, покрывая сетью протяженные фасады здания раздельно. Дополнительную базовую станцию мы разместили в подвале здания, обеспечивая прием в местах входа, выхода и первичного разветвления тепловых сетей.

5. Результаты и направление дальнейших исследований

В настоящее время реализован и опробован интенсивный сбор данных с небольших групп датчиков. Следующим шагом должна стать разработка гарантированного метода масштабирования таких сегментов до требуемых размеров. Мы хотим ориентироваться на объемы, заявляемые в рекламных проспектах, несколько сотен датчиков одновременно. Используя открытое ПО от ChirpStack и Semtech, удается обойти ограничения по частоте опроса датчиков, присущее коммерческим решениям, но развитие сети продолжается. Одним из направлений развития может быть совершенствование алгоритмов прогнозирования с использованием математических моделей исследуемых объектов. Если для нужд некоторых экспериментов мощности центрального сервера будет недостаточно, серверную часть можно перенести на более мощную машину, но до настоящего времени в этом не было необходимости. Более глубокое изучение серверного ПО от ChirpStack, мы надеемся, позволит оптимизировать предлагаемые подходы.

6. Заключение

В работе представлена модифицированная архитектура беспроводной сенсорной сети, позволяющая повысить интенсивность сбора данных по сравнению со стандартными реализациями LoRaWAN сетей. Основной функциональной особенностью сети, позволяющей достигнуть повышения интенсивности сбора данных при сохранении энергоэффективности на приемлемом уровне для автономной работы датчиков и экономного использования физического канала связи, является использование предложенного ранее алгоритма распределенного сжатия данных в сетях LPWAN. Алгоритм основан на предсказании данных на сервере, тогда как датчик не передает весь массив данных, а передает только корректирующую информацию в случае превышения заданной заранее ошибки предсказания. Подобный подход позволяет минимизировать использование передатчика автономных датчиков, что позволяет снизить нагрузку на физический канал связи. В свою очередь, это может понизить энергопотребление датчиков, продлив срок их службы до замены элементов питания. Вклад авторов в программную разработку сети лежит на уровне серверных приложений и прошивки датчиков устройств. Эффективность предложенного подхода будет исследована

при решении задачи интенсивного сбора данных температурных датчиков в задаче идентификации топологии тепловой системы офисного здания.

ЛИТЕРАТУРА

1. Mekki K., Bajic E., Chaxel F., Meyer F. Overview of Cellular LPWAN Technologies for IoT Deployment: Sigfox, LoRaWAN, and NB-IoT / 2nd IEEE International Workshop on Mobile and Pervasive Internet of Things - Athens. - 2018.
2. N. Pukrongta and B. Kumkhet, The relation of LoRaWAN efficiency with energy consumption of sensor node, 2019 International Conference on Power, Energy and Innovations (ICPEI), Pattaya, Chonburi, Thailand, 2019, pp. 90-93
3. N. I. Bazenkov et al., "Intensive data collection system for smart grid and smart building research,"2019 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA), Lipetsk, Russia, 2019, pp. 411-415.
4. S. Benninger, M. Magno, A. Gomez and L. Benini, "EdgeEye: A Long-Range Energy-Efficient Vision Node For Long-Term Edge Computing,"2019 Tenth International Green and Sustainable Computing Conference (IGSC), Alexandria, VA, USA, 2019, pp. 1-8.
5. Haykin, S. Adaptive filter theory (5-th edition) / S. Haykin — Prentice Hall, 2014. — 936p.
6. Jackson, L.B., Digital Filters and Signal Processing, Second Edition, Kluwer Academic Publishers, 1989. pp.255-257.
7. Dushin S.V., Frolov S.A. Distributed data compression algorithm for low-power wide-area networks. DCCN-2019
8. Official site of Semtech Corporation, https://www.semtech.com/uploads/documents/SX1272_DS_V4.pdf
9. Hasan A.H., Grachev A.N. On-Line Parameters Estimation Using Fast Genetic Algorithm // J. of Electrical and Control Engineering (JECE). –2014. – Vol. 4, No. 2. – P. 16–21.
10. https://www.st.com/en/microcontrollers-microprocessors/stm32l151-152.html
11. https://www.semtech.com/products/wireless-rf/lora-transceivers/sx1272

UDC: 004.052

Cesaro-heredity property in the shift register family

Sergey Yu. Melnikov and Konstantin E. Samouylov

Peoples' Friendship University of Russia (RUDN University)
6 Miklukho-Maklaya St, 117198, Moscow, Russia

melnikov@linfotech.ru, ksam@sci.pfu.edu.ru

Abstract

Non-autonomous binary automata from three classes are considered: shift registers, generalized shift registers, shift registers with internal XOR. We study the cesaro-heredity property of automata from these classes, that is, their ability to inherit the property of stability of relative word frequencies in growing initial segments of the input sequence. It is shown that shift registers always have this property. Conditions are obtained under which generalized shift registers and shift registers with internal XOR do not have this property.

Keywords: statistical properties of automata, cesaro sequences, shift register.

1. Introduction

Various randomness models are used to study the statistical properties of sequences generated by pseudo-random sequence generators. One of the possible models is the so-called cesaro sequences.

In [1] the issues were considered related with automaton transformations of cesaro sequences, that is, sequences that possess the property of stability of the relative frequencies of an arbitrary word in growing initial segments. An automaton is called a cesaro-hereditary if for any initial state it converts cesaro sequences in the input alphabet into cesaro sequences in the output alphabet. Ibidem sufficient conditions were obtained under which the automaton possesses and does not possess the property of cesaro-heredity.

In this paper we specify these results for three classes of automata: shift registers [2], generalized shift registers [3], shift registers with internal XOR [4]. Machines from these classes can be used as components of pseudo-random sequence generators [5]. We will consider only automata with binary input and output alphabets.

We will reformulate the basic definitions of [6] for the binary case.

The work is partially supported by RFBR grant No.16-01-20379.

Let $\{0,1\}^*$ – be the set of all binary words. Let $\Omega = \{\omega = w_1 w_2 ... | w_t \in \{0,1\}, t = 0,1,...\}$ – be the set of all binary infinite sequences. For every word $\alpha = a_0 a_1 ... a_{m-1}$, where $a_i \in \{0,1\}$, $i = 0,1,...,m$, $m = 1,2,...$ we define a cylinder

$$[\alpha] = [a_0 a_1 ... a_{m-1}] = \{\omega = w_0 w_1 ... | w_0 = a_0, w_1 = a_1, ..., w_{m-1} = a_{m-1}\} \subset \Omega.$$

The characteristic function of the arbitrary subset $F \subset \Omega$ will be denoted by I_F:

$$I_F = \begin{cases} 1, & \text{if} \quad \omega \in F \\ 0, & \text{if} \quad \omega \notin F \end{cases}.$$

Instead of $I_{[\alpha]}$ we will simply write I_α. Define a mapping T ("sequence shift") $T : \Omega \to \Omega$ by $T : \omega = w_0 w_1 ... \to \omega T = w_1 w_2 ....$ The equality $I_\alpha\left(\omega T^t\right) = 1$ means in such a way that

$$w_t = a_0,\ w_{t+1} = a_1, ...,\ w_{t+m-1} = a_{m-1}.$$

The number $\frac{1}{t}\sum_{j=0}^{t-1} I_\alpha\left(\omega T^{s+j}\right)$ is called the relative frequency of occurrence of the word α in the sequence ω on the segment from s to $s+t-1$. We say that the sequence ω is cesaro relative to the word α, if the limit $\lim\limits_{t\to\infty} \frac{1}{t}\sum_{j=0}^{t-1} I_\alpha\left(\omega T^j\right)$ exists. In this case the value of this limit

$$p_\alpha(\omega) = \lim_{t\to\infty} \frac{1}{t}\sum_{j=0}^{t-1} I_\alpha\left(\omega T^j\right),$$

can be interpreted as an average frequency of occurrence of the word α in the sequence ω. A sequence ω we will call l-cesaro if ω is cesaro relative to all words of length less than or equal to l, $l = 1,2,....$ We denote the class of l-cesaro sequences as $\Sigma^{(l)}$. A sequence ω we will call cesaro if ω is cesaro relative to the arbitrary word. We denote the class of cesaro binary sequences by Σ. The class of periodic binary sequences (both purely periodic and periodic with an initial section) we denote by T.

Let $A = (\{0,1\}, \{0,1\}, Q, h, f)$ be a strongly connected finite Moore machine with $\{0,1\}$ as input and output alphabets; Q as the set of states; $h : Q \times \{0,1\} \to Q$ as transition function; $f : Q \times \{0,1\} \to \{0,1\}$ as output function.

Following [6], we fix two sets of words

$$\{\alpha_i \in \{0,1\}^*, i = 1,2,...,t\} \text{ and } \{\beta_j \in \{0,1\}^*, j = 1,2,...,k\}, \quad t \geq 0,\ k \geq 1.$$

Let us suppose that an automaton A, starting to work from the state q_0, processes a sequence $\chi = (x_0, x_1, ...)$ into a sequence $\gamma = (y_0, y_1, ...)$. With sequence χ we associate the vector

$$z_{(A,q_0)}(\chi) = (p_{\alpha_1}(\chi), ..., p_{\alpha_t}(\chi), p_{\beta_1}(\gamma), ..., p_{\beta_k}(\gamma)), \tag{1}$$

if all quantities on the right-hand side exist.

The rule (1) defines a map

$$Z_A : T \to [0,1]^{t+k} \subset R^{t+k}.$$

It was shown in [6] that the closure (the set of all limit points) of the set $Z_A(T)$ is a convex polyhedron in $[0,1]^{t+k}$. This set will be denoted by R_A. The correctness of the used notation follows from the fact that automaton A is strongly connected.

An automaton A is called a cesaro-hereditary ([1]) if, starting to work from an arbitrary initial state $q \in Q$, it transforms an arbitrary sequence $\chi \in \Sigma$ into a sequence γ, wherein $\gamma \in \Sigma$.

2. Cesaro-inheritance of shift registers

Let V_n be the space of n-dimensional binary vectors, F_n be the set of all Boolean functions of n arguments, $n = 1, 2, ...$

Let $A_f = (X = \{0,1\}, V_n, Y = \{0,1\}, h, f)$ be the Moore machine (n-bit shift register [2]) with the states V_n, the transition function $h((a_1, ..., a_n), x) = (a_2, ..., a_n, x)$, $x, a_i \in \{0,1\}$, $i = 1, 2, ..., n$, the output function $f(x_1, x_2, ..., x_n) \in F_n$.

In [1] it is proved that A_f is a cesaro-hereditary automaton. It was shown that for the cesaro-hereditary automaton A the following is true $R_A = Z_A(\Sigma)$. We show that for the automaton A_f in the case when $k = t = 1$, $\alpha_1 = \beta_1 = 1$, this result can be refined. It turns out that it is enough to use for the input sequence only the requirement of stability of frequencies of n-grams, and the requirement of stability of frequencies of multigrams of greater length is unnecessary. Let us set:

$$R_A^{\Sigma^{(n)}} = Z_A\left(\Sigma^{(n)}\right) \subseteq [0,1]^{t+k}.$$

The correctness of this denotation, as before, follows from the strong connectivity of the automaton A_f.

To prove the corresponding result, we need a fundamental result from the theory of limits belonging to Toeplitz [7].

Theorem 1 (Toeplitz). Let us suppose that the coefficients ($1 \le m \le n$) of an infinite "triangular" matrix satisfy three conditions:

(a) Elements in any column tend to zero:

$$t_{nm} \to 0, \quad (m \text{ is fixed})$$

(b) The sums of the absolute values of the elements in any row are all limited by one constant:

$$|t_{n1}| + |t_{n2}| + ... + |t_{nn}| \le K$$

(c) $t_{n1} + t_{n2} + ... + t_{nn} \to 1$ at $n \to \infty$.

Then if for the set x_n, $n = 1, 2, ...$ $x_n \to a$ is true (a is finite), then

$$x_1 t_{n1} + x_2 t_{n2} + ... + x_n t_{nn} \to a.$$

Theorem 2. The following equality is true:

$$R_{A_f} = R_{A_f}^{\Sigma^{(n)}}.$$

Proof. We use the fact that $R_{A_f} = Z_{A_f}(\Sigma)$. The inclusion

$$Z_{A_f}(\Sigma) \subseteq Z_{A_f}\left(\Sigma^{(n)}\right)$$

follows from the inclusion $\Sigma \subseteq \Sigma^{(n)}$.

Since the vertices of the convex polygon R_{A_f} belong to the set $Z_{A_f}\left(\Sigma^{(n)}\right)$, to prove the reverse inclusion, it is sufficient to show the convexity of the set $Z_{A_f}\left(\Sigma^{(n)}\right)$.

Let $\chi_i = \left(x_0^{(i)}, x_1^{(i)}, ...\right) \in \Sigma^{(n)}$, $z(\chi) = \left(p^{(i)}, \pi^{(i)}\right)$, $i = 1, 2$. Let $\lambda \in (0, 1)$. The proof is to construct the sequence $\chi \in \Sigma^{(n)}$ such that $z(\chi) = \lambda z(\chi_1) + (1 - \lambda) z(\chi_2)$. Let m_k, l_k be the natural numbers, $m_k, l_k \to \infty$, $m_k / l_k \to \lambda$, at $k \to \infty$. In addition, we suppose that the sequence satisfies the conditions:

$$l_k \geq k, \quad \frac{l_k}{l_1 + l_2 + ... + l_k} \to 0. \tag{2}$$

The last condition is satisfied, in particular, if l_k is a polynomial from k.

Let us denote, $L(k) = \sum_{j=1}^{k} l_j$ ($L(0) = 0$) and suppose

$$x_i = \begin{cases} x_{i-L(k)}^{(1)}, & \text{if} \quad L(k) \leq i < L(k) + m_k; \\ x_{i-L(k)}^{(2)}, & \text{if} \quad L(k) + m_k \leq i < L(k). \end{cases}$$

Let us denote for brevity:

$$p_k^{(1)} = \frac{1}{m_k} \sum_{j=0}^{m_k - 1} x_j^{(1)},$$

$$p_k^{(2)} = \frac{1}{l_k - m_k} \sum_{j=0}^{l_k - m_k - 1} x_j^{(2)},$$

$$t_{kj} = \frac{l_j}{L(k)}, \quad \lambda_k = \frac{m_k}{l_k}.$$

Then we have:

$$\frac{1}{L(k)}\sum_{j=0}^{L(k)-1} x_j = \frac{1}{L(k)}\sum_{j=1}^{k}\left(\sum_{i=0}^{m_j-1} x_i^{(1)} + \sum_{i=0}^{l_j-m_j-1} x_i^{(2)}\right) =$$

$$= \frac{1}{L(k)}\sum_{j=1}^{k} l_j\left(\frac{m_j}{l_j}p_j^{(1)} + \frac{l_j-m_j}{l_j}p_j^{(2)}\right) =$$

$$= \sum_{j=1}^{k} t_{kj}\left(\lambda_j p_j^{(1)} + (1-\lambda_j)p_j^{(2)}\right).$$

Since $t_{kj} \to 0$ at $k \to \infty$, $\sum_{j=1}^{k} t_{kj} = 1$ and, by construction of λ_j

$$\lambda_j p_j^{(1)} + (1-\lambda_j)p_j^{(2)} \to \lambda p^{(1)} + (1-\lambda)p^{(2)}$$

is satisfied, Toeplitz theorem is applicable to the sum under consideration, according to which the following is satisfied at $k \to \infty$

$$\sum_{j=1}^{k} t_{kj}\left(\lambda_j p_j^{(1)} + (1-\lambda_j)p_j^{(2)}\right) \to \lambda p^{(1)} + (1-\lambda)p^{(2)}.$$

Now let $L(k) < N < L(k+1)$. Using (2), it is easy to show that at $k \to \infty$

$$\frac{1}{N}\sum_{j=0}^{N-1} x_j - \frac{1}{L(k)}\sum_{j=0}^{L(k)-1} x_j \to 0.$$

It means that

$$\lim_{N\to\infty}\frac{1}{N}\sum_{j=0}^{N-1} x_j = \lambda p^{(1)} + (1-\lambda)p^{(2)}.$$

The proof that $\chi \in \Sigma^{(n)}$ and

$$\lim_{N\to\infty}\frac{1}{N}\sum_{j=0}^{N-1} y_j = \lambda \pi^{(1)} + (1-\lambda)\pi^{(2)},$$

is carried out similarly. Since there is "gluing" on the segment $2k-1$ of the length $L(k)$, not more than $2n(2k-1)$ values of the function $f(x_1, x_2, ..., x_n)$ can be distorted at the junction of sequences. By virtue of (2), such a number of distortions will not affect the limit value.

The theorem is proved.

Example 1 (The shift register). Let us consider the automaton A_f in the case $n = 2$, $f = x_1 (x_2 \oplus 1)$. The output function takes on the value "1" only in the state "10", in all other states it is equal to zero.

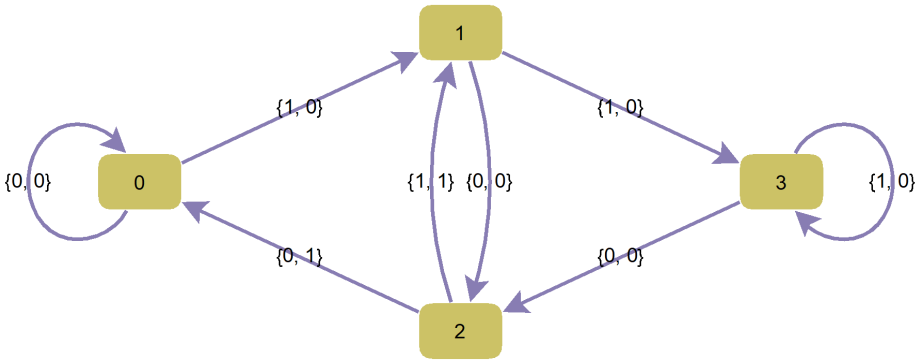


Fig. 1. The transition graph of the automaton A_f.

The transition graph of the automaton A_f is shown in the Fig. 1 1. It has 4 vertices "0", "1", "2", "3" (the lexicographical order is used) and the following 6 cycles:

1) the loop at the vertex "0" with marking (0,0),
2) the loop at the vertex "3" with marking (1,0),
3) the cycle of the length 2 with the vertices "1" and "2" and marking (10, 10),
4) the cycle of the length 3 with the vertices "0", "1", "2" and marking (100, 000),
5) the cycle of the length 3 with the vertices "1", "3", "2" and marking (101, 001),
6) the cycle of the length 4 with the vertices "0", "1", "3", "2" and marking (1100, 0001).

According to the Theorem 1 of [6], the following equality is correct

$$R_{A_f} = Conv\left\{\left(\frac{0}{1},\frac{0}{1}\right),\left(\frac{1}{1},\frac{0}{1}\right),\left(\frac{1}{2},\frac{1}{2}\right),\left(\frac{1}{3},\frac{0}{3}\right),\left(\frac{2}{3},\frac{1}{3}\right),\left(\frac{2}{4},\frac{1}{4}\right)\right\} =$$
$$= Conv\left\{(0,0),(1,0),\left(\frac{1}{2},\frac{1}{2}\right),\left(\frac{1}{3},0\right),\left(\frac{2}{3},\frac{1}{3}\right),\left(\frac{1}{2},\frac{1}{4}\right)\right\} =$$
$$= Conv\left\{(0,0),(1,0),\left(\frac{1}{2},\frac{1}{2}\right)\right\}.$$

The resulting polygon is shown in the Fig. 2.

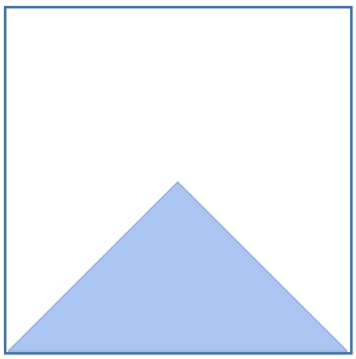


Fig. 2. The polygon of the automaton A_f. Its vertices are: $(0,0)$, $(1,0)$, $(1/2,1/2)$.

3. Cesaro-heredity shift register with internal XOR

By a shift register with internal XOR we mean an automaton, which, under the action of the input symbol $a_0 \in \{0,1\}$ from the state $(a_1, a_2, ..., a_n)$ passes to the state $(a_0 \oplus a_1, a_1 \oplus a_2, ..., a_{n-1} \oplus a_n)$, where $\oplus$ is a modulo 2 summation. We will denote such an automaton as $A_f^{\oplus}$.

The automaton-theoretic properties of registers similar to $A_f^{\oplus}$, in the autonomous case were considered in [2], and the issues of their hardware implementation in [8].

Theorem 3. If for the Boolean function $f(x_1, x_2, ..., x_n)$ the following condition is satisfied $f(0,0,...,0) \neq f(0,0,...,0,1)$, then $A_f^{\oplus}$ is not a cesaro-hereditary automaton.

Proof. It is easy to see that in the graph of the automaton A_f, there is exactly one cycle, the movement of which occurs when a sequence consisting of only zeros is input, this is a loop at the zero vertex. There are several such cycles for the automaton $A_f^{\oplus}$. In fact, since when the symbol at the input of the automaton is zero, the state $(a_1, a_2, ..., a_n)$ the state passes into the state $(a_1, a_1 \oplus a_2, ..., a_{n-1} \oplus a_n)$, the condition for the presence of a cycle of length 1 (loop) looks like this: $a_2 = a_1 \oplus a_2, a_3 = a_2 \oplus a_3, ..., a_n = a_{n-1} \oplus a_n$. Hence we obtain the equalities: $a_i = 0$, $i = 1, 2, ..., n-1$, $a_n = 0, 1$. Therefore, in the graph of the automaton $A_f^{\oplus}$ there are two loops, the movement of which occurs when the input symbol is 0: a loop at the vertex $(0,0,...,0)$ and a loop at the vertex $(0,0,...,0,1)$.

Let the statement condition be satisfied now. It follows from it that when moving along one of the above loops, the output sequence of the automaton $A_f^{\oplus}$ consists of units, and when moving along another one, this sequence consists of zeros. Since the diameter of the graph of the automaton $A_f^{\oplus}$ is equal to n it is possible to move from the state $(0,0,...,0)$ to the state $(0,0,...,0,1)$ and vice versa by not more than n steps. By ξ_{01} and ξ_{10} we denote the input sequences that provide these transitions.

Now let us consider the infinite input sequence

$$\chi = \xi_0 0^{k_1} \xi_{01} 0^{k_2} \xi_{10} 0^{k_3} \xi_{01} 0^{k_4} \xi_{10} ...,$$

where ξ_0 is an input sequence that transfers the automaton $A_f^{\oplus}$ from the initial state to the state $(0, 0, ..., 0)$, 0^{k_i} is a sequence of k_i zeros, k_i are integers, $i = 1, 2, ...$ The output sequence will obviously be

$$\gamma = \zeta_0 0^{k_1} \zeta_{01} 1^{k_2} \zeta_{10} 0^{k_3} \zeta_{01} 1^{k_4} \zeta_{10} ...,$$

where $\zeta_0, \zeta_{01}, \zeta_{10}$ are some binary sequences, the length of each of which does not exceed n, 1^{k_j} is a sequence consisting of k_j units.

It is easy to verify that at $k_i = 2^{2^i}$ the sequence ξ is cesaro, and for the sequence γ there is no limit on the relative frequency of occurrence of a unit in growing initial segments. Therefore, it is not cesaro and the automaton $A_f^{\oplus}$ is not cesaro-hereditary.

Example 2 (The shift register with internal XOR). Let us consider the automaton $A_f^{\oplus}$ in the case $n = 2$, $f = (x_1 \oplus 1)x_2$. The output function takes on the value "1" only in the state "01", in all other states it is equal to zero.

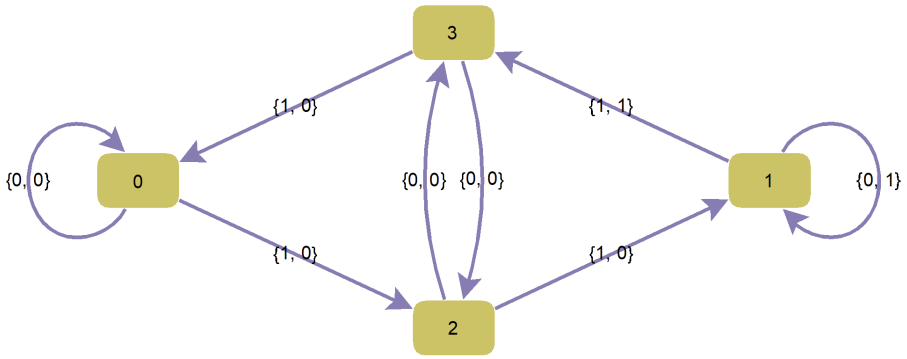


Fig. 3. The transition graph of the automaton $A_f^{\oplus}$.

The transition graph of the automaton $A_f^{\oplus}$ is shown in the Fig. 3. It has 4 vertices "0", "1", "2", "3" (the lexicographical order is used) and the following 6 cycles:

1) the loop at the vertex "0" with marking (0,0),
2) the loop at the vertex "1" with marking (0,1),
3) the cycle of the length 2 with the vertices "2" and "3" and marking (00, 00),
4) the cycle of the length 3 with the vertices "0", "2", "3" and marking (101, 000),

5) the cycle of the length 3 with the vertices “1”, “2”, “3” and marking (110, 010),
6) the cycle of the length 4 with the vertices “0”, “1”, “2”, “3” and marking (1111, 0010).

According to the Theorem 1 of [6], the following equality is correct

$$R_{A_f^{\oplus}} = Conv\left\{\left(\frac{0}{1},\frac{0}{1}\right),\left(\frac{0}{1},\frac{1}{1}\right),\left(\frac{0}{2},\frac{0}{2}\right),\left(\frac{2}{3},\frac{0}{3}\right),\left(\frac{2}{3},\frac{1}{3}\right),\left(\frac{4}{4},\frac{1}{4}\right)\right\} =$$
$$= Conv\left\{(0,0),(0,1),\left(\frac{2}{3},0\right),\left(\frac{2}{3},\frac{1}{3}\right),\left(1,\frac{1}{4}\right)\right\} =$$
$$= Conv\left\{(0,0),(0,1),\left(\frac{2}{3},0\right),\left(1,\frac{1}{4}\right)\right\}.$$

The resulting polygon is shown in the Fig. 4.

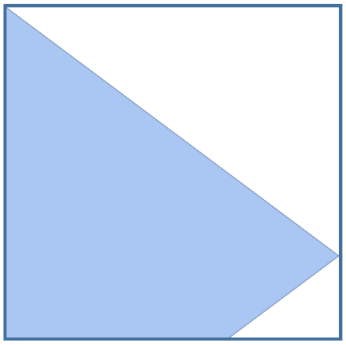


Fig. 4. The polygon of the automaton $A_f^{\oplus}$. Its vertices are: $(0,0)$, $(0,1)$, $(1,1/4)$, $(2/3,0)$.

4. Cesaro-inheritance of the generalized binary shift register

In [3] generalized shift registers (GSR) are defined which transition graphs are generalized de Bruijn graphs [9]. The binary GSR of order m, $m = 1, 2, ...$ is the Moore machine $A_f^{(m)} = (X, Y, Q, h, f)$, where the input and output alphabets are $X = Y = \{0, 1\}$, the set of states is $Q = \{0, 1, ..., m-1\}$, the transition function is defined by the rule $h(q, \varepsilon) = (2q + \varepsilon) \mathrm{mod} m$, $q \in Q$, $\varepsilon = 0, 1$, the output function is some mapping $f : Q \to \{0, 1\}$. At $m = 2^t$ binary GSR is a binary pass-through shift register with a accumulator of the capacity t.

It turns out that GSR, unlike ordinary registers, are not cesaro-hereditary.

Theorem 4. If $m = 2^t$, $t \geq 0$, then at any output function f the automaton $A_f^{(m)}$ is cesaro-hereditary.

If $m \neq 2^t$, then there is an output function f, for which the automaton $A_f^{(m)}$ is not cesaro-hereditary.

Proof. Let $m = 2^t$. GSR is the pass-through shift register with a accumulator of the capacity and its cesaro-inheritance proved in [1].

Let $m \neq 2^t$. Then $m = 2^k s$, $k \geq 0$, $s \geq 3$ is odd. Let us prove that in the GSR graph there are at least two cycles c_1 and c_2, which input markup consists of only zeros.

Consider a sequence $2^i \mathrm{mod} m$, $i = 1, 2, ...$ Since 2^i and s are coprime numbers, inequality

$$2^i \mathrm{mod} 2^k s \neq 0 \mathrm{mod} 2^k s$$

holds. Therefore, all elements of the sequence under consideration are nonzero.

Obviously, this is a periodic sequence, possibly with an initial non-periodic segment.

The set of different elements of this sequence forms a cyclic semigroup $\langle 2 \rangle$ generated by element 2. The length l of the period of this sequence is the period of element 2, and the length r of the initial non-periodic segment is the index of element 2 in semigroup $\langle 2 \rangle$ [10].

Thus, the sequence of states

$$c_1 = \left(2^r \mathrm{mod} m, 2^{r+1} \mathrm{mod} m, ..., 2^{r+l-1} \mathrm{mod} m\right)$$

is a nonzero cycle of the length l in the GSR graph with an input markup consisting of only zeros.

We can take as c_2 the loop at the zero vertex, $c_2 = (0)$.

So, we indicated two different cycles in the transition graph of the automaton, the input marking of which consists of zeros. We define the output function f so that at the states of the cycle c_1 it takes on only zero values, and at the states of the cycle c_2 (i.e., at state 0) $f = 1$. On other states, the function f can be set arbitrarily.

For convenience, we denote the state of the cycle c_i by $q_1^{(i)}, ..., q_{l_i}^{(i)}$, l_i is the length of the cycle c_i, $i = 1, 2$. Let $\chi^{(i)} = 0^{l_i}$ be the input sequence under the action of which $A_f^{(m)}$ sequentially passes the states of the cycle c_i, starting from $q_1^{(i)}$, $i = 1, 2$. By $\xi_{12}(\xi_{21})$ we denote the shortest sequence that transfers the considered GRS from the state $q_1^{(1)}$ to the state $q_1^{(2)}$ (the state $q_1^{(2)}$ to the state $q_1^{(1)}$).

Let us define the infinite binary sequence

$$\chi = \left(\chi^{(1)}\right)^{k_1} \wedge \xi_{12} \wedge \left(\chi^{(2)}\right)^{k_2} \wedge \xi_{21} \wedge \left(\chi^{(1)}\right)^{k_3} \wedge \xi_{12} \wedge \left(\chi^{(2)}\right)^{k_4} ...,$$

where the symbol $\wedge$ means concatenation of sequences. At $k_i = 2^{2^i}$ the sequence χ is cesaro, because, as it is easy to see, the limit of the relative frequency of occurrence in it of any word other than a series of zeros exists and is equal to zero. For words of the form 0^j, $j = 1, 2, ...$ such limits are equal to one. Let $Ext_{(A,q_0)}(\chi)$ be the output register sequence $A_f^{(m)}$. Let us consider its initial length segment $k_1 + |\xi_{12}| + k_2 + |\xi_{21}| + ... + k_N$. Since $|\xi_{12}|$ and $|\xi_{21}|$ do not exceed the diameter of the register graph, it is easy to show that the relative frequency of occurrence of units on this segment is equal to $0 + O\left(2^{-2^{N-1}}\right)$ at odd N and $1 + O\left(2^{-2^{N-1}}\right)$ at even N. This means that the sequence of relative frequencies of units in the growing initial segments does not have a limit and, therefore, the output sequence is not cesaro, which completes the proof.

Example 3 (Generalized binary shift register). Let us consider the automaton $A_f^{(m)}$ in the case $m = 6$, and the output function is defined as follows:

$$f(i) = \begin{cases} 0, & \text{if} \quad 0 \le i \le 2, \\ 1, & \text{if} \quad 3 \le i \le 5. \end{cases} .$$

The output function of the automaton takes on the value "1" exactly on half of the set of states and "0" on the remaining half.

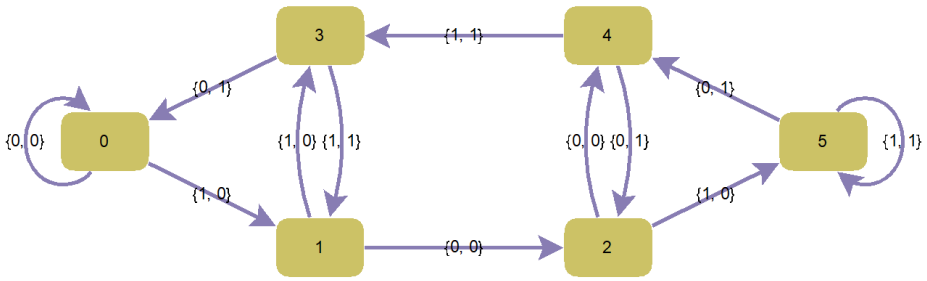


Fig. 5. The transition graph of the automaton $A_f^{(6)}$.

The transition graph of the automaton $A_f^{(6)}$ is shown in the Fig. 5. It has 6 vertices "0", "1", "2", "3", "4", "5" (the lexicographical order is used) and the following 10 cycles:

1) the loop at the vertex "0" with marking (0,0),
2) the loop at the vertex "5" with marking (1,1),
3) the cycle of the length 2 with the vertices "1" and "3" and marking (11, 01),

4) the cycle of the length 2 with the vertices “2” and “4” and marking (00, 01), the cycle of the length 3 with the vertices “0”, “1”, “3” and marking (110, 001),
5) the cycle of the length 3 with the vertices “2”, “5”, “4” and marking (100, 011),
6) the cycle of the length 4 with the vertices “1”, “2”, “4”, “3” and marking (0011, 0011),
7) the cycle of the length 5 with the vertices “1”, “2”, “5”, “4”, “3” and marking (01011, 00111), the cycle of the length 5 with the vertices “0”, “1”, “2”, “4”, “3” and marking (10010, 00011),
8) the cycle of the length 6 with the vertices “0”, “1”, “2”, “5”, “4”, “3” and marking (101010, 000111).

According to the Theorem 1 of [6], the following equality is correct

$$R_{A_f^{(6)}} = Conv\left\{\left(\frac{0}{1},\frac{0}{1}\right),\left(\frac{1}{1},\frac{1}{1}\right),\left(\frac{2}{2},\frac{1}{2}\right),\left(\frac{0}{2},\frac{1}{2}\right),\left(\frac{2}{3},\frac{1}{3}\right),\left(\frac{1}{3},\frac{2}{3}\right),\left(\frac{2}{4},\frac{2}{4}\right),\left(\frac{3}{5},\frac{3}{5}\right),\left(\frac{2}{5},\frac{2}{5}\right),\left(\frac{3}{6},\frac{3}{6}\right)\right\} = Conv\left\{(0,0),(1,1),\left(1,\frac{1}{2}\right),\left(0,\frac{1}{2}\right),\left(\frac{2}{3},\frac{1}{3}\right),\left(\frac{1}{3},\frac{2}{3}\right),\left(\frac{1}{2},\frac{1}{2}\right),\left(\frac{3}{5},\frac{3}{5}\right),\left(\frac{2}{5},\frac{2}{5}\right),\left(\frac{1}{2},\frac{1}{2}\right)\right\} = Conv\left\{(0,0),(0,1),\left(1,\frac{1}{2}\right),\left(0,\frac{1}{2}\right)\right\}.$$

The resulting polygon is shown in the Fig. 6.

Fig. 6. The polygon of the automaton $A_f^{(6)}$. Its vertices are: $(0,0)$, $(0,1/2)$, $(1/2,0)$, $(1,1)$.

5. Conclusion

We study whether binary automata of three classes possess the cesaro-heredity property: shift registers, generalized shift registers, shift registers with internal XOR. It is shown that shift registers always have this property, and automata from other classes under consideration, generally speaking do not have to possess it. The conditions for the output function of the generalized shift registers and shift registers with internal XOR which ensure the absence of the cesaro-heredity property are given. Examples of transition graphs of the considered automata are given and polygons are constructed that characterize their statistical properties.

REFERENCES

1. Melnikov S. Yu., Samouylov K. E. Cesaro Sequences and Cesaro Hereditary Automata // In: Galinina O., Andreev S., Balandin S., Koucheryavy Y. (eds) NEW2AN/ruSMART 2020 (to appear).
2. Golomb S. W. Shift Register Sequences, 3rd revised edition. Singapore: World Scientific, 2017.
3. Maksimovskiy A. Yu., Melnikov S. Yu. Spectral and Combinatorial Characteristics of the Reduced De Brijn Graphs
Voprosy kiberbezopasnosti. 2018. V. 4(28). P. 70–76.
4. Jabbari H., Muzio J. C., Sun L. A new class of cellular automata // In: Proceedings 10th Euromicro conference on digital system design architectures, methods and tools (DSD). 2007. P. 331–338.
5. Kyaw T. N. N., Tsuneda A. Generation of chaos-based random bit sequences with prescribed auto-correlations by post-processing using linear feedback shift registers // Nonlinear Theory and Its Applications, IEICE. 2017. V. 8(3). P. 224–234.
6. Melnikov S. Yu., Samouylov K. E. Polyhedra of Finite State Machines and their Use in the Identification Problem // In: Galinina O., Andreev S.,Balandin S., Koucheryavy Y. (eds) NEW2AN/ruSMART 2020 (to appear).
7. Boos J., Cass F.P. Classical and Modern Methods in Summability. Oxford University Press. 2000. 586 p.
8. Chen W.-K. The VLSI Handbook, Second Edition. CRC Press, Chicago. 2006. 2320 p.
9. Imase M., Itoh M. Design to minimize diameter on building-block network // IEEE Trans. Comput. 1981. V. 30. P. 439–442.
10. Clifford A. H., Preston G. B. The Algebraic Theory of Semigroups. Vol. 1. AMS Publ. 1964. 224 p.

UDC: 004.75

Wireless Sensor Network for Intensive Data Collection Based on LoRaWAN Technology and Distributed Data Compression Algorithm

Yury Rassadin[1] and Sergey Dushin[1]

[1]Trapeznikov Institute of Control Sciences of Russian Academy of Sciences,
Profsoyuznaya st. 65, Moscow, Russia

rassadin@ipu.ru, s.dushin@inbox.ru

Abstract

In the paper a way to build a wireless sensor network based on LoRaWAN technology is considered. The proposed approach allows to increase the intensity of data collection compared to standard implementations of LoRaWAN network and save energy efficiency of wireless sensors. The main feature of the network is the edge computing algorithm of distributed data compression. Higher data collection rates are achieved along with acceptable level of energy efficiency for autonomous operation of sensors and low load of the physical communication channel. The effectiveness of the proposed solution has been illustrated by the created network of temperature sensors used for intensive temperature measurement in the problem of identifying the topology of the office building heating system. The network is based on open source software, ChirpStack server, LoRaWAN end-device stack from Semtech and additional modules developed by the authors.

Keywords: wireless sensor network, high-intensity data collection, building microclimate, LoRaWAN, edge computing, IoT

1. Introduction

For buildings from XXth century automatization of energy, heating and climatic systems is related with various difficulties due to the fact that the necessary functionality was not provided at the design stage. Typical problems that affects the efficiency of management are the power supply, availability of data cables, direct access to key building objects. The exact schemes of the facilities are often missing, and the maintenance is held under conditions of uncertainty and varying parameters. Wireless autonomous sensors are most likely non-alternative when connection to reliable power networks is lacking and cable infrastructure of the communication

network is absent. Energy efficiency for this king of sensors is very important, so the use of high-speed wireless technologies, such as IEEE 802.11 (Wi-Fi) standards family, is most likely excluded.

There are avaliable solutions based on modern wireless LAN technologies, such as LoRaWAN, SigFox and NB-IoT. They solve typical problems of data collection and equipment management quite effectively [1]. A common disadvantage of this kind of technologies for building wireless sensor network is the low bandwidth of the wireless channels used and the intensity of data exchange limitations for keeping the sensors battery life at an acceptable level [2, 3]. At the same time in real IoT applications it is often necessary to collect much more data than LPWAN technologies provide and currently there is scientific interest in such problems [4]. An interesting example of this problem is the high frequency temperature measurement of heat pipes that is necessary for identification of the building's heating system topology. To register temperature waves a measuring and registration system with a frequency of at least 1 Hz is required. In the paper we propose a method for providing high-intensity data collection from wireless autonomous sensors, which solves the problem of intensive temperature measurement and can be used for identifying the topology of the building's heating subsystem, as well as in a variety of other applications that require intensive data collection by wireless sensors.

The paper is organized as follows. The second section describes the main idea of the work, the data compression algorithm for communication between peripherals and the central server, which allows to increase the polling frequency of LoRa channel. The third part describes the network architecture, server side, and software for end devices. The fourth section deals with network deployment issues, hardware lists and areas for further research.

2. The Intensive Data Collection Method

The data compression algorithms [5] are powerful tools for improving channel efficiency and optimizing power consumption in communication networks. However the use of traditional compression in LPWAN networks leads to unacceptable increases in data latency and power consumption on the sensor side as additional data processing is required. This is why the information is transmitted in uncompressed form and the development of such algorithms for LPWAN must take into account the requirements and specific conditions of these networks.

The essence of the proposed approach is synchronous prediction of the input signal both on the server side and on the sensor side [6]. In this case, information from the sensor is sent when the prediction error exceeds the predefined threshold. This new measurement is used to correct the prediction on the server side. If the sensor is silent, the server decides that the prediction is accurate enough and uses it

as real-time information. Prediction procedure uses linear algorithm that is based on the Levinson-Durbin recursion. It is also important that the predictor filter coefficients [7, 8] are calculated on the server side to save battery budget. They are transmitted to the sensor only if the error threshold value is exceeded. This algorithm is schematically shown in the figure 1.

Under these conditions the use of a physical transmitter is minimized (the most energy conmsuming functional part of an autonomous sensor). The server uses its input estimation as real-time data until corrective information from the sensor is recieved.

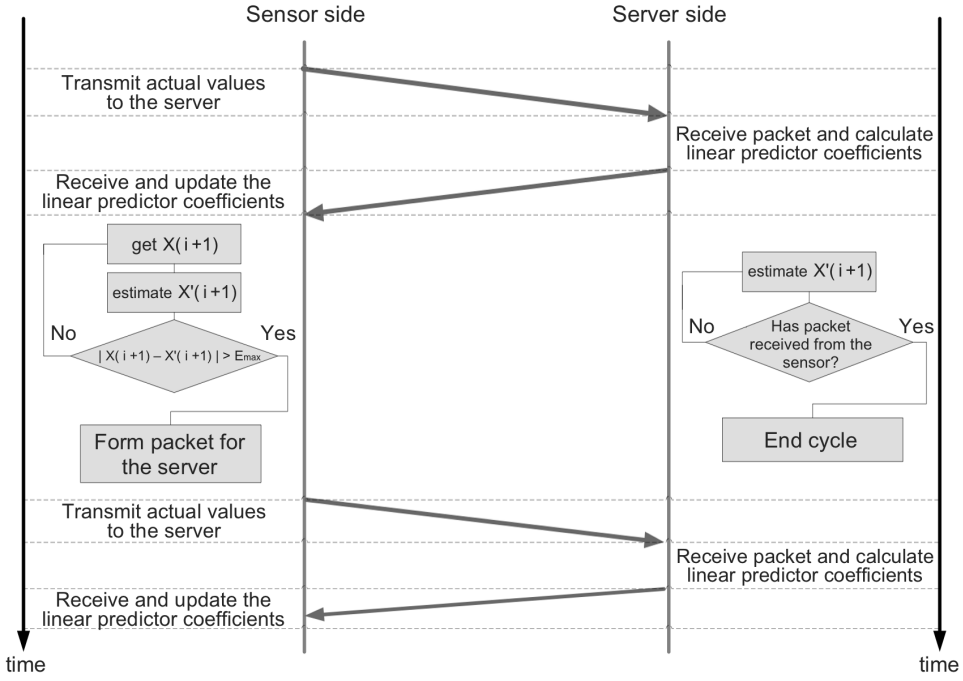


Fig. 1. Distributed data compression algorithm for LPWAN

3. System Architecture

Same as the standard LoRaWAN network architecture [9], the modified system has server and client sides. The main functional blocks are shown in the figure 2. The server part consists of standard LoraWAN components (network bridge, network server, application server) and additional modules that provide information compression, improve sensor energy efficiency and channel utilization efficiency. At

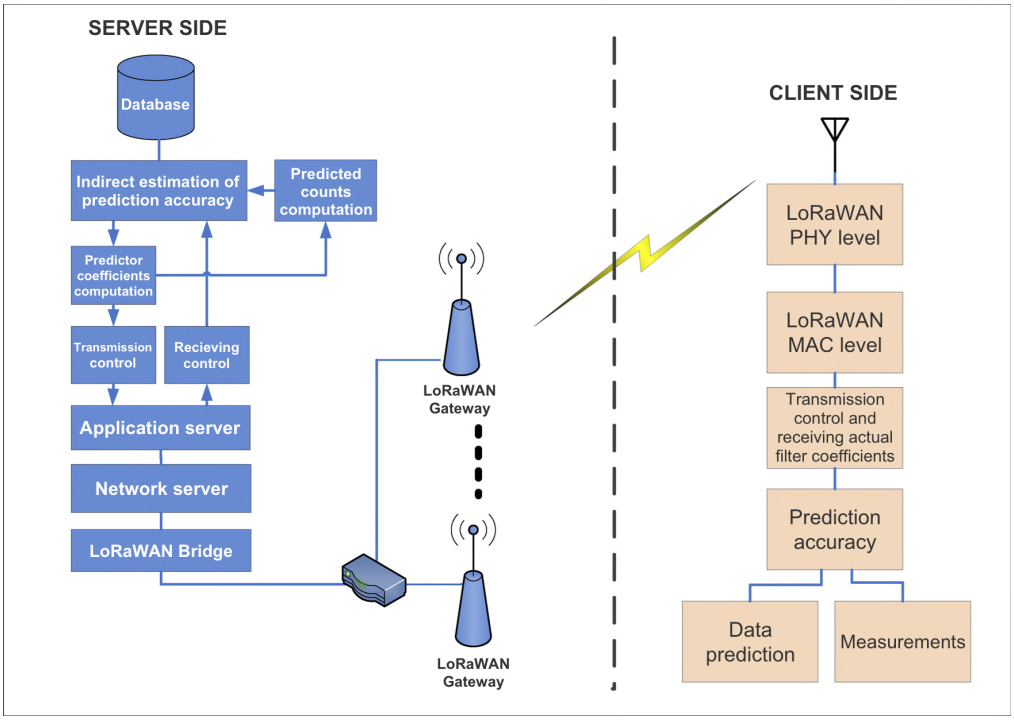


Fig. 2. Modified LoRaWAN network structure

each time interval the sampling module implements the measurement prediction if the data from the sensor did not come or transfer the data directly if it was received from the sensor. If data is received from the sensor, which means that specified error threshold is exceeded, the module for recalculating the predictor filter coefficients generates new coefficient values according to the Levenson-Durbin algorithm. The Transmission Control Module sends predictor filter coefficients to the sensor and receives and processes packets from the sensor.

The physical and channel levels required for LoRaWAN as well as additional modules are implemented on the sensor side. The prediction module estimates input value using a formula identical to the server one. The predicted value is then compared to the result of real measurements. The measurement module has quite common design. Since NTC thermistors change their characteristic nonlinearly, the temperature values are stored as a tabular function. We use binary search in this table, intermediate values are calculated linearly, $t(r) = t_1 + (t_2 - t_1)(r - r_1)/(r_2 - r_1)$. The resistance corresponding to the temperature of 25°C is used as the zero iteration. The control module provides transmitting data if the prediction error [5] is greater

than the predefined threshold, as well as receiving predictor filter coefficients from the server and updating them.

4. Intensive Temperature Measurement of the Building Heat Pipes

The process of deploying the described network can be divided into three stages. The first stage was the choice of equipment capable of solving the problems faced by the authors. The second stage was to create software for end devices and for the server part, testing and debugging their interaction. The third and final stage should include work on the deployment of the network in a particular building: scaling the results of the second stage on the entire building, providing reliable coverage by gateways, testing launches.

The required equipment list contains temperature sensors, server, gateways and network infrastructure for them. We use a virtual machine with OpenSUSE operating system as central node. The gateways were chosen to comply with Russian regional standards. An affordable solution is the MikroTik R11e-LR8 gateway based on the Semtech SX1301 processor. During the development process, we also used the Vega BS-1 gateway, which will be integrated into the working network later. The temperature sensor is controlled by the STM32L151 [10] microcontroller, which is declared by Semtech to be energy efficient. The choice of the LoRa transmitter manufacturer is obvious, we opted for Semtech SX1272 [11]. The temperature measuring module is a voltage divider with a NTC thermistor B57861. A battery with a voltage of 3.3–5 V (in our case, 3.6V) can be used for power supply.

4.1. Server Side. To build the server part, the open software package ChirpStack was used, the computational module of the compression algorithm is implemented by means of the provided API. We use Python programming language to implement the predictor filter, as well as the WebSocket client and server. We placed the ChirpStack gateway bridge, network server and application server components on the central server machine in order to shift the main computing load onto it in accordance with the main idea of work.

4.2. End Device Side. For creating the client firmware we used the Semtech development environment and open-source libraries. The Semtech company gives full freedom to developers, including the commercial use of the results. Sensor firmware development was implemented in C language in the ac6 System Workbench for STM32 IDE environment on the basis of the LoRaWAN end device stack open project. Energy efficiency is achieved not only by reducing the number of data transmission sessions, but also by controlling power supply to peripheral devices. The SX1272 modem is capable of independently controlling the high-frequency transmitter module, but we decided to control it directly from the microcontroller maximize energy savings.

4.3. Equipment Layout. There should be two types of devices distributed throughout the building, sensors and gateways. The LoRaWAN protocol allows great freedom and convenience in the arrangement of end devices, so in the paper we consider the placement of sensors to depend on the specific problem. In the case of identification of the heating system, the sensors should be located in various rooms on the heating pipes mainly near the windows. The LoRaWAN Protocol has a mechanism for managing multiple gateways connected to a single network server, so the most difficult problem when deploying such a network is to find an optimal non-excess location of the gateways. The building where the experimental network was deployed is elongated, so we decided to place the base stations near the windows, covering the extended facades of the building separately. We have placed an additional base station in the basement of the building for coverage of the main nodes there.

5. Results and Futher Research Area

Intensive data collection from small groups of sensors has been implemented and tested for today. The next step is to develop a guaranteed method to scale such segments to the required size. We want to orientate ourselves to the volumes declared in the LPWAN brochures, several hundreds of sensors simultaneously. Using open source software from ChirpStack and Semtech, we are able to bypass the limitations on sensor polling frequency that are inherent to commercial solutions, and the network is still evolving. One of the directions of development can be improvement of prediction algorithms using mathematical models of monitored objects. For some experiments the central server capacity may not be enough, so the server site can be transferred to a more powerful machine, but so far this has not been necessary. We hope that a deeper study of server software from ChirpStack will help us to optimize the proposed approaches.

6. Conclusion

A modified architecture of wireless sensor network is presented in the paper, which allows to increase the intensity of data collection in comparison with standard implementations of LoRaWAN networks. The main functional feature of the network is the use of the authors' algorithm for distributed data compression in LPWAN networks. The proposed approach achieves the main goal in terms of intensity while energy efficiency stays at an acceptable level for autonomous operation of sensors and load of the physical communication channel remains low. The algorithm involves prediction of input data on the server side, while the sensor only transmits corrective information in case of exceeding the predefined prediction error. This approach minimizes the use of standalone sensors and reduces the load on the physical

communication channel. Lower power consumption allows to extend sensors lifetime until the batteries are replaced. The authors' contribution to network software development is at the level of server applications and sensor firmware of the devices. The effectiveness of the proposed approach will be investigated when solving the problem of intensive data collection of temperature sensors in the identification of the topology of the heating system of an office building.

REFERENCES

1. Mekki K., Bajic E., Chaxel F., Meyer F. Overview of Cellular LPWAN Technologies for IoT Deployment: Sigfox, LoRaWAN, and NB-IoT / 2nd IEEE International Workshop on Mobile and Pervasive Internet of Things - Athens. - 2018.
2. N. Pukrongta and B. Kumkhet, The relation of LoRaWAN efficiency with energy consumption of sensor node, 2019 International Conference on Power, Energy and Innovations (ICPEI), Pattaya, Chonburi, Thailand, 2019, pp. 90-93
3. N. I. Bazenkov et al., "Intensive data collection system for smart grid and smart building research," 2019 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA), Lipetsk, Russia, 2019, pp. 411-415.
4. S. Benninger, M. Magno, A. Gomez and L. Benini, "EdgeEye: A Long-Range Energy-Efficient Vision Node For Long-Term Edge Computing," 2019 Tenth International Green and Sustainable Computing Conference (IGSC), Alexandria, VA, USA, 2019, pp. 1-8.
5. Haykin, S. Adaptive filter theory (5-th edition) / S. Haykin — Prentice Hall, 2014. — 936p.
6. Jackson, L.B., Digital Filters and Signal Processing, Second Edition, Kluwer Academic Publishers, 1989. pp.255-257.
7. Dushin S.V., Frolov S.A. Distributed data compression algorithm for low-power wide-area networks. DCCN-2019
8. Official site of Semtech Corporation, https://www.semtech.com/uploads/documents/SX1272_DS_V4.pdf
9. Hasan A.H., Grachev A.N. On-Line Parameters Estimation Using Fast Genetic Algorithm // J. of Electrical and Control Engineering (JECE). –2014. – Vol. 4, No. 2. – P. 16–21.
10. https://www.st.com/en/microcontrollers-microprocessors/stm32l151-152.html
11. https://www.semtech.com/products/wireless-rf/lora-transceivers/sx1272

UDC: 004.7

Flying Network for Emergency using Tethered Multicopters

V. Vishnevsky[1], T.D. Dinh[2], A. Vybornova[2], R. Kirichek[1,2]

[1]V.A. Trapeznikov Institute of Control Sciences of RAS, 65 Ulitsa Profsoyuznaya, Moscow, Russia
[2]The Bonch-Bruevich Saint-Petersburg State University of Telecommunications, 22 Prospekt Bolshevikov, St.Petersburg, Russia
vishn@inbox.ru, din.cz@spb.ru, a.vybornova@gmail.com, kirichek@sut.ru

Abstract

In recent years, the interest of tethered UAVs high-altitude platforms has been widely constantly increasing in many fields. The long-time operating possibility is one of the main advantages of tethered unmanned high-altitude platforms compared to autonomous UAVs. In the paper, a flying network for emergencies using tethered multicopters is proposed. The combination of tethered unmanned high-altitude platforms and groups of UAVs in flying network for emergencies is expected to enhance the effectiveness of search and rescue operation in the wilderness as well as after natural disasters.

Keywords: UAV, flying network, tethered multicopters, search and rescue

1. Introduction

Over the past decade, the emergence of new technologies as well as the development of science and technology has greatly assisted search and rescue operation in emergencies. The dissemination of UAVs for civilian purposes has turned them into a useful search and rescue (SAR) tool in different situations, such as for emergency prevention, monitoring emergencies, searching for missing people after natural disasters, or urgently delivering the necessary cargo to places where it is needed in an emergency. In addition, UAVs are used for environmental purposes, such as to protect beaches, study the melting of polar ice, monitor forests, monitor the coast and water areas, determine the effects of various pollutants, etc [1, 2].

Multifunctional complexes with UAVs in control and communication systems are utilized for relaying signals or in studies of the pattern of radio signals transmission, and for inspection of cell towers [3]. In some cases, UAVs can work as "network

The reported study was funded by RFBR according to the research project No.20-37-70059.

nodes" to connect the network to the Internet (Internet of Drones - IoD). Moreover, in order to expand the working area, cellular networks can also be employed as an additional communication channel to UAVs, along with conventional P2P (point to point) networks, for example, in automated air traffic control systems. In many scientific articles [4, 5, 6], using groups of UAVs had been proven to be much more effective than using only one UAV. However, the main disadvantage of UAVs is a limited time of operation due to the small battery resource of UAVs equipped with electric motors or the fuel reserve for internal combustion engines. In order to solve this problem, tethered UAVs high-altitude platforms are consider. They can support long-term operation with power supply of engines and payload equipment is provided from the ground-based energy sources.

Due to above-mentioned features of UAV groups, in this study, the paper provides a Wi-Fi network based on groups of UAVs and tethered UAVs high-altitude platforms, called flying network for emergencies using tethered multicopters, that can help rescuers to communicate with victims or find their locations using Wi-Fi signals generated from their phone. In addition, the deployment of rescue operations in difficult or dangerous areas for rescuers is also addressed with the help of flying network.

2. Tethered UAVs high-altitude platform

At present, research centers in leading countries of the world are carrying out intensive scientific work on the design and implementation of tethered UAVs high-altitude platforms [7, 8, 9, 10], given the wide spread of their practical application. The long-time operating possibility is one of the major advantages of tethered unmanned high-altitude platforms compared to autonomous UAVs. UAVs can be presented by two types: multicopter type and fixed wing type. Fixed wing type UAV has a high flight duration, maximum flight altitude, high speed, and high payload. On the other hand, multicopter type has the ability to stay stable in the air, as well as high maneuverability [13]. With these advantages, multicopter type is more suitable for tethered UAVs system due to its structural characteristics and missions.

Tethered UAVs high-altitude platform consists of terrestrial and flying modules. The terrestrial module contains a ground control station for a high-altitude platform (TCS), a ground voltage converter, a winch of a tethered cable of a high-altitude platform and a mooring device (Fig. 1).

A tethered UAVs system consists of a multicopter, cable and flight platform. A new energy transfer technology will provide the multicopter with the ability of lifting to a height of 300 m and with a payload of up to 50 kg, and a long working time (up to 24 hours) which is limited only by the reliability characteristics of the multicopter. The cable, of either copper wires or optical fiber, ensures the transfer of

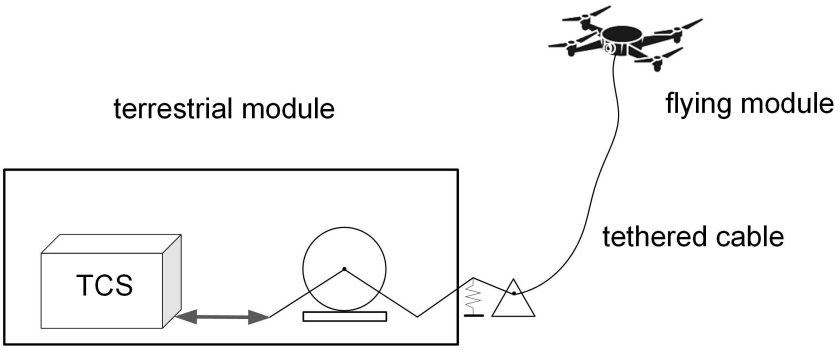


Fig. 1. Tethered UAVs high-altitude platform

large amounts of information from board to ground and vice versa. Local navigation systems equipped in high-altitude platforms, provide high positioning accuracy and increase noise immunity compared to satellite navigation systems.

3. Flying network for emergency using tethered multicopters

One of the important applications of UAVs in communication systems is the UAVs network or FANET (Flying Ad-Hoc Network). Nowadays, FANET is widely used in various fields: military, commercial, agricultural, etc. In particular, an application of FANET in the search and rescue operations was developed, named Flying network for emergencies [11].

Flying network for emergencies consists of two segments: a flying segment and a terrestrial segment. In the flying segment, UAVs are divided into groups, which are able to simultaneously communicate with each other and to the emergency services, victims or sensor nodes in the terrestrial segment without having any predefined and fixed infrastructure. In order to solve critical issues in FANET, such as communications and networking of the multiple UAVs, the modified of protocol IEEE 802.11p was presented in [12].

Nevertheless, one of the weaknesses of UAVs in flying network for emergencies is the short working time. For multicopter type, the flight time is about 30-60 minutes, and for fixed wing type, it can reach 1-2 hours. However, this is a relatively short time in the search and rescue missions, which often leads to inefficient operations or the need to replace UAVs many times. To increase the effectiveness of search and rescue operations, using tethered UAVs in flying network for emergencies is proposed, because of the following advantages:

- The effectiveness of the tethered UAVs system in various civilian areas, their mobility, compactness and cost-effectiveness compared to very expensive satellite systems;

- Super long working time, UAVs can operate up to 24 hours powered by ground;

- Possibility of lifting the platform to a height of up to 500 m with a payload of up to 50 kg;

- Ultra-wide bandwidth for data transmission through optical fiber inside the cable;

- The ability to shoot high-definition video and images acquired by the camera mounted on the UAV and then they are sent back to the ground through an optical fiber inside the cable;

- The system either can be freestanding or mounted on the rear of the vehicle. It is suitable for various industrial applications, such as television broadcasting, alarm relays, video surveillance, etc. When the system is installed on a car, the attached UAVs themselves can follow the car within a speed of 25 km per hour;

- Local navigation system based on tethered UAVs system provides high positioning accuracy and increased noise immunity compared to satellite navigation systems;

- A relatively short time of deployment of tethered UAVs system, approximately no more than 10 minutes;

- Tethered UAVs system provides the possibility of its operation at temperatures from -50 to +50 degrees Celsius, and the UAV itself can perform a flight with wind up to 15 meters per second;

- The ability to expand the operating range of the tethered UAVs system by using a chain of UAVs tethered one to the other. The first UAV in the chain is tethered to a ground station, while the last one serves as end effector.

Architectures of flying network for emergencies using tethered multicopters are considered in following scenarios:

•Collecting data from sensor fields in flying network for emergencies using tethered multicopters;

•Interactions within Flying Network for Emergencies using tethered multicopters;

•Multimedia transfers over the flying network for emergencies using tethered multicopters.

3.1. Collect data from sensor filed in flying network for emergencies using tethered multicopters. After a natural disaster, it is impossible for most telecommunication infrastructures to avoid from being destroyed, so the consequences and scale of the destruction must be assessed first. To do this, it is necessary to read data from sensory nodes, located in the destruction zone. Since sensor nodes can communicate with UAVs using various technologies, it is advisable to use a

heterogeneous gateway for data collection. Such a gateway, mounted on a UAV, will allow collecting data from sensor nodes and delivering them to a public communication network.

In a SAR operation, mobile base stations will deploy groups of UAVs, including tethered UAVs, to areas around MBS to gather information. All UAVs are equipped with a heterogeneous gateway, which is a network device or a relay system designed to ensure the interaction of two information networks that have different characteristics, using different sets of protocols and supporting different transmission technologies. Data can be collected by UAVs from nodes in sensor fields with technologies such as ZigBee, 6LoWPAN, LoRa, BLE, NB-IoT, etc... Therefore, these data can be transmitted through a chain of UAVs via IEEE 802.11p. By using tethered UAVs, which have a long working time, data can be transmitted to the MBS via IEEE 802.11p, LTE-A or LoRa, depending on a specific situation. An architecture of collecting data from sensor fields in flying network for emergencies using tethered multicopters is shown in Fig.2.

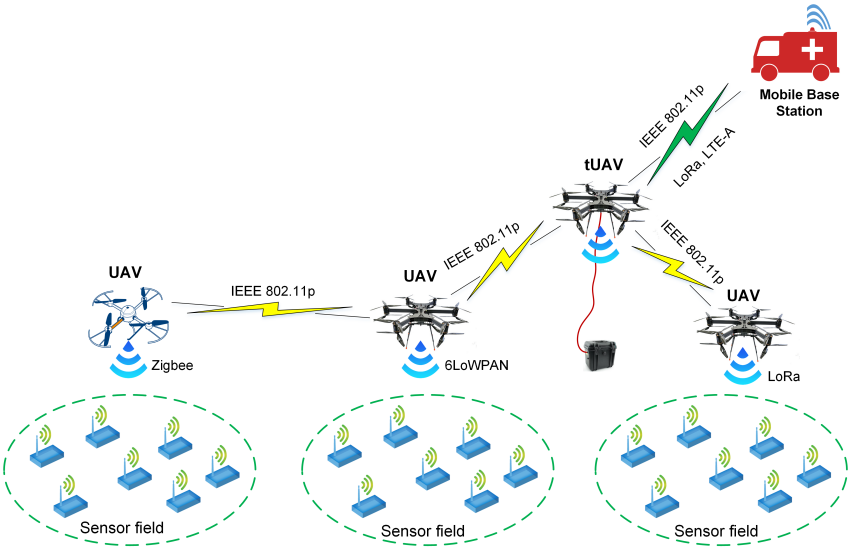


Fig. 2. Collect data from sensor filed in flying network for emergencies using tethered multicopters

3.2. Interaction of Flying Network for Emergencies using tethered multicopters. In the flying network for emergencies, communication among UAVs in a group and among groups of UAVs is of paramount importance. Technology IEEE 802.11p with the modified protocol CMMpp in [12] was developed to solve this issue. Moreover, tethered UAVs can be used in this network, which is presented in

Fig. 3. Tethered UAVs become super cluster nodes, which can receive information from cluster nodes of the groups or can replace cluster nodes when all of UAVs in the group can not be the cluster head. Furthermore, with its own advantages, tethered UAVs can carry on modules LTE to support the transmitting data with technology LTE-A. These tethered UAVs, therefore, can cover the areas which destroyed cellular base stations were supposed to support in the disaster. In addition, with tethered UAVs, the network will be more stable and reliable.

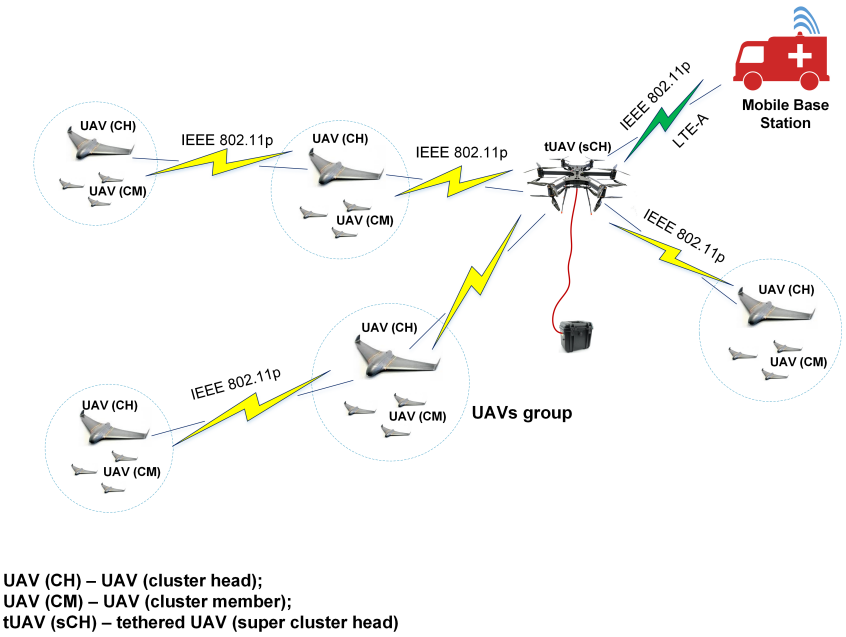


Fig. 3. Interaction of Flying Network for Emergencies using tethered multicopters

3.3. Multimedia transfers over the flying network for emergencies using tethered multicopters. Being equipped like any other UAVs, tethered UAVs can be joint in the multimedia transfers over flying network for emergencies. Assuming that there are subscribers wanting to call each other or rescuers trying to connect to missing people via VoWi-Fi using UAV groups in a destroyed area. According to functioning algorithms of mobile phones, in the absence of communication with the base station, the phones switch to scanning mode of available Wi-Fi networks. After a natural disaster occurred, scanning in the area will help discover subscribers who might be injured or buried under the rubble, waiting for help. A call between two subscribers will be performed through a chain of UAVs interacting with each other. UAVs can receive voice traffic by IEEE 802.11n or IEEE 802.11ac from subscribers,

transmit it through a chain of UAVs to the mobile base station, and connect to mobile operator to set up the call (Fig. 4).

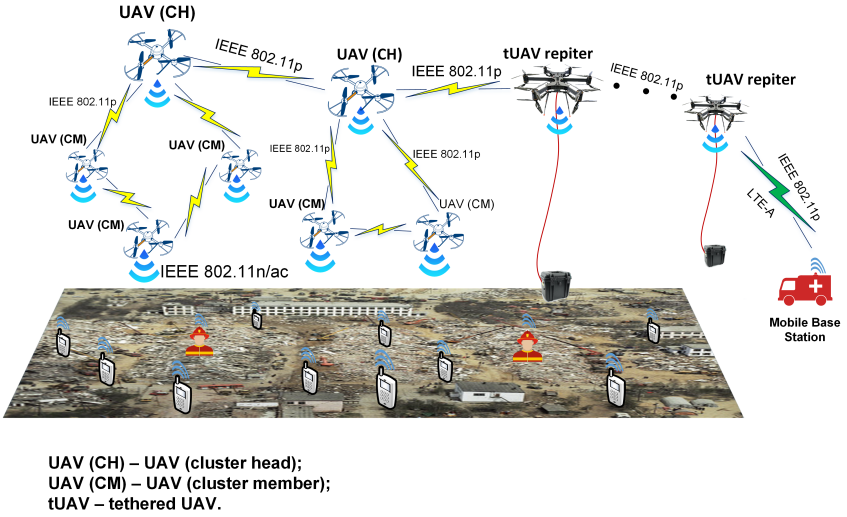


Fig. 4. Multimedia transfers over the flying network for emergencies using tethered multicopters

4. Conclusion

The paper provides a brief overview of tethered UAVs high-altitude platforms analyzing their advantages and disadvantages. With the benefits of the tethered UAVs high-altitude platform, flying network using tethered multicopters was proposed for emergency situations. Different architectures of this network were presented in order to enhance the effectiveness of search and rescue operations. In the future work, the research will prioritize in conducting a series of experiments and simulations to evaluate the performance of the proposed architectures.

REFERENCES

1. Scherer, J., Yahyanejad, S., Hayat, S., Yanmaz, E., Andre, T., Khan, A., ... & Rinner, B. An autonomous multi-UAV system for search and rescue // In Proceedings of the First Workshop on Micro Aerial Vehicle Networks, Systems, and Applications for Civilian Use. 2015. P. 33–38.
2. Cubber, G. D., Doroftei, D., Rudin, K., Berns, K., Matos, A., Serrano, D., ... & Silva, E // Introduction to the use of robotic tools for search and rescue. 2017. P. 1–17.

3. Shakhatreh, H., Sawalmeh, A. H., Al-Fuqaha, A., Dou, Z., Almaita, E., Khalil, I., ... & Guizani, M. Unmanned aerial vehicles (UAVs): A survey on civil applications and key research challenges // IEEE Access. 2019. V.7. P. 48572–48634.
4. Dinh, T. D., Pirmagomedov, R., Pham, V. D., Ahmed, A. A., Kirichek, R., Glushakov, R., & Vladyko, A. Unmanned aerial system–assisted wilderness search and rescue mission // International Journal of Distributed Sensor Networks. 2019. V.15(6). P. 1–15.
5. Koucheryavy, A., Vladyko, A., Kirichek, R.: State of the art and research challenges for public flying ubiquitous sensor networks // In: Balandin, S., Andreev, S., Koucheryavy, Y. (eds.) ruSMART 2015. LNCS. Springer, Cham. 2015. V. 9247. P. 299–308.
6. Kirichek, R., Paramonov, A., Koucheryavy, A.: Swarm of public unmanned aerial vehicles as a queuing network // In: Vishnevsky, V., Kozyrev, D. (eds.) DCCN 2015. CCIS. Springer, Cham. 2016. V 601. P. 111–120.
7. Fagiano, L. Systems of tethered multicopters: modeling and control design // IFAC-PapersOnLine. 2017. V.50(1). P. 4610–4615.
8. Al-Radaideh, A., & Sun, L. Self-localization of a tethered quadcopter using inertial sensors in a GPS-denied environment // In 2017 International Conference on Unmanned Aircraft Systems (ICUAS). IEEE. 2017. P. 271–277.
9. Vishnevsky, V., & Meshcheryakov, R. Experience of Developing a Multifunctional Tethered High-Altitude Unmanned Platform of Long-Term Operation // In International Conference on Interactive Collaborative Robotics. Springer, Cham. 2019. P. 236–244.
10. Vishnevsky, V., Tereschenko, B., Tumchenok, D., & Shirvanyan, A. Optimal method for uplink transfer of power and the design of high-voltage cable for tethered high-altitude unmanned telecommunication platforms // In International Conference on Distributed Computer and Communication Networks. Springer, Cham. 2017. P. 240–247.
11. Dinh, T. D., Kirichek, R., & Koucheryavy, A. Flying network for emergencies // In International Conference on Distributed Computer and Communication Networks. Springer, Cham. 2018. P. 58–70.
12. Dinh, T. D., Le, D. T., Tran, T. T. T., & Kirichek, R. Flying Ad-Hoc Network for Emergency Based on IEEE 802.11p Multichannel MAC Protocol // In International Conference on Distributed Computer and Communication Networks. Springer, Cham. 2019. P. 479–494.
13. Paredes, J. A., Saito, C., Abarca, M., & Cuellar, F. Study of effects of high-altitude environments on multicopter and fixed-wing UAVs' energy consumption and flight time // In 2017 13th IEEE Conference on Automation Science and Engineering (CASE). IEEE. 2017. P. 1645–1650.

Научное электронное издание
Номер госрегистрации 0322002892 в НТЦ «Информрегистр».

Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь (DCCN-2020)

МАТЕРИАЛЫ XXIII МЕЖДУНАРОДНОЙ НАУЧНОЙ КОНФЕРЕНЦИИ

(14–18 СЕНТЯБРЯ 2020 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 0322002892 в НТЦ «Информрегистр»
Мин. системные требования:
Pentium 4, Internet Explorer, Acrobat reader 4.0 и выше

Дата подписания к использованию: 01.09.2020
1 электронно-оптический диск (CD-R), 44,9 Мб, Тираж 210 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